

**Kodolányi János University**

# **THESIS**

**SHAGAI TURTOGTOKH**  
**COMPUTER SCIENCE OPERATIONAL ENGINEERING – NBCSIK (BA/BSC)**

**Budapest**  
**2026**

**Kodolányi János University**  
**Department of Informatics**

# **Objective Evaluation of Student Performance from Moodle Logs via Similarity Analysis**

A Tool for Log-Based, Data-Driven and Anti-Discriminative Student Assessment

**Supervisor: Dr. Pitlik László**

**AUTHOR: SHAGAI TURTOGTOKH**  
**DEPARTMENT OF INFORMATICS**

**Budapest**

**2026**

## **Abstract**

The rapid uptake of e-learning, accelerated during the COVID-19 period, has increased reliance on learning management systems (LMSs) for communication and assessment. In discussion-based courses, evaluation is often labour-intensive and partly impression-based, even though LMS platforms record interaction traces and student text that could support more transparent decision making. The literature reviewed in the thesis identified a gap: behavioural logs and semantic information from written responses are frequently treated as separate evidence streams, and multi-dimensional performance is often aggregated through implicit judgement or fragile weighting, with limited robustness checking. To address this, the thesis developed and demonstrated a direction-aware, multi-attribute workflow for teacher-facing decision support. A pseudonymised Moodle forum export (24 students, 576 messages; 17 October–17 November 2024) was cleaned and pseudonymised, transformed into student-level indicators, and organised as an Object–Attribute Matrix with explicit benefit/cost directionality. Topic alignment was operationalised as cosine similarity of sentence embeddings and used as one bounded attribute rather than a correctness proxy; an auxiliary AI-involvement score was included only as a probabilistic, supportive indicator under a non-punitive interpretation boundary. Evidence was rank-transformed and aggregated with COCO Y0 similarity analysis tool under a constant baseline ( $Y = 1000$ ), producing an explainable, anti-discriminative ordering and attribute-impact diagnostics, complemented by staircase-based exclusion runs. Empirically, the full run exhibited pronounced “sameness” with large tie groups; exclusion improved discrimination but altered the construct footprint, and objective rankings generally diverged from peer-based evaluations. Internal credibility was examined through inverse-run function-symmetry checks and distorted-input tests that produced expected failure modes under implausible inputs. The end-to-end pipeline was implemented in a Streamlit prototype and released with reproducibility artefacts for instructors and course leaders; the work remained bounded to this cohort and forum-export evidence, and future work prioritises replication across courses, sensitivity analysis of modelling choices, and tie-robust ranking and test automation, and platform integration into existing LMS reporting and analytics workflows.

Keywords: Moodle, learning analytics, student performance evaluation, Object–Attribute Matrix, COCO Y0, similarity analysis, semantic similarity, decision support

## **Absztrakt**

Az e-learning gyors térnyerése, amelyet a COVID–19 időszak tovább erősített, növelte a tanulásmenedzsment-rendszerekre (LMS-ekre) való támaszkodást a kommunikáció és az értékelés terén. A diszkusszióalapú kurzusokban az értékelés gyakran munkaigényes és részben benyomásokon alapul, noha az LMS-platformok olyan interakciós nyomokat és hallgatói szövegeket rögzítenek, amelyek átláthatóbb döntéshozatalt támogathatnának. Az értekezésben áttekintett szakirodalom egy hiányosságot azonosított: a viselkedési naplóadatokat és az írott válaszokból származó szemantikai információkat gyakran elkülönült bizonyítékforrásokként kezelik, miközben a többdimenziós teljesítményt sok esetben implicit megítélés vagy sérülékeny súlyozási megoldások alapján aggregálják, korlátozott robusztussági ellenőrzés mellett. Ennek kezelésére a dolgozat egy irányérzékeny, többattribútumos munkafolyamatot dolgozott ki és demonstrált, kifejezetten oktatói döntéstámogatási célokra.

Egy pszeudonimizált Moodle-fórumexport (24 hallgató, 576 üzenet; 2024. október 17. – 2024. november 17.) tisztításra és pszeudonimizálásra került, majd hallgatói szintű indikátorokká alakult, és explicit haszon-/költségirányokkal rendelkező Objektum–Attribútum Mátrixba (Object–Attribute Matrix, OAM) szerveződött. A tematikus illeszkedés a mondatbeágyazások koszinusz-hasonlóságaként került operacionalizálásra, és nem a helyesség közvetett mérőszámaként, hanem egy korlátos attribútumként szerepelt; egy kiegészítő AI-érintettség pontszám pedig kizárólag valószínűségi, támogató indikátorként került bevonásra, nem büntető értelmezési keret mellett. A bizonyítékok rangtranszformáción estek át, majd a COCO Y0 hasonlóságelemző eszközzel, konstans bázis ( $Y = 1000$ ) mellett kerültek aggregálásra, amely magyarázható, antidiszkriminatív sorrendet és attribútumhatás-diagnosztikát eredményezett, lépcsőfüggvény-alapú kizárási futtatásokkal kiegészítve.

Empirikusan a teljes futtatás kifejezett homogenitást mutatott, nagy holtversenycsoportokkal; a kizárás javította a megkülönböztethetőséget, ugyanakkor módosította a konstrukció tartalmi profilját, és az objektív rangsorok általában eltértek a társértékelésen alapuló értékelésektől. A belső megbízhatóság vizsgálata inverz futtatású függvény-szimmetria-ellenőrzések, valamint torzított bemeneti tesztek révén történt, amelyek valószínűtlen inputok esetén az elvárt hibamódokat eredményezték. A teljes, végponttól végpontig terjedő folyamat egy Streamlit-prototípusban valósult meg, és reprodukálhatósági artefaktumokkal együtt került közzétételre oktatók és kurzusfelelősök számára; a munka ugyanakkor erre a kohorszra és a fórumból exportált bizonyítékokra korlátozódott. A jövőbeli kutatások elsődleges irányai a különböző kurzusokon történő replikáció, a modellezési döntések érzékenységvizsgálata, a

holtversenyekkel szemben robusztus rangsorolás és tesztautomatizálás, valamint a platform integrálása a meglévő LMS-riportálási és tanulóelemzési munkafolyamatokba.

Kulcsszavak: Moodle, tanulóelemzés, hallgatói teljesítményértékelés, Objektum–Attribútum Mátrix, COCO Y0, hasonlóságelemzés, szemantikai hasonlóság, döntéstámogatás

## Contents

1	Introduction .....	1
1.1	Aims and Objectives .....	1
1.1.1	Explicit Promises .....	2
1.1.2	Research questions and expected outcomes .....	3
1.1.3	Hypothesis.....	4
1.1.4	Methodology .....	5
1.2	Tasks.....	5
1.3	Targeted groups .....	6
1.4	Utilities.....	7
1.4.1	Stakeholder groups and value types .....	7
1.4.2	Quantitative value estimation and assumptions .....	8
1.4.3	Analytical interpretation and overall impact .....	10
1.5	Motivation .....	10
1.6	Structure of the thesis.....	11
2	Literature .....	14
2.1	Learning Analytics .....	14
2.1.1	Conceptual foundations .....	15
2.1.2	Practical applications.....	16
2.2	E-Learning Systems and Moodle .....	17
2.2.1	Moodle as an LMS.....	17
2.2.2	Platform analytics vs. research/assessment analytics.....	18
2.2.3	Implications for educational measurement in Moodle-based courses .....	19
2.3	Log-Based Student Modelling .....	20
2.3.1	From LMS traces to student models.....	20
2.3.2	Modelling choices: what counts as “evidence”? .....	22
2.3.3	Validity and limitations: logs as imperfect proxies .....	22
2.3.4	Privacy, ethics, and GDPR constraints on log-based modelling .....	23
2.4	Assessment and Evaluation Theory in Education .....	24
2.4.1	Why evaluation is inherently value-laden .....	25
2.4.2	Limits of pre-digital evaluation under LMS conditions .....	25

2.4.3	Why multi-attribute and validated evaluation matters .....	25
2.4.4	Responsibility and governance in automated evaluation .....	26
2.5	Similarity Analysis for Educational Text .....	27
2.5.1	Why “similarity” is a usable signal in educational text .....	27
2.5.2	Lexical similarity: transparent but brittle .....	27
2.5.3	Semantic similarity: meaning-oriented comparison for open-ended responses ....	28
2.5.4	Validity limits .....	28
2.6	Multi-Attribute Evaluation .....	29
2.6.1	Why multi-attribute evaluation is necessary in LMS-based assessment.....	30
2.6.2	The Object–Attribute Matrix as the formal representation .....	30
2.6.3	Directionality and the unavoidable role of value choices .....	31
2.6.4	Aggregation without arbitrary weights: anti-discriminative optimisation.....	31
2.6.5	Attribute exclusion as a diagnostic tool for discrimination and importance.....	32
2.6.6	Verification and robustness as part of multi-attribute evaluation quality .....	32
2.6.7	Positioning COCO Y0 among MCDA methods .....	33
2.7	Automated/Algorithmic Decision Support in Education .....	34
2.7.1	What “decision support” means in learning analytics and EDM.....	35
2.7.2	Typical decision-support functions: dashboards, early warning, and recommendations .....	36
2.7.3	Decision support embedded in LMS ecosystems .....	36
2.7.4	Quality assurance: from “automated output” to defensible support.....	37
2.7.5	Responsibility, ethics, and GDPR as design constraints.....	38
2.7.6	Decision support vs automated judgement: why scope boundaries matter .....	39
2.8	AI Detection in Education.....	40
2.8.1	Why “AI detection” became an educational problem .....	40
2.8.2	How AI-writing detectors work.....	40
2.8.3	Empirical performance and the “evidence problem” .....	41
2.8.4	Vulnerability, adaptivity, and “detector evasion” .....	41
2.8.5	Fairness, language, and governance constraints .....	41
2.9	The Gap.....	42
2.9.1	What the literature already enables and what it does not.....	43
2.9.2	The specific gap addressed by this thesis .....	44

2.9.3 What the thesis contributes .....	46
2.10 BPROF Subjects and the Thesis .....	47
2.10.1 Networks & Computer Architectures.....	47
2.10.2 Introduction to Algorithms .....	47
2.10.3 Operating Systems .....	48
2.10.4 Introduction to Programming .....	48
2.10.5 Programming I, II .....	48
2.10.6 Databases .....	49
2.10.7 Data Visualization.....	49
2.10.8 Electronics & Circuits.....	49
2.10.9 System Modelling .....	50
2.10.10 System Operation .....	50
2.10.11 System Planning .....	50
2.10.12 Software Architectures .....	50
2.10.13 Software Testing .....	51
2.10.14 Business Process Management .....	51
2.10.15 Business Law and Regulation.....	51
2.10.16 IT Security .....	52
2.10.17 ICT in IT-Security.....	52
2.10.18 Intercultural Communication.....	52
2.11 Use of LLM in the Thesis Writing and Development Process .....	53
2.11.1 Scope of LLM Assistance.....	53
2.11.2 Boundaries of LLM Use.....	54
2.11.3 Critical Posture Toward LLM Output.....	54
2.12 Comparison with Other Running COCO-Oriented Theses .....	55
2.12.1 Shared methodological core.....	56
2.12.2 What differentiates the present thesis from the peer works.....	57
2.12.3 Lessons for the present Moodle-based evaluation model.....	58
3 Own Developments .....	60
3.1 Analytical process .....	61
3.1.1 Moodle Data & Structure .....	62

3.1.2	Log Data Extraction & Pre-Processing .....	64
3.1.3	Semantic Similarity for Descriptive Text Responses.....	66
3.1.4	AI Detection in Text Responses .....	69
3.1.5	OAM Construction/Ranking & Attributes.....	71
3.1.6	COCO Y0 Analysis & Attribute Exclusion .....	77
3.1.7	Function-Symmetry Validation .....	81
3.1.8	Results .....	83
3.1.9	Input Verification Test.....	94
3.2	Demo Software Development.....	100
3.2.1	Purpose and Scope of the Demo Tool.....	101
3.2.2	System Architecture .....	102
3.2.3	Data-Pipeline Implementations .....	104
3.2.4	Similarity Analysis and AI Detection in Text Responses .....	106
3.2.5	Ranking & Scoring Logic .....	106
3.2.6	User Interface Interaction Design .....	107
3.2.7	Visual Outputs and AI-Assisted Interpretive Summaries .....	108
3.2.8	Testing.....	108
3.2.9	IT Security, Privacy, and Ethics in Implementation .....	110
3.2.10	System Requirements .....	111
3.2.11	Prototype Limitations .....	112
3.3	Benchmarking against an independent LLM solution .....	113
3.3.1	Summary of the LLM benchmark method.....	113
3.3.2	Comparative evaluation of the human-authored and LLM-authored solutions...	114
4	Discussion .....	116
4.1	Restatement of study focus .....	116
4.2	Interpretation of key findings .....	116
4.3	Resolution of method-bound interpretive issues.....	119
4.4	Comparison with literature.....	122
4.5	Theoretical contributions .....	123
4.6	Practical implications .....	123
4.7	Strengths of the study.....	124

4.8	Limitations .....	124
4.9	Research implications and directions .....	125
4.10	Interpretation of the LLM benchmark .....	126
4.11	Take home message .....	128
5	Conclusion.....	129
5.1	Answers to the Research Questions .....	131
6	Future work.....	132
7	Summary .....	132
8	Annexes .....	133
8.1	List of Abbreviations .....	133
8.2	List of Figures .....	134
8.3	List of Tables.....	135
8.4	List of Equations .....	137
8.5	List of Definitions .....	139
8.6	Reproducibility Package.....	141
8.6.1	Excel file.....	141
8.6.2	Model source code directory.....	141
8.6.3	GitHub repository .....	142
8.6.4	Peer evaluation Excel file.....	142
8.6.5	Database .....	142
8.6.6	Working principle of COCO.....	142
8.6.7	Demo test results .....	142
8.6.8	T01–T16 Reference-Coverage Workbook.....	142
8.6.9	Shared LLM Conversation .....	142
8.7	LLM: prompt, output, and artefacts .....	143
8.7.1	Verbatim benchmark prompt and full LLM output.....	146
8.7.2	Benchmark-generated artefacts and metadata .....	156
8.8	Tutorial on operation.....	157
8.9	Licenses and ethics .....	157
8.10	References.....	158
8.11	Reference-Coverage Table .....	162

8.12 Repository of parallel theses consulted for the comparative synthesis.....	163
---	-----

# 1 Introduction

The rapid institutional uptake of e-learning—accelerated during the COVID-19 period—has intensified the use of Learning Management Systems (LMSs) as routine infrastructures for course delivery, communication, and assessment. In such environments, instructors increasingly face a practical tension: student evaluation must remain pedagogically meaningful and fair, yet it is often time-consuming and partly impression-based, especially when courses rely on discussion tasks and open-ended responses. At the same time, LMS platforms such as Moodle generate detailed interaction traces (e.g., logs), including timestamps, reply relations, and student-produced text, which can serve as bounded evidence for decision support if transformed through transparent, auditable rules rather than ad hoc interpretation. This thesis therefore treats Moodle as an evidence environment and develops a reproducible workflow that translates forum-export records into a multi-attribute evidence model for student performance evaluation. The approach integrates behavioural indicators (e.g., activity regularity, response latency, interaction structure) and text-related indicators (e.g., semantic topic alignment) into an Object–Attribute Matrix (OAM), applies explicit directionality rules ( $0 = \uparrow$  more ideal;  $1 = \downarrow$  more ideal), and aggregates the resulting evidence using the COCO Y0 optimisation engine under a constant baseline ( $Y = 1000$ , COCO index unit; fixed baseline), complemented by internal verification and robustness tests. The academic problem addressed here is therefore not merely how to process Moodle logs, but how to justify a defensible, multi-attribute, non-punitive evaluation logic from imperfect behavioural and textual proxies.

## 1.1 Aims and Objectives

The key definitions used in this thesis are as follows. An LMS is an institutional platform that supports course delivery and records interaction traces; Moodle is the LMS used as the study’s data source; learning analytics (LA) is understood as decision support based on learner data; an Object–Attribute Matrix (OAM) is the formal representation where each row is a student (object) and each column is an indicator (attribute); COCO Y0 is a component-based object comparison method operating under a constant target producing an explainable ordering without relying on manually assigned weights; semantic similarity is operationalised as cosine similarity between sentence embeddings to quantify topical alignment of responses.

- **Aim A1 (methodological):** To develop a transparent, reproducible, log-based evaluation workflow that transforms Moodle forum exports into a construct-explicit

indicator system represented as an OAM with explicit directionality rules. Implemented in: §§3.1.1–3.1.5.

- **Aim A2 (aggregation):** To generate an explainable multi-attribute ordering using COCO Y0 under a constant baseline, including staircase-based diagnostics for identifying weakly informative attributes. Implemented in: §§3.1.6–3.1.8.
- **Aim A3 (verification):** To strengthen defensibility by integrating internal consistency validation (inverse-run function-symmetry) and robustness probing through deliberately distorted inputs. Implemented in: §§3.1.7–3.1.9.
- **Aim A4 (decision-support engineering):** To operationalise the pipeline as a teacher-facing prototype that executes the workflow end-to-end, preserves intermediate artefacts for auditability, and documents governance constraints (privacy/ethics/IT security) as design boundaries. Implemented in: §§3.2.1–3.2.11 and §3.2.9.

### Operational objectives:

- **O1.** Define a cohort-level dataset model and pre-processing rules that preserve traceability from raw Moodle rows to student-level indicators (e.g., timestamps, thread structure, reply relations). Implemented in: §§3.1.1–3.1.2.
- **O2.** Compute and document multiple indicator families (plural → examples): (1) participation intensity indicators (e.g., total posts, active days), (2) temporal indicators (e.g., response-time measures), (3) interaction-structure indicators (e.g., unique discussions), and (4) text-related indicators (e.g., topic relevance, citation count, auxiliary AI signal). Implemented in: §§3.1.3–3.1.5.
- **O3.** Compare objective COCO-based rankings with a peer-based subjective ranking to characterise agreement/divergence under explicit tie-handling. Implemented in: §3.1.8.4.

#### 1.1.1 Explicit Promises

- **P1. Multi-attribute evidence model:** the thesis will transform Moodle forum-export data into an Object–Attribute Matrix (OAM) with explicit directionality rules for A1–A29. Grounded in §§2.3–2.6, this is realised in §§3.1.1–3.1.5, reported in §§3.1.8–3.1.9, and discussed in Chapter 4.
- **P2. Weight-avoiding aggregation:** the thesis will aggregate heterogeneous student indicators through COCO Y0 under a constant baseline ( $Y = 1000$ ), rather than by

manually assigned weights. This is grounded in §§2.6.6–2.6.7 and §2.9.2, realised in §3.1.6, reported in §3.1.8, and discussed in §§4.2–4.5.

- **P3. Semantic topic-alignment evidence:** the thesis will operationalise descriptive-response alignment through sentence-embedding cosine similarity using all-MiniLM-L6-v2. Grounded in §§2.5.3–2.5.4 and §2.9.2, this is realised in §3.1.3, incorporated into the OAM in §3.1.5, and discussed in §§4.3–4.4.
- **P4. Bounded AI-involvement signal:** the thesis will include a RoBERTa-based AI-involvement indicator only as a supportive, low-stakes attribute rather than as proof of misconduct. Grounded in §2.8 and §2.9.2, this is realised in §§3.1.4–3.1.5 and discussed in §§3.2.9 and 4.3–4.4.
- **P5. Internal validation:** the thesis will test whether the COCO-based evaluation behaves coherently under deterministic inversion of the ranked input. Grounded in §§2.6.7 and 2.9.2, this is realised in §3.1.7, reported in §3.1.8, and discussed in §§4.2–4.5.
- **P6. Robustness probing:** the thesis will test whether distorted or implausible inputs produce unstable or degenerate outputs rather than artificially persuasive rankings. Grounded in §§2.6.6–2.6.7 and §2.9.2, this is realised in §3.1.9 and discussed in §§4.2–4.5.
- **P7. Objective–subjective comparison:** the thesis will compare objective log-based rankings with peer-based subjective evaluation using tie-aware rank harmonisation and Spearman correlation. Grounded in §§2.3.3, 2.4.4, and 2.9.2, this is realised in §3.1.8.4, discussed in §§4.2–4.4, and concluded in Chapter 5.
- **P8. Prototype feasibility under governance constraints:** the thesis will operationalise the workflow in a Streamlit prototype that preserves auditability while documenting privacy, ethics, and implementation boundaries. Grounded in §§2.3.3, 2.8, and 2.9.2, this is realised in §3.2 and discussed in §§4.6–4.8.

### 1.1.2 Research questions and expected outcomes

- **RQ1:** How can Moodle discussion-forum logs and text-based signals be transformed into a transparent, reproducible Object-Attribute Matrix for teacher-facing evaluation?
- **RQ2:** How does COCO Y0 behave on the full, diligence-only, and understanding-only matrices, and how does staircase-based exclusion change discrimination and construct representation?

- **RQ3:** To what extent do COCO-based objective rankings agree with peer-based subjective evaluation in the same cohort?
- **RQ4:** How internally coherent and robust is the proposed pipeline when assessed through inverse-run function-symmetry validation and distorted-input tests?

### Expected outcomes

- The expected outcome for RQ1 is a fully traceable workflow from Moodle export to student-level OAM attributes under explicit directionality rules.
- The expected outcome for RQ2 is that the full configuration may exhibit low discrimination, while exclusion-driven re-runs improve separability at the cost of changing the evaluated construct footprint.
- The expected outcome for RQ3 is limited monotonic agreement between objective and peer-based rankings, because the two procedures rely on partly different evidence channels.
- The expected outcome for RQ4 is bounded internal defensibility: the pipeline should remain coherent under inverse-run checks and should fail in interpretable ways under distorted inputs rather than generating artificial structure.

### 1.1.3 Hypothesis

The thesis is developed in an R&D format where hypotheses are treated as auditably testable expectations linked to the sections that evaluate them.

- **H1 (construct divergence):** Objective rankings derived from the log-based, multi-attribute pipeline will exhibit limited monotonic agreement with peer-based subjective rankings, reflecting partially different performance constructs and evidence visibility. Evaluated in: §3.1.8.4 and interpreted in §4.2.4.
- **H2 (diagnostic exclusion):** Staircase-based exclusion of weakly informative attributes will increase discrimination (i.e., reduce “sameness”/large tie groups) but will also change the construct footprint, so excluded-run outputs must be interpreted as a complementary variant rather than as a definitive correction. Evaluated in: §§3.1.6.3–3.1.8.1 and interpreted in §§4.2.1–4.2.2.

### 1.1.4 Methodology

The thesis follows a research-and-development (R&D) methodology that combines (i) an analytical evaluation workflow and (ii) an engineering realisation as a reference prototype. Empirically, the analysis uses a pseudonymised Moodle discussion-forum export in which each row is treated as a forum-post event and aggregated into student-level indicators (24 students, 12 discussion threads, 576 messages within a defined time window). Methodologically, the pipeline proceeds through (1) extraction, cleaning, and database structuring (SQLite), (2) computation of log-derived behavioural indicators, (3) computation of semantic topic alignment for descriptive responses using sentence embeddings (all-MiniLM-L6-v2) and cosine similarity, (4) computation of an auxiliary AI-involvement signal using a RoBERTa-based detector (used only as bounded decision-support evidence), (5) construction of the OAM with explicit directionality and rank transformation for cross-attribute comparability, and (6) COCO Y0 execution under a constant baseline with staircase diagnostics and re-runs. Reliability is treated as internal consistency and sensitivity, not supervised predictive accuracy: inverse-run function symmetry provides a coherence check, while deliberately distorted inputs probe failure modes (unstable vs degenerate outputs). Finally, the pipeline is operationalised in a Streamlit prototype that exposes intermediate artefacts (OAM, ranked OAM, COCO tables, validation outputs) to support instructor-facing auditability.

## 1.2 Tasks

- **Task T1 (data and governance preparation):** Acquire, anonymise, and structure the Moodle discussion export for traceable processing (including consistency checks and local storage). Performed in: §§3.1.1–3.1.2.
  - **Rationale:** without a traceable and governance-aware data foundation, subsequent indicators are neither reproducible nor defensible.
- **Task T2 (indicator construction):** Define and compute the indicator catalogue (A1–A29) and construct the OAM with explicit directionality rules (0↑/1↓). Performed in: §3.1.5.
  - **Rationale:** explicit construct meaning and directionality are required to avoid implicit value choices hidden in ad hoc scoring.
- **Task T3 (text evidence extraction):** Implement semantic similarity for descriptive answers and compute student-level topic relevance; implement the auxiliary AI-involvement indicator as bounded evidence. Performed in: §§3.1.3–3.1.4.

- **Rationale:** discussion tasks require content-sensitive evidence (e.g., topical alignment), but such signals must be bounded and integrated as one attribute among many.
- **Task T4 (multi-attribute aggregation):** Rank-transform the OAM and execute COCO Y0 under a constant baseline; perform staircase diagnostics and exclusion-driven re-runs. Performed in: §§3.1.5.3–3.1.6.3 and consolidated in §3.1.8.
- **Task T5 (quality assurance):** Perform inverse-run function-symmetry validation and distorted-input verification tests to characterise internal consistency and failure modes. Performed in: §§3.1.7–3.1.9.
- **Task T6 (human comparison):** Compare objective rankings with peer-based subjective evaluation using tie-aware harmonisation and Spearman correlation. Performed in: §3.1.8.4.
  - **Rationale:** the thesis explicitly investigates whether and how objective evidence diverges from subjective judgement in the same cohort.
- **Task T7 (prototype implementation):** Implement the end-to-end workflow as a Streamlit-based demo tool with persisted artefacts and documented governance constraints. Performed in: §3.2 (incl. §3.2.9).

### 1.3 Targeted groups

This thesis targets the following stakeholder categories:

1. **Course-level instructors and lecturers (Moodle forum-based teaching):** They face time pressure and interpretive uncertainty when evaluating discussion participation and open-ended responses; the thesis outputs support structured review and auditability rather than automated judgement.
2. **Course leaders/programme coordinators (quality assurance and consistency):** They require defensible, reproducible evaluation procedures across cohorts and teaching teams; the OAM + COCO framing supports transparent criteria documentation and comparability within defined contexts.
3. **Institutional e-learning/learning-analytics support units (implementation and governance):** They must operationalise analytics under GDPR/ethics constraints and ensure traceability of outputs; the prototype documents local-first storage, artefact retention, and governance boundaries as design elements.

4. **Educational researchers and method developers (evaluation methodology):** They may reuse the reproducible pipeline artefacts (OAM definitions, validation logic, robustness tests) as a reference implementation for similar log-based evaluation contexts.

## 1.4 Utilities

This thesis operationalises Moodle forum exports as auditable decision-support evidence for performance evaluation: log- and text-derived indicators are formalised in an Object–Attribute Matrix (OAM) with explicit directionality rules and aggregated using COCO Y0, while a Streamlit prototype executes the workflow end-to-end and preserves intermediate artefacts for inspection and reporting (§§3.1–3.2).

The core utility is therefore not “automated judgement”, but measurable reductions in evaluation workload and increased transparency and defensibility for stakeholders who must justify assessment decisions.

### 1.4.1 Stakeholder groups and value types

The project creates value for the following stakeholder groups (aligned with §1.3 and the system scope of §§3.1–3.2):

- **Instructors/lecturers (primary users):** operational and economic value via time savings in evidence extraction, indicator computation, and documentation; educational value via faster, more consistent feedback.
- **Course leaders/programme coordinators (quality assurance):** operational value via comparability and audit trails across courses; economic value via reduced moderation and dispute-handling time.
- **Institutional e-learning/learning-analytics support units:** technological and operational value via a reproducible pipeline that reduces ad hoc scripting and supports governance-by-design (local-first storage, artefact retention).
- **Students (affected group):** educational and social value via increased transparency (clearer evidence basis), faster feedback cycles, and reduced administrative friction in clarifications/appeals.
- **Educational researchers/method developers:** technological value via a reusable workflow (OAM construction, validation, robustness logic) that shortens repeated data-preparation cycles.

## 1.4.2 Quantitative value estimation and assumptions

The following estimates use a generalised “typical course” scale:

Moodle-based course with  $N = 100$  enrolled students and discussion-based evaluation over one semester. Values scale approximately linearly with  $N$  and with the number of courses.

A consolidated reference table for recurring units, scales, and coding conventions used across the thesis is provided in Annex §8.5, Table 8.5- II.

### **Key parameters:**

- Labor cost basis (Hungary): €14.1/hour (Eurostat European Commission, 2025).
- Instructor time baseline anchor: online teaching studies report substantial grading and discussion-forum time investments (e.g., discussion facilitation and grading/feedback components).
- Student time valuation (minimum-wage proxy): 1,856 HUF/hour (Magyarország Kormánya, 2025), converted using ECB EUR/HUF reference rate (25 Feb 2026: 1 EUR = 375.88 HUF (European Central Bank, 2026)): €4.94/hour.
- Optional hosting overhead (if not fully local): AWS LightSail entry instance \$5/month (Amazon Web Services, 2026).

Stakeholder group	Benchmark (before) Costs	Benchmark (before) Economic value*	Post-project (after) Costs	Post-project (after) Economic value	Net added value (after–before)
Instructors lecturers /	47.7 h × €14.10 = €672	€ 672	29.0 h × €14.10 = €409	€ 672	€ 263
Course leaders / QA	2.0 h × €14.10 = €28	€ 28	1.0 h × €14.10 = €14	€ 28	€ 14
E-learning / LA support	2.0 h × €14.10 = €28	€ 28	0.8 h × €14.10 + €1.70 hosting share = €13	€ 28	€ 15
Students (cohort of 100)	admin friction 0.10 h/student × €4.94 = €49	productive use of feedback 1.00 h/student × €4.94 = €494	admin friction (0.07 h/student incl. brief orientation) × €4.94 = €35	improved feedback uses 1.20 h/student × €4.94 = €593	€ 114
Researchers (per study/course reuse)	40.0 h × €14.10 = €564	€ 564	22.0 h × €14.10 = €310	€ 564	€ 254

Table 1.4- I Two-layer economic analysis per stakeholder group (per semester; per typical course; currency: EUR) (Source: Own Presentation)

### Notes (hour derivations):

- 47.7 h (instructors, benchmark) = (100 students × 10 graded posts × 2.5 min)/60 + 6.0 h overhead (export, evidence compilation, clarifications).
- 29.0 h (instructors, post-project) = (100 × 10 × 1.5 min)/60 + 4.0 h reduced overhead due to automated OAM/COCO reporting.
- 2.0 h → 1.0 h (QA) = 4 checkpoints/semester × (30 min → 15 min) for sampling, consistency checks, and brief dispute review.
- 2.0 h → 0.8 h (LA support) = 4 support cases/semester × (30 min → 12 min) for export/configuration help under a repeatable workflow.

### Assumption logic:

- Instructor time decreases because the prototype automates log ingestion, indicator extraction, OAM construction, rank transformation, COCO execution, and exportable reporting (§§3.1–3.2).
- QA and support time decreases because intermediate artefacts and outputs are preserved and exportable, reducing manual evidence reconstruction and ad hoc processing (§3.2.1, §3.2.9).

- Student value increases modestly because faster and clearer evidence improves the efficient use of feedback and reduces clarification overhead (modelled via small time shifts valued at a minimum-wage proxy).

Item	Per-course added value (€)	Courses/semester (count)	Semester added value (€)
Instructor time savings	263	50	13,150
QA time savings	14	50	705
E-learning/LA support savings (incl. one shared small hosting instance)	15	50	750
Student time/value gain	114	50	5,700
Total (order-of-magnitude)			~20,300

Table 1.4- II Example scaling to an institution (illustrative; 50 similar courses) (Source: Own Presentation)

Table 1.4-II illustrates how the estimated per-course added value scales to an institutional scenario under the thesis’s simplifying assumptions.

If the tool is run fully locally, the small hosting overhead can be removed; if the institution has larger cohorts or more discussion-based courses, totals increase proportionally.

### 1.4.3 Analytical interpretation and overall impact

Across stakeholder groups, the dominant measurable benefit is labour-time reallocation: instructors and support roles spend fewer hours on evidence extraction and documentation while preserving (or improving) traceability of the evaluation basis through stored intermediate artefacts and exportable outputs (§§3.1–3.2).

The quantitative tables therefore represent a conservative lower bound of economic value because they monetise only time effects under publicly reported labour-cost proxies, not downstream gains such as reduced conflict escalation, improved trust, or improved governance readiness.

At societal level, the main impact is procedural fairness and transparency: by treating Moodle as an evidence environment and keeping evaluation outputs auditable (rather than opaque or purely impression-based), the approach supports less discriminatory assessment practices and clearer accountability boundaries for algorithmic decision support in education (§2 and §3).

## 1.5 Motivation

This thesis is motivated by a concrete educational evaluation situation encountered in an IT-course discussion task: students solved structured forum-based tasks and then conducted peer

evaluation in the classroom, after which the corresponding Moodle export was used to perform objective log-based analysis and compare outcomes to subjective judgement. This sequence exposed both the potential of trace evidence (repeatability, auditability) and the limitations of impression-based evaluation when the evidence basis is not explicit. The project therefore pursues a pragmatic goal: to reduce evaluation friction and increase transparency by converting Moodle discussion traces into a documented, reproducible workflow and an executable prototype, while maintaining strict scope boundaries so that outputs remain decision support rather than automated judgement.

## 1.6 Structure of the thesis

The thesis is organised to keep methodological justification logically prior to implementation and interpretation. Chapter 2 defines the theoretical and methodological foundations needed to read the pipeline correctly: learning analytics as decision support (§2.1), Moodle as evidence environment (§2.2), log-based modelling under validity and GDPR constraints (§2.3), evaluation theory and governance boundaries (§2.4, §2.7), similarity analysis for descriptive text with explicit validity limits (§2.5), multi-attribute evaluation and weight-avoidance via COCO (§2.6), AI detection as bounded auxiliary evidence (§2.8), the integrated gap addressed by the thesis (§2.9) and the transparent documentation of LLM-assisted writing and development support (§2.11). Chapter 3 presents the own developments: the analytical process (§3.1) and the automation/prototype implementation (§3.2), including validation and robustness procedures. In addition, §3.3 introduces an independent LLM benchmark addressing the same overall goal through a method-nonidentical solution, whose interpretation is developed in §4.10 and whose full prompt, output, and related artefacts are documented. Chapter 4 interprets results and explicitly revisits promised limitations and governance risks, including “sameness,” directionality assumptions, and misuse boundaries. Annexes (Chapter 8) provide reproducibility artefacts (spreadsheets, exports, documentation, full prompt, output, and related artefacts), consistent with the thesis’ claim that the workflow is auditable and repeatable.

**Volume-limited items:** Some production-level features are intentionally out of scope for a bachelor thesis prototype (e.g., full institutional authentication/authorisation, consent-management workflows, and local COCO execution without external endpoint dependency); these are documented as limitations (§3.2.11) and positioned for future work. A deeper

comparative analysis of retained versus staircase-excluded attributes was left outside the scope of the present thesis due to space limitations.

## Formatting and symbol conventions used throughout

Convention	Meaning
<b>Bold</b>	4 uses: structural labels (A1, H1, T1...), inline headings, table headers, results section anchors
<i>Italics</i>	3 uses: direct quotes, database field names, section names within cited works
§x.y and §§x.y-z	Single vs. range section references
V1, V2, V3	Verification test shorthand (§3.1.9)
Equation I, II...	Roman numeral equation sequence
Table x.y- Z / Figure x.y- Z	Chapter.subsection-RomanNumeral numbering
(Source: Own / Adapted / Extracted)	Three caption source types
(my translation)	Flags non-English source translations
RANK(), RANK.AVG()	Excel function names in all-caps
A1-A29, 0↑/1↓	Attribute codes and direction symbols

Table 1.6- I Formatting and typographic conventions. (Source: Own Presentation)

Table 1.6- I summarises the formatting and typographic conventions applied throughout the thesis.

## Units, scales, and coding notes used in the quantitative parts of the thesis

A consolidated reference table for recurring units, scales, and coding conventions used across the thesis is provided in Annex §8.5, Table 8.5- II.

## 2 Literature

Chapter 2 establishes the theoretical and methodological foundations required to understand and justify the evaluation pipeline implemented in Chapter 3. It first introduces learning analytics as a decision-support paradigm and clarifies the requirements of interpretability and traceability (§2.1), then positions Moodle as the socio-technical evidence environment from which trace data are extracted (§2.2). Building on this, the chapter explains how LMS traces can be modelled as student-level indicators under validity and GDPR/ethics constraints (§2.3), and situates these indicators within educational evaluation theory to motivate transparent, multi-dimensional measurement rather than single-score judgement (§2.4). It then reviews similarity analysis for open-ended educational text (§2.5) and formalises how heterogeneous evidence can be represented and aggregated through multi-attribute evaluation without arbitrary weighting (§2.6). The chapter further frames the work as algorithmic decision support that must remain auditable and governance-compliant (§2.7), addresses AI-detection as a bounded auxiliary signal under fairness and due-process risks (§2.8), and synthesises the resulting gap and thesis contributions (§2.9). Finally, §2.10 maps programme subjects to the concrete competencies operationalised in the thesis artefact, strengthening curriculum traceability and scope transparency.

### 2.1 Learning Analytics

This section introduces learning analytics (LA) as a decision-support approach for using learner data to improve learning and learning environments. It first outlines the conceptual foundations of LA, including the learning analytics cycle and the requirements of interpretability and traceability (§2.1.1). It then reviews two practical strands—early-warning interventions and didactic reflection—showing how LA outputs become educationally meaningful when they support action rather than isolated scoring (§2.1.2). This framing establishes why the thesis treats Moodle traces as bounded evidence and implements a reproducible workflow in §3. Clow (2013) notes that “*Learning analytics, the analysis and representation of data about learners in order to improve learning, is a new lens through which teachers can understand education.*” (p. 683).

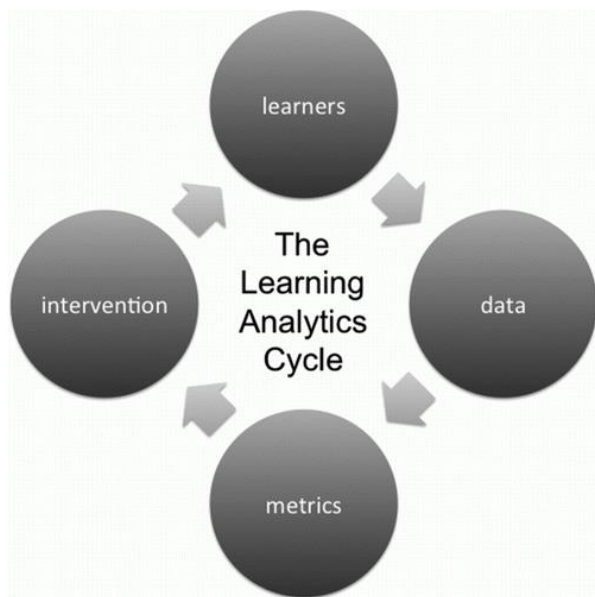


Figure 2.1- I The learning analytics cycle. (Source: Adapted from (Clow, 2013). February, 01, 2026, from <https://doi.org/10.1080/13562517.2013.827653>)

**Legend:** circular nodes = core elements of the cycle; arrows = iterative feedback flow.

As shown in Figure 2.1- I, the learning analytics cycle links learners, data, metrics, and intervention in a feedback loop. Accordingly, this thesis requires indicators to remain traceable to Moodle activity to support instructor review.

### 2.1.1 Conceptual foundations

Learning analytics (LA) is commonly defined as the measurement, collection, analysis, and reporting of data about learners and their contexts, with the purpose of understanding and improving learning and the environments in which it occurs. In this view, LA is not merely “data reporting”; it is a decision-support approach that becomes educationally meaningful only when analytics outputs can inform reflection or action. Clow (2013) emphasizes that *“the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs.”* (p. 685).

A central conceptual model in LA is the learning analytics cycle, which frames analytics as a feedback loop: learners generate data, data are processed into metrics, and metrics inform interventions that subsequently influence learner behaviour and outcomes. This cycle highlights two implications that are directly relevant for system design: (i) metrics should be interpretable enough to support decisions, and (ii) the analytic workflow must preserve traceability so that stakeholders can connect an output back to its data-generating activities.

Clow (2013) also notes that LA draws on a “toolbox” of methods (e.g., web log analysis, predictive modelling, natural language processing), which supports practical adoption but requires explicit assumptions and validation logic to remain interpretable and auditable.

### 2.1.2 Practical applications

In practice, a widely deployed LA use case is early warning / early intervention, where institutions combine performance and engagement signals to identify students who may be at risk and then communicate supportive actions. “Course Signals” illustrates this approach: the system integrates multiple data streams—including course performance, prior academic history, and effort measured through interaction with the learning system (LMS)—to generate risk statuses that are communicated to students via simple, interpretable signals (e.g., traffic-light indicators) and instructor messaging (Arnold & Pistilli, 2012). Importantly, this example shows LA as intervention-oriented: analytics is valuable because it enables feedback loops, not because it produces a score in isolation (Clow, 2013). Arnold & Pistilli (2012) note that “*Course Signals relies not only on grades to predict students’ performance, but also demographic characteristics, past academic history, and students’ effort as measured by interaction with Blackboard Vista, Purdue’s learning management system.*” (p. 267).

A second practical strand concerns didactic reflection using platform-generated traces, especially in blended learning contexts where meaningful learning activities occur across online and face-to-face components. Yang (2023) states that “*However, the LA results are rarely used for didactic reflection in foreign language teaching practice.*” (Abstract, para. 1). Using a blended language course as the context, the study highlights how platform traces (e.g., login frequency, resource access, forum/discussion contributions, response times) can be interpreted to characterize online learning behaviour and to inform instructional redesign. This perspective is relevant for theses that rely on LMS logs because it legitimizes trace data as evidence for analysing participation patterns—while also implying that interpretation must remain cautious: trace intensity does not automatically equal learning quality.

### **Implications for this thesis**

Based on the above, the thesis operationalises learning analytics as a reproducible, log-based feedback workflow by extracting Moodle discussion traces, transforming them into interpretable indicators, and presenting results as transparent outputs for educator review, implemented in §§3.1.1–3.1.2 and automated in §3.2.3.

The expected benefit is a consistent and auditable evidence pipeline aligned with the LA cycle (data → metrics → interpretation). The main limitation is constructing risk: activity traces and derived indicators may reflect participation volume more strongly than learning quality, which must be revisited critically in §4 (Discussion).

## 2.2 E-Learning Systems and Moodle

This section introduces e-learning systems—most commonly implemented as Learning Management Systems (LMS)—as socio-technical infrastructures that integrate content delivery, communication, assessment, and course administration within a single platform. Murtaza et al. (2022) note that “*The intrinsic features of adaptivity and adaptability require robust and continuous collection of data. In addition to assessments’ results from learners, other attributes such as usage patterns and learning trends may also be important for determining similarities among learners.*” (p. 81325). Accordingly, an LMS is treated here not only as a delivery channel but as an evidence environment that records observable learner–system interactions and enables analysis and decision support (Murtaza et al., 2022).

### 2.2.1 Moodle as an LMS

Moodle is a widely used open-source Learning Management System (LMS), typically deployed at institutional scale to support multiple courses, user roles, and learning workflows; “*Moodle is a Learning Management System (LMS) designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalised learning environments.*” (Moodle, 2025, *What is Moodle?*, para. 1). At the course level, Moodle provides structured spaces for learning materials, assignment submissions, quizzes, grading, and discussion forums (Moodle, 2025). As Bates (2015) notes, “*By 1995, the Web had enabled the development of the first learning management systems (LMSs), such as WebCT, which later became Blackboard*” (“6.2.4.3 Çevrimiçi öğrenme ortamları,” para. 1; my translation). Historically, this shift consolidated course communication and activity management into online environments and thereby enabled systematic collection of interaction traces as a normal feature of digital learning. From an analytics perspective, these tools matter because they generate event-level traces that can later be aggregated into student-level indicators (e.g., participation intensity, response timing, interaction structure, and text production). This aligns with the general learning analytics assumption that as learning shifts into digital environments, interactions can, in principle, be recorded and used to understand and optimize learning processes (Clow, 2013).

In the Kodolányi János University context, Moodle is described as the central e-learning platform, with course/user synchronization connected to the university registry and integrated delivery functions (e.g., online learning materials, tasks, exams, and webinar integration). Kodolányi János Egyetem (2025) states that “*Kodolányi University uses the Moodle LMS as its e-learning system, where each course has its own interface with additional content uploaded by instructors (e.g., notes, assignments)*” (“*Moodle LMS,*” para. 1; my translation). Such institutional integration implies that Moodle becomes a default infrastructure for learning activity, not merely an optional add-on (Kodolányi János Egyetem, 2025). This institutional role became especially salient during emergency remote teaching periods. As Virágh (2020) notes, for studio-based classes “*we cannot hold lessons via Skype*” (para. 10, my translation), which shows that online continuity was maintained under real practical constraints rather than frictionless platform substitution. Kodolányi János University (2023) notes that “*Following the online part of the program, in May we met with the students from Portugal, Poland, and Bulgaria face-to-face at our university and instantly had exciting conversations with them*” (“*Involvement of Hungarian KJU students in the Erasmus+ Blended Intensive Program,*” para. 2). KJU also engages in blended learning formats that combine online and face-to-face phases (e.g., Erasmus+ blended intensive programmes), reinforcing that digital platforms and trace-generating activities form a routine part of students’ learning experience (Kodolányi János University, 2023).

### 2.2.2 Platform analytics vs. research/assessment analytics

A key distinction for this thesis is the difference between (a) platform-native analytics designed for operational monitoring inside Moodle, and (b) external, auditable analytics performed by exporting data and applying explicit models outside the platform. Moodle provides a dedicated Analytics subsystem intended to move beyond descriptive reporting toward predictive and actionable insights; “*Moodle provides a variety of built-in reports based on log data, but they are primarily descriptive in nature—they tell participants what happened, but not why, and they don't predict outcomes or advise participants how to improve outcomes.*” (Moodle, 2024, *Analytics vs. reporting*, p. 1). According to Moodle’s documentation, the Analytics framework supports both machine-learning and static (rule-based) models, including built-in options such as identifying “students at risk of dropping out” and detecting “upcoming activities due,” and it can trigger notifications to support timely intervention. The framework is also presented as extensible, allowing customization and new indicators through an API (Moodle, 2024). At the same time, Moodle’s documentation notes that machine-learning analytics models must be

trained on site data before predictions are possible, and that model selection should align with institutional educational priorities (Moodle, 2024).

For a thesis that aims to justify evaluation decisions, these characteristics have methodological consequences: platform analytics may be useful for day-to-day support, but thesis-level evaluation typically requires (i) transparent feature definitions, (ii) reproducible pre-processing, and (iii) outputs that can be validated and critiqued independently of the LMS user interface or configuration. Therefore, a common research strategy is to treat the LMS as the data source while conducting the evaluation logic externally, where data transformations and model behaviour can be documented and tested.

### 2.2.3 Implications for educational measurement in Moodle-based courses

Using Moodle traces for educational evaluation is attractive because discussion forums and online tasks combine behavioural engagement signals (e.g., timing, frequency, reply structure) with artifact-based signals (e.g., the content of the text itself). Motz et al. (2019) note that “*it remains unclear whether LMS activity indeed reflects behavioural properties of student engagement, and it also remains unclear how to deal with variability in LMS usage across a diversity of courses.*” (p. 300). In practice, however, the presence of rich trace data does not automatically guarantee valid measurement: platform logs reflect what the system can record, and what the course design incentivizes students to do. For example, gamified or highly structured courses may increase measurable activity without necessarily increasing conceptual understanding, while minimal platform usage may reflect pedagogy rather than disengagement. Majorosi et al. (2016) state that “*Gamification should be regarded rather as an effective tool and not as entertainment*” (p. 54; my translation). This is consistent with KJU-oriented discussions emphasizing that digitalization and platform-based learning elements influence motivation and satisfaction, but they must be treated as professional pedagogical tools rather than “just for fun” engagement mechanics (Majorosi et al., 2016).

In addition, Moodle activity levels can be shaped by learners’ external digital conditions, not only by motivation or learning quality. Malatyinszki (2022) notes that, “*With the onset of the virus situation, families were forced into their homes, so family members had to share assets, both in space and time*” (p. 68); accordingly, participation in online work and education during the pandemic was shaped not only by personal engagement but also by the availability of technological resources, domestic space, shared time, and digital awareness. Therefore, low

trace intensity should not automatically be interpreted as low effort; it may also reflect socio-technical constraints that the LMS cannot observe directly.

In short, Moodle can support data-informed education, but turning Moodle traces into defensible evaluation requires careful design: (i) define indicators that correspond to educational constructs (e.g., diligence vs. understanding), (ii) document pre-processing so indicators are reproducible, and (iii) interpret outcomes as decision support rather than as automated “final judgement.”

### **Implications for this thesis**

Based on the above, the thesis operationalises the LMS “evidence environment” by extracting Moodle forum export data and transforming event-level records into a structured database, implemented in §3.1.1–§3.1.2 (raw Moodle dataset → cleaning → SQLite). The expected benefit is a transparent and reproducible basis for computing student-level indicators before applying multi-attribute evaluation and validation. The main limitation is that LMS traces are shaped by platform configuration and course design (what gets recorded, and what students are asked to do), which must be treated as an interpretive risk and revisited in §4.

## **2.3 Log-Based Student Modelling**

This section introduces log-based student modelling as the transformation of LMS interaction traces into interpretable indicators of learner behaviour. Motz et al. (2019) note that “*Learning management system (LMS) web logs provide granular, near-real-time records of student behaviour as learners interact with online course materials in digital learning environments.*” (p. 300). Because Moodle logs record time-stamped actions rather than learning itself, the central methodological task is to define construct-aligned variables (e.g., timeliness, participation regularity, interaction structure) while acknowledging that such variables remain proxies. The literature reviewed here motivates why pre-processing decisions (event selection, aggregation level, missingness rules) must be explicit to ensure reproducibility and defensible interpretation. The section concludes by positioning log-derived indicators as one evidence stream that later becomes formalised into an Object–Attribute representation. §§3.1.2–3.1.5 (Motz et al., 2019).

### **2.3.1 From LMS traces to student models**

Log-based student modelling refers to constructing interpretable indicators about learners from the digital traces they leave while interacting with a Learning Management System (LMS), e.g.,

Blackboard, Canvas, or Moodle (Motz et al., 2019). Motz et al. (2019) note that “*Learning management system (LMS) web logs provide granular, near-real-time records of student behaviour as learners interact with online course materials in digital learning environments*” (p. 2). As Kadoic & Oreski (2018) explain, “*A Moodle log consists of the time and date it was accessed, the Internet Protocol (IP) address from which it was accessed, the name of the student, each action completed (i.e., view, add, update, or delete), the activities performed in different modules (e.g., the forum, resources, or assignment sections), and additional information about the action*” (p. 731). Accordingly, typical traces include timestamps, tool-use events, forum actions (posting, replying, reading), and navigation or submission patterns. Agudo-Peregrina et al. (2014) state that “*The most basic unit of learning data in virtual learning environments for learning analytics is the interaction, but there is no consensus yet on which interactions are relevant for effective learning*” (p. 542). This directly supports the thesis’s position that LMS traces must be classified and interpreted before they can function as defensible indicators. Importantly, log events are not “learning” by themselves; they are observable proxies that can be aggregated into variables intended to approximate constructs such as participation, persistence, timeliness, and interaction structure (Motz et al., 2019).

A practical advantage of Moodle-based environments is that the LMS log infrastructure records fine-grained, time-stamped events that can be transformed into student-level variables. For example, Moodle logs can be used to derive behaviour indicators and examine their association with academic outcomes, although such relationships are shaped by course design and therefore require cautious interpretation (Kadoic & Oreski, 2018).

A recurring finding is that the meaning of trace activity is context dependent. Using log data from multiple courses, (Agudo-Peregrina et al., 2014) show that some interaction categories relate to academic performance in fully online settings, while the same categories may have weak or non-significant relations in VLE-supported face-to-face contexts. This implies that “activity” is not a universal construct: the same interaction count may indicate learning engagement in one setting, but routine compliance or administrative behaviour in another (Agudo-Peregrina et al., 2014).

Large-scale validation work reinforces this caution. Motz et al. (2019) examine LMS activity logs as measures of engagement and conclude that log features can be useful, but there is no one-size-fits-all mapping from “clicks” to “engagement,” because course design and tool-use norms shape what counts as behavioural evidence. Consequently, log-based student modelling

should justify indicator meaning within the specific course and task context rather than importing generic dashboards as universal measures (Motz et al., 2019).

### 2.3.2 Modelling choices: what counts as “evidence”?

Because log variables are proxies, modelling requires explicit decisions about (i) construct definitions, (ii) feature engineering, and (iii) aggregation level (event → session → student → cohort). As Park & Jo. (2016) explain, “*Through data pre-processing, general indicators reflecting login frequencies [...] and activity-based indicators presenting the activation [...] of diverse functions provided by Moodle were derived*” (p. 2). This demonstrates a pragmatic approach in which general indicators (e.g., logins) and activity-based indicators (patterns across LMS functions) are constructed first and then interpreted through activity theory to explain how institutional rules, tools, and divisions of labour shape observable behaviour. Their work illustrates a key methodological point for this thesis: log variables only become evidence after transparent pre-processing and a defensible interpretation frame are specified.

Similarly, Agudo-Peregrina et al. (2014) emphasises that even the basic unit of log data—an interaction—requires classification, because platforms differ in tools and event semantics. Their system-independent categorisation approach still yields context-driven differences across modalities, supporting a conservative modelling stance: the model should document what is counted, how it is aggregated, and why it is theoretically relevant, rather than treating raw counts as self-explanatory (Agudo-Peregrina et al., 2014).

In Moodle-specific work, the feasibility of modelling student behaviour directly from Moodle logs is demonstrated through extraction and aggregation strategies that connect trace variables to course outcomes, while also acknowledging that results may not generalise across different course structures (Kadoic & Oreski, 2018). This further motivates the thesis design choice to define indicators explicitly and interpret them within the local instructional context, rather than treating log-derived metrics as universally valid.

### 2.3.3 Validity and limitations: logs as imperfect proxies

A central methodological risk is constructing validity. High activity may reflect engagement, but it can also reflect confusion, inefficient navigation, or platform-driven compliance. In a foundational overview, Clow (2013) warns that learning analytics can become “data-led” in an uncritical way—treating available traces as automatically meaningful—unless indicator selection is grounded in theory and linked to plausible mechanisms of learning. For log-based

student modelling, this implies that each indicator should (a) state the construct being approximated, (b) justify directionality (what is “more ideal”), and (c) avoid overclaiming that traces directly measure learning. Motz et al. (2019) note that “*there is not a single generic relationship between activity and engagement, and what constitutes the behavioural components of “engagement” will be contingent on course structure.*” (p. 2).

Kadoic & Oreski (2018) explicitly caution that “*However, this cannot be generalized because the research was conducted in only one course*” (p. 734). This caution is consistent with Moodle log studies that relate trace variables to student success: log–outcome associations can be informative but remain dependent on context and course implementation details, which limits generalization and increases the need for careful interpretation (Kadoic & Oreski, 2018).

A second limitation concerns process versus outcome. Traditional assessment often focuses on final products, whereas logs can capture intermediate learning processes. As Buckley et al. (2006) explain, “*We use students’ log files as performance assessments of model-based inquiry on inquiry tasks called ‘hot spots’ and track changes in inquiry skills over time both within and across domains*” (p. 1). This illustrates how log files can support in-situ performance assessment of inquiry processes by tracking learner interactions within modelling tools. At the same time, their work highlights a practical constraint: meaningful process evidence requires careful task design and instrumentation, not merely exporting a log table.

#### 2.3.4 Privacy, ethics, and GDPR constraints on log-based modelling

Because LMS logs are typically personal data (or become personal data once linkable), log-based modelling must operate under explicit governance constraints. As Teresa & McGrath (2021) note, “*The top three ethical areas most often addressed in the selected literature are transparency, privacy, and informed consent*” (p. 123). In their systematic review of empirical ethics research in learning analytics, these areas recur alongside the view that analytics-driven interventions require additional scrutiny. National Forum for the Enhancement of Teaching and Learning in Higher Education (2018) state that “*Processing means working with data in any way. It includes collecting, recording and storing data*” (p. 1).

From a student perspective, privacy concerns shape trust and willingness to participate. Mutimukwe et al. (2021) frame these concerns in learning analytics through perceived vulnerability and perceived control, and warn that “*The consequences of such concerns could range from individuals declining or refusing to disclose personal information, and/or mistrust*

*in online services*” (p. 2); this indicates that governance is not merely a matter of legal compliance, but also a condition for legitimacy.

For GDPR-aligned practice, National Forum for the Enhancement of Teaching and Learning in Higher Education (2018) highlights operational requirements such as a lawful basis, purpose limitation, transparency, and data minimisation. For this thesis, this implies: (i) using only variables necessary for the defined evaluation purpose, (ii) anonymising or pseudonymising early in the pipeline, and (iii) documenting retention and access boundaries (National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018).

### **Implications for this thesis**

Based on the above, the thesis operationalises log-based student modelling by extracting Moodle discussion/activity traces at the student level, transforming them into a defined attribute catalogue (A1–A29) with explicit meaning and directionality, and using these indicators as the evidence base for the evaluation pipeline in §3.1.1–§3.1.2 and the automated workflow in §3.2.3. Moodle log studies demonstrate that such traces can be aggregated into indicators that relate to student success, but they also highlight context dependence and limited generalizability (Kadoic & Oreski, 2018), therefore, the thesis treats indicators as proxies and reinforces interpretation through robustness checks in §3.1.9 and governance safeguards in §3.2.9, with limitations revisited in §4.

## **2.4 Assessment and Evaluation Theory in Education**

This section frames student evaluation as both a measurement problem and a judgement problem, where “objectivity” depends on making value choices explicit rather than claiming to eliminate them; “*We recognize there is no single determination of worth for any educational endeavour. Worth is complex and personal.*” (Page & Stake, 1979, p. 47). It establishes why performance in LMS-supported contexts is typically multidimensional and why single-score evaluation is fragile, especially when trace data can amplify superficial activity. The review emphasises traceability and accountability: an evaluation is academically defensible only if its criteria, directionality, and aggregation logic can be inspected and critiqued (Page & Stake, 1979). These principles directly motivate the thesis approach in §3, where evaluation is implemented as a transparent multi-attribute workflow with validation steps, rather than as an opaque scoring system.

### 2.4.1 Why evaluation is inherently value-laden

Educational evaluation is not only measurement; it is a judgement of worth under an explicit or implicit value system. (Page & Stake, 1979) frame this as a structural tension: objective measurement supports comparability and disciplined reasoning, yet subjective judgement remains unavoidable because evaluators must decide which properties matter and what standards apply. The implication for this thesis is direct: an automated workflow does not eliminate subjectivity; it relocates it into (i) attribute selection, (ii) directionality settings, (iii) aggregation rules, and (iv) interpretation of results. Therefore, evaluation quality depends on making those value choices explicit and auditable rather than claiming “pure objectivity”. Page & Stake (1979) note that “*The act of ‘evaluation’ would presuppose some prior system of values; yet any technology of values, or even their place in rational judgment, is given scant attention in the field*” (p. 46).

### 2.4.2 Limits of pre-digital evaluation under LMS conditions

Traditional evaluation approaches (tests, instructor impressions, peer evaluation, end-of-course surveys) often provide limited temporal granularity and weak traceability. In LMS-supported learning, these limitations become more visible because the platform provides fine-grained behavioural records that may diverge from perceptions. Cantabella et al. (2018) state that “*In the rest of faculties, we detect that the real number of audio-visual resources uploaded to the LMS is considerably lower than the number stated by lecturers [...]*” (p. 918). Cantabella et al. (2018) explicitly compare lecturers’ subjective perceptions of LMS activity with objective log-derived indicators and report systematic discrepancies between the two perspectives. Methodologically, this supports analytics-informed evaluation as a corrective lens when decisions rely heavily on impressions rather than recorded evidence.

In the empirical setting of this thesis, §3 reports a parallel logic on the student side: objective log-based rankings are compared against peer-based subjective evaluations, enabling an explicit analysis of divergence and agreement (see §3.1.8.4). This comparison motivates the thesis’s emphasis on transparent indicators and reproducible evaluation rather than purely impression-based judgement.

### 2.4.3 Why multi-attribute and validated evaluation matters

Most educational performance constructs are multidimensional: a student may be active but off-topic, timely but shallow, or original but inconsistent. This makes single-metric evaluation

fragile and increases fairness risks “*It means: the importance of the attributes is a relative phenomenon. The importance of a particular attribute is depending on the set of the given attributes*” (Bán et al., 2024, Introduction, para. 1). A defensible response is structured evaluation: separating dimensions (attributes), defining directionality, and combining evidence through transparent rules—paired with validation procedures that test internal consistency and sensitivity to distortions (Bán et al., 2024).

This logic aligns with the thesis methodology in §3, where the evaluation workflow is not treated as a black box: attributes are explicitly defined, results are validated using internal consistency checks, and robustness is examined using deliberately inconsistent/extreme inputs. In evaluation-theoretic terms, the goal is decision support with auditability—structuring judgement—rather than claiming to replace judgement.

#### 2.4.4 Responsibility and governance in automated evaluation

When evaluation outputs influence grades, interventions, or student reputation, responsibility increases. (Clow, 2013) stresses that learning analytics should produce actionable intelligence, but also implies that analytics can shape behaviour through the metrics it makes visible, creating unintended consequences if indicators are poorly chosen or interpreted. Therefore, automated evaluation requires both methodological transparency and governance: clear purpose statements, minimisation of collected data, and justification of why specific indicators are necessary. (Clow, 2013) note that “*If an educational system is designed to optimise metrics that do not encompass learning, it is likely that learning will be optimised away*” (p. 691).

#### **Implications for this thesis**

Based on the above, the thesis operationalises educational evaluation as a multi-attribute, auditable decision-support process that contrasts subjective peer evaluation with log-based evidence. This is implemented in §3.1.5–§3.1.8 (OAM construction, COCO processing, and result consolidation) and explicitly compared in §3.1.8.4, while verification/robustness procedures are documented in §3.1.7 and §3.1.9. The expected benefit is improved traceability and reduced reliance on impression-based judgement; the main limitation is that attribute choice and directionality embed normative assumptions and may raise fairness/privacy concerns, revisited in §4 and governed in implementation in §3.2.9.

## 2.5 Similarity Analysis for Educational Text

This section introduces text similarity as a practical indicator for analysing descriptive student responses in Moodle forums, where multiple valid phrasings may express the same intended concept. A core motivation for similarity-based support in descriptive-answer contexts is automation-oriented evaluation; “*This paper demonstrates a computational intelligence-based method for automatic evaluation of descriptive answers.*” (Bahel & Thomas, 2021, p. 2). It distinguishes lexical overlap from semantic similarity and motivates why embedding-based methods can better capture topical alignment under paraphrase. At the same time, the section sets strict interpretation boundaries: similarity is treated as probabilistic evidence of alignment, not proof of correctness or learning quality. This positioning is essential for reasoning because it clarifies how the thesis later uses similarity as one attribute among many in §3.1.3, rather than a standalone grading mechanism (Bahel & Thomas, 2021).

### 2.5.1 Why “similarity” is a usable signal in educational text

In LMS-supported teaching, many learning activities require descriptive student responses (e.g., forum replies, explanations, reflections). In such settings, evaluation often shifts from selecting a single correct option to assessing whether a response meaningfully addresses the intended topic and expected content Bahel & Thomas (2021) note that “*Sometimes evaluators are often caught being biased towards some students while evaluating the answers.*” (p. 2). This motivates the use of text similarity as a practical operationalisation: estimating how close a student’s response is to a reference text (e.g., an instructor prompt, exemplar answer, or expected concept set) in order to support scalable evaluation of descriptive answers. (Bahel & Thomas, 2021).

### 2.5.2 Lexical similarity: transparent but brittle

A classical family of approaches represents texts as bags of words (or n-grams) and compares them using measures such as cosine similarity, Jaccard similarity, or n-gram overlap. As Rahman & Siddiqui (2018) explain, “*The cosine similarity between two documents generates a metric which tells how two documents are related by looking at the angle as a substitute of magnitude*” (p. 2). Within this classical approach, texts are therefore represented as bags of words or n-grams and compared through measures such as cosine similarity, Jaccard similarity, or n-gram overlap.

They also include set-based comparisons (e.g., Jaccard) and structural signals (e.g., bigram similarity), illustrating that similarity can be operationalized along different “angles” of resemblance.

However, lexical methods are sensitive to paraphrase: two responses can be semantically equivalent yet share few words, especially in authentic student writing. This limits their use when the educational goal allows multiple valid phrasings (common in discussion forums and open-ended prompts). This motivates moving beyond lexical overlap toward semantic similarity.

### 2.5.3 Semantic similarity: meaning-oriented comparison for open-ended responses

Semantic similarity approaches aim to capture meaning even when wording differs. A common operationalisation is to represent each response as a dense sentence embedding and then compute similarity using cosine distance in vector space. Reimers & Gurevych (2019) state that “we present *Sentence-BERT (SBERT)*, a modification of the pretrained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity.” (p. 3982). In automatic descriptive-answer evaluation, (Rahman & Siddiqui, 2018) summarize multiple lines of work showing that the most effective similarity technique depends on the question type, and that methods such as Doc2Vec-based cosine similarity or LSA-based similarity can be useful in different contexts. This is a key lesson for this thesis: similarity is not a universal “grading rule”; it is a context-dependent indicator whose validity depends on task design and interpretation boundaries.

(Bahel & Thomas, 2021) propose an automated descriptive-answer evaluation architecture using text analysis and a similarity model (MaLSTM), motivated partly by workload and potential bias in manual scoring. They also frame evaluation as multi-feature rather than purely one-dimensional, using multiple components that contribute to scoring. This supports a cautious thesis stance: semantic similarity can be valuable evidence for topical alignment, but robust educational evaluation typically requires additional dimensions (e.g., timeliness, participation structure, correctness checks where possible).

### 2.5.4 Validity limits

A critical boundary condition is that high similarity does not necessarily imply correctness, and low similarity does not necessarily imply poor understanding. Similarity can be inflated by

superficial paraphrase, template-like answers, or overly generic responses; conversely, a correct response may be expressed with uncommon wording, added context, or creative but valid reasoning. (Rahman & Siddiqui, 2018) explicitly note that different similarity techniques vary in effectiveness across question types, reinforcing that “similarity” must be interpreted as probabilistic evidence, not proof. Their prior work explicitly warns that similarity effectiveness is context dependent: *“the most effective similarity measure technique depends on the type of question. Based on the question, the effectiveness of similarity measurement techniques is varied.”* (Rahman & Siddiqui, 2018, p. 37)

Therefore, thesis-quality use of similarity analysis requires (i) transparent definition of the construct (e.g., topical alignment rather than correctness), (ii) careful pairing of student text to the appropriate reference text, and (iii) embedding similarity into a broader, auditable evaluation design rather than treating it as a standalone grading mechanism.

### **Implications for this thesis**

Based on the above, the thesis operationalises topical alignment of descriptive forum responses by computing embedding-based semantic similarity score between each student reply and its linked instructor prompt, using cosine similarity as the comparison rule, implemented in §3.1.3.

The expected benefit is a paraphrase-robust indicator that can scale across many messages and support objective, repeatable analysis in forum-based tasks; the main limitation is validity risk (semantic similarity captures closeness in meaning, not correctness or quality), revisited in §4 and mitigated by treating similarity as one attribute within a multi-attribute evaluation pipeline (see §3.1.5) rather than as a single decisive metric.

## **2.6 Multi-Attribute Evaluation**

This section motivates multi-attribute evaluation as the appropriate response to the inherently multi-dimensional nature of LMS-based performance evidence. It introduces the idea that students (objects) must be compared through a structured set of indicators (attributes), requiring explicit directionality rules (benefit vs. cost) and a defensible aggregation logic. The review also highlights why naïve weighted scoring can embed hidden biases and why anti-discriminative aggregation approaches are attractive when the goal is transparent comparison rather than prediction of an external label (Bán et al., 2024). These arguments prepare the methodological core of §3.1, where evidence is organised into an Object–Attribute Matrix and processed using a reproducible, validation-aware pipeline. *“This study proposes a*

*comprehensive model that leverages 29 distinct attributes extracted from Moodle (e-learning platform) log data to provide a multifaceted/objective evaluation of student performance.”* (Turtogtokh et al., 2025, *Abstract*, para. 1)

### 2.6.1 Why multi-attribute evaluation is necessary in LMS-based assessment

In LMS-supported courses, “performance” is rarely one-dimensional. Students may differ simultaneously in participation intensity, timeliness, interaction structure, and the topical relevance of written contributions, meaning that single-indicator evaluation risks oversimplifying complex learning behaviour (Clow, 2013). In a Moodle discussion context, multi-dimensional profiling is practically unavoidable because the available evidence naturally produces multiple indicators (e.g., post/reply activity, time-based patterns, and text-related measures), which motivates multi-attribute evaluation rather than a single composite score (Turtogtokh et al., 2025)<sup>1</sup>. One recent Moodle-based multi-attribute framework states that *“This research seeks to address these challenges by utilizing Moodle log data, advanced natural language processing (NLP) models, and an anti-discriminative engine to analyze student performance with 29 different attributes that reveal multiple dimensions of student activity.”* (Turtogtokh et al., 2025, *Introduction*, para. 3).

### 2.6.2 The Object–Attribute Matrix as the formal representation

Pitlik (2014) defines the OAM as *“object–attribute matrix, in other words a learning sample, in which the rows are the objects (cases) and the columns are their characteristics (attributes, aspects, variables, etc.)”* (*“A hasonlóság-élemezés alapfogalmai,”* para. 29, my translation). A standard representation for multi-attribute evaluation is to define objects (here: students) and attributes (measurable indicators describing each student), then represent the dataset as an Object–Attribute Matrix (OAM) (Pitlik, 2014). In this representation, the methodological focus is not prediction of an external outcome variable, but the derivation of a coherent comparative evaluation from heterogeneous descriptors that jointly characterize each object (Pitlik, 2014).

In evaluation-model terms, this representation is commonly stated as follows: *“If an OAM (object-attribute-matrix) is given, where the objects are e.g., human beings and the attributes*

---

<sup>1</sup> **Self-citation note.** Turtogtokh et al. (2025) is a peer-reviewed conference proceedings paper co-authored by the thesis author. It is cited only to trace the origin of selected implementation elements reused here (e.g., indicator definitions and validation routines). This bachelor thesis extends that work with a complete end-to-end pipeline, an automated prototype, governance measures, and additional testing evidence (Chapter 3).

*are their properties, then it can become necessary to derive an aggregated index value.*” (Bán et al., 2024, para. 12).

This OAM-based framing is also consistent with evaluation-model literature that treats attribute sets as context-dependent and interpretable only relative to the defined descriptor system; in such settings, analysis of attribute relevance naturally operates on the OAM itself (Bán et al., 2024).

### 2.6.3 Directionality and the unavoidable role of value choices

Any multi-attribute evaluation must specify directionality, i.e., whether higher values represent “more ideal” performance (benefit-type) or “worse” performance (cost-type). Page & Stake (1979) note that *“The act of “evaluation” would presuppose some prior system of values; yet any technology of values, or even their place in rational judgment, is given scant attention in the field.”* (p. 46). This is not a purely technical step: directionality encodes educational values, because it formalizes what the evaluator treats as desirable or undesirable behaviour (Page & Stake, 1979).

In Moodle-log-based performance profiling, directionality decisions become explicit when indicators such as timeliness, responsiveness, or alignment measures must be interpreted consistently before aggregation (Turtogtokh et al., 2025). Therefore, a thesis-appropriate stance is that multi-attribute evaluation can reduce arbitrariness by forcing explicit rules, but it cannot eliminate normative judgement entirely, because attribute selection and directionality remain value-laden design choices (Page & Stake, 1979).

### 2.6.4 Aggregation without arbitrary weights: anti-discriminative optimisation

A common approach to multi-attribute aggregation is weighted scoring, where each attribute is assigned a manual or statistically estimated weight. However, attribute “importance” is not stable across contexts because the effect of any one attribute depends on the presence and configuration of the other attributes (e.g., overlap, compensation, or contradiction). (Bán et al., 2024) explicitly argue that attribute importance is relative to the attribute set, and that simple correlation-based intuitions can be misleading—particularly when the objective is to derive an aggregated evaluation index from heterogeneous descriptors rather than to predict an external outcome variable. In the COCO Y0 framing, the baseline condition is described explicitly: *“The anti-discriminative optimization in our used example does not have two real attributes, because the hypothetical Y0-variable has only a constant value.”* (Bán et al., 2024, para. 11).

To address this, anti-discriminative optimisation methods aim to compute an aggregated index and ranking without imposing an arbitrary “pre-set” weight structure. In the COCO-family tradition, evaluation is framed as component-based object comparison for objectivity, where aggregation is derived from the multi-attribute evidence itself rather than from a manually chosen weighting scheme (Pitlik, 2014).

In practical terms, this approach seeks an evaluation result that is consistent with the descriptor system while remaining transparent about the dependence of outcomes on the selected attributes.

### 2.6.5 Attribute exclusion as a diagnostic tool for discrimination and importance

A known challenge in anti-discriminative multi-attribute evaluation is the low-discrimination (“sameness”) problem: the model may legitimately assign identical or near-identical scores to many objects when the attribute system does not separate them strongly, or when several attributes contribute in a highly uniform manner. In such cases, attribute exclusion is not an ad-hoc manipulation; it is a diagnostic step that tests whether the evaluation outcome is driven by a small subset of attributes and whether meaningful distinctions are hidden by overly dominant descriptors (Bán et al., 2024). Bán et al. (2024) also note that “*The question is: how may we interpret, if always one single attribute is excluded from the entire OAM and the aggregated index values will be calculated on these partial OAMs?*” (Para. 11)

(Bán et al., 2024) operationalize attribute importance analysis through systematic exclusion: the evaluation is repeatedly re-run on reduced OAMs, and the changes in the aggregated output are analysed to infer relevance or dominance under the current descriptor configuration.

This logic aligns with a thesis-quality interpretability requirement: the evaluation should not only output a ranking, but also provide evidence about why objects differ and which attributes most strongly contribute to separability.

### 2.6.6 Verification and robustness as part of multi-attribute evaluation quality

Because multi-attribute evaluation outputs may influence educational judgement, a thesis-level methodology should include verification beyond “the model produced a ranking.” In Moodle-log based performance evaluation, reliability can be examined through inverse-run symmetry validation, where the input ranking is reversed, the evaluation is rerun, and consistency is checked using a delta-based validity condition (e.g., sign-consistency constraints). (Turtohtokh et al., 2025) state that “*If the product of the two delta values is zero or less, the model’s results*

*are confirmed to be valid and reliable. If the product is greater than zero, it indicates potential inconsistencies.” (Validation of Results, para. 7).*

In addition, robustness can be examined through sensitivity testing, where deliberately inconsistent or extreme input values are injected to observe whether the evaluation outcome reacts in the expected direction (i.e., whether the method is fragile to distortions and thus highlights data-quality risks). Such tests support a defensible methodological claim: the evaluation pipeline is not only computed, but also audited for internal consistency and sensitivity to data distortion (Turtoqtokh et al., 2025).

### 2.6.7 Positioning COCO Y0 among MCDA methods

Figueira et al. (2005) note that MCDA “*helps making decisions mainly in terms of choosing, ranking or sorting the actions*” (p. xiv). Accordingly, multi-criteria decision analysis comprises established method families for aggregating heterogeneous criteria into comparative decisions, including value/utility-based models, distance-to-ideal approaches, and outranking methods. Within this broader framework, COCO Y0 is used in this thesis as an optimisation-based, context-free similarity-analysis approach that derives attribute-level “staircase” functions through linear programming instead of applying pre-set constant weights. This orientation aligns with Pitlik (2004) view that “*Complete objectivity is characterized by the fact that an analysis can be carried out without any individual value determinations*” (p. 199, my translation).

More specifically, Pitlik & Varga (2015) explain that “*Similarity analysis delivers as solution optimized staircase-function, where a set of descending rank/stair-values will be calculated for each known attribute-level based on linear programming*” (“*Similarity Analysis,*” “*Theory,*” para. 1). In parallel, Bánkuti & Pitlik (2010) describe COCO as “*Component-based Object Comparison for Objectivity*” and as “*a recently developed Hungarian, Linear Programming based context-free similarity analysis method*” (p. 1). Together, these characteristics make COCO Y0 suitable for transparent benchmarking from an Object–Attribute Matrix without requiring an externally validated outcome variable or a stakeholder-elicited weight vector.

A widely used alternative is TOPSIS. As Behzadian et al. (2012) note, “*the standard TOPSIS method attempts to choose alternatives that simultaneously have the shortest distance from the positive ideal solution and the farthest distance from the negative-ideal solution.*” (p. 2). In standard TOPSIS, the decision matrix is normalised and then weighted before distance computations; thus, the resulting ranking depends on the weight specification and

normalisation choices. In the present thesis context—where stakeholder-validated weights are unavailable and the design goal is to minimise ad hoc weighting—TOPSIS serves as a useful baseline comparator but is less consistent with the thesis’s weight-avoidance objective.

A second major alternative is PROMETHEE, which Brans & Vincke (1985) describe as “*a new family of methods for multicriteria decision-making*” (p. 647). As an outranking approach, it compares alternatives pairwise using criterion-level preference functions and weights to generate a ranking. PROMETHEE is particularly suitable when decision-makers can meaningfully specify such preference functions and weights; however, this also means that subjectivity is relocated into parameter elicitation. Accordingly, TOPSIS and PROMETHEE are used here as comparators because they represent two widely taught and widely applied MCDA families—distance-to-ideal and outranking—thereby providing clear baselines for contrasting COCO Y0’s optimisation-based, weight-avoiding stance (Figueira et al., 2005).

### **Implications for this thesis**

Based on the above, the thesis operationalises multi-attribute evaluation by (i) representing student evidence as an Object–Attribute Matrix (OAM), (ii) defining explicit attribute directionality, and (iii) aggregating the resulting evidence via the COCO Y0 approach to reduce reliance on manually assigned weights, implemented in §§3.1.5–3.1.6. COCO Y0 is selected because it supports an optimisation-based, weight-avoiding aggregation stance consistent with transparent benchmarking from heterogeneous indicators (Bánkuti & Pitlik, 2010). TOPSIS and PROMETHEE are reviewed as established MCDA baselines (distance-to-ideal and outranking), but they typically require explicit weight and/or preference-function elicitation; therefore, they are not adopted as the primary aggregation logic in this thesis demo setting (Behzadian et al., 2012; Brans & Vincke, 1985). The expected benefit is an auditable aggregation of heterogeneous indicators that supports decision support rather than automated judgement. The main limitation is that attribute selection and directionality still embed normative assumptions, and low discrimination “sameness” may occur in some configurations, motivating attribute-exclusion diagnostics and repeated runs (revisited in §4).

## **2.7 Automated/Algorithmic Decision Support in Education**

This section positions algorithmic decision support as the production of interpretable, actionable signals that assist—rather than replace—educational judgement. It reviews why dashboards, early-warning signals, and analytic reports are valuable only when their indicators

are transparent, connected to plausible interventions, and accompanied by uncertainty awareness; “Techniques like dashboards and/or OLAP services could be used by each person (for benchmarking) and especially for further modelling.” (Pitlik et al., 2017, p. 1). The section also introduces quality assurance expectations for analytics pipelines (verification, robustness, traceability), which become particularly important when outputs may influence assessment decisions (Pitlik et al., 2017). These decision-support principles directly inform the thesis implementation strategy in §3, where intermediate artifacts and validation results are surfaced to enable auditability.

### 2.7.1 What “decision support” means in learning analytics and EDM

In educational contexts, automated or algorithmic decision support refers to computational methods that transform educational data into actionable information for human stakeholders, including teachers, students, and programme managers. Clow (2013) captures this process briefly: “The cycle starts with learners, who generate data, which is processed into metrics, which are used to inform interventions, which in turn affect learners” (p. 685). Dormezil et al. (2019) conclude that “Following keyword analysis, we conclude it is more accurate to describe what appears to be two domains (i.e., Educational Data Mining and Learning Analytics) as one domain (i.e., Learning Analytics) with one prominent subset (i.e., Educational Data Mining)” (p. 17). In this thesis context, that framing supports the move from raw activity logs toward indicators, predictions, recommendations, and other structured outputs that can inform interventions or reflective redesign. Accordingly, decision support in education is not limited to prediction, but also encompasses descriptive dashboards, benchmarking reports, and evaluation pipelines that render otherwise hidden patterns visible and auditable (Pitlik et al., 2017).

A key quality criterion is whether the analytics closes the loop—i.e., whether the outputs are interpretable enough to guide a concrete action (feedback, support, redesign), rather than merely producing metrics (Clow, 2013). This matters for usefulness: if indicators cannot be connected to a plausible educational response, they risk becoming “data for data’s sake,” with unclear pedagogical value (Clow, 2013).

## 2.7.2 Typical decision-support functions: dashboards, early warning, and recommendations

A common form of algorithmic decision support is early-warning signalling: using behavioural traces to flag students who may be at risk and to enable timely support. “Course Signals was developed to allow instructors the opportunity to employ the power of learner analytics to provide real-time feedback to a student.” (Arnold & Pistilli, 2012, p. 2). A well-known example is Course Signals, which operationalised learning analytics as a student-success intervention mechanism, framing analytics explicitly as a support tool rather than as an assessment replacement (Arnold & Pistilli, 2012). This illustrates a general pattern: decision support becomes meaningful when (i) a signal is tied to a clear intervention logic and (ii) uncertainty is communicated responsibly.

Decision support is also increasingly framed from the learner perspective. As Schumacher & Ifenthaler (2018) report, “*The findings show that students expect learning analytics features to support their planning and organization of learning processes, provide self-assessments, deliver adaptive recommendations, and produce personalized analyses of their learning activities*” (Abstract, para. 1). These expectations position analytics not merely as activity-visualisation tools, but as mechanisms for planning, self-assessment, and personalised support aligned with self-regulated learning. However, this also creates an important risk, since (Schumacher & Ifenthaler, 2018) note that students “*might fear losing autonomy in managing their learning activities, which is a key component for motivation*” (Introduction, para. 3). Therefore, decision support must be designed as assistive and transparent, not as hidden automation.

## 2.7.3 Decision support embedded in LMS ecosystems

Modern LMS platforms increasingly include built-in analytics features, meaning that algorithmic decision support is no longer an optional add-on but part of the learning infrastructure. Moodle’s analytics framework explicitly combines model-based approaches (including machine-learning and rule-based models) to generate indicators and predictions intended to support educational decisions (Moodle, 2024). Ouadoud et al. (2021) note that “*There are a large number of distance learning platforms on the international market, around more than 600 including around forty under free licenses.*” (p. 54). This is important for thesis positioning: it shows that the “baseline expectation” in current e-learning practice is not just data storage, but platform-integrated analytic support.

At the same time, platform variety and architecture shape what evidence is available and how reliably it can be interpreted. Surveys of e-learning platforms show that LMS features differ substantially in data granularity, logging semantics, and tool availability, directly constraining the design space of decision support (Ouadoud et al., 2021). Adoption research further suggests that organisational and cultural conditions affect whether users actually trust and use analytic support, even when it is technically available; indeed, Katoua et al. (2016) conclude that “*a successful e-learning system should consider the personal, social, cultural, technological, organizational and environmental factors*” (p. 754). This highlights a practical usefulness constraint: decision support that is not accepted or not understood cannot improve outcomes.

#### 2.7.4 Quality assurance: from “automated output” to defensible support

A bachelor-thesis-level analytics pipeline should treat quality assurance as part of the decision-support concept, not as an afterthought. In log-based educational management research, dashboards and OLAP-style reporting are understood as tools for benchmarking and for preparing or validating subjective decisions rather than replacing human judgement. As Pitlik et al. (2017) note, “*Model-based support processes can validate and/or prepare subjective/intuitive decisions*” (p. 1). Similarly, Balogh et al. (2018) describe this cybernetic framing as “*the complex handling of a real (intuitive/subjective) situation in frame of a cybernetic process chain (before, during and/or after the decision making as such)*” (Abstract, para. 1), suggesting that analytics can support decision making before, during, and after an intuitive decision by making data assets interpretable and by enabling systematic checks. Methodologically, this implies that decision support must include explicit verification steps, such as consistency checks, sensitivity tests, and diagnostic reruns; otherwise, outputs remain black-box suggestions with weak accountability. This position is reinforced by Pitlik (2009) argument that defensible interpretation requires formalizable logic and repeatable explanatory procedures rather than unchecked associative intuition. As Pitlik (2009) puts it, “*It is not enough to create something once, it shall be automated! And it is not enough to produce associations, they have to be proven!*” (“*Thought experiments I: Objectivity, sustainability, equilibrium,*” para. 3). In this view, the central danger lies not in computation itself but in unverified inference, which is why quality assurance must include explicit consistency checks and counter-verification steps whenever outputs are used to support decisions.

This thesis’s broader evaluation direction (multi-attribute, direction-aware aggregation and validation) fits that expectation: the decision-support value is strongest when the pipeline

remains auditable and when the model's fragility under distortion is tested (Clow, 2013; Pitlik et al., 2017). Here, the emphasis is not only on producing an index/ranking, but on producing one that can be defended under scrutiny.

### 2.7.5 Responsibility, ethics, and GDPR as design constraints

Automated decision support in education operates under heightened responsibility because outputs may influence learning opportunities, support allocation, or reputational outcomes; *"The top three ethical areas most often addressed in the selected works are transparency, privacy, and informed consent."* (Teresa & McGrath, 2021, p. 124). From the student perspective, privacy concerns are linked to perceived vulnerability and perceived ability to control disclosure, which affects trust and willingness to participate in analytics-enabled learning environments (Mutimukwe et al., 2021).

A widely cited practitioner-policy framework that operationalises these ethical requirements is Sclater's Code of Practice for Learning Analytics. As Sclater (2015) states, *"Institutions must decide who has overall responsibility for the legal, ethical and effective use of learning analytics"* (*"Responsibility,"* para. 1). This is methodologically relevant for thesis work because it makes "defensibility" concrete: an analytics output is not justified solely by being computed; it must be explainable, valid for its intended purpose, and embedded in an appropriate intervention and accountability process.

Under GDPR-aligned practice, lawful basis, purpose limitation, transparency, and data minimization are recurring operational requirements for learner data processing (National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018). As Ismayilzada et al. (2025) note, *"The introduction of GDPR has forced EDM practitioners to re-assess how student data is collected, stored, and analysed"* (p. 608). In EDM contexts, GDPR therefore creates both constraints and opportunities: it motivates stronger security and governance while encouraging privacy-conscious analytics practices. Together with Sclater's code, this implies that governance is not a "legal appendix" but a design boundary condition on (a) what indicators may be collected and retained, (b) who can access analytics outputs, (c) how interventions are documented and audited, and (d) how uncertainty and limitations are communicated to learners and staff (National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018; Sclater, 2015).

### 2.7.6 Decision support vs automated judgement: why scope boundaries matter

A recurring problem in educational technology is scope creep: tools built for support can become used for high-stakes judgement. The literature on embedded computer-based assessment shows the value of log data for capturing process evidence at scale, but also illustrates that meaningful inference requires careful design choices and infrastructure—not merely exporting logs (Buckley et al., 2006). This distinction matters because decision support outputs often look “objective,” which can encourage over-trust. Buckley et al. (2006) state that “*The xml log files have been validated at each stage of processing and data reduction.*” (p. 4)

A contemporary example is AI content detection used for academic integrity. As Elkhatat et al. (2023) report, “*when applied to human-written control responses, the tools exhibited inconsistencies, producing false positives and uncertain classifications*” (p. 1). This raises fairness risks if such tools are used as punitive evidence rather than contextual signals. More broadly, (Elkhatat et al., 2023) caution that these tools “*should not be used as the sole determinant in academic integrity cases*” and that “*a more holistic approach that includes manual review and consideration of contextual factors should be adopted*” (p. 13). This illustrates the broader principle: decision support should not be treated as an automated verdict; it should be framed as structured evidence with uncertainty, plus a clear human review process.

#### **Implications for this thesis**

Based on the above, the thesis operationalises automated/algorithmic decision support as a teacher-facing, auditable evaluation workflow that transforms Moodle-derived evidence into interpretable indicators and structured comparisons, while preserving human accountability. Concretely, the implementation focuses on (i) making evidence explicit through a formal indicator catalogue and directionality decisions, (ii) generating decision-support outputs (rankings/diagnostics) that are reproducible and explainable, and (iii) integrating quality assurance (verification and robustness checks) before presenting results to stakeholders, implemented in §3.1.5–§3.1.9 and operationalised in the automated prototype workflow in §3.2.3. Governance requirements (GDPR, privacy, transparency) are treated as design constraints rather than add-ons and are consolidated in §3.2.9, reflecting the ethical responsibility of analytics-based support (Ismayilzada et al., 2025; National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018; Teresa & McGrath, 2021). The expected benefit is useful and traceable decision support (not automated judgement) that can reduce reliance on unverified impressions; the main limitation is that any algorithmic

support still embeds normative choices (indicator selection/directionality) and may be misused as high-stakes judgement—risks revisited in §4.

## 2.8 AI Detection in Education

This section introduces AI-content detection as a recent and debated response to generative AI use in student writing, with particular attention to validity and governance risks. It notes that detector outputs are model-based estimates that can vary across tools, genres, and language proficiency, so false positives are a serious fairness concern. Accordingly, the thesis treats AI detection only as a supportive contextual signal and never as decisive evidence of misconduct §3.1.4. Elkhatat et al. (2023) state that “*Therefore, while AI-detection tools can help to spot potentially suspicious work, they should not be used as the sole determinant in academic integrity cases.*” (p. 13).

### 2.8.1 Why “AI detection” became an educational problem

The public availability of large language models (LLMs) has created a new integrity dilemma in higher education: student text can be produced (fully or partially) by generative AI, while instructors often lack reliable ground-truth evidence about the writing process. In response, a market of AI-writing detectors emerged that output a probability-like judgement about whether a text “looks AI-generated”. Ardito (2025) states that “*The resultant academic environment, where students and academics are left to interpret ‘AI scores’ without a coherent policy framework, has led to inconsistent and unjust outcomes.*” (p. 17). However, the literature consistently frames this area as contested: detection is attractive as a quick control mechanism, yet its measurement validity and governance fit are uncertain (Ardito, 2025).

In this thesis context, the issue matters because a subset of the analysed forum answers is free-text, and the evaluation framework explicitly includes an “AI involvement” indicator as one qualitative dimension among multiple attributes, rather than as a disciplinary verdict §3.1.4.

### 2.8.2 How AI-writing detectors work

AI-writing detectors typically operationalize “AI-likeness” by comparing statistical patterns in a text against what is expected from human writing versus machine-generated writing (e.g., token-probability patterns, stylometric regularities, and other classifier-derived features). Ardito (2025) summarizes that “*AI detectors operate on a fundamentally different principle. They assess the degree to which text aligns with the most common outputs of supported LLMs and do not provide any means to independently verify their assessment.*” (p. 17).

A key methodological implication is that detector outputs are not direct observations of authorship. They are model-based inferences that can change when detectors are updated, when student writing is edited, or when the writing context shifts (Ardito, 2025).

### 2.8.3 Empirical performance and the “evidence problem”

Empirical evaluation suggests that detection accuracy varies substantially across tools and conditions. (Elkhatat et al., 2023) evaluated multiple AI-content detection tools on human vs AI-generated texts and report notable limitations: performance is inconsistent across detectors and can be sensitive to text type and experimental setup. Elkhatat et al. (2023) note that “*to a certain extent, their performance is inconsistent, making them unreliable tools in determining whether to take disciplinary actions against students in high-stakes contexts such as academic integrity investigations.*” (p. 13)

Beyond average accuracy, the educational risk is asymmetric: false positives can harm students and damage trust. (Ardito, 2025) argues that detectors are problematic in real assessment settings because they are difficult to falsify from the student side (i.e., a student cannot easily “prove” human authorship if a detector flags their work), making high-stakes use procedurally risky.

### 2.8.4 Vulnerability, adaptivity, and “detector evasion”

A recurring critique is that detectors are vulnerable in adversarial settings: if students can test against detectors and iteratively paraphrase, then the detector becomes a moving target and incentivizes “humanization” services rather than authentic learning “*The combination of this rapid dissemination with commercial incentives to identify such weaknesses makes reliance on AI detection tools for maintaining academic integrity a fundamentally flawed strategy.*” (Ardito, 2025, p. 14). In practical terms, detection tools can be undermined by (i) paraphrasing, (ii) mixed authorship (human + AI), and (iii) rapid evolution of both generators and detectors (Ardito, 2025).

### 2.8.5 Fairness, language, and governance constraints

AI detection intersects with fairness because detector scores may correlate with writing style, language proficiency, or genre conventions rather than authorship. (Ardito, 2025) explicitly highlights concerns around bias and non-English speakers, which is especially relevant in multilingual educational environments “*false positives may still not be uniformly distributed*

*across text samples but instead disproportionately affect certain subgroups of human writers.”* p. 15).

From an ethics and governance perspective, learning analytics research stresses transparency, privacy, and informed consent as recurring baseline requirements (Teresa & McGrath, 2021). Applying this to AI detection implies that (i) students should understand what is being measured, (ii) the decision consequences should be proportionate, and (iii) the institutional process must be auditable and contestable.

### **Implications for this thesis**

Based on the above, the thesis operationalises AI detection as an auxiliary, low-stakes indicator of potential AI involvement in student forum answers, rather than as an authorship proof or sanction trigger. Concretely, the implementation computes an `avg_AI_involvedMsg_score` using a RoBERTa-based detector and aggregates message-level scores to the student level as one attribute within the qualitative dimension set, implemented in §3.1.4.

Expected benefit: A consistent, documentable signal that can support instructor review when interpreted alongside other evidence (topic relevance, citation use, correctness, and log-based diligence indicators), strengthening transparency within the multi-attribute framework.

Main limitation (risk): Detector outputs have contested validity, are vulnerable to evasion, and can create fairness and due-process risks if treated as decisive evidence (Elkhatat et al., 2023). Therefore, this thesis explicitly frames AI detection as decision support only, with governance constraints and interpretation safeguards revisited in §3.2.9 and critically discussed in §4.

## **2.9 The Gap**

This section synthesises the preceding literature to identify the precise research and implementation gap that motivates this thesis. A key synthesis point is that analytics can only improve what the assessment model actually measures and rewards; therefore, aligning indicators and aggregation rules with the intended learning values is a prerequisite for meaningful decision support. Clow (2013) puts it *“If assessment does not reflect and reward those aspects of learning that are valued, a learning analytics system that improves assessment scores will not improve those aspects of learning.”* (p. 692). It first delineates what current work already enables across learning analytics, log-based modelling, text similarity assessment, multi-attribute evaluation, and AI-era integrity debates (§2.9.1). It then specifies the thesis-level gaps that remain—particularly the lack of an integrated, auditable, direction-aware, and

governance-compliant workflow that combines behavioural traces and semantic evidence without arbitrary weighting (§2.9.2). Finally, the section states the thesis contributions as an explicit set of design and methodological commitments that are operationalised in the end-to-end pipeline documented in Chapter 3 (§2.9.3).

### 2.9.1 What the literature already enables and what it does not

Across learning analytics and educational data mining, a broad consensus is that LMS environments can serve as “evidence ecosystems” where learner traces support monitoring, feedback, and decision-making at scale (Clow, 2013; Dormezil et al., 2019). Early-warning interventions (e.g., course signalling) demonstrate that analytics can be operationalised in ways that are actionable for student support, not only descriptive reporting (Arnold & Pistilli, 2012). At the same time, students’ expectations indicate that analytics should support planning, self-assessment, and personalised guidance—i.e., interpretability and usefulness matter at least as much as predictive power (Schumacher & Ifenthaler, 2018).

Within LMS practice, Moodle provides a mature platform context and an increasingly analytics-oriented feature set, including ML-assisted analytics components, dashboards, and configurable indicators (Moodle, 2024). However, platform analytics are not automatically aligned with institution-specific evaluation constructs; they are constrained by available features, event semantics, and how “meaning” is assigned to activity (Agudo-Peregrina et al., 2014; Park & Jo, 2016). This is amplified by the well-established validity concern that log activity is a proxy rather than direct evidence of learning: the same interaction count can reflect engagement, confusion, compliance, or routine navigation depending on course design and modality (Agudo-Peregrina et al., 2014). In line with this concern, Yavuz (2019) reports that *“This finding shows that the effects of RT and the number of actions were not associated linearly with the abilities of students and difficulties of items”* (p. 383), reinforcing the point that behavioural traces should not be interpreted as straightforward measures of learning performance.

Parallel to log-based modelling, automatic assessment research shows that text similarity and NLP methods can support evaluation of descriptive answers and written responses (Bahel & Thomas, 2021; Rahman & Siddiqui, 2018). Yet much of this work focuses on text-only scoring, often without integrating behavioural context (e.g., interaction sequences, posting patterns), and without treating “evaluation” as a multi-dimensional construct requiring explicit directionality and aggregation rules (Page & Stake, 1979). Multi-attribute evaluation

literature—including KJU/MIAÚ-oriented similarity and COCO-type approaches—explicitly targets this aggregation problem by aiming for object comparison without ad-hoc weighting, and by using attribute exclusion/importance analyses to interpret what truly differentiates cases under a given OAM configuration (Bán et al., 2024).

Finally, recent work on generative AI in education highlights that “AI detection” is methodologically fragile and institutionally risky when treated as a primary evaluation tool; detection can create false positives, incentive distortions, and procedural unfairness if used as a shortcut for assessment integrity (Ardito, 2025; Elkhatat et al., 2023). Motz et al. (2019) make this limitation explicit: *“However, it remains unclear whether LMS activity indeed reflects behavioural properties of student engagement, and it also remains unclear how to deal with variability in LMS usage across a diversity of courses.”* (p. 300). This directly intersects with ethics and GDPR: if analytics or automated judgments are used in ways that affect students, transparency, purpose limitation, minimisation, and governance become non-negotiable design constraints (Ismayilzada et al., 2025; Mutimukwe et al., 2021; National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018; Teresa & McGrath, 2021).

### 2.9.2 The specific gap addressed by this thesis

Gap A — Fragmentation across evidence types (logs vs. text): Existing studies typically treat log indicators or text similarity or platform dashboards as separate evidence streams (Bahel & Thomas, 2021; Moodle, 2024; Motz et al., 2019). There is comparatively little thesis-level, reproducible work that integrates Moodle trace evidence and semantic similarity evidence into one auditable evaluation model, while keeping constructs explicit and interpretable (Page & Stake, 1979; Park & Jo, 2016).

Gap B — Aggregation and fairness without arbitrary weights: In educational evaluation, multi-dimensional performance is common, but aggregation is often handled by implicit instructor judgement or by fragile weighted scoring schemes (Page & Stake, 1979). Anti-discriminative, similarity-driven evaluation provides a principled alternative, yet it is rarely operationalised in a way that is (i) direction-aware, (ii) verifiable, and (iii) demonstrably usable as decision support in an LMS context (Bán et al., 2024; Pitlik, 2014).

Gap C — Verification/robustness as a first-class requirement: Many analytics pipelines end at “the model outputs a dashboard / score / classifier.” The literature supports stronger standards: validity concerns in logs (Agudo-Peregrina et al., 2014; Motz et al., 2019) and risks of

automated judgement. (Ardito, 2025) imply that evaluation pipelines should include internal verification and stress testing, not only results reporting.

Gap D — Moving beyond “AI detection” toward evidence-grounded evaluation: AI detection debates show why detector output should not be treated as a proxy for learning quality or authorship integrity (Ardito, 2025; Elkhataat et al., 2023). In this context, Kishore et al. (2023) argue that “*While ChatGPT and other GAI models present both opportunities and challenges for education, an outright ban may not be the most effective solution*” (p. 14). The practical gap, therefore, lies in designing an evidence-based evaluation pipeline that remains useful even when AI-generated text is possible, by triangulating behavioural traces, contribution patterns, and similarity signals under explicit governance.

Gap E — Governance operationalisation, not only ethical statements: Ethics/GDPR literature defines requirements, but many implementations remain vague about operational steps (e.g., minimisation choices, early anonymisation, retention boundaries) (Ismayilzada et al., 2025; Mutimukwe et al., 2021; National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018). This thesis addresses the gap by embedding governance constraints directly into the pipeline design assumptions rather than treating them as a purely narrative appendix. This operational constraint is stated bluntly in GDPR guidance: “*adequate, relevant and limited to what is necessary. In other words, only data that is required for the explicit purpose detailed above should be gathered and stored.*” (National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018, p. 1).

Institutionally, these gaps are especially relevant in distance and blended contexts where Moodle-mediated activity becomes a core part of learning and evaluation practice (Kodolányi János Egyetem, 2025; Virágh, 2020). Student experience research points in the same direction: Sukhbaatar et al. (2023) emphasise that “*Students’ evaluation of classes taken online during the lockdown is valuable for improving the quality and effectiveness of e-learning and making it more accessible to learners*” (p. 42, my translation). This strengthens the case for systematic evaluation of e-course quality and participation conditions.

At the same time, Below (2021) explicitly warns that “*The analysed data, compressed and displayed on dashboards, should not simply be accepted, but should be critically questioned*” (“3. Herausforderungen,” “*Kompetenzentwicklung*,” para. 2, my translation), underscoring that those digital traces must be translated into defensible meaning rather than treated as self-evident evidence. KJU/MIAÚ case-oriented work already argues for decision support that can

validate or prepare subjective judgement using logs and modelling logic (Balogh et al., 2018; Pitlik et al., 2017), but the remaining gap is a single, end-to-end, verifiable workflow that combines these components at bachelor-thesis level with explicit constructs and governance.

### 2.9.3 What the thesis contributes

Contribution 1 — A construct-explicit, multi-source evidence model: The thesis contributes an operationalisation where Moodle traces and contribution/text similarity signals are formalised as a transparent attribute catalogue with stated meaning and directionality, rather than treated as “raw activity” proxies (Agudo-Peregrina et al., 2014; Park & Jo, 2016).

Contribution 2 — Anti-discriminative multi-attribute aggregation with interpretability tools: Building on similarity-analysis traditions and attribute-importance logic, the thesis contributes a direction-aware, anti-discriminative aggregation strategy (COCO-type) and interpretable diagnostics via attribute exclusion to reveal hidden differences when “sameness” appears (Bán et al., 2024; Pitlik, 2014).

Contribution 3 — Verification and robustness as part of evaluation quality: The thesis contributes a workflow standard where evaluation is accompanied by internal verification logic (e.g., consistency expectations across transformed inputs) and stress tests, addressing well-known validity risks in learning analytics (Motz et al., 2019; Page & Stake, 1979). Motz et al. (2019) state that *“there is not a single generic relationship between activity and engagement, and what constitutes the behavioural components of “engagement” will be contingent on course structure.”* (p. 300).

Contribution 4 — A decision-support stance compatible with AI-era constraints: Instead of substituting evaluation with AI detectors, the thesis positions algorithmic support as triangulated evidence to assist judgement, consistent with critiques of detection-only approaches (Ardito, 2025; Elkhataat et al., 2023; Kishore et al., 2023).

Contribution 5 — Governance operationalisation aligned with GDPR/ethics: The thesis contributes a GDPR-aware design framing: minimisation, transparency, and purpose limitation are treated as design constraints on which attributes are extracted and how outputs are used (Ismayilzada et al., 2025; National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018; Teresa & McGrath, 2021). It is described *“Personal data must be: processed lawfully, fairly and in a transparent manner. collected for specified, explicit and*

*legitimate purposes and only used for these purposes.*” (National Forum for the Enhancement of Teaching and Learning in Higher Education, 2018, p. 1).

### **Implications for this thesis**

Based on the above, the thesis operationalises the identified gap by implementing an end-to-end Moodle-based evaluation pipeline that (i) extracts student-level log and contribution indicators, (ii) represents them as an Object–Attribute Matrix with explicit directionality, (iii) aggregates evidence via anti-discriminative multi-attribute evaluation rather than manual weighting, and (iv) includes verification and robustness procedures under GDPR/ethics constraints. This is implemented in the data extraction and modelling workflow, the evaluation and aggregation steps (§3.1), and the automation/prototype layer (§3.2), with governance constraints documented in §3.2.9 and critical limitations revisited in §4.

## 2.10 BPROF Subjects and the Thesis

This section connects each BSc course to specific parts of the thesis. For each course, the applied knowledge elements, their contribution to the implemented tool and workflow, their location in the document (mainly §3 and the Annexes), and the deliberately excluded elements are specified in order to keep the project small, clear, and feasible at bachelor level.

### 2.10.1 Networks & Computer Architectures

- **Applied knowledge/components:** Practical constraints of running the full pipeline on a standard laptop (CPU/RAM trade-offs).
- **How it supports the tool/workflow:** Motivates lightweight, local-first execution choices in the prototype.
- **Where:** §3.2.1 (scope: local-first), §3.2.10 (system requirements).
- **Left out:** Hardware acceleration and low-level optimization beyond what is required for a thesis-grade demo.

### 2.10.2 Introduction to Algorithms

- **Applied knowledge/components:** Deterministic step-by-step workflow design (transformations, aggregation, ranking, diagnostic re-runs).

- **How it supports the tool/workflow:** Enables reproducible transformation from Moodle rows into student-level indicators (OAM) and rank-based COCO inputs.
- **Where:** §3.1.2–§3.1.6 (pre-processing to OAM to COCO), §3.2.3 (pipeline stages).
- **Left out:** Advanced optimisation claims; the focus remains on correctness, traceability, and reproducibility.

### 2.10.3 Operating Systems

- **Applied knowledge/components:** Local versus hosted execution awareness; file-based reproducibility and stable run conditions.
- **How it supports the tool/workflow:** Justifies why persistence and governance controls are reliable in local mode, while hosted mode can be session-scoped.
- **Where:** §3.2.2 (deployment-dependent persistence), §3.2.9–§3.2.10 (local-first, requirements).
- **Left out:** Kernel-level profiling and container hardening; kept to simple cross-platform steps.

### 2.10.4 Introduction to Programming

- **Applied knowledge/components:** Readable functions and deterministic computation (load, compute, export) with basic error handling.
- **How it supports the tool/workflow:** Implements repeatable pre-processing, attribute computation, and report generation without hidden steps.
- **Where:** §3.1 (process tools: Excel/Python/SQLite), §3.2.3 (implementation stages).
- **Left out:** Complex framework engineering; the prototype prioritises transparency over abstraction.

### 2.10.5 Programming I, II

- **Applied knowledge/components:** Modular implementation and separation of concerns (computation, configuration, persistence).
- **How it supports the tool/workflow:** Keeps the prototype maintainable and auditable (attribute computation and COCO parsing are explicit components).

- **Where:** §3.2.2 (architecture and module map), §3.2.3 (orchestration).
- **Left out:** Performance micro-optimisation and advanced packaging beyond what is needed for the demo scope.

### 2.10.6 Databases

- **Applied knowledge/components:** Relational thinking via SQLite for traceable storage and queryable processing.
- **How it supports the tool/workflow:** Supports reproducible pre-processing and persistence of datasets, runs, and artifacts (audit trail).
- **Where:** §3.1.2 (SQLite for cleaned logs), §3.2.8 (runs database and stored artifacts).
- **Left out:** Full RDBMS deployment and SQL performance tuning; the dataset scale is thesis-bounded.

### 2.10.7 Data Visualization

- **Applied knowledge/components:** Clear outputs and explainable reporting intended for instructor inspection.
- **How it supports the tool/workflow:** Supports interpretation of rankings and diagnostics as decision support rather than opaque scoring.
- **Where:** §3.2.1–§3.2.3 (inspectable outputs and intermediate artifacts).
- **Left out:** Advanced interactive analytics beyond the Streamlit demo scope.

### 2.10.8 Electronics & Circuits

- **Applied knowledge/components:** General engineering constraint awareness (keep the solution software-only and portable).
- **How it supports the tool/workflow:** Reinforces the 'runs on a standard machine' constraint and avoids hardware dependence.
- **Where:** §3.2.1 (scope constraints), §3.2.10 (system requirements).
- **Left out:** Hardware acceleration (GPU-as-requirement, FPGA, dedicated devices).

### 2.10.9 System Modelling

- **Applied knowledge/components:** Stage-based pipeline modelling with clear inputs and outputs per stage.
- **How it supports the tool/workflow:** Keeps the workflow auditable: Moodle export to pre-processing to OAM to ranking to COCO to validation to reporting.
- **Where:** §3.1.1–§3.1.9 (analytical pipeline), §3.2.3 (implemented pipeline stages).
- **Left out:** Heavy formal modelling notation; emphasis remains on runnable, inspectable steps.

### 2.10.10 System Operation

- **Applied knowledge/components:** Reproducible run practice: clear requirements, stable execution steps, and local-first operation.
- **How it supports the tool/workflow:** Enables consistent execution and repeated runs, including persistence of run artifacts.
- **Where:** §3.2.10 (system requirements), §3.2.8 (stored runs and evidence).
- **Left out:** Production DevOps (CI/CD, monitoring) because the tool is a thesis prototype.

### 2.10.11 System Planning

- **Applied knowledge/components:** Scope control through explicit aims and tasks within an R&D methodology structure.
- **How it supports the tool/workflow:** Keeps development aligned with the analytical pipeline and validation obligations (avoids feature creep).
- **Where:** Chapter 1 (aims/tasks) and §3 (own developments as R&D realisation).
- **Left out:** Full project management tooling; planning is documented in text and bounded limitations.

### 2.10.12 Software Architectures

- **Applied knowledge/components:** Layered architecture (UI, processing, persistence) and explicit module responsibilities.

- **How it supports the tool/workflow:** Keeps computations traceable (attribute calculations and COCO parsing remain explicit components).
- **Where:** §3.2.2 (architecture), §3.2.3 (page flow and orchestration).
- **Left out:** Microservices and institutional integration; intentionally excluded in the demo scope.

#### 2.10.13 Software Testing

- **Applied knowledge/components:** Scenario-driven end-to-end pipeline testing and evidence persistence.
- **How it supports the tool/workflow:** Demonstrates that the workflow completes and stores artifacts and logs for auditing.
- **Where:** §3.2.8 (end-to-end runs; stored artifacts).
- **Left out:** Full unit and integration test suite; acknowledged as a prototype limitation.

#### 2.10.14 Business Process Management

- **Applied knowledge/components:** Repeatable procedure and explicit hand-offs between stages (data to indicators to ranking to outputs).
- **How it supports the tool/workflow:** Supports practical instructor use: a guided workflow with exportable artifacts.
- **Where:** §3.2.1–§3.2.3 (workflow), §3.2.2 (page structure).
- **Left out:** Enterprise workflow tooling; unnecessary for the thesis-sized target use.

#### 2.10.15 Business Law and Regulation

- **Applied knowledge/components:** Governance framing (privacy, ethics, GDPR) as design constraints.
- **How it supports the tool/workflow:** Drives anonymisation, minimisation, local-first handling, and the 'decision support' boundary.
- **Where:** §2.7.5–§2.7.6 (responsibility and boundaries), §3.2.9 (implementation safeguards).

- **Left out:** Formal legal review and institutional policy integration; beyond prototype scope.

#### 2.10.16 IT Security

- **Applied knowledge/components:** Basic security-by-design practices appropriate for a demo analytics tool (local-first, minimization, auditability).
- **How it supports the tool/workflow:** Reduces exposure of educational log data and improves accountability via stored run artifacts.
- **Where:** §3.2.9 (security, privacy, ethics), §3.2.8 (auditability via stored runs).
- **Left out:** Production identity management and hardened multi-user deployment (explicit limitation).

#### 2.10.17 ICT in IT-Security

- **Applied knowledge/components:** Secure handling of inputs and outputs and explicit acknowledgement of external dependency risk.
- **How it supports the tool/workflow:** Frames COCO execution as an external endpoint dependency with availability and privacy implications.
- **Where:** §3.2.3 (COCO request and parsing), §3.2.9 (endpoint dependency risk).
- **Left out:** Formal threat modelling and compliance certification; beyond bachelor scope.

#### 2.10.18 Intercultural Communication

- **Applied knowledge/components:** Clear terminology and a reader-oriented explanation style (interpretability emphasis).
- **How it supports the tool/workflow:** Supports transparent decision support: intermediate artifacts are inspectable and contestable.
- **Where:** §3.2.1 (Explainability/auditability), §2.7 (decision support framing).
- **Left out:** Formal discourse analysis; applied pragmatically through documentation and UI design.

## 2.11 Use of LLM in the Thesis Writing and Development Process

This section documents the role of LLM assistance used to support selected writing and development tasks during the preparation of this thesis. The LLM fulfilled a strictly supporting and editorial function; the research design, empirical analysis, implementation logic, and all substantive intellectual content are the author's own work.

### 2.11.1 Scope of LLM Assistance

The LLM was consulted for the following categories of task:

#### **Writing and editorial support:**

- **Language refinement:** Improving grammatical correctness, sentence-level flow, and phrasing in non-native English academic writing, particularly in technical sections where precise formulation is required.
- **Writing style consistency:** Suggesting adjustments to maintain a consistent academic register across chapters and to avoid colloquial or imprecise phrasing.
- **Structural advice:** Reviewing the logical ordering of sections and subchapters, and advising on whether the argumentation proceeds coherently from the literature foundations (Chapter 2) through the analytical pipeline (Chapter 3) to interpretation (Chapter 4).
- **Internal coherence and golden thread:** Checking whether forward and backward cross-references (e.g., §1.1 promises revisited in §4.3) were consistently maintained, and flagging passages where the internal logic of the argument was weakened or disconnected.
- **Abbreviation extraction:** Assisting in the compilation of the list of abbreviations (§8.1) by identifying defined acronyms and initialisms across the manuscript.
- **Formatting and typographic error detection:** Identifying inconsistencies in heading numbering, table and figure caption conventions, equation sequencing, and citation formatting across the document.

#### **Development support:**

- **Code debugging:** Assisting in identifying logic errors in Python scripts during prototype development (§3.2), including data-pipeline functions, COCO output parsing routines, and Streamlit component configurations.

- Problem solving during R&D: Advising on specific technical challenges encountered during implementation, such as SQLite query construction, Streamlit layout issues, and formatting of intermediate artefact exports.

**Benchmark comparison:**

- §3.3 and §8.7 document a separate benchmark experiment included for comparative purposes.

### 2.11.2 Boundaries of LLM Use

**The LLM was explicitly not used for:**

- generating, computing, or interpreting the thesis’s own human-authored core empirical results, including the OAM construction, semantic-similarity calculations, COCO Y0 runs, validation checks, and the corresponding substantive claims;
- generating, modifying, or pre-processing any dataset used in the empirical analysis;
- verifying bibliographic source details or retrieving final reference metadata; the reference list was compiled and checked manually, because LLMs may confabulate bibliographic details and DOIs;
- replacing the author’s own analytical judgement in the core interpretive sections of the thesis, especially the gap analysis (§2.9), the discussion chapter (§4), and the conclusion (§5).

All citations, numerical results, and final argumentative claims included in the thesis were manually verified by the author prior to submission.

A separate exception applies only to the benchmark experiment documented in §3.3 and Annex §8.7, where an independent LLM-generated solution was produced for comparative purposes and was subsequently interpreted critically by the author in §4.10.

### 2.11.3 Critical Posture Toward LLM Output

All LLM outputs were treated as preliminary drafts and subjected to manual review before incorporation. This posture was maintained consistently throughout the writing and development process for the following reasons:

- LLMs exhibit a tendency to hallucinate in factually strict contexts, including reference details, statistical claims, and specific technical assertions; accordingly, any factual content proposed by the LLM was independently verified by the author.
- Suggested paragraph reformulations were checked against the original intended meaning to prevent distortion of technical content or loss of precision.
- Grammatical and punctuation corrections were accepted only after manual proofreading confirmed they did not alter meaning.
- Structural suggestions were evaluated against the thesis's own argumentation logic before any reorganisation was adopted.

Representative examples of prompts submitted to the LLM and the corresponding responses are documented in [Annex §8.7](#), illustrating the nature of interactions across the writing and development phases.

## 2.12 Comparison with Other Running COCO-Oriented Theses

The parallel COCO-oriented theses reviewed in the same academic environment confirm that the COCO/OAM methodological family is highly transferable, but they also show that transferability does not mean conceptual uniformity. The shared element is not the research topic itself, but the decision-support logic: heterogeneous observations are transformed into an Object–Attribute Matrix, the directions of the attributes are declared explicitly, and the final ranking or estimation is derived without relying on arbitrary manual weighting. For the present thesis, this comparison is important because it clarifies both the strengths and the methodological limits of using COCO Y0 for Moodle-based student evaluation. The repository pages used for this comparative positioning are listed in [Annex §8.12](#). Table 2.12- I summarises the comparative positioning of the present thesis relative to selected parallel COCO-oriented theses reviewed in the same methodological family.

Thesis	COCO application	Compared thesis	My Thesis
Boldsukh Ganzorig	Sector ranking from workplace-efficiency indicators	Numeric sector data	Student log and text evidence
Munkhjargal Ariunbold	Ranking of 298 Formula 1 race entries with validation reruns	Technical race objects	Educational performance
Byekbolat Nurbol	Country comparison combining objective and subjective well-being rankings	Macro-level indicators	Moodle-based evidence
Tsetsegsuren Namjiljav	Wine-quality estimation from physicochemical attributes	Stable laboratory data	Indirect learning proxies
Amarsanaa Amgalanbatar	Password-strength Risk Index from structural attributes	Technical artefacts	Human learners
Sukh-Ochir Dulguun	Multi-criteria benchmark of sentiment-classification algorithms	Algorithms	Students
Munkh-Orgil Batbayar	Validation of cross-platform sentiment interpretation	Platform-level interpretation risk	Course performance
Kenyó Kristóf István	Anomaly prioritization from aggregated log windows	Cyber-risk	Educational assessment

Table 2.12- I Comparative positioning of the present thesis among selected parallel COCO-oriented theses (Source: Own Presentation)

**Legend:** COCO application = how the compared thesis uses COCO or related OAM-based logic; Compared thesis = the main object type and evidence context of the peer thesis; My Thesis = the corresponding object and evidence context of the present thesis, highlighting the main comparative difference.

The broader set of reviewed theses extends the same family further into price-performance and risk-oriented applications. Ganbat Bayanmunkh develops a country-level comparative model around HDI, education, and internet-usage trajectories; Kosdi Gábor applies COCO logic to attribute-based product comparison; Kosárszki Tamás to spray-drone price-performance evaluation; Lehrer Zoltán to USB-related risk assessment; Váradi Viktor to news-source objectivity within a larger media-automation system; and Leitner Patrik to wage-feasibility modelling. These works broaden the comparative background, but they also reinforce the same methodological point: COCO is not tied to one topic area; its validity depends on how carefully the object set, attribute design, and interpretation rules are specified.

### 2.12.1 Shared methodological core

Across the reviewed theses, the most important common feature is the explicit construction of the evaluation space. The objects to be compared are never treated as self-explanatory; they must first be defined, described with attributes, and placed into a transparent OAM structure. This is true whether the objects are sectors, countries, race entries, passwords, algorithms, network-log time windows, or student profiles. In all stronger peer works, COCO is therefore not presented as a magic ranking engine, but as the last step of a prior modelling decision in which attribute meaning and directionality are made explicit. This insistence on explicit descriptor construction is methodologically important because it makes the later aggregation step auditable and prevents the final ranking from being detached from the modelling assumptions that generated it.

Another common element is the rejection of single-metric judgement. The reviewed theses repeatedly demonstrate that apparently simple evaluation tasks become misleading when reduced to one indicator only. Password quality cannot be reduced to length alone, Formula 1 performance cannot be reduced to one raw timing variable, country-level development cannot be reduced to one index, and classifier quality cannot be reduced to accuracy alone. The same lesson applies even more strongly to the present thesis, where student performance in a Moodle forum clearly cannot be represented by posting frequency alone. The comparative review therefore supports the decision to integrate behavioural, temporal, interactional, and text-based evidence into one multi-attribute model.

The third shared element is that COCO-based outputs are generally framed as decision support rather than unquestionable truth. The stronger parallel theses interpret rankings as model-relative outputs whose meaning depends on the chosen descriptor system. This interpretive modesty is crucial in the present thesis as well. A COCO-based ranking of students is not a direct measure of true learning in an absolute sense, but a structured comparative result derived from explicitly selected indicators under clearly bounded assumptions.

A further shared characteristic of the stronger peer works is a reproducibility-oriented posture. Rank transformation, inverse-run checking, attribute exclusion, spreadsheet- or software-supported execution, and the preservation of intermediate artefacts appear not as cosmetic additions but as part of the methodological argument itself. The intensity of this reproducibility logic differs across the reviewed theses, yet the common lesson remains clear: COCO-based

evaluation becomes substantially more defensible when the transformation path from raw evidence to ranked output is preserved and inspectable.

### 2.12.2 What differentiates the present thesis from the peer works

The main difference lies in the nature of the evaluated objects. In many peer theses, the objects are comparatively stable technical, economic, or macro-level units: passwords, products, algorithms, countries, sectors, or race entries. By contrast, the present thesis evaluates students, that is, human actors whose observed Moodle traces are only partial proxies of the underlying construct of academic performance. This makes the modelling task more sensitive, because the analysis must remain aware of the gap between observable digital behaviour and the broader educational phenomenon it is supposed to approximate.

The evidence structure is also more heterogeneous in the present work than in most peer theses. Several reviewed theses rely mainly on numeric and structurally cleaner inputs. The present thesis, however, combines log-derived activity indicators, timing measures, interaction-related variables, semantic similarity of text responses, and an auxiliary AI-related signal. This places the thesis at the intersection of learning analytics, text analysis, and anti-discriminative multi-attribute evaluation. As a consequence, the present model is methodologically closer to the hybrid logic of the polarization-risk and log-anomaly theses than to purely numeric price-performance or country-ranking applications.

Another major difference concerns interpretation risk. In product comparison, sector benchmarking, algorithm selection, or password scoring, the ranking primarily supports a technical or managerial choice. In educational evaluation, however, the ranking touches human judgement, fairness, and possible consequences for students. For this reason, the present thesis must impose stronger interpretation boundaries than most parallel works. The semantic similarity score cannot be treated as a correctness proxy, the AI-related signal cannot be used as a punitive indicator, and the final output must remain teacher-facing decision support rather than automated judgement.

### 2.12.3 Lessons for the present Moodle-based evaluation model

The comparison with the other running COCO-oriented theses strengthens four methodological conclusions for the present thesis. First, the use of COCO Y0 is justified because the thesis

addresses a genuinely multi-attribute problem for which one-dimensional scoring would be reductionist. Second, the comparison shows that the interpretability of the final ranking depends primarily on the quality of the OAM design; therefore, the attribute construction in §§3.1.3–3.1.5 is not a technical preliminary step, but the substantive core of the model.

The peer comparison also shows that methodological credibility depends not only on whether a COCO model can produce an ordering, but on whether that ordering is tested under controlled methodological stress. In this thesis, that stronger standard is realised through staircase-based exclusion, inverse-run function-symmetry checks, and distorted-input verification. These procedures should not be interpreted as attempts to force a preferred result, but as ways of making the behaviour of the evaluation framework inspectable under changing conditions.

Third, the peer works confirm that validation is what separates a merely executable COCO model from a defensible one. This directly supports the inclusion of staircase-based exclusion, inverse-run function-symmetry checks, and distorted-input verification in §§3.1.6–3.1.9. Among the reviewed theses, the most convincing works do not stop at producing a ranking, but also test how that ranking behaves under controlled methodological stress. The present thesis follows this stronger line and extends it further by comparing the objective ranking with peer-based subjective evaluation.

Fourth, the peer comparison clarifies the distinctive contribution of the present thesis within the class-wide COCO family. The thesis does not simply transfer COCO Y0 to another numeric ranking task. Instead, it applies the method to a more delicate educational setting in which behavioural logs and semantic signals must be interpreted as bounded evidence about student performance. In this sense, the present thesis belongs to the same methodological family as the other COCO-oriented works, but it addresses a comparatively demanding application context by combining hybrid evidence, human-centred interpretation, validation requirements, and ethical caution within a single evaluation framework.

## 3 Own Developments

This chapter presents the work completed during the research-and-development phase of the thesis, from the initial analytical experiments and their results to the implementation and testing of the prototype software. The experiment is based on real-world pseudonymised datasets collected from Moodle forum discussions in an IT-related course context, where students were required to provide descriptive and timely responses. These discussion logs form the empirical foundation for constructing measurable indicators of student activity and interaction, and for evaluating performance in a transparent and reproducible manner.

The work originated from a course assignment in which students solved a task through structured forum discussions, followed by a peer-evaluation session conducted in the classroom. In a subsequent assignment, the instructor provided the corresponding Moodle log data, and the cohort performed an objective log-based analysis on the data and compared the results with the earlier subjective peer evaluations. This comparison highlighted both the potential and the limitations of traditional assessment practices and motivated the research presented in this thesis (see §1.5).

The primary contribution of this project is a log-based evaluation framework that derives multiple performance-related attributes from Moodle activity data, organizes them into an Object Attribute Matrix (OAM) with explicitly defined directionality, and applies the Component-Based Object Comparison for Objectivity (COCO) engine to estimate attribute impacts and get anti-discriminative, multi-attribute results. To ensure practical applicability/automation, the analytical workflow is operationalized in a Streamlit-based prototype tool that executes the pipeline end-to-end and generates outputs for educators in both web-accessible and portable executable formats.

Chapter 3 presents the thesis's original developments, progressing from the analytical workflow to its operational realization as a usable prototype. §3.1 documents the full analytical process applied to pseudonymised Moodle discussion logs: §3.1.1 introduces the Moodle log structure and object definition, §3.1.2 describes extraction, preprocessing, anonymization, and SQL-based preparation, §3.1.3 operationalizes semantic similarity via embedding-based cosine similarity for topic relevance, and §3.1.4 adds an auxiliary AI-involvement signal as a cautious contextual indicator. §3.1.5 constructs the Object-Attribute Matrix (OAM), defines the A1-A29 attribute catalogue with explicit directionality, and applies direction-aware ranking to

produce a standardized COCO input. §3.1.6 executes COCO Y0 under a constant baseline and applies staircase diagnostics to identify non-informative attributes for exclusion and re-run analysis, while §3.1.7 validates internal consistency through inverse-run function-symmetry checks. Results are consolidated in §3.1.8 across the full attribute set and the diligence-only and understanding-only subsets, including excluded-run variants and validation flags; §3.1.8.4 compares objective COCO-based rankings with peer-based subjective rankings using tie-aware rank harmonization and Spearman correlation. Finally, §3.1.9 performs an input verification test using deliberately inconsistent/extreme inputs to assess the robustness and reliability of the COCO-based evaluation logic. Building on these validated analytical components, §3.2 documents the demo software development that operationalizes the §3.1 pipeline, detailing the tool’s purpose and scope (§3.2.1), system architecture (§3.2.2), deterministic data-pipeline implementations (§3.2.3), the integrated similarity and AI-signal modules (§3.2.4), ranking and scoring logic including inverse-run validation integration (§3.2.5), instructor-facing user-interface interaction design (§3.2.6), visual and AI-assisted interpretation outputs (§3.2.7), pipeline-level testing evidence (§3.2.8), implementation-level IT security, privacy, and ethics (§3.2.9), system requirements for reproducible execution (§3.2.10), and prototype limitations (§3.2.11). In addition, §3.3 introduces a separate benchmark against an independent LLM-generated solution addressing the same overall problem, thereby enabling a structured comparison between the human-authored and LLM-authored approaches.

### 3.1 Analytical process

This subchapter presents the analytical workflow developed during the course, describing each step from data pre-processing to results and validation. The workflow is designed to be transparent and reproducible, and it is implemented using a combination of Excel (for structured inspection and intermediate calculations), Python-based models (for semantic similarity §3.1.3 and AI-generated text signals §3.1.4), SQLite (for structured storage and queryable processing of log records), and the COCO Y0 engine for anti-discriminative multi-attribute analysis. The outputs of this process are the constructed Object–Attribute Matrix (OAM), similarity and auxiliary indicators derived from text responses, and the final COCO-based rankings with attribute-level impact information, which are subsequently accomplished using the procedures and methods described in §§3.1.1–3.1.9.

Figure 3- I presents the full pipeline of the analytical process §§3.1.1–3.1.9.

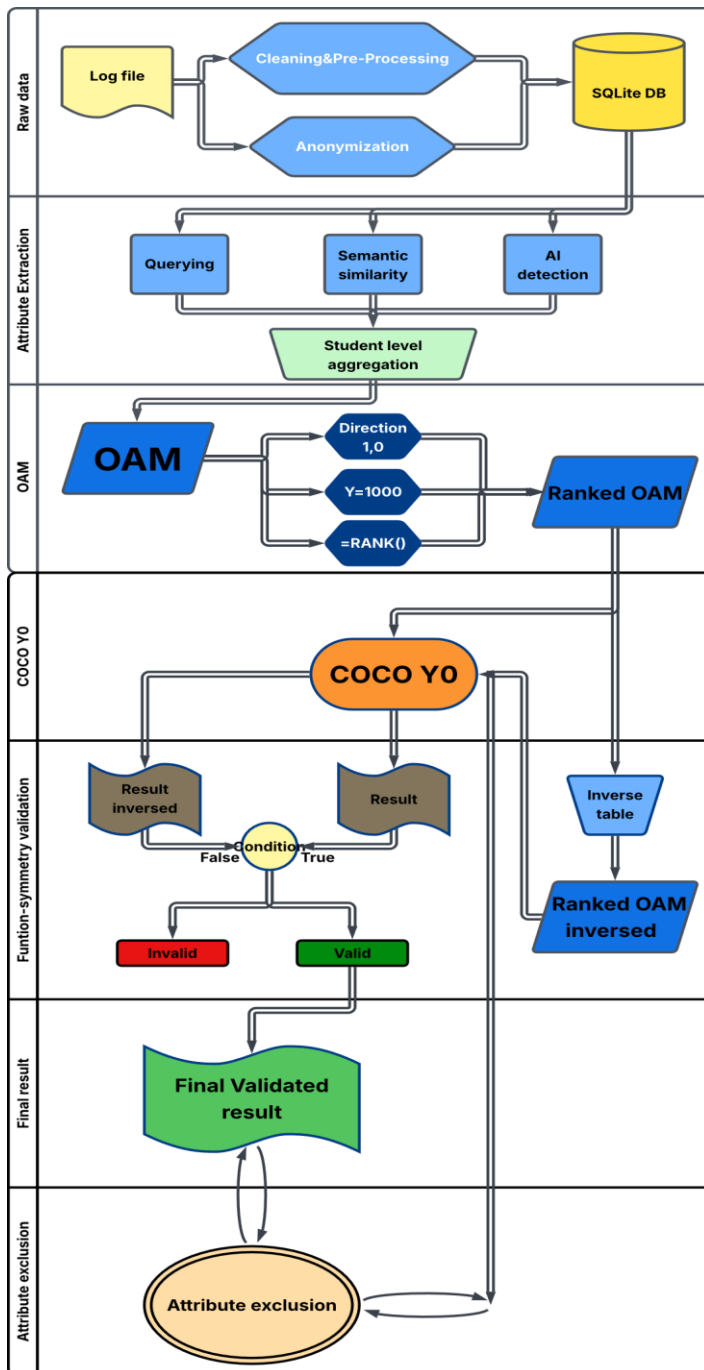


Figure 3- 1 End-to-end analytical pipeline of the proposed evaluation framework. (Source: Own Presentation)

**Legend:** yellow = source/database elements; blue = processing steps; green = aggregation and validated output; orange = COCO YO analysis; red = invalid outcome; beige = exclusion step; arrows = workflow.

### 3.1.1 Moodle Data & Structure

The analysis uses a pseudonymised dataset collected from a Moodle discussion forum. The analysis is based on forum log data from an IT-course discussion task in which students were

required to provide descriptive responses within specified time constraints, enabling both behavioural indicators (e.g., activity patterns and response timing) and text-based indicators to be derived from the same interaction records.

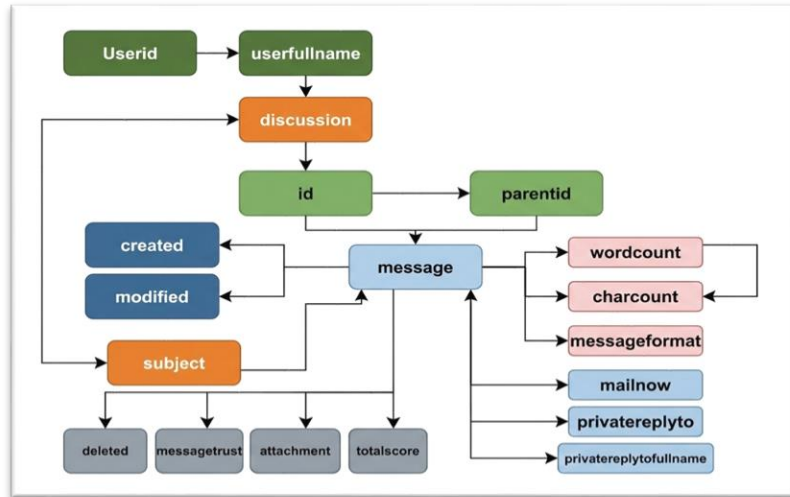


Figure 3.1.1- I Structure of the raw Moodle log dataset used in this study. (Source: Adapted from the author’s conference presentation slides (Turtogtokh et al., 2025), retrieved February, 01, 2026, from [https://miau.myx.hu/miau/319/performances/Conference\\_PPT.pptx](https://miau.myx.hu/miau/319/performances/Conference_PPT.pptx))

**Legend:** green = identifiers and relational keys; orange = discussion-context fields; blue = message, time, and system fields; pink = text-length/format fields; grey = auxiliary metadata; arrows = field relations.

Figure 3.1.1- I presents the structure of the raw log data extracted from the Moodle platform and used in this project. Given the complexity and granularity of these records, the dataset is processed using structured queries and supporting data-handling technologies to isolate and compute student-level indicators from event-level log entries. Based on the defined evaluation framework, 29 attributes are extracted and organized into two conceptual categories: diligence (24 attributes), capturing behavioral engagement patterns (e.g., activity regularity and response timing), and understanding (5 attributes), capturing indicators related to content quality and topical alignment. All attribute definitions are explained in detail in §3.1.5 below.

### Dataset and unit of analysis

The dataset comprises Moodle-style higher-education discussion forum posts collected in a thesis/automation/quasi-exam teaching context. It spans 17 October 2024, 12:13 to 17 November 2024, 08:10 (creation timestamps) and includes 24 student pseudonyms (*student\_\*\**) and two professor pseudonyms. The export contains 12 threads (12 root posts where *parent* = 0) and 576 messages, including 564 replies (*parent* ≠ 0). No separate event-

log table is available; therefore, each row is treated as a forum-post event (576 row-events) for aggregation into student-level indicators.

### **Governance, privacy, and storage**

Identifiers are pseudonymised (stable labels without direct personal identifiers), enabling within-dataset linkage while reducing re-identification risk. Pre-processing checks found 0 duplicate IDs (deduplication key: *id*), no empty threads, and one empty message body, which is retained in totals but treated as missing for text-dependent indicators. The working dataset is stored locally in an SQLite database ([Annex §8.6.5](#)) to support traceable, queryable processing within the thesis scope; outputs are used only for decision-support analysis, not automated judgement.

#### **3.1.2 Log Data Extraction & Pre-Processing**

As in most data-driven studies, the analysis begins with data cleaning and pre-processing, since raw datasets are rarely ready for direct analysis. Moodle log exports are no exception: the CSV-formatted forum data contains heterogeneous fields and timestamp representations that must be standardized before reliable attribute extraction is possible. Therefore, the raw CSV data are transformed and loaded into a structured SQL database, which enables efficient querying and consistent computation of student-level indicators.

The dataset covers a single academic semester and consists of Moodle discussion-forum records. The key fields retained for analysis are: *userid*, *discussion* (thread identifier), *parent* (reply-tracker), *created*, *modified*, *message*, *subject*, and derived word/character counts (see §3.1.1). These fields provide the information required to analyse participation intensity, temporal dynamics (e.g., response latency), interaction structure (via parent–child relations), and text-based indicators on the platform.

To ensure privacy, direct personal identifiers (e.g., student names) were removed from the export before any processing steps were executed. As (Shanthi & Karthikeyan, 2012) note, anonymization can “*reduce re-identification risk by making individual records indistinguishable within a group*” (pp. 1–4), commonly using generalization and suppression. The cleaned dataset is then inspected for basic consistency issues. Duplicate records are identified and removed, and empty threads (discussions without meaningful content) are excluded to prevent noise in subsequent aggregations. A key pre-processing step is the

conversion of the original timestamps from UNIX time to an ISO 8601 datetime representation to ensure correct ordering, filtering, and aggregation of events within the SQL environment.

**UNIX timestamp** 1729167180 → **ISO 8601:** 2024-10-17T12:13:00

After time normalization and basic consistency checks, the processed data are imported into an SQL database. The database structure follows the cleaned raw-data schema (row and column layout preserved, with pseudonymised identifiers), which supports direct traceability from source records to derived indicators. SQLite is selected due to its portability and suitability for reproducible analysis workflows, while still enabling efficient query-based extraction of attributes. The structure of the raw data is presented §3.1.1.

Table 3.1.2- I below presents raw data headers and its content example.

#	Header	Content
1	id	Post/message unique ID
2	discussion	Discussion thread ID
3	parent	Parent post ID (0/NULL if top-level)
4	userid	Author's user ID
5	userfullname	Author's full name
6	created	Created timestamp (Unix time)
7	modified	Last-modified timestamp (Unix time)
8	mailed	Whether mailed/notified (flag)
9	subject	Post subject/title
10	message	Post body/content
11	messageformat	Message format code (e.g., HTML/plain)
12	messagetrust	Trust flag for message content
13	attachment	Attachment (flag)
14	totalscore	Aggregate rating/score (if enabled)
15	mailnow	Force mail now (flag)
16	deleted	Deleted flag
17	privatereplyto	Private reply target user ID
18	privatereplytofullname	Private reply target full name
19	wordcount	Word count of message
20	charcount	Character count of message

Table 3.1.2- I Raw dataset schema: header fields and contents used in the discussion-level analysis. (Source: extracted from Moodle export (Annex §8.6.5))

Most of the 29 attributes are computed using SQL queries (e.g., counts, time-based aggregations, and interaction patterns), while a smaller subset requires additional mathematical computations performed outside the database (e.g., derived ratios and composite measures). In addition to query-based extraction of quantitative indicators, the workflow also incorporates text-oriented analysis to capture aspects of response quality. Specifically, semantic similarity is computed using cosine similarity to measure topical alignment between student responses and the corresponding instructor prompts, and an AI-generated text detector is applied to provide an auxiliary signal of potential AI usage in written answers. These text-based components complement the log-derived attributes and are explained in detail in §§3.1.3- 3.1.4.

### 3.1.3 Semantic Similarity for Descriptive Text Responses

To evaluate the topical alignment of students' descriptive answers with the instructor's prompts, this study computes a `topic_relevance_score` based on semantic text similarity. The central idea is that responses addressing the intended topic should be semantically close to the corresponding instructor post, even when different wording is used. This is especially important in forum-based tasks where multiple valid phrasings and paraphrases may occur, making purely lexical (keyword-based) matching insufficient.

#### **Sentence embeddings and model choice**

Semantic similarity is computed using sentence embeddings generated by the all-MiniLM-L6-v2 model from the Sentence-Transformers framework. The model maps each text (prompt or response) into a fixed-length vector representation intended to capture semantic meaning. This model was selected due to its favorable trade-off between embedding quality and inference efficiency, which enables scalable processing of many student messages within a course dataset.

#### **Similarity metric**

Given an instructor prompt embedding  $p$  and a student reply embedding  $s$ , similarity is computed using cosine similarity:

$$\text{cosine\_sim}(p, s) = \frac{p * s}{\|p\| \|s\|}$$

*Equation 1 Cosine Similarity Formula*

The resulting score ranges from -1 to 1 (in practice typically 0 to 1 for this task), where higher values indicate stronger semantic alignment between the student’s response and the instructor’s post.

### Data extraction and pairing logic

The computation operates on Moodle forum data stored in SQLite (see §3.1.2). Only replies whose parent value matches an instructor post ID are retained, preventing invalid comparisons and ensuring that each computed score has a well-defined reference prompt.

### Processing workflow

The implementation follows these steps:

1. Load data from SQLite into a processing environment (Python/pandas).
2. Preprocess text fields by handling missing or empty messages (e.g., excluding empty content from embedding).
3. Generate embeddings for each instructor post and each student reply using all-MiniLM-L6-v2.
4. Compute cosine similarity between each reply and its corresponding instructor post embedding, producing a reply-level Topic Relevance Score.
5. Persist results by exporting a report file and storing the numeric score back into SQLite (e.g., a `relevance_scores` table keyed by message ID).

The output dataset contains, for each reply, the student identifier, the linked instructor prompt, and the computed similarity score. Table 3.1.3- I illustrate the reply-level structure of the topic-relevance output.

userid	userfullname	student_message	prof_message	topic_relevance_score	topic_relevance_score_0_100
3	student_3	...	...	0.12	11.58
5	student_5	...	...	0.31	31.16
3	student_3	...	...	0.49	48.69
8	student_8	...	...	0.69	69.16
5	student_5	...	...	0.21	20.97
...	...	...	...	...	...

Table 3.1.3- I Example structure of reply-level topic relevance results. (Source: Own Presentation)(Unit: `topic_relevance_score`: model score, `topic_relevance_score_0_100`: aggregated model score 1-100)

**Legend:** `topic_relevance_score` = raw model score; `topic_relevance_score_0_100` = rescaled score on a 0–100 scale; in this example, raw scores span 0.12–0.69 and rescaled scores span 11.58–69.16 on the shown example.

A consolidated reference table for recurring units, scales, and coding conventions used across the thesis is provided in Annex §8.5, Table 8.5- II.

### **Aggregation to a student-level OAM attribute**

Because the OAM requires a single value per student, reply-level similarity scores are aggregated at the student level. First, each reply-level cosine similarity score  $r_i$  is mapped to a 0–100 scale:

$$r_i^{(0..100)} = 100 * r_i$$

*Equation II Similarity Score Scaling*

Then, for each student  $u$ , the `topic_relevance_score` attribute is computed as the arithmetic mean across the student’s valid replies:

$$TRS(u) = \frac{1}{u_u} \sum_{i=1}^{u_u} r_i^{(0..100)}$$

*Equation III Topic Relevance Mean*

where  $u_u$  is the number of replies authored by student  $u$  that are linked to an instructor post and non-empty. This average score is used as the student-level `topic_relevance_score` attribute in the Object Attribute Matrix (see §3.1.5).

### **Interpretation and limitations**

While semantic similarity provides a useful proxy for topical alignment, it does not directly measure correctness. High similarity can occur in superficial paraphrases or generic responses, while low similarity may occur in correct answers expressed using unusual wording or additional context. Therefore, the `topic_relevance_score` is treated as one indicator among multiple attributes, rather than as a standalone assessment metric. This design choice supports the thesis goal of multi-attribute and anti-discriminative evaluation.

### **Reproducibility**

The complete implementation (SQL queries, preprocessing logic, and similarity computation script) is included in the reproducibility package and referenced (see [Annex §8.6.2](#)). In the main thesis, only the workflow and design decisions are documented to ensure clarity and traceability.

The semantic similarity score captures topical alignment but does not indicate how the text was produced. To complement this indicator, the next section (§3.1.4) introduces an auxiliary analysis that estimates the likelihood of AI-assisted generation in student responses, which is used only as a supportive signal within the multi-attribute evaluation framework.

### 3.1.4 AI Detection in Text Responses

#### **Purpose and role in the framework**

In addition to topical alignment (§3.1.3) and log-derived behavioral indicators, this study includes an auxiliary indicator related to suspected AI-assisted text generation in student responses. The motivation is that AI-assisted writing has become increasingly common in higher education, and many institutions treat it as relevant contextual information for academic integrity and assessment practices as mentioned in §2.4. However, because automated detectors are imperfect and may produce false positives, the detector output is not interpreted as proof of misconduct. Instead, it is used as a supportive signal that complements other attributes in the multi-attribute evaluation framework.

#### **Model and output definition**

AI detection is performed using the pretrained transformer model `roberta-base-openai-detector`, applied to each student message. For each reply, the model produces a probability-like score that reflects how strongly the text resembles AI-generated content according to the detector's learned patterns. In this thesis, the reply-level detector output is transformed into a 1–10 rating, where higher values indicate higher suspected AI involvement.

#### **Preprocessing and inference procedure**

Before inference, message text is minimally cleaned by normalizing whitespace and removing punctuation to reduce noise in the input representation. Messages are processed with truncation to the model's maximum sequence length (512 tokens) to ensure consistent inference behavior for longer responses. Messages that are empty or missing are excluded from scoring.

#### **Scaling and aggregation to a student-level OAM attribute**

Let  $q_i \in [0,1]$  denote the detector score for reply  $i$ . The score is mapped to a 1–10 scale for interpretability:

$$a_i = \max(1, \text{round}(10 \cdot q_i))$$

*Equation IV AI Score Scaling*

A student-level attribute is then calculated by averaging the reply-level ratings across all valid replies written by student  $u$ :

$$AI(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} a_i$$

*Equation V Student AI Score Mean*

where  $n_u$  is the number of non-empty replies scored for student  $u$ . This aggregated value is stored as the OAM attribute `avg_AI_involvedMsg_score`.

### **Directionality and interpretation (“higher = worse”)**

Within the OAM, the AI-related attribute is defined with negative directionality (higher values are treated as less favorable). This design choice reflects a conservative assessment stance in which higher suspected AI involvement is treated as a potential risk factor for authenticity rather than as evidence of learning quality. Importantly, due to the known limitations of detector reliability, this attribute is interpreted cautiously and only in combination with other indicators (e.g., topic relevance, participation patterns, and correctness checks). It is therefore used to support transparency and review, not to justify automatic punitive decisions.

### **Limitations**

AI-detection models can be sensitive to domain shift, language proficiency, and writing style, which may lead to false positives or false negatives. For this reason, the detector score is treated as an auxiliary signal; results are presented with appropriate warnings, and the overall ranking relies on multiple attributes rather than the AI score alone. Further validation across cohorts, tasks, and writing conditions is identified as future work §6.

**Implementation note:** The full script (data cleaning, model inference, and database export) is included in the reproducibility package and referenced in the ([Annex §8.6.1](#)).

### 3.1.5 OAM Construction/Ranking & Attributes

To represent student performance in a reproducible and machine-processable form, the extracted indicators are organised into an Object–Attribute Matrix (OAM). In the OAM, each row corresponds to a student (object), and each column corresponds to a derived indicator (attribute). This structure supports multi-attribute comparison and provides the standardised input for the subsequent COCO Y0 analysis (§3.1.6). The OAM design follows the similarity-analysis logic where object-level comparison is performed through a direction-aware attribute set.

A total of 29 attributes (A1–A29) are included and grouped into two conceptual perspectives: diligence/behavioural engagement (e.g., participation frequency, activity regularity, response latency, editing behaviour) and understanding/content-related indicators (e.g., topic alignment, task correctness, citation usage, and an auxiliary AI-detection signal). The complete attribute dictionary is shown in Table 3.1.5- I.

Attr_ID	Attribute	Group	Type	Direction	Unit
A1	totalPosts	Diligence (Quantitative)	integer	0 ↑	count (posts)
A2	activeDays	Diligence (Quantitative)	integer	0 ↑	count (days)
A3	total_replies_to_prof	Diligence (Quantitative)	integer	0 ↑	count (replies)
A4	total_characters	Diligence (Quantitative)	integer	0 ↑	count (characters)
A5	total_words	Diligence (Quantitative)	integer	0 ↑	count (words)
A6	avg_words	Diligence (Quantitative)	decimal	0 ↑	words per post
A7	unique_interactions	Diligence (Quantitative)	integer	0 ↑	count (unique users)
A8	unique_discussions	Diligence (Quantitative)	integer	0 ↑	count (threads)
A9	engagement_rate	Diligence (Quantitative)	decimal	0 ↑	ratio (unitless)
A10	normalized_score	Diligence (Quantitative)	decimal	0 ↑	score (unitless)
A11	deadline_exceeded_posts(Quasi exam I)	Diligence (Quantitative)	integer	1 ↓	count (late posts)
A12	deadline_exceeded_posts(Quasi exam II)	Diligence (Quantitative)	integer	1 ↓	count (late posts)
A13	deadline_exceeded(Quasi exam III)	Diligence (Quantitative)	integer	1 ↓	count (late posts)
A14	Pattern_followed(quasi exam i)	Understanding (Qualitative)	integer	0 ↑	count (0–2)
A15	avg_AI_involvedMsg_score	Understanding (Qualitative)	integer	1 ↓	rating (1–10)
A16	consistency_score	Diligence (Quantitative)	integer	0 ↑	score (unitless)
A17	topic_relevance_score	Understanding (Qualitative)	integer	0 ↑	score (0–100)
A18	citation_count	Understanding (Qualitative)	integer	0 ↑	count (citations/links)
A19	max_streak	Diligence (Quantitative)	integer	0 ↑	count (days)
A20	avg_charcount	Diligence (Quantitative)	decimal	0 ↑	chars per post
A21	max_charcount	Diligence (Quantitative)	integer	1 ↓	count (characters)
A22	min_charcount	Diligence (Quantitative)	integer	0 ↑	count (characters)
A23	avg_reply_time	Diligence (Quantitative)	decimal	1 ↓	hours
A24	valid_response	Understanding (Qualitative)	binary	0 ↑	binary (0/1)
A25	modification_count	Diligence (Quantitative)	integer	0 ↑	count (edits)
A26	avg_modified_time_minutes	Diligence (Quantitative)	decimal	0 ↑	minutes
A27	response_time_in_hours(Task-I)	Diligence (Quantitative)	integer	1 ↓	hours
A28	response_time_in_hours(Task-II)	Diligence (Quantitative)	integer	1 ↓	hours
A29	average_posts_per_day	Diligence (Quantitative)	decimal	0 ↑	posts per day

Table 3.1.5- I Attribute dictionary and direction rules. (Source: Own Presentation)

**Legend:** Direction = where 0 means the more, the more ideal and 1 means the less, the more ideal; the table contains 29 attributes: 24 diligence and 5 understanding.

In addition to constructing a single combined OAM containing all 29 attributes, the indicators were also separated into two category-specific matrices—one containing only diligence-related attributes and one containing only understanding-related attributes. This separation supports sensitivity-oriented interpretation by allowing the ranking outcomes of each dimension to be examined independently and compared against the composite ranking produced by the full

OAM. The resulting rankings (combined OAM, diligence-only OAM, and understanding-only OAM see [Annex §8.6.1 sheet: diligence, understanding](#)) were then compared with the peer-review evaluation conducted within the classroom exclusively among students, enabling an empirical check of how closely log-derived, objective rankings align with cohort-based subjective judgements across both the full attribute set and the category-specific subsets. Table 3.1.5- II presents the table header rows and their uses.

Field / row	Meaning	Where used
<b>userid</b>	Pseudonymised student identifier	OAM row identity
<b>attribute_name</b>	attribute column labels	documentation + traceability
<b>attribute_id (A1...A29)</b>	stable attribute identifiers	COCO input + reproducibility
<b>type</b>	integer / decimal	parsing + validation
<b>direction</b>	0=↑ more ideal, 1=↓ more ideal	ranking logic + COCO
<b>Y</b>	constant target/label column (e.g., 1000)	COCO engine input format requirement

Table 3.1.5- II Description of dataset header fields and their analytical roles in the proposed workflow. (Source: Own Presentation)

The attribute set (A1–A29) was selected according to four design principles to ensure methodological coherence and practical usability. First, observability: each attribute must be derivable from Moodle forum logs or from clearly defined text-processing outputs, ensuring that measurement is grounded in recorded behaviour. Second, interpretability and actionability: indicators are chosen to support pedagogically meaningful interpretation and to enable actionable feedback (e.g., regularity of participation, response latency, topical alignment). Third, coverage of complementary performance dimensions: the diligence–understanding split operationalises multi-faceted assessment and avoids reliance on a single proxy variable. Fourth, redundancy control: where multiple indicators capture related aspects (e.g., counts vs rates), attributes are retained only when they add distinct informational value or robustness to noise in any single measure.

Because LMS log data may contain incomplete records (e.g., students who did not participate in a specific thread, tasks with no reply, or absent events such as edits), a consistent rule for missingness is required. Missing attribute values are imputed with zero when the absence of an event has a clear behavioural interpretation (e.g., zero late posts, zero citations, zero modifications, or no submission for a task). This decision preserves the semantics of non-participation as observable evidence and ensures that the OAM remains rectangular (one value per student per attribute), which is necessary for deterministic ranking and COCO processing. To reduce the risk of conflating true zeros with data artefacts, integrity checks described in §3.1.2 are applied (including duplicate handling and removal of empty threads). For message-level indicators (e.g., per-reply topic relevance, per-message AI score), values are aggregated

to the student level using defined aggregation rules (e.g., arithmetic means over available messages); if a student has no eligible messages for a given indicator, the student-level value is set to 0 under the same non-event interpretation. This imputation choice is treated as a modelling assumption particularly for students with minimal participation.

### 3.1.5.1 Attribute Direction

A key requirement for objective ranking is that each attribute has an explicit direction of improvement (direction vector). In this thesis, direction = 0 denotes a benefit-type attribute (higher values are more favourable), while direction = 1 denotes a cost-type attribute (lower values are more favourable). In line with the methodological decision described in §3.1.4, the AI-related attribute (A15) is treated as cost-type (higher = worse) because it is interpreted as a risk signal rather than a direct indicator of learning quality. For transparency, non-obvious direction assignments are explicitly motivated. First, the AI-related signal (A15) is encoded as cost-type to prevent a scenario where a high AI-likelihood score could improve ranking; the indicator is used to support interpretation and potential review, not automated punishment. Second, maximum post length (A21) is treated as cost-type in this context because excessive single-message length can reflect verbosity or low conciseness in time-constrained, descriptive tasks, and should not be assumed to be beneficial as a performance proxy. These choices are context-dependent modelling decisions and are revisited in the discussion §4.

Code in OAM	Meaning	Interpretation
0	Benefit-type (↑ more ideal)	Higher values indicate better performance
1	Cost-type (↓ more ideal)	Lower values indicate better performance

*Table 3.1.5- III Direction codes and the corresponding decision rules used to interpret attribute influence within the evaluation framework. (Source: Own Presentation)*

### 3.1.5.2 Baseline Target Value

In the exported OAM, the final column  $Y$  is set to a constant value for all students.

$$Y = 1000$$

*Equation VI Baseline Target Value*

This constant functions as a baseline performance index required by the downstream COCO Y0 workflow, which applies optimisation-based comparison logic and expects a stable, positive reference level. Setting a strictly positive baseline avoids representational problems that may arise when negative values occur (e.g., sign changes that are difficult to interpret

consistently in linear/optimisation settings). In this thesis, the constant target  $Y = 1000$  is interpreted as a fixed baseline expressed in COCO index units rather than as an observed grade or externally measured outcome see Table 8.5- II.

The choice of 1000 is pragmatic: it is sufficiently large to provide numerical headroom for increases/decreases in derived scores while remaining easy to interpret. Larger baselines (e.g., 10,000 or 1,000,000) could increase numeric granularity in intermediate computations, but would not materially change the ordinal comparison structure; they primarily scale magnitudes rather than the comparative ordering.

Importantly,  $Y$  is not used as an external ground-truth label (such as exam grades). Instead, it is a fixed reference value that enables consistent processing of the OAM within the COCO engine. See Table 3.1.5- IV

id	userid	username/attribute_name	direction	0	0	0	0	1	1	0	Y
			type	x	x	x	x	x	x		
		Value	integer	integer	integer	integer	integer	integer	integer	decimal	
		attribute_id	A1	A2	A3	A4	A27	A28	A29		
		unit	count (posts)	count (days)	count (replies)	count (characters)	hours	hours	hours	count (posts per day)	
		totalPosts	activeDays	total_replies_to_prof	total_characters	response_time_in_hours(Task-I)	response_time_in_hours(Task-II)	response_time_in_hours(Task-III)	average_posts_per_day		
1	1	student_1	16	3	0	8697	0	0	0	5.3	1000
2	2	student_2	16	5	13	5493	34	34	34	3.2	1000
3	3	student_3	10	3	8	1841	32	32	32	3.3	1000
4	4	student_4	22	8	19	14790	12	16	16	2.8	1000
5	5	student_5	8	4	7	3666	15	0	0	2.0	1000
6	6	student_6	37	10	23	10778	14	14	14	3.7	1000
7	7	student_7	14	6	11	3696	12	12	12	2.3	1000
8	8	student_8	18	5	15	7140	15	15	15	3.6	1000
9	9	student_9	11	4	8	3015	16	16	16	2.8	1000
10	10	student_10	17	5	10	2191	13	13	13	3.4	1000
11	11	student_11	3	1	2	383	0	0	0	3.0	1000
12	12	student_12	14	7	9	3422	0	0	0	2.0	1000
13	13	student_13	2	1	2	1452	0	0	0	2.0	1000
14	14	student_14	3	3	2	1308	0	0	0	1.0	1000
15	15	student_15	27	12	15	5429	34	35	35	2.3	1000
16	16	student_16	17	6	14	8260	10	15	15	2.8	1000
17	17	student_17	21	8	17	6484	10	16	16	2.6	1000
18	18	student_18	12	4	9	2852	16	13	13	3.0	1000
19	19	student_19	16	6	11	7209	18	18	18	2.7	1000
20	20	student_20	20	6	17	7590	11	15	15	3.3	1000
21	21	student_21	17	6	14	4696	14	15	15	2.8	1000
22	22	student_22	20	4	17	9436	7	8	8	5.0	1000
23	23	student_23	4	3	1	414	0	0	0	1.3	1000
24	24	student_24	1	1	1	51	0	0	0	1.0	1000

Table 3.1.5- IV Object attribute matrix. (Source: Own Presentation Annex §8.6.1 sheet: OAM\_ALL\_ATT)

**Legend:** Direction = where 0 means the more, the more ideal and 1 means the less, the more ideal;  $Y = 1000$  denotes the fixed COCO baseline.

Before processing, the OAM is transformed into a rank-based representation to improve comparability across heterogeneous measurement scales (counts, hours, ratios, NLP-derived scores) using the built-in =RANK() function. For each attribute  $A_j$ , students are ordered according to the attribute's predefined directionality (benefit-type vs cost-type), yielding for each student  $i$  a rank  $R_{ij}$  that indicates the student's relative position within the cohort for that attribute.

id	userid	direction	type	Value	attribute_id	unit	username/attribute_name
1	1	student_1	11	18	24	4	
2	2	student_2	11	11	10	10	
3	3	student_3	18	18	16	19	
4	4	student_4	3	3	2	1	
5	5	student_5	19	14	18	14	
6	6	student_6	1	2	1	2	
7	7	student_7	14	6	11	13	
8	8	student_8	7	11	6	8	
9	9	student_9	17	14	16	16	
10	10	student_10	8	11	13	18	
11	11	student_11	21	22	19	23	
12	12	student_12	14	5	14	15	
13	13	student_13	23	22	19	20	
14	14	student_14	21	18	19	21	
15	15	student_15	2	1	6	11	
16	16	student_16	8	6	8	5	
17	17	student_17	4	3	3	9	
18	18	student_18	16	14	14	17	
19	19	student_19	11	6	11	7	
20	20	student_20	5	6	3	6	
21	21	student_21	8	6	8	12	
22	22	student_22	5	14	3	3	
23	23	student_23	20	18	22	22	
24	24	student_24	24	22	22	24	

Table 3.1.5- V Excel formula used to generate the ranked Object-Attribute Matrix (ranked OAM) from the base OAM by applying the project’s rule-based scoring and ordering logic. (Source: Own Presentation, Annex §8.6.1 sheet: Rank\_All)(Unit: Ranked ordinal values)

**Legend:** the red highlighted formula cell illustrates the Excel rule used to derive the ranked OAM from the base OAM; purple cells denote ranked ordinal values generated by that transformation.

Formally, for a given attribute  $A_j$  with student values  $\{x_{1j}, x_{2j}, \dots, x_{nj}\}$ , the rank is computed as:

$$r_{ij} = Rank(x_{ij}, \{x_{1j}, \dots, x_{nj}\}, d_j)$$

Equation VII Direction-Aware Rank Formula

where  $d_j$  is the direction code stored in the OAM header. In the Excel implementation, direction = 0 applies descending ranking (largest value receives rank 1), while direction = 1 applies ascending ranking (smallest value receives rank 1). This direction-aware ranking provides two methodological advantages. First, it ensures that attributes remain interpretable under a uniform “lower rank = more ideal” convention across all columns, despite differing units and scales. Second, it reduces the influence of magnitude difference between attributes (e.g., hours vs counts) by converting them into a common ordinal representation while preserving the intended direction of improvement.

The resulting rank matrix creates a structured, standardised input that supports the next stage of analysis: COCO Y0 anti-discriminative evaluation, where students are assessed individually and comparatively based on the multi-attribute evidence encoded in the OAM (§3.1.6). See Table 3.1.5- VI.

id	userid	direction	0				1		1		0	Y
		type	x	x	x	x	x	x	x	x		
Value		integer	integer	integer	integer	integer	integer	integer	integer	decimal		
attribute_id		A1	A2	A3	A4	A27	A28	A29	A29			
username/attribute_name		totalPosts	activeDays	total_replies_to_prof	total_characters	response_time_in_hours(Task-I)	response_time_in_hours(Task-II)	average_posts_per_day				
1	1	student_1	11	18	24	4	1	1	1	1000		
2	2	student_2	11	11	10	10	23	23	8	1000		
3	3	student_3	18	18	16	19	22	22	8	1000		
4	4	student_4	3	9	2	1	18	12	13	1000		
5	5	student_5	19	14	16	14	18	7	19	1000		
6	6	student_6	1	2	1	2	15	13	3	1000		
7	7	student_7	14	6	11	13	12	10	17	1000		
8	8	student_8	7	11	6	8	15	14	4	1000		
9	9	student_9	17	14	16	16	18	18	13	1000		
10	10	student_10	8	11	13	18	14	11	8	1000		
11	11	student_11	21	22	19	23	1	1	9	1000		
12	12	student_12	14	5	14	15	1	1	19	1000		
13	13	student_13	23	22	19	20	1	1	19	1000		
14	14	student_14	21	18	19	21	1	1	23	1000		
15	15	student_15	2	1	6	11	23	24	18	1000		
16	16	student_16	8	6	8	5	9	14	11	1000		
17	17	student_17	4	3	3	9	9	18	16	1000		
18	18	student_18	16	14	14	17	20	11	8	1000		
19	19	student_19	11	6	11	7	21	21	15	1000		
20	20	student_20	8	6	8	6	11	14	6	1000		
21	21	student_21	8	6	8	12	15	14	11	1000		
22	22	student_22	5	14	3	3	8	9	9	1000		
23	23	student_23	20	18	22	22	1	1	22	1000		
24	24	student_24	24	22	22	24	1	1	23	1000		

Table 3.1.5- VI Ranked OAM. (Source: Own Presentation, Annex §8.6.1) (Unit: Ranked ordinal values)

**Legend:** purple cells denote ranked ordinal values in the ranked OAM; Y = 1000 denotes the fixed COCO baseline.

All computations, intermediate steps, and derived OAM values are fully documented and verifiable in the accompanying Excel workbook (see [Annex §8.6.1 sheet: OAM\\_ALL\\_ATT](#)), which contains the complete pipeline outputs and the formulas used to reproduce each attribute and ranking result.

### 3.1.6 COCO Y0 Analysis & Attribute Exclusion

This study applies the COCO Y0 variant to evaluate student performance profiles under a constant reference target rather than an externally observed dependent variable. The Y0 setting is appropriate in contexts where a meaningful “true” outcome variable (e.g., exam grade, price) is unavailable or intentionally not used, and the analytical goal is to assess general similarity, coherence, and comparability of objects described by multiple attributes. In the present case, the rank-transformed OAM constructed in §3.1.5 provides the model input, and COCO is used as an optimisation-based, direction-aware mechanism to derive a consistent ordering of students from heterogeneous indicators while avoiding direct dependence on raw measurement units. This aligns with the similarity-analysis framing where objects are compared through direction-controlled attributes and stepwise monotonic structures (Pitlik, 2014).

### 3.1.6.1 COCO Y0 inputs, configuration, and execution

COCO Y0 was executed using the web-based interface provided by the MYX/MIAU environment (see [Annex §8.6.1 sheet: Model\\_All](#)). The input to COCO consisted of the ranked OAM matrix (all attributes, understanding attributes, diligence attributes, and excluded attributes), submitted as a numeric matrix without header labels (attribute names), using the same attribute ordering as the OAM catalogue (A1–A29) see Figure 3- I. The dataset contained 24 objects (students) and 29 attributes, and the target column was set to the constant baseline  $Y = 1000$  for every object, following the baseline rationale defined in §3.1.5.

Exact input and its outputs can be found from EXCEL file in ([Annex §8.6.1 sheet: Rank\\_All](#)).

In the configuration, the number of staircase steps (“stairs”) was equal to the number of objects (24 stairs for 24 students). Using one step per object follows the default tool logic and maximises the flexibility of the fitted stepwise (staircase) representation. In this run, negative steps were disabled, and the offset parameter remained at the default value (offset = 0). The additional function-symmetry and inverse-run checks (used to assess stability and directional consistency) are reported separately in §3.1.7 and therefore are not expanded here.

For reproducibility, each COCO execution is uniquely traceable via the run identifier generated by the web tool, and the corresponding ranked input matrix and exported output tables are stored in ([Annex §8.6.1 sheet: Model\\_All](#)).

### 3.1.6.2 Output structure and interpretation

COCO Y0 run produces a structured output with four main components (see [Annex §8.6.1 sheet: Model\\_All](#)):

- Object ranking table: the model-consistent ordering of objects, accompanied by each object’s attribute-side positions under the fitted structure.
- Staircase-function tables (stairs(1) and stairs(2)): diagnostic representations of the fitted step structure per attribute under the chosen number of stairs.
- COCO estimation table: the core results table reporting the estimated values (“Estimation”) and related diagnostics (e.g., the difference between estimation and the constant target, shown as delta-type quantities).
- Global diagnostic summary: aggregate indicators of model fit and internal consistency (e.g., step totals and discrepancy summaries).

In this thesis, the primary quantitative output used for comparative evaluation is the COCO-estimated value reported in the estimation table. These estimated values represent a model-consistent performance index produced under the constant baseline ( $Y = 1000$ ) and the fitted staircase structure. Interpretation is therefore comparative: students are evaluated by the ordering induced by COCO (ranking and estimated indices), rather than by any single raw attribute.

Because the  $Y$  vector is constant in  $Y_0$ , COCO is not performing supervised prediction and does not attempt to reproduce external grades. Instead, the informative content lies in how well each object's ranked attribute profile can be represented under a common monotone, stepwise structure. Put differently, COCO  $Y_0$  operationalises compatibility with the cohort's multi-attribute pattern: objects whose profiles can be represented with smaller inconsistencies relative to the constant target are positioned accordingly in the output ordering (Pitlik, 2014). The role of the staircase tables is diagnostic: they provide a structured lens for identifying attributes that behave informatively or degenerately under the fitted step structure, which is used directly in the exclusion procedure described next. The full description of the use of the COCO  $Y_0$  engine can be found in [Annex §8.6.6](#) of the thesis.

### 3.1.6.3 Attribute exclusion based on staircase diagnostics

To support interpretability and reduce redundancy, the COCO analysis was complemented with an attribute exclusion step driven by the model's staircase diagnostics. The COCO output includes the "stairs(2)" table, which summarises the fitted step structure per attribute. In the present workflow, attributes were flagged as exclusion candidates when the staircase diagnostics indicated a degenerate / weakly informative configuration relative to the cohort size.

Operationally, the applied rule was: an attribute is marked as candidate for exclusion if the staircase diagnostic yields a top-step value  $S_1 = n - 1$ , where  $n$  is the number of objects (here  $n = 24$ , hence  $S_1 = 23$ ). In this dataset, this pattern was treated as evidence that the attribute does not introduce meaningful separation beyond what is already implied by the constant-baseline setting and the rank constraints, and therefore contributes limited additional structure to the  $Y_0$  evaluation.



evaluation output and supports principled attribute reduction through staircase diagnostics. This section therefore forms the analytical bridge between OAM construction and ranking (§3.1.5) and the validation/robustness analysis (§3.1.7) and consolidated results (§3.1.8).

### 3.1.7 Function-Symmetry Validation

Because COCO Y0 operates under a constant target value ( $Y = 1000$  for all objects) and is used here for aimless similarity rather than supervised prediction, validation must focus on internal consistency rather than goodness-of-fit to external labels. One such internal check is function-symmetry validation via an inverse run, which examines whether the model’s deviation pattern around the constant baseline behaves consistently when the ranked input is systematically inverted.

#### **Why inverse-run check is meaningful in Y0**

In the Y0 setting, informative structure is not carried by variability in the Y-vector (since it is constant), but by how each object’s ranked attribute profile can be represented by the stepwise (staircase) model under the monotonic constraints. The COCO output therefore includes object-level diagnostics that indicate whether the model’s estimated value falls slightly above or below the constant baseline. A symmetry check becomes meaningful under the following expectation: if an object’s profile is transformed into its rank-reversed counterpart, the direction of its deviation from the baseline should tend to reverse as well (up to discretisation effects, ties, and constraints). This provides a practical consistency test for the optimisation-based scoring logic, aligned with the similarity-analysis framing (Pitlik, 2014).

#### **Constructing the inverse ranked matrix**

The inverse run uses an inverted representation of the ranked OAM (the same object set and the same attribute set, but with ranks mirrored). Let  $n$  denote the number of objects (here  $n = 24$ ). For each ranked cell  $r_{ij}$  (object  $i$ , attribute  $j$ ), the inverse value is computed by:

$$r_{ij}^{inv} = (n + 1) - r_{ij}$$

*Equation VIII Rank Matrix Inversion*

With  $n = 24$ , this reduces to  $r_{ij}^{inv} = 25 - r_{ij}$ . The Y column remains constant at  $Y = 1000$  to preserve the Y0 setting. The inverted ranked matrix is then submitted to COCO Y0 using the same configuration principle as the baseline run (same number of objects and attributes; stairs equal to the object count).

### Validation signal used: Delta/Fact from the estimation table

The primary diagnostic used for this validation is the Delta/Fact column from the COCO estimation table. In the COCO output, the constant baseline appears as “Fact+0” (=1000), and the model produces an “Estimation” per object. The reported deviation is:

- $Fact + 0 = 1000$
- $\Delta = Fact - Estimation$
- $\Delta/Fact = (Fact - Estimation)/Fact$

*Equation IX Delta-to-Fact Ratio*

Accordingly,  $\Delta/Fact > 0$  indicates an estimated value below the baseline (under-baseline), while  $\Delta/Fact < 0$  indicates an estimate above the baseline (over-baseline). Values equal to zero indicate an exact match to the baseline in the output resolution of the model.

### Decision rule: sign-reversal consistency under inversion

For each object  $i$ , the baseline run produces a deviation ratio  $(\frac{\Delta}{Fact})_i$ , and the inverse run produces  $(\frac{\Delta}{Fact})_i^{inv}$ . The inverse-run symmetry condition is operationalised as a sign-reversal or neutralisation rule:

$$valid_i = \begin{cases} 1, & \text{if } (\frac{\Delta}{Fact})_i * (\frac{\Delta}{Fact})_i^{inv} \leq 0 \\ 0, & \text{Otherwise.} \end{cases}$$

*Equation X Symmetry Validity Rule*

This rule treats an object as consistent (valid = 1) if the deviation sign flips between the baseline and inverse runs, or if at least one of the deviations is exactly zero (neutral). Conversely, valid = 0 indicates a same-sign deviation under inversion, interpreted as a local inconsistency relative to the expected mirror behaviour.

Figure 3.1.7- I shows the high-level operation of the function-symmetry check.

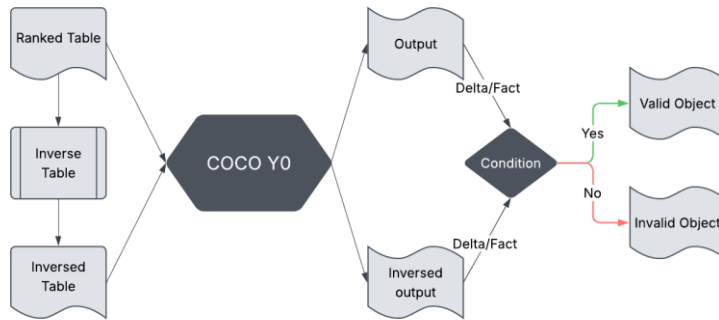


Figure 3.1.7- 1 Function-symmetry validation logic. (Source: Own Presentation).

**Legend:** grey = transformation and validation steps; green arrow = valid output; red arrow = invalid output;

### Interpretation and limitations of the symmetry test

This inverse-run check is an internal consistency diagnostic, not a guarantee of correctness. Several practical factors can prevent perfect sign-reversal even when the model is operating coherently: (i) ties in ranked attributes compress information, potentially stabilising the solution on the same side of the baseline; (ii) the staircase constraint imposes discrete step structures that may yield coarse deviation values; and (iii) Y0 optimisation can admit multiple near-equivalent solutions when the constant target does not constrain object separation strongly. For these reasons, the validation is interpreted as a robustness indicator: a higher proportion of valid objects supports the claim that the COCO Y0 evaluation behaves consistently under a deterministic inversion of the ranked input, while isolated invalid cases motivate closer inspection in the subsequent verification and results chapters.

### 3.1.8 Results

This chapter reports the results produced by the COCO Y0 engine across three configurations: (1) the full attribute set, (2) the diligence-only attribute subset, and (3) the understanding-only attribute subset. For each configuration, results are presented for the initial run and for the reduced (excluded-attribute) re-run, where the exclusion set is derived from the staircase diagnostics described in §3.1.6.3. In addition, each run is accompanied by the function-symmetry validation outcome (see §3.1.7), reported in the summary tables through the delta/fact inverse and validation columns.

For consistency across subchapters, each results block begins with a run summary table containing the COCO-estimated values (“estimation”), the relative deviation from the constant

target (delta/fact), the corresponding quantity from the inverse-run (delta/fact inverse), the derived validity flag (validation), and the resulting object ordering (rank). All outputs and intermediate computations are archived in the accompanying reproducibility workbook ([Annex §8.6.1 sheet: Model\\_All](#)). In the following result tables,  $Y$  and estimation are reported in COCO index units, delta/fact measures in %, validation as a binary flag (0/1), and rank as ordinal position; see Annex §8.5, Table 8.5- II.

#### 3.1.8.1 Full attribute set (A1–A29)

Table 3.1.8- I reports the COCO  $Y_0$  output obtained from the ranked OAM with 24 objects (students) and 29 attributes, under a constant baseline target ( $Y = 1000$ ). The resulting estimated values are highly saturated around the baseline (only 999, 1000, and 1001 appear), which yields very coarse discrimination and large tie groups in the final ordering (e.g., multiple students share rank 1, rank 9, and rank 20). This pattern should not be interpreted as a “failure” of COCO; rather, it indicates that the chosen attribute set—after rank transformation and under the monotone staircase constraints of  $Y_0$ —does not separate the cohort strongly. In practical terms, many students exhibit near-equivalent rank profiles, or several attributes become effectively identical/low-variance after ranking, so COCO can represent many objects with almost the same model-consistent estimate. This saturation is likely reinforced by the relatively high attribute-to-object ratio, since 29 ranked attributes are used to distinguish only 24 students, which can intensify low separability under the  $Y_0$  staircase constraints.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	delta/fact (deviation %)	delta/fact inverse (deviation %)	Validation (Binary 1/0)	Rank (ordinal Position)
1	student 1	1000	1000.0	0.00	0.00	1	9
2	student 2	1000	1001.0	-0.10	0.00	1	1
3	student 3	1000	999.0	0.10	0.00	1	20
4	student 4	1000	1001.0	-0.10	-0.10	0	1
5	student 5	1000	1000.0	0.00	0.00	1	9
6	student 6	1000	1000.0	0.00	0.00	1	9
7	student 7	1000	999.0	0.10	0.00	1	20
8	student 8	1000	1001.0	-0.10	0.00	1	1
9	student 9	1000	1000.0	0.00	0.00	1	9
10	student 10	1000	999.0	0.10	0.00	1	20
11	student 11	1000	1000.0	0.00	0.00	1	9
12	student 12	1000	999.0	0.10	0.00	1	20
13	student 13	1000	1001.0	-0.10	0.00	1	1
14	student 14	1000	1000.0	0.00	0.00	1	9
15	student 15	1000	1000.0	0.00	0.10	1	9
16	student 16	1000	1001.0	-0.10	-0.10	0	1
17	student 17	1000	1000.0	0.00	0.00	1	9
18	student 18	1000	999.0	0.10	0.00	1	20
19	student 19	1000	1001.0	-0.10	0.00	1	1
20	student 20	1000	1001.0	-0.10	-0.10	0	1
21	student 21	1000	1000.0	0.00	0.00	1	9
22	student 22	1000	1001.0	-0.10	0.00	1	1
23	student 23	1000	1000.0	0.00	0.00	1	9
24	student 24	1000	1000.0	0.00	0.00	1	9

Table 3.1.8- I Full attribute set. Result summary. (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 999.0–1001.0; observed direct deviation spans –0.10 to 0.10, inverse deviation –0.10 to 0.10; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

Validation outcome. In the inverse-run symmetry check, most objects satisfy the expected sign-change condition (validation = 1). A small number of cases fail the check (validation = 0), which is consistent with the limited numeric spread of the estimates in the initial full-attribute run: when the solution is highly saturated (near-constant estimates), small rounding effects and tie structures can make the inverse-run discrepancy behave non-ideally for a subset of objects.

Staircase-driven exclusion and re-run. To test whether the saturation is caused by attributes that are weakly informative under the staircase diagnostic (§3.1.6.3), the model was re-run after excluding 13 attributes flagged by the staircase rule. The excluded attributes are listed in Table 3.1.8- III (Excluded IDs: A3, A4, A7, A9, A10, A13, A14, A19, A23, A24, A25, A27, A28).

### Attribute exclusion and re-run

Table 3.1.8- II shows the results after applying the staircase-based exclusion rule and re-running COCO Y0 on the reduced attribute set. Compared to the initial run, the reduced model produces a substantially wider spread of estimated values and therefore a much more informative ordering. In other words, after removing attributes flagged as non-informative

under the staircase diagnostic, the remaining attribute set yields a sharper and more discriminative similarity structure among students.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	delta/fact (deviation %)	delta/fact inverse (deviation %)	Validation (Binary 1/0)	Rank (ordinal Position)
1	student 1	1000	1001.3	-0.13	0.13	1	11
2	student 2	1000	947.7	5.23	-5.21	1	22
3	student 3	1000	923.2	7.68	-7.66	1	24
4	student 4	1000	1053.4	-5.34	5.32	1	2
5	student 5	1000	974.8	2.52	-2.52	1	19
6	student 6	1000	1063.4	-6.34	6.32	1	1
7	student 7	1000	1022.8	-2.28	2.28	1	7
8	student 8	1000	993.3	0.67	-0.67	1	13
9	student 9	1000	947.2	5.28	-5.26	1	23
10	student 10	1000	1022.8	-2.28	2.28	1	7
11	student 11	1000	981.8	1.82	-1.07	1	18
12	student 12	1000	1041.3	-4.13	3.37	1	4
13	student 13	1000	974.3	2.57	-2.57	1	20
14	student 14	1000	1017.8	-1.78	1.78	1	10
15	student 15	1000	1001.3	-0.13	0.13	1	11
16	student 16	1000	987.3	1.27	-1.27	1	15
17	student 17	1000	984.3	1.57	-1.57	1	16
18	student 18	1000	958.7	4.13	-4.12	1	21
19	student 19	1000	1026.3	-2.63	2.63	1	6
20	student 20	1000	1047.9	-4.79	4.77	1	3
21	student 21	1000	993.3	0.67	-0.67	1	13
22	student 22	1000	1033.3	-3.33	3.32	1	5
23	student 23	1000	1019.3	-1.93	1.93	1	9
24	student 24	1000	983.3	1.67	-1.67	1	17

Table 3.1.8- II Excluded attributes result. (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 923.2–1063.4; observed direct deviation spans –6.34 to 7.68, inverse deviation –7.66 to 6.32; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

Excluded ID	Excluded attribute name
A3	total_replies_to_prof
A4	total_characters
A7	unique_interactions
A9	engagement_rate
A10	normalized_score(...)
A13	deadline_exceeded(Quasi exam III)
A14	Pattern_followed(quasi exam i)
A19	max_streak
A23	avg_reply_time
A24	valid_response
A25	modification_count
A27	response_time_in_hours(Task-I)
A28	response_time_in_hours(Task-II)

Table 3.1.8- III List of excluded attributes. (Source: Own Presentation).

Effect on ranking. The shift from the saturated initial ranking to the more differentiated excluded-run ranking should be interpreted primarily as an increase in resolution rather than a contradiction. In Y0, the COCO estimate is not a supervised prediction of an external criterion; it is a model-consistent performance index derived from the multi-attribute rank profile under the monotone staircase structure (Pitlik, 2014). When the attribute set produces limited structural separation, COCO will naturally yield near-baseline estimates and large ties; when weakly informative attributes are removed, the remaining attributes can induce a clearer monotone structure, producing wider estimate dispersion and a more interpretable ordering. For transparency, both variants (full and excluded-run full) are retained in the results, because they represent complementary views of the same cohort under (i) maximal attribute inclusion and (ii) a staircase-diagnostic parsimony constraint.

#### 3.1.8.2 Diligence attribute subset

Table 3.1.8- IV reports the COCO Y0 output using only the diligence attribute subset. Although the estimates remain centred around the constant baseline ( $Y = 1000$ ), the ordering is highly tie-dominated: 20 of 24 students share rank 1, indicating that the diligence-only rank profiles are largely homogeneous under the monotone staircase constraints, and therefore do not separate the cohort strongly in this configuration. The inverse-run function-symmetry check is satisfied for 22/24 objects; two cases are flagged as inconsistent (student\_5 and student\_12, validation = 0), which is consistent with the tie-heavy solution structure.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	delta/fact (deviation %)	delta/fact inverse (deviation %)	Validation (Binary 1/0)	Rank (ordinal Position)
1	student 1	1000	1001.0	-0.10	0.00	1	1
2	student 2	1000	1000.0	0.00	0.00	1	19
3	student 3	1000	1001.0	-0.10	0.00	1	1
4	student 4	1000	1000.0	0.00	0.00	1	18
5	student 5	1000	1001.0	-0.10	-0.10	0	1
6	student 6	1000	1001.0	-0.10	0.00	1	1
7	student 7	1000	1001.0	-0.10	0.00	1	1
8	student 8	1000	1001.0	-0.10	0.00	1	1
9	student 9	1000	1001.0	-0.10	0.00	1	1
10	student 10	1000	1001.0	-0.10	0.10	1	1
11	student 11	1000	1001.0	-0.10	0.00	1	1
12	student 12	1000	1001.0	-0.10	-0.10	0	1
13	student 13	1000	1000.0	0.00	0.00	1	10
14	student 14	1000	1001.0	-0.10	0.00	1	1
15	student 15	1000	1001.0	-0.10	0.00	1	1
16	student 16	1000	1001.0	-0.10	0.10	1	1
17	student 17	1000	1001.0	-0.10	0.00	1	1
18	student 18	1000	1001.0	-0.10	0.10	1	1
19	student 19	1000	1000.0	0.00	0.00	1	5
20	student 20	1000	1001.0	-0.10	0.00	1	1
21	student 21	1000	1001.0	-0.10	0.10	1	1
22	student 22	1000	1001.0	-0.10	0.00	1	1
23	student 23	1000	1001.0	-0.10	0.00	1	1
24	student 24	1000	1000.0	0.00	0.00	1	1

Table 3.1.8- IV Diligence attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 1000.0–1001.0; observed direct deviation spans –0.10 to 0.00, inverse deviation –0.10 to 0.10; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

### Attribute exclusion and re-run

After applying the staircase-based exclusion rule and re-running COCO (Table 3.1.8- V), the ranking becomes more informative and less saturated by a single tie-group. The best diligence profile is student\_6 (rank 1, estimation 1042.7), followed by student\_11 (rank 2, 1019.6) and student\_4 (rank 3, 1016.1), while the lowest estimates occur for student\_2 (rank 24, 964.4) and student\_19 (rank 23, 969.9). A tie-group still remains (rank 7 shared by 10 students at estimation 1003.6), but overall discrimination improves relative to the initial run, and the symmetry validation improves to 23/24 valid (only student\_16 flagged as invalid).

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	delta/fact (deviation %)	delta/fact inverse (deviation %)	Validation (Binary 1/0)	Rank (ordinal Position)
1	student 1	1000	1003.6	-0.36	0.35	1	7
2	student 2	1000	964.4	3.56	-3.53	1	24
3	student 3	1000	1003.6	-0.36	0.35	1	7
4	student 4	1000	1016.1	-1.61	0.35	1	3
5	student 5	1000	1003.6	-0.36	0.35	1	7
6	student 6	1000	1042.7	-4.27	4.64	1	1
7	student 7	1000	1003.6	-0.36	0.35	1	7
8	student 8	1000	995.5	0.45	-0.45	1	18
9	student 9	1000	1003.6	-0.36	0.35	1	7
10	student 10	1000	1003.6	-0.36	0.35	1	7
11	student 11	1000	1019.6	-1.96	2.39	1	2
12	student 12	1000	1003.6	-0.36	0.35	1	7
13	student 13	1000	1010.6	-1.06	1.05	1	5
14	student 14	1000	1003.6	-0.36	0.35	1	7
15	student 15	1000	971.4	2.86	-2.84	1	22
16	student 16	1000	998.5	0.15	0.35	0	17
17	student 17	1000	1003.6	-0.36	0.35	1	7
18	student 18	1000	1003.6	-0.36	0.35	1	7
19	student 19	1000	969.9	3.01	-2.99	1	23
20	student 20	1000	994.5	0.55	-1.04	1	19
21	student 21	1000	972.9	2.71	-2.69	1	21
22	student 22	1000	1012.6	-1.26	1.65	1	4
23	student 23	1000	1007.1	-0.71	0.70	1	6
24	student 24	1000	988.5	1.15	-1.14	1	20

Table 3.1.8- V Excluded diligence attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 964.4–1042.7; observed direct deviation spans –4.27 to 3.56, inverse deviation –3.53 to 4.64; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

Excluded ID	Excluded attribute name
A1	totalPosts
A5	total_words
A8	unique_discussions
A9	engagement_rate
A11	deadline_exceeded_posts(Quasi exam I)
A13	deadline_exceeded(Quasi exam III)
A25	modification_count
A27	response_time_in_hours(Task-I)
A28	response_time_in_hours(Task-II)
A29	average_posts_per_day

Table 3.1.8- VI List of excluded attributes for the diligence run. (Source: Own Presentation)

### 3.1.8.3 Understanding attribute subset

In contrast to diligence, the understanding-only run produces a naturally wider estimate range and a more structured ordering already in the initial configuration (Table 3.1.8- VII). The best-ranked case (student\_24) is clearly above the baseline, whereas the lowest-ranked case (student\_1) falls substantially below it. The inverse-run symmetry check is satisfied for nearly

the entire cohort, with a single exception (student\_7), indicating high internal consistency overall in this subset.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	delta/fact (deviation %)	delta/fact inverse (deviation %)	Validation (Binary 1/0)	Rank (ordinal Position)
1	student 1	1000	975.2	2.48	-2.49	1	24
2	student 2	1000	992.6	0.74	-1.00	1	19
3	student 3	1000	1012.5	-1.25	0.98	1	2
4	student 4	1000	1007.5	-0.75	0.58	1	6
5	student 5	1000	996.1	0.39	-0.51	1	14
6	student 6	1000	1009.5	-0.95	0.58	1	4
7	student 7	1000	1000.6	-0.06	-0.21	0	11
8	student 8	1000	1004.6	-0.46	0.09	1	8
9	student 9	1000	1007.5	-0.75	0.48	1	5
10	student 10	1000	1005.6	-0.56	0.29	1	5
11	student 11	1000	1009.0	-0.90	0.78	1	4
12	student 12	1000	993.1	0.69	-0.80	1	9
13	student 13	1000	979.2	2.08	-1.89	1	12
14	student 14	1000	986.2	1.38	-1.40	1	11
15	student 15	1000	1005.1	-0.51	0.48	1	4
16	student 16	1000	999.1	0.09	-0.21	1	6
17	student 17	1000	1002.1	-0.21	0.09	1	5
18	student 18	1000	1003.1	-0.31	0.19	1	4
19	student 19	1000	995.1	0.49	-1.25	1	4
20	student 20	1000	1011.5	-1.15	0.78	1	2
21	student 21	1000	1011.5	-1.15	0.78	1	2
22	student 22	1000	989.6	1.04	-1.30	1	2
23	student 23	1000	989.1	1.09	-1.10	1	2
24	student 24	1000	1015.0	-1.50	1.48	1	1

Table 3.1.8- VII Understanding attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 975.2–1015.0; observed direct deviation spans –1.50 to 2.48, inverse deviation –2.49 to 1.48; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

### Attribute exclusion and re-run

Excluding the two staircase-flagged understanding attributes (A17 topic relevance and A18 citation count) yields a moderate reordering rather than a complete reshuffle (Table 3.1.8- VIII). The top position shifts to student\_16, while student\_1 remains the lowest-ranked case, indicating partial stability at the extremes. Notably, all objects pass the inverse-run validation in the excluded understanding run (valid throughout), suggesting that the reduced understanding set forms a slightly more symmetry-consistent representation under Y0. This excluded-run is therefore best interpreted as a sensitivity/robustness variant of the understanding ranking, rather than a necessary step to “make the model work.”

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	delta/fact (deviation %)	delta/fact inverse (deviation %)	Validation (Binary 1/0)	Rank (ordinal Position)
1	student 1	1000	984.8	1.52	-1.52	1	24
2	student 2	1000	994.8	0.52	-0.52	1	17
3	student 3	1000	993.8	0.62	-0.62	1	19
4	student 4	1000	1014.8	-1.48	1.48	1	2
5	student 5	1000	1000.8	-0.08	0.08	1	12
6	student 6	1000	1000.8	-0.08	0.08	1	12
7	student 7	1000	995.8	0.42	-0.42	1	16
8	student 8	1000	1007.8	-0.78	0.78	1	5
9	student 9	1000	1005.8	-0.58	0.58	1	6
10	student 10	1000	985.8	1.42	-1.42	1	23
11	student 11	1000	1001.8	-0.18	0.18	1	10
12	student 12	1000	994.8	0.52	-0.52	1	17
13	student 13	1000	1001.8	-0.18	0.18	1	10
14	student 14	1000	986.8	1.32	-1.32	1	22
15	student 15	1000	1004.8	-0.48	0.48	1	7
16	student 16	1000	1015.8	-1.58	1.58	1	1
17	student 17	1000	1004.8	-0.48	0.48	1	7
18	student 18	1000	998.8	0.12	-0.12	1	15
19	student 19	1000	1002.8	-0.28	0.28	1	9
20	student 20	1000	1009.8	-0.98	0.98	1	4
21	student 21	1000	993.8	0.62	-0.62	1	19
22	student 22	1000	999.8	0.02	-0.02	1	14
23	student 23	1000	987.8	1.22	-1.22	1	21
24	student 24	1000	1011.8	-1.18	1.18	1	3

Table 3.1.8- VIII Excluded understanding attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 984.8–1015.8; observed direct deviation spans –1.58 to 1.52, inverse deviation –1.52 to 1.58; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

Table 3.1.8- IX lists the attributes excluded in the understanding-only exclusion run.

Excluded ID	Excluded attribute name
A17	topic_relevance_score
A18	citation_count

Table 3.1.8- IX List of excluded attributes for the understanding run. (Source: Own Presentation)

### 3.1.8.4 Objective vs Subjective

To contextualize the COCO-derived student ordering against human judgement, the objective rankings produced by the COCO Y0 workflow were compared with a peer-based subjective evaluation conducted within the classroom (subjective rank; source: [Annex §8.6.4](#)). The objective side comprises six rank outputs derived from the ranked OAM: (i) full attribute run, (ii) full attribute run after staircase-based exclusion, (iii) diligence-only run, (iv) diligence-only run after exclusion, (v) understanding-only run, and (vi) understanding-only run after exclusion (see §3.1.6–§3.1.8.3 for the corresponding run logic and exclusion rule).

### Rank harmonization and correlation procedure (tie-aware)

Several COCO outputs contained ties (multiple students receiving the same integer rank). Because correlation on ranks is sensitive to tie treatment, all rank columns were transformed using average ranking (Excel RANK.AVG) before comparison. This step assigns the mean of the occupied rank positions to each tied case, yielding a tie-consistent ordinal representation. Subsequently, Spearman rank correlation was computed by applying Excel CORREL directly to the tie-adjusted rank vectors (i.e., Pearson correlation over ranks), which is equivalent to Spearman’s  $\rho$  when the inputs are ranks.

Table 3.1.8- X presents the ranking values turned into tie-handled ranked values using Excel built-in =RANK.AVG() function.

username	rank full	rank full excluded	rank diligence	rank diligence excluded	rank understanding	rank understanding excluded	rank subjective
student 1	14	12	11	12	24	24	9
student 2	5	22	24	24	23	18	5
student 3	22	24	11	12	4	20	10
student 4	5	2	23	3	16	2	11
student 5	14	19	11	12	22	13	21
student 6	14	1	11	1	9	13	16
student 7	22	8	11	12	20	16	20
student 8	5	14	11	18	17	5	23
student 9	14	23	11	12	13	6	3
student 10	22	8	11	12	13	23	15
student 11	14	18	11	2	9	11	2
student 12	22	4	11	12	18	18	7
student 13	5	20	22	5	21	11	12
student 14	14	10	11	12	20	22	24
student 15	14	12	11	22	9	8	22
student 16	5	15	11	17	16	1	18
student 17	14	16	11	12	13	8	4
student 18	22	21	11	12	9	15	13
student 19	5	6	21	23	9	9	19
student 20	5	3	11	19	4	4	17
student 21	14	14	11	21	4	20	6
student 22	5	5	11	4	4	14	8
student 23	14	9	11	6	4	21	1
student 24	14	17	11	20	1	3	14

Table 3.1.8- X Tie-handled rank matrix (RANK.AVG-transformed ranks) for objective and subjective comparisons. (Source: Own Presentation, Annex §8.6.1)(Unit: Ranked ordinal values)

**Legend:** rank\_full, rank\_full\_excluded, rank\_diligence, rank\_diligence\_excluded, rank\_understanding, and rank\_understanding\_excluded = RANK.AVG-transformed ordinal positions under each model variant; rank\_subjective = subjective comparison rank; lower rank values indicate stronger positions (green cells), while higher rank values indicate weaker positions (red cells).

### Empirical agreement patterns

The resulting correlations indicate limited monotonic agreement between the objective (COCO-based) rankings and the peer-based subjective ordering Table 3.1.8- XI:

- Full attribute (initial) vs subjective:  $\rho = -0.097$  (near-zero association).
- Full attribute (excluded-run) vs subjective:  $\rho = -0.244$  (weak negative association).

- Understanding-only vs subjective:  $\rho = 0.207$  (weak positive association).
- Understanding-only (excluded-run) vs subjective:  $\rho = -0.202$  (weak negative association after exclusion).
- Diligence-only vs subjective:  $\rho = -0.077$  (near-zero association).
- Diligence-only (excluded-run) vs subjective:  $\rho = 0.307$  (weak-to-moderate positive association; the strongest observed alignment with subjective ranks in this dataset).

Two additional results help interpret these outcomes as construct differences rather than simple “success/failure” of the model:

Understanding vs diligence (objective–objective):  $\rho = 0.325$ , suggesting that the two objective perspectives share some common ordering signal, but remain meaningfully distinct. This is consistent with the design choice in §3.1.5, where diligence and understanding attributes intentionally represent different behavioral/quality dimensions.

Stability of rankings before vs after exclusion Table 3.1.8- XII:

- Full vs full(excluded):  $\rho = 0.132$
- Diligence vs diligence(excluded):  $\rho = 0.084$
- Understanding vs understanding(excluded):  $\rho = 0.168$

Comparison	Spearman's $\rho$ (unitless)
All attributes (full) vs Subjective	-0.09
All attributes (excluded-run) vs Subjective	-0.24
Understanding vs Subjective	0.20
Understanding (excluded-run) vs Subjective	-0.20
Diligence vs Subjective	-0.07
Diligence (excluded-run) vs Subjective	0.30
Understanding vs Diligence	0.32
All attributes (full) vs Understanding	-0.07
All attributes (full) vs Diligence	-0.55

Table 3.1.8- XI Spearman correlations between objective (COCO) and subjective rank variants. (Source: Own Presentation, Annex §8.6.1)

Comparison	Spearman's $\rho$ (unitless)
Full vs Full(excluded-run)	0.13
Diligence vs Diligence(excluded-run)	0.08
Understanding vs Understanding(excluded-run)	0.16

Table 3.1.8- XII Correlation between Full attribute set and excluded attribute set. (Source: Own Presentation, Annex §8.6.1)

These low correlations imply that the staircase-based exclusion step can materially reshape the ordering. Therefore, excluded-run results should be interpreted as a parsimonious alternative view (driven by the diagnostic rule), not as a minor refinement of the original ranking.

### **Interpretation and methodological implications**

The weak associations between objective and subjective rankings are plausible given that the two assessments are not guaranteed to operationalize the same concept of “performance.” In the COCO Y0 setting, the output ordering reflects how compatible each student’s multi-attribute profile is with a monotonic stepwise representation under a constant baseline, given the chosen attribute set and direction rules (§3.1.6). In contrast, peer evaluation may prioritize perceived content quality, collaboration, or contextual knowledge that is only partially captured by log-derived indicators. Consequently, low or mixed correlations do not automatically invalidate either approach; rather, they indicate that the objective pipeline captures a performance construct that is only partially overlapping with human judgement in this cohort.

Finally, because the ranking changes notably after exclusion, the study reports both the full and excluded-run variants in §3.1.8 as complementary outputs. The next section (§3.1.9) shifts from comparative validity (objective vs subjective) to technical reliability, by testing the COCO pipeline under deliberately inconsistent/extreme inputs to verify whether the system behaves robustly when the input context is distorted.

The comparative results in §3.1.8 establish the methodological behavior of the proposed evaluation logic under full and reduced attribute configurations, including stability checks through inverse-run validation. However, methodological validity alone does not guarantee operational usability in educational practice. For this reason, §3.2 transitions from analytical outcomes to software realization: it documents how the validated pipeline was implemented as a reproducible demo application, how run-level evidence is persisted in a structured database, and how instructors can execute, inspect, and export each stage of the workflow in a transparent, auditable manner.

#### **3.1.9 Input Verification Test**

Because COCO Y0 is applied in a Y0 (constant-target) setting, evaluation quality cannot be assessed against an external ground-truth label (since Y is fixed). Instead, credibility depends on whether the input matrix is internally consistent and behaviourally plausible. For this reason, an input verification (sensitivity) test was conducted by intentionally distorting the COCO

input and checking whether the output becomes unstable or non-informative. The underlying expectation is that (i) plausible, exclusion-refined inputs yield interpretable rankings, while (ii) implausible or structurally inconsistent inputs lead to recognizable failure modes (e.g., erratic estimations or collapsed/flat results).

All tests were executed using the excluded (staircase-refined) input configuration as the baseline, and distortions were applied relative to that baseline to ensure comparability across cases. The verification tables use the same result-scale conventions as §3.1.8, namely *Y* and estimation in COCO index units, validation as a binary flag (0/1), and rank as ordinal position; see Annex §8.5, Table 8.5- II.

### Test V1: Random excessively high values (range distortion)

A subset of input cells was replaced with unrealistically large values (“too high” inputs), violating the intended cohort-scale comparability of the ranked matrix (Table 3.1.9- I).

Expected behaviour. The output should become unstable, reflected by irregular estimations and inconsistent separation, indicating that the input no longer represents a realistic multi-attribute cohort structure.

Observed behaviour (Table 3.1.9- II). The COCO output became erratic and did not yield a clean, interpretable result, confirming sensitivity to range distortions.

	direction	0	0	0	0	0	1	0	
type	x	x	x	x	x	x	x	x	
Value	integer	integer	integer	integer	integer	integer	integer	decimal	
attribute_id	A1	A2	A3	A4	A5	A28	A29		
unit	count (posts)	count (days)	count (replies)	count (characters)	count (words)	hours	count (posts per day)		
userid	username/attribute_name	totalPosts	activeDays	total_replies_to_prof	total_characters	total_words	response_time_in_hours(Task-II)	average_posts_per_day	Y
1 student_1		0	0	2222222	8697	1473	0	5.3	1000
2 student_2		16	0	13	5493	234234234	34	3.2	1000
3 student_3		10	0	8	1841	352	32	3.3	1000
4 student_4		22	0	19	14790	2432	16	2.8	1000
5 student_5		8	0	7	3666	528	0	2.0	1000
6 student_6		10	10	10	10	10	10	10.0	1000
7 student_7		0	0	11	3696	610	12	2.3	1000
8 student_8		18	0	15	7140	1163	15	3.6	1000
9 student_9		11	0	8	3015	433	16	2.8	1000
10 student_10		222323232	0	10	2191	352	13	3.4	1000
11 student_11		3	0	222222	383	74	0	3.0	1000
12 student_12		14	0	9	3422	518	0	2.0	1000
13 student_13		2	0	2	1452	275	0	2.0	1000
14 student_14		3	0	2	1308	118	0	1.0	1000
15 student_15		2324342	0	15	5429	911	35	2.3	1000
16 student_16		17	0	14	8260	1373	15	66666.0	1000
17 student_17		21	0	17	6484	1216	16	2.6	1000
18 student_18		4432323	0	9	2852	364	13	3.0	1000
19 student_19		16	0	11	7209	1123	18	2.7	1000
20 student_20		10000000	10000000	10000000	10000000	10000000	10000000	10000000.0	1000
21 student_21		17	0	14	4696	758	15	2.8	1000
22 student_22		20	0	17	9436	1551	8	5.0	1000
23 student_23		23232333	0	1	414	84	0	1.3	1000
24 student_24		1	0	1	51	12	0	1.0	1000

Table 3.1.9- I Distorted input matrix. (Source: Own Presentation, Annex §8.6.1)

**Legend:** direction: 0 means the more, the more ideal and 1 means the less, the more ideal; Y = 1000 denotes the fixed COCO baseline; the table illustrates a deliberately distorted input matrix with extreme outlier values used to test model robustness.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	Validation (Binary 0/1)	Rank (Ordinal Position)
1	student 1	1000	1002.5	1	5
2	student 2	1000	958.4	1	24
3	student 3	1000	1002.5	1	5
4	student 4	1000	1002.5	1	5
5	student 5	1000	1002.5	1	5
6	student 6	1000	1046.6	1	2
7	student 7	1000	1002.5	1	5
8	student 8	1000	973.9	1	22
9	student 9	1000	1002.5	1	5
10	student 10	1000	1051.1	1	1
11	student 11	1000	1013.0	1	4
12	student 12	1000	1002.5	1	5
13	student 13	1000	992.0	1	18
14	student 14	1000	1002.5	1	5
15	student 15	1000	1002.5	1	5
16	student 16	1000	1002.5	1	5
17	student 17	1000	1002.5	1	5
18	student 18	1000	981.5	1	20
19	student 19	1000	970.4	1	23
20	student 20	1000	1002.5	1	5
21	student 21	1000	975.4	1	21
22	student 22	1000	1002.5	1	5
23	student 23	1000	1021.6	1	3
24	student 24	1000	983.5	1	19

Table 3.1.9- II COCO output (erratic estimates / unstable separation). (Source: Own Presentation, Annex §8.6.1)

**Legend:** Estimation spans 958.4–1051.1, corresponding to –41.6 to +51.1 COCO units from the baseline; Validation: 1 = valid, 0 = invalid; Rank = ordinal position.

**Test V2: One-object constant extreme advantage (single-object dominance)**

Design. One student was assigned a constant extreme advantage pattern across the input, while other students remained near-zero/low (Table 3.1.9- III).

Expected behaviour. The output should collapse toward a degenerate solution (loss of meaningful differentiation), because the matrix no longer represents a plausible distribution of multi-attribute behaviour.

Observed behaviour (Table 3.1.9- IV). The COCO output converged toward identical (or near-identical) estimates, producing a flat, non-informative ordering, consistent with a dominance-driven degeneracy.

	direction	0	0	0	0	0	1	0	
type	x	x	x	x	x	x	x		
Value	integer	integer	integer	integer	integer	integer	integer	decimal	
attribute_id	A1	A2	A3	A4	A5	A28	A29		
unit	count (posts)	count (days)	count (replies)	count (characters)	count (words)	hours	count (posts per day)		
userid	username/attribute_name	totalPosts	activeDays	total_replies_to_prof	total_characters	total_words	response_time_in_hours(Task-II)	average_posts_per_day	Y
1	student_1	0	0	0	0	0	0	0.0	1000
2	student_2	0	0	0	0	0	0	0.0	1000
3	student_3	0	0	0	0	0	0	0.0	1000
4	student_4	0	0	0	0	0	0	0.0	1000
5	student_5	0	0	0	0	0	0	0.0	1000
6	student_6	0	0	0	0	0	0	0.0	1000
7	student_7	0	0	0	0	0	0	0.0	1000
8	student_8	0	0	0	0	0	0	0.0	1000
9	student_9	0	0	0	0	0	0	0.0	1000
10	student_10	0	0	0	0	0	0	0.0	1000
11	student_11	0	0	0	0	0	0	0.0	1000
12	student_12	0	0	0	0	0	0	0.0	1000
13	student_13	0	0	0	0	0	0	0.0	1000
14	student_14	0	0	0	0	0	0	0.0	1000
15	student_15	0	0	0	0	0	0	0.0	1000
16	student_16	0	0	0	0	0	0	0.0	1000
17	student_17	0	0	0	0	0	0	0.0	1000
18	student_18	0	0	0	0	0	0	0.0	1000
19	student_19	0	0	0	0	0	0	0.0	1000
20	student_20	10000000	10000000	10000000	10000000	10000000	10000000	10000000.0	1000
21	student_21	0	0	0	0	0	0	0.0	1000
22	student_22	0	0	0	0	0	0	0.0	1000
23	student_23	0	0	0	0	0	0	0.0	1000
24	student_24	0	0	0	0	0	0	0.0	1000

Table 3.1.9- III Distorted input (single-object dominance). (Source: Own Presentation, Annex §8.6.1)

**Legend:** Direction: 0 means the more, the more ideal and 1 means the less, the more ideal; Y = 1000 denotes the fixed COCO baseline; the table illustrates a deliberately distorted input with single-object dominance, where one object contains extreme values while the remaining objects are zero.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	Validation (Binary 0/1)	Rank (Ordinal Position)
1	student 1	1000	1000.0	1	1
2	student 2	1000	1000.0	1	1
3	student 3	1000	1000.0	1	1
4	student 4	1000	1000.0	1	1
5	student 5	1000	1000.0	1	1
6	student 6	1000	1000.0	1	1
7	student 7	1000	1000.0	1	1
8	student 8	1000	1000.0	1	1
9	student 9	1000	1000.0	1	1
10	student 10	1000	1000.0	1	1
11	student 11	1000	1000.0	1	1
12	student 12	1000	1000.0	1	1
13	student 13	1000	1000.0	1	1
14	student 14	1000	1000.0	1	1
15	student 15	1000	1000.0	1	1
16	student 16	1000	1000.0	1	1
17	student 17	1000	1000.0	1	1
18	student 18	1000	1000.0	1	1
19	student 19	1000	1000.0	1	1
20	student 20	1000	1000.0	1	1
21	student 21	1000	1000.0	1	1
22	student 22	1000	1000.0	1	1
23	student 23	1000	1000.0	1	1
24	student 24	1000	1000.0	1	1

Table 3.1.9- IV COCO output (flat/degenerate separation). (Source: Own Presentation, Annex §8.6.1)

**Legend:** All estimation values equal 1000.0, with 0.0-unit deviation from the baseline for all cases; Validation: 1 = valid, 0 = invalid; identical Rank values indicate flat/degenerate separation.

### Test V3: Artificial deterministic structure (non-behavioural pattern)

Design. A purely synthetic input pattern was constructed (e.g., systematically increasing values by object order), ensuring “all objects differ” by rule rather than by evidence (Table 3.1.9- V).

Expected behaviour. The output should be weakly interpretable or degenerate, because the structure reflects the construction rule rather than behavioural signals derived from Moodle logs.

Observed behaviour. The COCO output was non-informative (flat/degenerate), indicating that interpretable COCO results require realistic, log-derived structure rather than arbitrary deterministic sequences (Table 3.1.9- VI).

	direction	0	0	0	0	0	1	0	
type	x	x	x	x	x	x	x	x	
value	integer	integer	integer	integer	integer	integer	integer	decimal	
attribute_id	A1	A2	A3	A4	A5	A28	A29		
unit	count (posts)	count (days)	count (replies)	count (characters)	count (words)	hours	count (posts per day)		
userid	username/attribute_name	totalPosts	activeDays	total_replies_to_prof	total_characters	total_words	response_time_in_hours(Task-II)	average_posts_per_day	Y
1	student_1	1	2	3	1	1	1	1.0	1000
2	student_2	2	2	2	2	2	2	2.0	1000
3	student_3	3	3	3	3	3	3	3.0	1000
4	student_4	4	4	4	4	4	4	4.0	1000
5	student_5	5	5	5	5	5	5	5.0	1000
6	student_6	6	6	6	6	6	6	6.0	1000
7	student_7	7	7	7	7	7	7	7.0	1000
8	student_8	8	8	8	8	8	8	8.0	1000
9	student_9	9	9	9	9	9	9	9.0	1000
10	student_10	10	10	10	10	10	10	10.0	1000
11	student_11	11	11	11	11	11	11	11.0	1000
12	student_12	12	12	12	12	12	12	12.0	1000
13	student_13	13	13	13	13	13	13	13.0	1000
14	student_14	14	14	14	14	14	14	14.0	1000
15	student_15	15	15	15	15	15	15	15.0	1000
16	student_16	16	16	16	16	16	16	16.0	1000
17	student_17	17	17	17	17	17	17	17.0	1000
18	student_18	18	18	18	18	18	18	18.0	1000
19	student_19	19	19	19	19	19	19	19.0	1000
20	student_20	20	20	20	20	20	20	20.0	1000
21	student_21	21	21	21	21	21	21	21.0	1000
22	student_22	22	22	22	22	22	22	22.0	1000
23	student_23	23	23	23	23	23	23	23.0	1000
24	student_24	24	24	24	24	24	24	24.0	1000

Table 3.1.9- V Distorted input (synthetic deterministic structure). (Source: Own Presentation, Annex §8.6.1)

**Legend:** Direction: 0 means the more, the more ideal and 1 means the less, the more ideal; Y = 1000 denotes the fixed COCO baseline; the table illustrates a deliberately distorted input with a synthetic deterministic structure, where attribute values increase in a fixed sequential pattern across objects.

User ID	Username	Y (COCO index units)	Estimation (COCO index units)	Validation (Binary 0/1)	Rank (Ordinal Position)
1	student 1	1000	1000.0	1	1
2	student 2	1000	1000.0	1	1
3	student 3	1000	1000.0	1	1
4	student 4	1000	1000.0	1	1
5	student 5	1000	1000.0	1	1
6	student 6	1000	1000.0	1	1
7	student 7	1000	1000.0	1	1
8	student 8	1000	1000.0	1	1
9	student 9	1000	1000.0	1	1
10	student 10	1000	1000.0	1	1
11	student 11	1000	1000.0	1	1
12	student 12	1000	1000.0	1	1
13	student 13	1000	1000.0	1	1
14	student 14	1000	1000.0	1	1
15	student 15	1000	1000.0	1	1
16	student 16	1000	1000.0	1	1
17	student 17	1000	1000.0	1	1
18	student 18	1000	1000.0	1	1
19	student 19	1000	1000.0	1	1
20	student 20	1000	1000.0	1	1
21	student 21	1000	1000.0	1	1
22	student 22	1000	1000.0	1	1
23	student 23	1000	1000.0	1	1
24	student 24	1000	1000.0	1	1

Table 3.1.9- VI COCO output (flat/degenerate result). (Source: Own Presentation, Annex §8.6.1)

**Legend:** All estimation values equal 1000.0, with 0.0-unit deviation from the baseline for all cases; Validation: 1 = valid, 0 = invalid; identical Rank values indicate flat/degenerate results.

### Implication for methodological reliability

Overall, the verification cases indicate that COCO Y0 does not “manufacture” meaningful performance structure from inconsistent inputs. When the ranked cohort representation is distorted, outputs become either unstable (V1) or degenerate/flat (V2–V3). Conversely, the excluded (staircase-refined) input configuration supports interpretable separation and can be combined with the internal consistency checks of §3.1.7. Therefore, input correctness and structural plausibility are necessary conditions for defensible COCO-based interpretation in the proposed workflow.

## 3.2 Demo Software Development

§3.1 defines the evaluation method, while §3.2 presents its Streamlit-based reference implementation. Therefore, §3.2 reports operational engineering choices (UI, modules, persistence, deployment) rather than additional methodological justification. Following Knuth’s observation that “*Science is what we understand well enough to explain to a computer. Art is everything else we do*” (Knuth, 1996 pp. vii–viii), the demo software translates the thesis methodology into an executable, auditable, and repeatable pipeline for Moodle discussion logs.

The implementation covers data input, configuration, attribute computation (OAM construction), rank transformation, COCO execution, automated staircase-based exclusion, validation, visualisation, and interpretation.

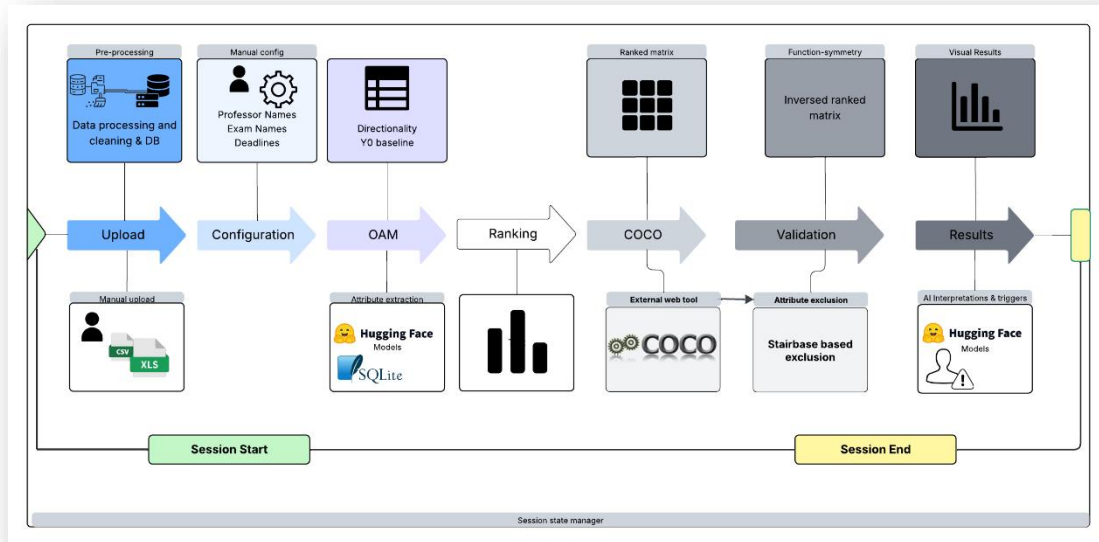


Figure 3.2- I Full operational flow of the demo application. (Source: Own Presentation)

**Legend:** blue = data-processing modules; grey = control and analysis modules; purple = OAM-related setup; green = session start; yellow = session end; arrows = operational flow.

Figure 3.2- I describes the full operational flow of the demo application.

### 3.2.1 Purpose and Scope of the Demo Tool

The demo tool has three primary purposes:

- Operationalization: execute the pipeline from raw Moodle log input to COCO-based comparative evaluation in a single coherent workflow aligned with §3.1.
- Explainability and auditability: expose intermediate artifacts (OAM table, rank matrix, COCO output tables, validation results) to enable traceable interpretation rather than black-box scoring
- Practical instructor workflow: provide an instructor-facing interface for repeatable analysis of class datasets, with exportable evidence suitable for annex documentation.

The scope is intentionally constrained to a thesis-demonstration profile:

- Single dataset per run: one uploaded export is analysed in a controlled run context.
- Local-first operation: the intended mode is local execution to preserve data control and reproducibility. A hosted demo instance is also provided for accessibility and quick evaluation: <https://moodleloganalysisdemo-2025.streamlit.app/>
- The hosted instance is treated as a convenience deployment; it may run with limited resources and session-scoped persistence.

Transparency over automation complexity: intermediate tables and decisions are surfaced rather than hidden.

Out-of-scope or partial in the current demo: full operationalization of the entire A1–A29 catalogue as a single standardized UI flow, and production governance features (authentication, institutional integration, consent workflows).

### 3.2.2 System Architecture

The prototype follows a layered architecture with separation between UI, processing utilities, and persistence.

#### **Presentation layer (Streamlit):**

- Guided workflow pages (data upload, configuration, attribute calculation, ranking, COCO analysis, visualization, AI insights, database view).
- Stage gating: pages that depend on upstream artifacts (e.g., visualization and interpretation) remain inactive until prerequisites are computed.

#### **Processing layer:**

- Attribute computation module.
- COCO request/response parsing module.
- Run/session orchestration and configuration management.

#### **Persistence layer (local SQL):**

- SQLite database accessed via an Object-Relational Mapping (ORM) layer to store dataset metadata, raw posts, computed metrics, ranking outputs, parsed COCO artifacts, and run events.

- Persistence behaviour differs by deployment:
  - Local mode: database persists on disk and can store multiple runs.
  - Hosted demo mode: database is typically initialized per session and may not persist beyond session lifetime.

Figure 3.2.2- I shows the directory structure of the demo application repository.

```

Moodle_Log_Analysis_DEMO
|
| Analysis.py
| README.md
| requirements.txt
| USER_GUIDE_2025.docx
|
|---assets
|   styles.css
|   ui_components.py
|
|---data
|   runs.db
|
|---pages
|   1_Data_Upload.py
|   2_Configuration.py
|   3_Attribute_Analysis.py
|   4_Ranking.py
|   5_COCO_Analysis.py
|   6_Visualizations.py
|   7_AI_Insights.py
|   8_Student_Insights
|
|---sample_data
|   discussion_demo.csv
|   discussion_demo.xlsx
|
|---utils
|   attribute_calculations.py
|   coco_utils.py
|   config_manager.py
|   db.py
|   session_data_manager.py
|   ui_steps.py

```

Figure 3.2.2- I Directory structure of the demo application repository. (Source: Own Presentation)

**Legend:** indentation and tree lines = directory hierarchy.

Note: The repository is structured to separate core implementation modules from configuration and data resources, enabling transparent execution, easier maintenance, and consistent reproduction of results.

Table 3.2.2- I summarises the module-to-responsibility mapping of the demo application.

Component	File / Module	Responsibility
App entry point	Analysis.py	Streamlit bootstrap; initializes managers; navigation baseline
Data persistence	utils/db.py	SQLite schema + ORM session; datasets/runs/artifacts/events
Session orchestration	utils/session_data_manager.py	Run state, caching, dataset/run lifecycle
Configuration	utils/config_manager.py	Stores and retrieves professor/exam configuration
Attribute computation	utils/attribute_calculations.py	Computes student-level metrics from raw posts
COCO integration	utils/coco_utils.py	Builds COCO input, calls Y0 endpoint, parses HTML tables, performs attribute exclusion
UI pages	pages/*.py	Per-stage UI workflow (upload → compute → rank → COCO → outputs)

Table 3.2.2- 1 Module-to-responsibility mapping of the demo application, summarising the core components, their implementation files, and their functional roles within the evaluation workflow. (Source: Own Presentation).

Note: The structure follows a layered design: UI pages form the presentation layer, manager modules orchestrate application state and configuration, computation modules implement domain logic, and the persistence layer ensures reproducible storage of datasets and run artifacts.

### 3.2.3 Data-Pipeline Implementations

The implemented pipeline is deterministic and consists of the following stages:

1. Ingestion and validation

The user uploads a CSV/XLSX Moodle log export. The tool verifies the minimum schema required for processing.

2. Run creation and traceability

A run is created with run-level metadata (dataset identity, time, and configuration snapshot). Events are recorded to preserve traceability (e.g., “uploaded”, “attributes computed”, “COCO executed”).

3. Configuration-driven filtering and aggregation

Instructor/professor posts are separated from student objects using configured identities; exam/task deadlines are used to derive deadline-related measures. Student-level aggregation produces the OAM table.

4. Ranking transformation

Selected attributes are converted into ordinal ranks with explicit directionality rules (benefit vs cost) as described in §3.1.5, forming the COCO input.

## 5. COCO execution and parsing

The ranked matrix is serialized into the COCO Y0 input format and submitted to the external endpoint. Returned HTML tables are parsed into structured data frames and stored as run artifacts. Staircase-diagnostic exclusion workflow. After the initial COCO pass, the application inspects the stairs(2) output and identifies candidate attributes according to the analytical rule defined in §3.1.6.3, namely the case where the S1 value equals  $n - 1$ , with  $n$  denoting the number of objects in the run. When such candidates are detected, the system constructs an excluded-attribute re-run matrix from the flagged rank columns and executes a second COCO pass on this reduced representation. If no candidates are detected, or if the second pass fails to return usable tables, the initial COCO result remains the active result. This logic ensures that the implementation follows the same diagnostic sequence as the analytical method rather than requiring manual intervention between stages.

## 6. Validation and interpretation

The inverse-ranking run is executed on the active COCO result, that is, on the excluded-attribute re-run when such a re-run succeeds, and otherwise on the initial pass. Sign-consistency rules generate validity indicators aligned with §3.1.7, after which the visual diagnostics and interpretation layers consume the stored artifacts. See Figure 3.2.3- I.

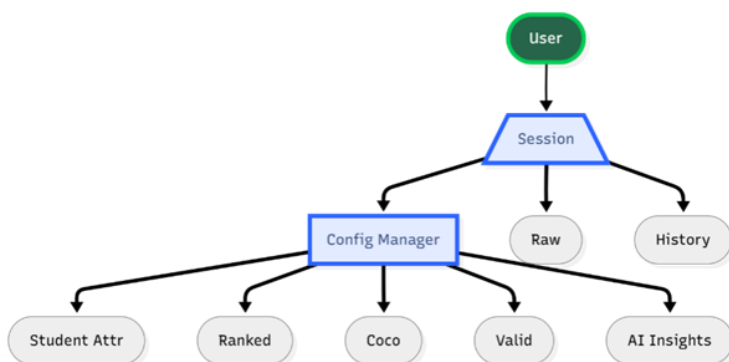


Figure 3.2.3- I Session State Manager. (Source: Own Presentation)

**Legend:** green = user actor; blue = session/control components; grey = stored state objects; arrows = state transitions.

Session State Manager Figure 3.2.3- I maintains the application’s session-level state (active dataset, run context, caching, and stage outputs) and orchestrates the dataset/run lifecycle across the UI layers to ensure consistent, reproducible execution.

### 3.2.4 Similarity Analysis and AI Detection in Text Responses

Two text-based indicators from §3.1 are operationalized as auxiliary attributes and interpretation inputs.

Topic relevance / similarity analysis

- Student replies are linked to instructor prompts using discussion structure (parent–child relations where available).
- Text embeddings are computed using a sentence-transformer model; cosine similarity is computed per eligible reply.
- A student-level topic relevance score is aggregated (e.g., mean similarity), consistent with §3.1.3.

AI-involvement signal

- A detector model provides message-level likelihood estimates.
- Message-level values are aggregated into a bounded student-level score (cf. §3.1.4).

These indicators are not interpreted as ground truth of “quality” or “misconduct”; they serve as contextual signals inside a broader multi-attribute evaluation.

### 3.2.5 Ranking & Scoring Logic

The ranking step implements explicit, deterministic logic:

- Users select a subset of computed attributes for inclusion.
- Directionality is applied (§3.1.5.1):
  - Cost-type (lower is more ideal): e.g., deadline-exceeded counts, response-time metrics.
  - Benefit-type (higher is more ideal): most contribution/engagement indicators.
- Per-attribute ranks are computed consistently (including the selected tie-handling rule).
- The ranked matrix is converted into COCO-compatible input:

- identifiers separated from rank columns,
- matrix serialized as tab-separated numeric content,
- object labels provided in parallel.

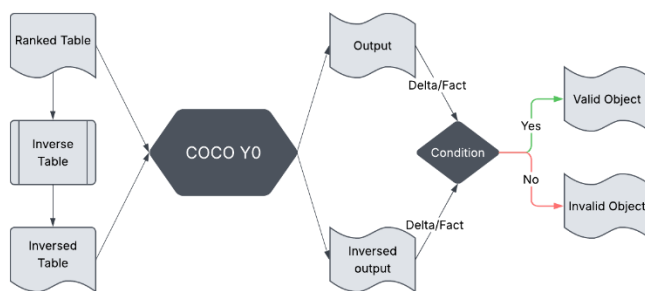
Function-symmetry / inverse-run validation (bridge to §3.1.7)

- The rank matrix is inverted using the cohort size N:

$$r^{inv} = (N + 1) - r$$

*Equation XI Cohort Rank Inversion*

- COCO is executed on the inverted matrix.
- Sign consistency between Delta/Fact (original vs inverse) is evaluated via a product rule, yielding valid/invalid flags for run-level diagnostics see Figure 3.2.5- I.



*Figure 3.2.5- I High-level validation logic of the demo application. (Source: Own Presentation)*

**Legend:** grey = transformation and validation steps; green = valid output; red = invalid output; arrows = validation flow.

### 3.2.6 User Interface Interaction Design

The UI is designed as a guided analytical journey see Figure 3.2.6- I:

- Workflow gating: later stages are disabled until prerequisites exist.
- Each stage provides:
  - summary metrics with brief interpretation guidance,
  - Preview of raw/intermediate tables for auditability.
- Configuration is separated from computation to reduce parameter drift.

- Exports are available at multiple stages (OAM, ranking, COCO, validation, figures)

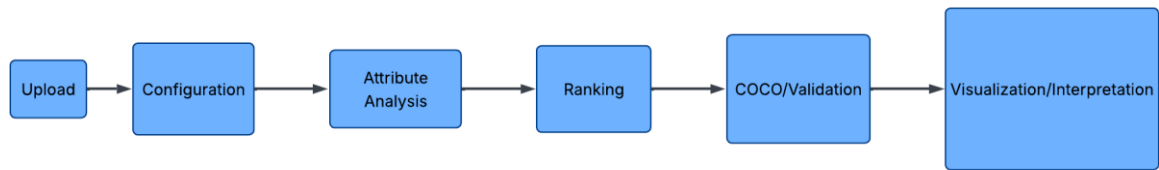


Figure 3.2.6- 1 Page navigation flow of the demo application. (Source: Own Presentation)

**Legend:** blue blocks = application pages/stages; arrows = navigation sequence.

### 3.2.7 Visual Outputs and AI-Assisted Interpretive Summaries

The demo provides diagnostic and reporting visualizations aligned with each stage:

- Attribute analytics: distributions, comparisons, top/bottom patterns
- Ranking analytics: rank distributions and rank heatmaps across selected attributes.
- COCO/validation analytics: estimated score plots, delta diagnostics, validity summaries.
- Student-centric views: profile summaries and cohort comparisons.

The AI interpretation subsystem is separated from objective scoring and functions as an assistant layer:

- produces run summaries and “watchlist” style prompts for instructor review,
- highlights potential outliers and notable patterns,
- provides student connectedness indications,
- supports comparative Q&A over computed metrics.

Safeguard principle: AI interpretation does not overwrite ranking/COCO outputs and must not be treated as an automated decision system. It produces supplementary narratives for human academic judgement.

### 3.2.8 Testing

Testing in the prototype is scenario-driven and pipeline-centric, focusing on complete workflow execution and persistent evidence rather than isolated unit tests.

To demonstrate operational scalability in the thesis context, two independent end-to-end runs were executed on heterogeneous inputs (CSV and XLSX) and stored in runs.db.

- Both runs completed with status validation\_completed and produced the full chain of outputs (raw records, computed attributes, rankings, COCO/validation artifacts, and event logs).
- Table 3.2.8- I consolidated evidence from the two persisted runs in runs.db, showing completion status, data volume, and traceability indicators required for structured post-run analysis.
- The database evidence confirms that the application is functioning end-to-end and that results are stored in a relational structure suitable for later data mining and auditing.
- The structured database file used for verification is included as annex material ([Annex §8.6.5](#)).

Evidence Unit	Verification Target	Observed Value in runs.db	Assessment
Run A (CSV)	End-to-end completion	discussion_demo.csv (576 rows), status=validation_completed, 24 students validated, duration≈267.6 s	Pass
Run B (XLSX)	End-to-end completion	discussion_demo001.xlsx (172 rows), status=validation_completed, 17 students validated, duration≈102.0 s	Pass
Per-run metric persistence	Computed metric storage	metric_observations: 456 (Run A) + 272 (Run B) = 728	Pass
Per-run ranking persistence	Ranking storage	ranking_observations: 456 (Run A) + 272 (Run B) = 728	Pass
Artifacts and logs	Traceability across pipeline stages	run_artifacts: 5 + 5 = 10; run_events: 7 + 7 = 14	Pass
Raw data retention	Event-level source preservation	raw_posts=748 linked to two datasets via foreign keys	Pass
Relational integrity	Data-mining readiness	7 linked tables: datasets, analysis_runs, raw_posts, metric_observations, ranking_observations, run_artifacts, run_events	Pass
Lifecycle clean-up	Consistency under deletion	Cascade delete is defined from datasets to dependent run tables	Pass

Table 3.2.8- I Consolidated evidence from the two persisted runs in runs.db, showing completion status, data volume, and traceability indicators required for structured post-run analysis. (Source: Own Presentation, Annex §8.6.7)

### Evaluation of testing results:

Across both runs, the pipeline executed end-to-end—upload, attribute computation, ranking, COCO processing, validation, and AI-based interpretation—without interruption, and each stage was persistently recorded in the relational store. The stored results are internally consistent (Run A: 456 metric and 456 ranking observations; Run B: 272 metric and 272 ranking observations), supporting traceability and reproducibility of the computed outputs. Runtime increased with input size (576 rows: ~267.6 s; 172 rows: ~102.0 s), which is consistent with expected operational scaling in this thesis-demonstration setting. As only two datasets

were processed, the evidence is necessarily preliminary and should be extended in future work using larger and more diverse datasets. Structured database of the test results can be found in [Annex §8.6.7](#).

### 3.2.9 IT Security, Privacy, and Ethics in Implementation

Because the prototype processes educational interaction logs that may contain personally identifiable information and sensitive behavioural signals, the implementation was designed with baseline safeguards appropriate to a demo-stage learning-analytics system. The primary technical protection is local-first operation: in the recommended local deployment mode, uploaded datasets, derived metrics, and generated artifacts are stored in a local SQLite database, reducing unnecessary network exposure and supporting data minimization in practice (only the fields needed for the analysis pipeline are processed and persisted). To support storage limitation and controlled lifecycle management, the application implements a structured deletion workflow: removing a dataset triggers cascading deletion of associated runs and derived artifacts, enabling the operator to comply with retention expectations and to reset analyses without leaving residual traces.

To ensure accountability and reproducibility, the tool stores run-level logs and artifacts (e.g., uploaded dataset metadata, configuration snapshots, computed tables, ranking matrices, COCO outputs, and validation summaries). This creates an auditable trace that allows results to be re-checked and exported as annex-level evidence, and it reduces “black-box” interpretive risk by making intermediate transformations inspectable. In hosted-demo mode, persistence is treated as session-scoped; therefore, evidence retention and governance controls are considered reliable only under local execution.

Ethical safeguards are enforced primarily at the interpretation layer. The AI-involvement indicator is presented explicitly as an auxiliary signal with known limitations, not as proof of misconduct or a proxy for learning quality. Likewise, AI-generated narrative outputs (summaries, watchlists, comparisons) are designed to support instructor review and pedagogical intervention planning, not to automate grading decisions or sanctions. The final responsibility for academic judgement remains with the instructor, consistent with the thesis’s methodological stance that automated signals should strengthen transparency and consistency without replacing human accountability.

Several governance limitations remain relative to a production-grade institutional deployment. The prototype does not currently implement a dedicated authentication/authorization model,

role-based access control, or formal consent management workflows. In addition, the COCO Y0 step depends on an external web endpoint; therefore, ranked matrices are transmitted during COCO execution. This external dependency introduces both availability risk and a privacy boundary that must be acknowledged in any real-world adoption: if used with non-pseudonymised data, the operator must ensure that institutional policy and legal requirements permit the transfer, or alternatively replace the external call with a locally hosted COCO component.

The prototype is provided as thesis-grade decision-support software without any warranty of institutional fitness, grading suitability, or error-free operation. Responsibility and liability for any real-world assessment decision remain with the human evaluator and the governing institution; therefore, the tool should be used only as bounded support evidence and not as a self-sufficient decision authority.

### 3.2.10 System Requirements

This subsection defines the operational conditions under which the prototype can be executed reproducibly. The tool supports two deployment modes: a hosted demonstration instance for rapid evaluation and a local installation mode for research-grade use. The distinction matters because persistence, performance, and evidence retention differ substantially between these environments.

In hosted demo (evaluation mode), the application is accessible via the public Streamlit deployment (<https://moodleloganalysisdemo-2025.streamlit.app/>). This mode is suitable for quick UI walkthroughs and first-contact demonstrations, but it should be treated as a convenience environment. Resource limitations may apply, and any persistence should be considered session-scoped (i.e., artifacts are not guaranteed to remain available across sessions).

For local installation (recommended for research use), the system runs on the instructor's machine, providing maximal control over privacy, repeatability, and storage of run evidence. This mode is preferred when the analysis outputs must be retained for annex-level documentation, when datasets must remain local for privacy reasons, or when performance is required for larger datasets and NLP-heavy computation.

For local execution, the required software prerequisites are: Python 3.10 or newer, the pip package manager, and a modern web browser (Chrome, Firefox, Safari, or Edge). The minimum hardware requirement is 8 GB RAM and a multi-core CPU. For ML-heavy

features—especially transformer-based embeddings and detector inference—16 GB RAM is recommended, and a GPU is beneficial because it significantly accelerates inference runtime.

Network access is required for two cases: (i) the initial download of optional Hugging Face models (if they are not already cached locally), and (ii) COCO Y0 web requests. Once dependencies and models are available locally, the majority of the pipeline can operate offline; however, COCO execution remains dependent on external web access in the current implementation.

Local installation follows a standard reproducible workflow: cloning the repository, installing dependencies from requirements.txt, and launching the Streamlit app via Streamlit run Analysis.py. If the browser does not open automatically, the application can be accessed through the local URL printed by Streamlit (typically <http://localhost:8501>). This setup ensures that configurations, intermediate artifacts, and exported outputs can be regenerated consistently under the same pipeline structure described in §3.2. A step-by-step operational guide for installation, execution, and basic use of the prototype is provided in [Annex §8.8](#).

### 3.2.11 Prototype Limitations

The prototype is intentionally positioned as a thesis-grade demonstration system, and several limitations therefore reflect scope boundaries rather than weaknesses of the underlying analytical method. Although the thesis defines an A1–A29 attribute catalogue (§3.1.5), the demo does not yet provide a single standardized interface that fully operationalises the entire catalogue under one unified configuration profile.

A further limitation concerns context-specific metrics. Certain indicators depend on task- or course-specific assumptions—for example, the `valid_response` logic relies on fixed parent identifiers and predefined answer patterns. Such implementations are valid for demonstration and for the analysed dataset, but they are not yet generalized to arbitrary course structures without additional configuration layers. In addition, the system’s COCO evaluation relies on an external web endpoint, meaning runtime stability can be affected by service availability, network conditions, or changes in the endpoint’s HTML response structure.

From a methodological-implementation perspective, the ranking stage currently uses a simple tie-handling approach. This is sufficient for deterministic execution and transparency, but it does not yet integrate more advanced tie-robust strategies that may be preferable when large tie groups occur. Finally, the prototype is not equivalent to a production, multi-user institutional

platform: it does not provide authentication/authorization, role management, or institutional integration features, and the testing approach is primarily scenario-driven pipeline verification rather than a fully automated unit and integration testing suite.

These limitations are treated explicitly as engineering and operational development targets and are therefore carried forward as concrete directions in §6.

### 3.3 Benchmarking against an independent LLM solution

This section introduces an independent LLM-generated solution addressing the same overall goal as the human-authored thesis: the objective, transparent, and bounded evaluation of student performance from Moodle discussion-forum log data in a teacher-facing decision-support context. For the comparison to remain valid, the benchmark and the thesis share the same empirical domain, the same dataset scope, and the same decision-support boundary, while differing in analytical design. In both cases, one student across the observed time window is treated as the object of analysis, and Moodle forum data are interpreted as bounded evidence rather than as direct proof of learning quality. However, the benchmark was generated under an explicit independence constraint: it was instructed not to reproduce the human-developed OAM–COCO workflow, but instead to construct an alternative solution from first principles. The comparison is therefore goal-equivalent but method-nonidentical, which makes it suitable as a genuine benchmark rather than a disguised replication.

#### 3.3.1 Summary of the LLM benchmark method

The benchmark solution developed a **Triangulated Percentile-Geometric Profile (TPGP)** built on three evidence pillars: **participation**, **interaction**, and **content**. It used nine indicators in total, scaled them by cohort-relative percentile positions, aggregated them within pillars by the median, and combined the three pillar scores by a geometric mean. Final outputs were student-level overall scores, ranks, and quartile-based profile bands (*Leading*, *Strong*, *Developing*, *Low-evidence*). As its main robustness procedure, the benchmark applied a leave-one-discussion-out jack-knife and reported high ranking stability across re-estimations. Table 3.3.1- I shows the results generated by ChatGPT.

rank	student	posts	active_days	thread_coverage_pct	discussions_with_instructor_reply	instructor_uptake_breadth_pct	peer_interaction_breadth	median_reply_latency_hours	reply_promptness_index	median_wordcount	mean_thread_similarity	mean_parent_similarity	participation_score_pct	interaction_score_pct	content_score_pct	overall_score_pct	band
1	student_15	27	12	75	58.333	7	3	24.4666667	3.927	59	0.08679408	0.18849019	95.833	85.667	72.917	86.323	Leading
2	student_6	37	10	75	66.667	8	2	14.7666667	6.342	71	0.068428877	0.149434792	95.833	85.417	68.75	82.561	Leading
3	student_19	16	6	58.333	41.667	5	5	31.475	3.079	209	0.102998233	0.181989167	68.75	68.75	85.417	73.909	Leading
4	student_16	17	6	58.333	50	1	1	39.0666667	2.496	76	0.131731373	0.193131946	68.75	58.333	89.583	71.409	Leading
5	student_20	20	6	50	50	1	49.9416667	1.963	77.5	0.094308008	0.097946637	68.75	58.333	77.083	67.616	Leading	
6	student_4	22	8	50	50	0	46.6916667	2.097	88.5	0.063209142	0.139105852	87.5	47.917	68.75	66.028	Leading	
7	student_1	16	3	50	16.667	8	1.15	46.512	3.71	0.187979482	0.137196647	52.083	97.917	56.25	65.952	Strong	
8	student_22	4	4	58.333	58.333	7	1	67.35	1.463	61	0.070395053	0.114870586	79.167	58.333	50	61.349	Strong
9	student_10	17	5	66.667	41.667	3	13.3333333	6.865	20	0.068213452	0.115309566	64.583	89.583	36.583	61.181	Strong	
10	student_7	14	6	50	50	1	12.0333333	7.673	54.5	0.066634312	0.117074518	60.417	77.083	35.417	54.841	Strong	
11	student_9	11	4	50	50	0	16.4166667	5.742	63	0.04889534	0.152914947	37.5	77.083	56.25	54.581	Strong	
12	student_13	14	7	50	41.667	5	55.875	1.794	61	0.079844371	0.062711136	60.417	47.917	50	51.595	Strong	
13	student_17	21	8	50	41.667	1	16.3666667	5.758	66	0.06245331	0.093194015	85.417	58.333	27.083	51.292	Developing	
14	student_8	18	5	41.667	68.0833333	1.448	81.5	0.081164929	0.120205386	52.083	0.120205386	22.917	64.583	42.559	Developing		
15	student_5	8	4	41.667	69.2016667	1.432	73	0.079829753	0.040044485	38.417	0.040044485	22.917	56.25	35.74	Developing		
16	student_14	3	3	25	16.667	2	36.25	2.685	59	0.096907672	0.096907672	18.75	58.333	41.667	35.719	Developing	
17	student_2	16	5	41.667	41.667	0	57.675	1.704	69.5	0.053796688	0.065993734	52.083	35.417	22.917	34.835	Developing	
18	student_23	4	3	8.333	8.333	1	38.95	2.503	56.5	0.026959888	0.148776633	18.75	60.417	35.417	34.244	Developing	
19	student_21	17	6	41.667	41.667	0	60.9333333	1.615	49	0.050955959	0.038764941	64.583	31.25	18.75	33.573	Low-evidence	
20	student_18	12	4	41.667	41.667	0	43.5	2.255	35.5	0.036160411	0.109408115	35.417	47.917	14.583	29.142	Low-evidence	
21	student_3	10	3	33.333	33.333	4	118.95	0.834	44	0.064791725	0.052754343	22.917	22.917	22.917	22.917	Low-evidence	
22	student_13	2	1	16.667	287.6083333	0.346	137.5	0.160582254	0.116920215	6.25	0.116920215	6.25	16.667	19.75	21.375	Low-evidence	
23	student_11	3	1	16.667	188.5666667	0.559	29	0.070010765	0.306551844	12.5	0.306551844	12.5	16.667	22.917	16.859	Low-evidence	
24	student_24	1	1	8.333	96.8833333	1.022	12	0.045231559	0.299049088	4.667	0.299049088	4.667	14.583	10.417	8.586	Low-evidence	

Table 3.3.1- I Full student-level benchmark results generated by the independent LLM solution. (Source: ChatGPT benchmark output; see Annex §8.7.2)

**Legend:** rows = students; rank = final benchmark order; posts = total number of posts; active\_days = number of active posting days; thread\_coverage\_pct = share of discussion threads in which the student participated (%); discussions\_with\_instructor\_reply = number of discussions containing a direct reply to an instructor post; instructor\_uptake\_breadth\_pct = breadth of instructor-prompt uptake (%); peer\_interaction\_breadth = number of distinct peer-interaction links; median\_reply\_latency\_hours = median reply delay (hours); reply\_promptness\_index = inverse promptness indicator derived from reply latency; median\_wordcount = median post length (words); mean\_thread\_similarity = mean semantic similarity to the thread opener; mean\_parent\_similarity = mean semantic similarity to the parent post; participation\_score\_pct, interaction\_score\_pct, content\_score\_pct = pillar scores on a 0–100 scale; overall\_score\_pct = final benchmark composite score on a 0–100 scale; band = quartile-based profile category (Leading, Strong, Developing, Low-evidence); annotation = brief LLM-generated interpretive note for instructor review.

### 3.3.2 Comparative evaluation of the human-authored and LLM-authored solutions

The human-authored thesis is methodologically broader and more thesis-specific. It constructs a 29-attribute OAM, defines explicit benefit/cost directionality, applies COCO Y0 under a constant baseline, performs staircase-based exclusion, inverse-run function-symmetry validation, distorted-input verification, objective–subjective comparison, and prototype implementation. By contrast, the LLM benchmark is methodologically narrower but more compact: it uses a smaller indicator set, avoids COCO entirely, and produces a simpler rank-and-band decision-support profile.

The human solution’s main strength is depth. It makes the modelling assumptions explicit, tests what happens when separability fails, and shows that full-information evaluation can collapse into pronounced sameness (999/1000/1001), after which exclusion improves discrimination

but changes the construct footprint. It also demonstrates that agreement with peer-based subjective evaluation is generally weak, with only modest best alignment. In validation and reproducibility terms, the thesis is also stronger: it combines multiple internal checks with deterministic pre-processing, stored intermediate artefacts, and an implemented prototype, thereby supporting end-to-end auditability.

The LLM benchmark's main strength is parsimony. Its three-pillar structure is easy to read, easy to explain, and operationally stable: the benchmark reports a mean jackknife Spearman correlation of 0.979, a minimum of 0.950, and strong top/bottom overlap across re-estimations. Its reproducibility is transparent in a narrower sense, because the full prompt, full output, and exported CSV artefacts make the benchmark re-runnable and easy to inspect. Thus, the benchmark sacrifices methodological richness for clarity and stability. It is useful as a practical comparison case, but it does not address the more demanding anti-discriminative aggregation problem that is central to the human thesis.

Accordingly, the benchmark confirms that an LLM can generate a coherent and practically usable alternative solution, but it also shows that the human-authored thesis remains superior as an academic contribution because it offers deeper methodological justification, richer validation, and a more fully reproducible research-and-development realisation.

## 4 Discussion

This chapter discusses the main findings and explains what they mean in the context of the thesis. It clarifies the study's scope, interprets key patterns in the results, and relates them to the literature and methodological choices. It also highlights practical implications for teaching and summarises strengths, limitations, and future directions.

### 4.1 Restatement of study focus

This thesis investigates how Moodle discussion-forum exports can be transformed into auditable, multi-attribute evidence that supports objective and reproducible student performance evaluation in an LMS-supported course setting. The implemented workflow treats Moodle as an evidence environment and operationalises student “performance” not as a single score, but as a direction-aware attribute system organised into an Object–Attribute Matrix (OAM) with a deliberate separation between diligence/behavioural engagement indicators and understanding/content-related indicators. The evaluation step applies COCO Y0 under a constant baseline ( $Y = 1000$ ) to generate an explainable, anti-discriminative ordering and associated diagnostics, complemented by internal consistency checking through inverse-run function symmetry and by robustness-oriented distortion tests (see §§3.1.1–3.1.9; §3.2).

The study scope is explicitly bounded by the analysed cohort and task context: the dataset contains 24 students, 12 discussion threads, and 576 forum messages within a defined time window, and it is processed through deterministic pre-processing, ranking, COCO execution, validation, and reporting steps, with an R&D layer operationalised via a Streamlit prototype (see §3.1.1). Consequently, the central discussion obligation is not to “prove” a universal grading model, but to interpret what the observed COCO behaviours, stability patterns, and objective–subjective divergences imply about (i) the adequacy of the evidence model, (ii) the sensitivity of the evaluation logic to modelling choices, and (iii) the governance boundaries required for defensible educational decision support.

### 4.2 Interpretation of key findings

#### **Sameness in the full attribute run**

The most salient result is the near-constant outcome in the full attribute set run (A1–A29), where the COCO estimates collapse to 999/1000/1001 (See Table 3.1.8- I), producing large tie groups. This pattern indicates low cohort separability under the combined effect of (i) rank

transformation (which compresses magnitude differences into ordinal positions) and (ii) the monotone staircase constraints in the Y0 setting, where the target is constant and structure must therefore come from how well each student’s ranked profile fits a shared stepwise representation. Accordingly, the model signals that many students are nearly interchangeable given the chosen descriptor set and its ordinal representation; it is not a computational failure. In the same run, inverse-run validation passes 21/24 objects. One contributing factor is the relatively high attribute-to-object ratio: with 29 attributes describing only 24 students, the representation may become too dense for stable separation after rank compression, thereby amplifying the observed sameness effect.

This matters for interpretation because it reframes “low discrimination” as a validity-relevant property of the evidence model. When a multi-attribute evaluation yields sameness, the inference is that the current attribute configuration (after ranking and under Y0 constraints) does not encode sufficient distinctive structure to separate students in a stable way. The full-run saturation therefore functions as a diagnostic: either the cohort is genuinely similar under the selected behavioural and text indicators, or the indicator system—especially after discretisation into ranks—does not preserve the differences that educators intend to evaluate.

### **Staircase-based exclusion and construct change**

Re-running COCO after staircase-driven exclusion yields a substantially wider spread of estimates (approximately 923.2–1063.4) and a more informative ordering, demonstrating that the “sameness” phenomenon is not inevitable (see Table 3.1.8- II). The interpretive cost of this resolution gain is that the excluded-run is not merely a “cleaner” version of the original ranking; it is an evaluation under a reduced construct representation. The exclusion rule removes attributes identified as weakly informative in the staircase diagnostics (e.g., A3, A4, A7, A9, A10, A13, A14, A19, A23, A24, A25, A27, A28 in the full run), so the ordering is produced from different evidence bundle than the maximal-attribute configuration (see Table 3.1.8- III).

Staircase-based exclusion should therefore be treated as a parsimony-and-interpretability variant that can reveal a sharper monotone structure when the full descriptor set yields near-constant solutions. Discrimination improves because attributes that do not contribute separability under the imposed structure are removed. However, because the output is conditional on the attribute set, the excluded-run should be reported as a complementary view rather than as the “true” ordering. Consistently, stability between full and excluded variants is low ( $\rho = 0.132$  for full vs full (excluded);  $\rho = 0.084$  for diligence vs diligence(excluded);  $\rho =$

0.168 for understanding vs understanding(excluded); see Table 3.1.8- XII), confirming that exclusion is a substantive modelling choice rather than a minor refinement.

### **Diligence and understanding as distinct dimensions**

The correlation between the objective diligence and objective understanding rankings is positive but modest ( $\rho = 0.325$ ; see Table 3.1.8- XI), supporting the thesis decision to treat these as related but distinct dimensions. This is consistent with the OAM catalogue: diligence is dominated by behavioural regularity, timing, and activity structure, whereas understanding emphasises topic relevance, citation use, correctness/rubric adherence, and the auxiliary AI-related signal. The modest association implies that students can plausibly rank differently depending on whether participation dynamics or content-alignment evidence is prioritised; consequently, “overall performance” cannot be reduced to a single latent dimension without introducing stronger assumptions than the thesis currently adopts.

### **Objective–subjective divergence**

Across comparisons, agreement between COCO-derived ranks and peer-based subjective ordering is generally weak or near-zero (e.g., all attributes (full) vs subjective  $\rho = -0.097$ ; all attributes (excluded-run) vs subjective  $\rho = -0.244$ ; diligence vs subjective  $\rho = -0.077$ ; understanding vs subjective  $\rho = 0.207$ ; understanding (excluded-run) vs subjective  $\rho = -0.202$ ), with the strongest observed alignment occurring for diligence (excluded-run) vs subjective ( $\rho = 0.307$ ) (see Table 3.1.8- XI). This should not be read as either the objective pipeline or peer evaluation being “wrong.” Rather, the two procedures operationalise partially different performance constructs and draw on different evidence channels. The objective pipeline privileges what is recorded and formalised (timing, frequency, interaction structure, and defined text indicators), whereas peer judgement can incorporate context and qualitative impressions that are not captured (or not captured reliably) in Moodle logs.

The fact that diligence (excluded-run) aligns best with subjective ranks suggests a plausible mechanism: peer evaluation may weight visible participation effort and interaction responsiveness more consistently than content-quality indicators derived from embedding similarity and citation counts, especially when the latter are treated (as in this thesis) as bounded proxies rather than correctness measures (see Table 3.1.8- XI). Weak correlations are therefore informative: they indicate that log-based decision support is not a computational replica of human judgement, but a distinct evidential lens that can complement subjective impressions depending on which construct is emphasised.

## Verification and robustness

Inverse-run validation passes 21/24 objects in the full run (Table 3.1.8- I) and 24/24 in the full excluded-run (Table 3.1.8- II). The diligence-only runs yield 22/24 (Table 3.1.8- IV) and 23/24 after exclusion (Table 3.1.8- V), while the understanding-only runs yield 23/24 (Table 3.1.8- VII) and 24/24 after exclusion (Table 3.1.8- VIII). These results support a bounded reliability claim: under deterministic inversion of ranked inputs, COCO Y0 behaves coherently, while remaining sensitive to discretisation and tie structure. The distortion tests further indicate that COCO does not create structure from implausible inputs: range distortions yield erratic outputs and dominance/synthetic patterns yield degenerate flat solutions. Overall, the main runs are interpretable only insofar as the OAM and ranked OAM encode plausible cohort structure; input plausibility and descriptor design remain necessary conditions for defensible interpretation.

### 4.3 Resolution of method-bound interpretive issues

First, Moodle traces are shaped by platform configuration and course design, so derived indicators must be interpreted as evidence of interaction with a particular pedagogical instrumentation rather than as direct measurement of learning. In this thesis, the evidence environment is a discussion-based task with time constraints and structured prompts; therefore, high activity or timely replies plausibly indicate diligence within that design, but they may not generalise to other designs where the LMS is used differently. Accordingly, the evaluation outputs should be framed as decision support for instructor review rather than as an automated judgement of learning quality (see §2.2.3; §§2.3.3–2.3.4).

Second, the thesis explicitly treats logs as imperfect proxies and therefore accepts context dependence and limited generalisability as fundamental constraints. This is operationally reflected in two design choices: (i) the indicator catalogue is defined with explicit construct intention and directionality, and (ii) internal robustness is tested through validation and distortion cases rather than assuming that computed metrics are inherently valid. Nevertheless, these checks do not convert proxies into ground truth; they only demonstrate that the pipeline is internally coherent and sensitive in expected ways. Therefore, the scope claim must remain local: the results support conclusions about the analysed cohort and workflow, not universal claims about student performance across courses (see §2.3.4; §3.1.9).

Third, the thesis correctly emphasises that evaluation is value-laden, and that “objectivity” comes from making value choices explicit rather than pretending they do not exist. In this pipeline, normative choices appear at least in attribute selection, directionality settings, imputation rules (missingness as zero under “non-event” semantics), and the decision to separate diligence from understanding. These choices are academically defensible insofar as they are transparent and auditable, but they also carry fairness implications (see §2.7.6): if a course design privileges certain behaviours (e.g., frequent posting) then any indicator system that treats those behaviours as positive can systematically advantage students whose constraints allow such participation. The governance sections therefore function as interpretive boundaries, and the model should be used to support review and dialogue rather than to implement sanctions (see §3.1.5 on directionality; §3.2.9 governance constraints).

Fourth, the thesis’ boundary for semantic similarity must be maintained in interpretation: similarity estimates topical alignment, not correctness. Consequently, topic relevance should not be treated as an automated grading surrogate. Its defensible role is to provide a probabilistic signal that can flag responses that are unusually distant from the prompt, particularly when combined with other indicators such as response timing and participation structure. The multi-attribute design is precisely the safeguard here: similarity is one attribute among many, reducing the risk that a single uncertain measure dominates the decision (see §2.5.4; §3.1.4; §3.1.8.3).

Fifth, the thesis promised to revisit low discrimination (“sameness”), and the results show that sameness is an expected outcome when the rank-based descriptor set does not separate objects strongly under Y0 constraints. The correct practical response is diagnostic rather than apologetic: sameness motivates scrutiny of the descriptor system, tie density, and the consequences of ranking compression. Staircase-based exclusion is then justified as a principled attempt to reveal separability hidden by non-informative attributes under the fitted structure, while still requiring explicit reporting of how exclusion changes the evaluated construct footprint (see §3.1.8.1).

Sixth, the thesis explicitly distinguishes decision support from automated judgement, and the discussion must enforce that boundary. The tool produces structured evidence (rankings, attribute tables, diagnostics) that can support instructor decisions such as targeted review, formative feedback, or triage of cases needing attention. It does not—and should not—produce self-justifying verdicts about student ability, integrity, or grades. The governance risk is that

outputs look “objective” and thus invite over-trust; therefore, the defensible usage model is one in which the instructor remains accountable and uses the outputs as auditable inputs into human judgement (see §2.7.6; §3.2.9).

Seventh, the thesis committed to critically discuss AI detection validity, fairness, and evasion risks. Given the literature reviewed in §2.8 and the methodological choices in §3.1.4, the only defensible interpretation of the detector output is as an auxiliary, low-stakes signal that may prompt manual review—never as proof of misconduct or as a direct proxy for learning quality. This is reinforced by the pipeline architecture: the AI score is aggregated and direction-coded as “higher = worse” to prevent a scenario where high AI-likeness improves ranking, but it remains one attribute within a broader evidence system and should be treated as uncertain and contestable (see §2.8; §3.1.4; §3.2.9).

Eighth, the thesis promised to revisit the context-dependence of directionality decisions, explicitly noting A15 and A21 as non-obvious examples. For A15 (AI involvement), the cost-type choice is defensible under a conservative integrity framing and the stated non-punitive scope: it prevents rewarding suspected AI-likeness while maintaining the “supportive signal only” boundary. For A21 (maximum post length), treating very long posts as cost-type is defensible only within the explicitly stated task context (time-constrained descriptive responses where conciseness can be valued); however, the discussion must acknowledge the plausible alternative: in other contexts, longer posts may indicate elaboration and deeper explanation. The correct thesis-level stance is therefore conditional: directionality is a modelling assumption justified for this setting, and future work should test sensitivity to alternative direction coding (see §3.1.5; Table 3.1.5- III).

Ninth, the thesis promised that prototype testing evidence would be interpreted in Chapter 4. The two end-to-end runs stored in the prototype database demonstrate feasibility and traceable persistence of artifacts across heterogeneous inputs, supporting the claim that the pipeline can be operationalized in a reproducible instructor-facing workflow. However, because the number of tested datasets is limited, this evidence must be framed as baseline operational scalability, not as general performance validation across courses, languages, or institutional environments (see §3.2.8).

Finally, the thesis includes a global promise that “critical limitations” would be revisited in §4. The discussion therefore consolidates the limitations explicitly (see §4.8) rather than leaving

them dispersed, ensuring that the interpretive boundaries implied throughout Chapters 2–3 are stated in one coherent evaluative section.

#### 4.4 Comparison with literature

The empirical behaviour of the pipeline aligns with the literature’s caution that LMS traces are context-dependent proxies rather than direct measures of learning, and that analytics becomes defensible only when indicators remain interpretable and tied to plausible mechanisms. The present results operationalize this caution concretely: the full-run sameness illustrates how “more data” and “more indicators” do not automatically yield better evaluation if the descriptor system fails to encode separability under the chosen aggregation logic. This supports the literature’s broader argument that validity is not a property of logs alone, but of the modelling and interpretive framework imposed on them (see §2.2.3; §2.3.3; §2.9.1).

The divergence between objective rankings and peer subjective evaluations mirrors the literature’s observation that subjective perceptions and trace-based indicators can disagree systematically, because they privilege different evidence and may operationalize different performance constructs. In this thesis, that divergence is not treated as an embarrassment but as a methodological result: it demonstrates that log-based decision support is a distinct evidential perspective whose value lies partly in making evaluation assumptions explicit and auditable rather than relying on impression-based judgement alone (see §2.7.2–§2.7.6; §2.9.1).

The thesis’ treatment of semantic similarity as a bounded signal is consistent with the reviewed position that similarity can support evaluation of descriptive responses but cannot substitute for correctness or quality assessment. The fact that the understanding subset yields more structured discrimination than diligence in the initial configuration further motivates the literature-derived claim that content-related signals may carry separability in discussion tasks, while behavioural indicators can become homogeneous depending on course design and participation norms (see §2.5.4; §2.9.1; and Table 3.1.8- IV, Table 3.1.8- V, Table 3.1.8- VII, and Table 3.1.8- VIII).

Finally, the explicit scepticism toward AI detection as decisive evidence is directly aligned with the reviewed literature’s concerns about contested validity, fairness risks, and evasion. The pipeline’s design choice—integrating detector output only as an auxiliary attribute within a multi-attribute model and explicitly restricting interpretive scope—constitutes a governance-

aware response that treats detection as a contextual signal rather than an authorship verdict (see §2.8; §2.9.2).

## 4.5 Theoretical contributions

Within its bachelor-thesis scope, the work advances a coherent theoretical position: objectivity in educational evaluation is operational, not metaphysical—it is achieved by making evidence definitions, directionality, and aggregation rules explicit, reproducible, and open to critique. The implemented OAM formalism and direction-aware ranking instantiate this position as a computable evidence model rather than a purely narrative argument.

A second contribution is the empirical demonstration that “sameness” under anti-discriminative multi-attribute optimisation is not merely a nuisance but an interpretive signal about the descriptor system and the cohort’s separability under monotone constraints. The staircase-based exclusion procedure is thereby framed not as post-hoc tuning but as a diagnostic mechanism that clarifies which attributes become non-informative within the fitted structure and how parsimony can yield interpretability gains.

A third contribution is methodological: the thesis models evaluation quality as requiring explicit internal validation and robustness probing (inverse-run symmetry and distortion tests) when external ground truth is not used (Y0). This positions the pipeline within a defensible “audit logic” where interpretive confidence is built from transparency and controlled perturbation rather than from hidden parameters or unexamined outputs.

## 4.6 Practical implications

Practically, the pipeline supports instructors by converting discussion activity and text interactions into a structured evidence base that can guide triage and targeted review. For example, dimension-specific outputs enable different pedagogical responses: diligence-oriented patterns can inform participation interventions (e.g., supporting time management or engagement regularity), while understanding-oriented patterns can prioritise feedback on topical alignment and response quality. Because the workflow stores intermediate artifacts and rank/validation outputs, it also supports auditability, enabling an instructor to justify why a student was flagged for review using traceable indicators rather than opaque impressions.

However, the practical value depends on strict scope boundaries. The thesis’ own governance framing implies that outputs should be used to support educational judgement, not to automate

grading or disciplinary action. Accordingly, outputs require instructor accountability and contestability, and detector outputs must not be used as proof of misconduct (see §2.7.6; §2.8; §3.2.9). This is especially critical for the AI-detection attribute and for similarity scores: both are bounded indicators and must be interpreted alongside other evidence and the instructional context. The prototype’s design—exposing intermediate tables and retaining artifacts—supports this responsible use model by encouraging inspection rather than blind reliance on a single final score.

## 4.7 Strengths of the study

A principal strength is end-to-end traceability: the workflow is specified from raw Moodle forum exports through pre-processing, attribute construction, ranking transformation, COCO evaluation, validation checks, and comparative analysis against subjective ranks. This makes the evaluation logic inspectable at each stage and reduces the common risk that analytics pipelines function as black boxes producing unchallengeable outputs.

A second strength is the explicit inclusion of quality assurance logic appropriate to the Y0 setting. Instead of relying on goodness-of-fit to external labels, the thesis introduces internal consistency expectations (inverse-run symmetry) and stress tests (distorted input cases) that demonstrate meaningful sensitivity to implausible representations. This strengthens the defensibility of the pipeline as a decision-support tool by showing that outputs are not arbitrary artifacts of computation but depend on plausible cohort structure.

A third strength is the coherent integration of multi-source evidence under a single formal representation (OAM) without introducing arbitrary manual weights (see §2.6; §3.1.6). This integration—behavioural traces plus bounded text indicators—directly addresses the literature-derived gap articulated in Chapter 2 and is operationalized concretely in Chapter 3.

## 4.8 Limitations

### **Methodological limitations**

The most fundamental limitation is constructing validity under platform dependence. Moodle logs and discussion exports measure what the platform records and what course design incentivizes; therefore, indicators derived from them cannot be interpreted as direct measures of learning and should not be expected to generalise across substantially different course designs or LMS usage norms. This affects both behavioural indicators (which may reflect

compliance or resource constraints) and text-based indicators (which depend on how prompts and expected responses are formulated).

A second limitation is the dependence of COCO Y0 outcomes on descriptor design and discretisation. Rank transformation improves cross-attribute comparability but compresses magnitude information and increases the impact of tie structures; consequently, evaluation resolution can collapse (as observed in the full run) when the ranked descriptor system does not preserve separability under the monotone staircase constraints. The exclusion procedure improves discrimination but also changes the evidence bundle being evaluated, and low stability between full and excluded runs shows that exclusion is a consequential modelling decision rather than a neutral optimisation.

### **Governance and operational limitations**

A third limitation concerns bounded indicators with governance risk. Semantic similarity estimates topical alignment and may misrepresent correctness; AI detection is contested and vulnerable to false positives and evasion. Even when these signals are used conservatively as auxiliary attributes, their presence can invite misuse if stakeholders treat “probability-like” scores as proof. Therefore, the thesis’ decision-support framing is not optional rhetoric; it is a necessary condition for ethical and defensible interpretation.

Finally, the prototype testing evidence is preliminary. Two end-to-end runs demonstrate feasibility, persistence, and basic scalability across heterogeneous inputs, but they do not establish external validity across diverse cohorts, disciplines, writing genres, or institutional governance settings. Consequently, the current thesis can claim a working, auditable prototype and a validated internal pipeline logic, but it cannot claim that the tool yields stable or fair evaluation outcomes across contexts without further replication and governance embedding (see §3.2.8).

## **4.9 Research implications and directions**

Future work should first prioritise replication across courses and task designs to test portability of the attribute catalogue and to examine when diligence indicators become homogeneous versus discriminative. This would directly address the platform-and-design dependence limitation and clarify which indicators are robust across instructional modalities.

Second, the thesis should be extended with systematic sensitivity analyses over modelling choices that are explicitly value-laden, particularly directionality assignments (e.g., A21) and

missingness semantics (imputation as zero). Running controlled variants and reporting how rankings shift would transform “context-dependent” assumptions into empirically characterised design parameters.

Finally, any move toward broader deployment requires deeper governance operationalisation: explicit consent/notice patterns, access control, retention policy tooling, and an option to run COCO locally to avoid transmitting ranked matrices to an external endpoint. These steps follow directly from the thesis’ own ethics/GDPR framing and would be necessary before any high-stakes use could be considered even in principle (see §2.3.4; §2.7.6; §3.2.9).

## 4.10 Interpretation of the LLM benchmark

The benchmark comparison clarifies that the difference between the human-authored thesis and the LLM-authored solution is not simply a difference in output quality, but a difference in methodological ambition. The human solution was designed as a multi-attribute, validation-aware, and prototype-supported research-and-development workflow: it transforms Moodle traces into a 29-attribute OAM, applies COCO Y0 under a constant baseline, subjects the results to staircase-based exclusion, inverse-run function-symmetry validation, distorted-input verification, and objective–subjective comparison, and operationalizes the pipeline in a Streamlit prototype with stored intermediate artefacts for auditability. The benchmark, by contrast, deliberately adopts a narrower architecture: a three-pillar TPGP model built from participation, interaction, and content indicators, aggregated through percentile scaling, median-by-pillar logic, and a geometric mean. This makes the benchmark methodologically compact and easy to inspect, but also means that it engages a less demanding analytical problem than the thesis itself.

The added value of the human-authored solution becomes especially visible when the actual Chapter 3 findings are considered. The full COCO run produced the characteristic sameness pattern, with estimates collapsing to near-baseline values and large tie groups, which the thesis interprets not as computational failure but as a diagnostic signal about limited separability under the chosen descriptor set and ordinal representation. After staircase-based exclusion, ranking resolution improved substantially, but the thesis explicitly shows that this came at a conceptual cost: stability between full and excluded variants remained low ( $\rho = 0.132$  for full vs. full excluded;  $\rho = 0.084$  for diligence vs. diligence excluded;  $\rho = 0.168$  for understanding vs. understanding excluded), confirming that exclusion is a substantive modelling choice rather than a minor technical refinement. In other words, the human thesis does not merely produce a

ranking; it demonstrates how ranking behaviour changes when the evidential design itself changes.

A second strength of the human solution is that it interprets disagreement rather than hiding it. The comparison between objective COCO-based rankings and peer-based subjective ordering is generally weak or near-zero across configurations, with values such as  $\rho = -0.097$  for the full run versus subjective ranking,  $\rho = -0.244$  for the full excluded run, and only a modest best alignment of  $\rho = 0.307$  for diligence excluded versus subjective ranking. The thesis uses these weak correlations productively: they are read as evidence that the objective model and peer judgement capture partially different constructs and rely on different evidence channels. This is a stronger scholarly move than simply reporting agreement or disagreement, because it turns divergence into an interpretive result about construct structure, evidence visibility, and the limits of computational replication of human judgement.

The human thesis is also stronger in validation depth. Across the reported COCO configurations, inverse-run validation passes 21/24 objects in the full run, 24/24 in the full excluded run, 22/24 and 23/24 in the diligence runs, and 23/24 and 24/24 in the understanding runs. Together with the distortion tests, which showed erratic outputs under implausible range distortions and degenerate flat solutions under dominance-style synthetic inputs, these findings support a bounded reliability claim: the method behaves coherently under deterministic inversion, yet remains appropriately sensitive to discretisation, tie structure, and poor-quality inputs. The thesis therefore advances a richer understanding of robustness than simple stability alone.

By contrast, the benchmark's strength lies in parsimony and local stability. Its three-pillar design is easier to explain to a reader or instructor, and its main robustness procedure—the leave-one-discussion-out jackknife—produces strong stability indicators: mean Spearman correlation with the full ranking is 0.979, the minimum is 0.950, mean median absolute rank shift is 0.71 positions, and top-5 and bottom-5 overlap remain high across re-estimations. The benchmark also has a clear bounded-interpretation logic: it explicitly frames itself as a decision-support profile rather than a verdict and states that it cannot prove mastery, authorship authenticity, or causal learning effects. These are real strengths, and they show that the LLM can generate a coherent, transparent, and practically usable comparison case from the same dataset.

However, the benchmark's stability should not be confused with overall superiority. Its robustness is achieved partly by methodological simplification: fewer indicators, pillar-level aggregation, and a more compact evidential design reduce the chances of tie-saturation and construct-fragmentation, but they also reduce the analytical burden of the method. The human-authored thesis confronts a harder problem: how to aggregate a broader and more heterogeneous evidence structure while preserving interpretability, non-punitive boundaries, and auditability. For that reason, the benchmark is best interpreted as a useful counter-model that reveals the benefits of compactness, but not as a replacement for the human solution's deeper methodological contribution.

### **Comparative judgement on the relative strengths of the human-authored and LLM-authored solutions**

The human solution is not always better in every local criterion. The LLM benchmark performs better in compactness, immediate readability, and discussion-level stability. However, the human-authored thesis remains stronger as an academic contribution because it offers broader construct coverage, richer validation, clearer treatment of methodological tensions, and fuller end-to-end reproducibility through its pre-processing workflow, stored artefacts, and implemented prototype. Thus, the benchmark shows that an LLM can be competitive as a compact decision-support solution, but the thesis is superior as a scholarly work because it contributes not only a result, but also a critically examined methodology.

#### **4.11 Take home message**

Within one Moodle-based cohort, the results show that an auditable, direction-aware, and weight-avoiding evaluation workflow can be built from behavioural and semantic evidence. However, the full-information runs also show that aggregation may over-compress differences, whereas the excluded runs reveal how discrimination changes once low-informative staircase regions are removed. Overall, objective evaluation is feasible here as transparent and reproducible decision support, but its outputs remain bounded comparative evidence rather than self-sufficient judgement.

## 5 Conclusion

This thesis investigated whether Moodle discussion-forum exports can be transformed into auditable, multi-attribute evidence that supports objective and reproducible student performance evaluation in an LMS-supported course, under an explicit decision-support framing. The developed workflow treats Moodle as an evidence environment and operationalises “performance” as a direction-aware Object–Attribute Matrix (OAM) rather than a single score, with a deliberate separation between diligence (behavioural engagement) and understanding (bounded content-related indicators). Evaluation is performed via COCO Y0 under a constant baseline ( $Y = 1000$ ), complemented by staircase diagnostics, inverse-run function-symmetry validation, and robustness-oriented distortion tests.

The thesis objectives are met in the sense intended by the R&D methodology: (A1) a transparent log-based workflow and indicator system are defined and documented; (A2) a direction-aware, weight-avoiding aggregation is executed via COCO Y0 with diagnostic exclusion; (A3) defensibility is strengthened through internal consistency and failure-mode probing; and (A4) the method is operationalised as a teacher-facing prototype that preserves intermediate artefacts and documents governance constraints as design boundaries.

H1 (construct divergence) is supported. Agreement between COCO-derived rankings and peer-based subjective ordering was generally weak, indicating that the two approaches captured different aspects of student performance.

H2 (diagnostic exclusion) is supported. The full-attribute run showed pronounced sameness, which should be interpreted as low separability under the chosen representation rather than as computational failure.

Three findings are central for interpretation within the thesis boundaries.

First, the “sameness” effect in the full run is itself informative: it signals that, after ordinal ranking (including ties), many students become nearly interchangeable given the chosen descriptor set and the staircase constraints of Y0. In other words, the method provides a diagnostic about descriptor adequacy and separability under the current representation.

Second, diligence and understanding are empirically related but distinct dimensions. Their objective rankings show a positive but modest association (rank correlation  $\rho$  (rho) = 0.325), supporting the thesis decision to keep them separate and to avoid collapsing “overall

performance” into a single latent dimension without introducing stronger assumptions than the thesis adopts.

Third, reliability is defended as internal coherence and sensitivity rather than predictive accuracy. Inverse-run function-symmetry validation passes most objects across runs (e.g., 21/24 in the full run and 24/24 in the full excluded-run), and the distortion tests show that implausible inputs yield unstable or degenerate outputs rather than artificial structure. Consequently, interpretability depends on the plausibility and separability encoded in the OAM and ranked OAM.

The thesis contributes theoretically by framing objectivity as auditability and contestability, methodologically by demonstrating a direction-aware OAM-to-COCO Y0 workflow with diagnostic and robustness checks, and practically by implementing the process as a teacher-facing prototype that preserves intermediate artefacts for traceability.

The thesis’ conclusions are bounded by limitations that follow directly from the study design. Moodle logs measure what the platform records and what course design incentivises; therefore, indicators remain proxies and do not directly measure learning. Rank transformation improves comparability across heterogeneous indicators but compresses magnitude information, while ties reduce discriminative power and can propagate into the COCO input. Text-based evidence is bounded: semantic similarity estimates topical alignment rather than correctness, and the AI-involvement signal is explicitly auxiliary and must not be treated as proof. Finally, the prototype is thesis-grade: it lacks production governance features (authentication/authorisation, consent workflows), and COCO execution relies on an external endpoint, creating both availability risk and a privacy boundary.

Within these boundaries, the thesis demonstrates that Moodle discussion-forum exports can be converted into a reproducible, auditable multi-attribute evidence model and executed end-to-end as a teacher-facing prototype. The system’s primary value is structured decision support—reducing evidence-extraction workload while increasing transparency of the evaluation basis—rather than replacing instructor judgement. The thesis’ stakeholder utility and quantitative added-value estimation are provided separately in §1.4, consistent with the requirement that usefulness be stated with explicit assumptions and conservative interpretation.

## 5.1 Answers to the Research Questions

**Answer to RQ1:** Moodle discussion-forum logs and text-based signals can be transformed into a transparent and reproducible Object–Attribute Matrix through a traceable pipeline of pseudonymization, cleaning, attribute construction, and direction-aware formalization. In this thesis, the resulting student-level OAM enabled comparable multi-attribute evaluation under COCO Y0 and provided an auditable teacher-facing evidence structure rather than an opaque score or automated judgement.

**Answer to RQ2:** COCO Y0 behaved differently across the three matrix variants and thereby revealed an important diagnostic property of the workflow. The full and diligence-focused runs were strongly tie-dominated, whereas the understanding-focused variant showed greater separability. Staircase-based exclusion improved discrimination, but because it altered the evaluated evidence bundle, the excluded runs must be interpreted as complementary construct-variants rather than definitive corrections.

**Answer to RQ3:** Agreement between COCO-based objective rankings and peer-based subjective evaluation was limited. The generally weak correlations indicate that the objective pipeline and peer judgement captured different aspects of student performance. The findings therefore suggest that log-based evaluation may complement subjective assessment, but does not reproduce it.

**Answer to RQ4:** The proposed pipeline can be regarded as internally coherent and methodologically robust in a bounded sense. Validation and distortion testing supported internal consistency, reproducibility, and auditability by showing stable behaviour under plausible inputs and unstable or degenerate behaviour under implausible ones. At the same time, these findings remain within-sample and do not establish external validity or broader generalizability.

## 6 Future work

### **Replication for external validity and portability**

Future work should prioritise replication across courses, cohorts, and task designs to test portability of the attribute catalogue and to clarify when diligence indicators become homogeneous versus discriminative. This directly addresses the platform-and-design dependence limitation and is a prerequisite for any broader generalisation claims.

### **Sensitivity analysis of value-laden modelling choices**

A second extension is systematic sensitivity analysis over modelling decisions that are explicitly value-laden in this thesis—particularly (i) directionality assignments ( $0\uparrow/1\downarrow$ ) and (ii) missingness semantics (imputation as zero when absence has behavioural meaning). Controlled variants should be reported alongside ranking shifts to convert “context dependence” from a qualitative caveat into an empirically characterised design parameter.

### **Governance hardening for defensible adoption**

Any move toward broader deployment requires deeper governance operationalisation: explicit consent/notice patterns, access control, retention policy tooling, and—critically—an option to run COCO locally to avoid transmitting ranked matrices to an external endpoint. These steps follow directly from the thesis’ ethics/GDPR framing and from the principle that outputs remain decision support under human accountability.

### **Tie-robust ranking and stronger test automation (engineering-aligned extensions)**

Because large tie groups are empirically central to “sameness,” future work should integrate tie-robust ranking strategies and evaluate their impact on discrimination and stability. In parallel, the prototype should be extended from scenario-driven verification toward a more automated unit/integration testing suite, reflecting the thesis’ own identification of current demo-stage testing limitations.

## 7 Summary

This thesis addresses a recurring problem in LMS-supported teaching: evaluation of discussion-based participation and open-ended responses can become time-consuming and partly impression-driven, while Moodle exports provide trace evidence that remains

underutilised unless transformed into transparent and auditable indicators. The work follows a research-and-development structure and positions learning analytics outputs as decision support rather than automated judgement.

The methodological core is an end-to-end pipeline that transforms Moodle discussion-forum rows into student-level indicators and formalises them in an Object–Attribute Matrix (OAM) with explicit directionality rules (0 = ↑ more ideal; 1 = ↓ more ideal). The indicator catalogue (A1–A29) is intentionally split into diligence (behavioural engagement) and understanding (bounded content-related proxies). Text-oriented evidence is operationalised through semantic similarity (sentence embeddings with cosine similarity) and an auxiliary AI-involvement signal, each treated as bounded evidence rather than direct learning measures.

For evaluation, the ranked OAM is processed using COCO Y0 under a constant baseline ( $Y = 1000$ ), enabling a direction-aware, weight-avoiding multi-attribute ordering together with staircase diagnostics for identifying weakly informative attributes. Reliability is treated as internal coherence and sensitivity: inverse-run function symmetry provides a consistency diagnostic for the Y0 setting, and deliberately distorted inputs probe failure modes.

Empirically, the full attribute run exhibits a pronounced “sameness” pattern (999/1000/1001), interpreted as low separability under ordinal compression and Y0 staircase constraints rather than method failure. Staircase-based exclusion improves discrimination while changing the construct footprint, so excluded-run results are interpreted as complementary construct-variants. Diligence and understanding are related but distinct (rank correlation  $\rho$  (rho) = 0.325), and objective rankings generally diverge from peer-based subjective evaluation (typically weak or near-zero rank correlations, with only modest best alignment at  $\rho$  (rho) = 0.307). The pipeline is operationalised in a Streamlit prototype that preserves intermediate artefacts and documents governance boundaries, while explicitly acknowledging demo-stage limitations and future work requirements.

## 8 Annexes

Chapter 8 contains the supplementary materials and supporting artefacts of the thesis.

## 8.1 List of Abbreviations

Abbreviation	Full Form
A1–A29	Attribute catalogue identifiers
AI	Artificial Intelligence
API	Application Programming Interface
BA/BSc	Bachelor of Arts / Bachelor of Science
BPROF	BPROF curriculum
COCO Y0	Component-Based Object Comparison for Objectivity (Y0 variant)
CORREL	CORREL (Excel function)
COVID-19	Coronavirus Disease 2019
CPU	Central Processing Unit
CSV	Comma-Separated Values
EDM	Educational Data Mining
EXCEL	Microsoft Excel
GB	Gigabyte
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HTML	HyperText Markup Language
ICT	Information and Communication Technologies
IDF	Inverse Document Frequency
ISO-8601	International Organization for Standardization 8601 format
IT	Information Technology
K+F	Kutatás-fejlesztés
KJU	Kodolányi János University
LA	Learning Analytics
LMS	Learning Management System
LSA	Latent Semantic Analysis
MCDCA	Multi-Criteria Decision Analysis
NLP	Natural Language Processing
OAM	Object–Attribute Matrix
OLAP	Online Analytical Processing
ORM	Object-Relational Mapping
PROMETHEE	Preference Ranking Organization METHod for Enrichment Evaluations
Q&A	Question and Answer
R&D	Research and Development
RAM	Random Access Memory
RANK.AVG	RANK.AVG (Excel function)
S1	S1 (staircase diagnostic value)
SQL	Structured Query Language
TF	Term Frequency
TF-IDF	Term Frequency–Inverse Document Frequency
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
UI	User Interface
UNIX	UNiplexed Information and Computing System time
URL	Uniform Resource Locator
V1–V3	Verification tests V1–V3
VLE	Virtual Learning Environment
XLSX	Office Open XML Spreadsheet (.xlsx)
Y0	Y0 (constant target variable)

Table 8.1- I Abbreviations (Source: Own Presentation)

## 8.2 List of Figures

Figure 2.1- I The learning analytics cycle. (Source: Adapted from (Clow, 2013). February, 01, 2026, from <https://doi.org/10.1080/13562517.2013.827653>)

Figure 3- I End-to-end analytical pipeline of the proposed evaluation framework. (Source: Own Presentation)

[Figure 3.1.1- I Structure of the raw Moodle log dataset used in this study. \(Source: Adapted from the author’s conference presentation slides \(Turtogtokh et al., 2025\), retrieved February, 01, 2026, from \[https://miau.myx.hu/miau/319/performances/Conference\\\_PPT.pptx\]\(https://miau.myx.hu/miau/319/performances/Conference\_PPT.pptx\)\)](#)

Figure 3.1.7- I Function-symmetry validation logic. (Source: Own Presentation).

Figure 3.2- I Full operational flow of the demo application. (Source: Own Presentation)

Figure 3.2.2- I Directory structure of the demo application repository. (Source: Own Presentation)

Figure 3.2.3- I Session State Manager. (Source: Own Presentation)

Figure 3.2.5- I High-level validation logic of the demo application. (Source: Own Presentation)

Figure 3.2.6- I Page navigation flow of the demo application. (Source: Own Presentation)

### 8.3 List of Tables

Table 1.4- I Two-layer economic analysis per stakeholder group (per semester; per typical course; currency: EUR) (Source: Own Presentation)

Table 1.4- II Example scaling to an institution (illustrative; 50 similar courses) (Source: Own Presentation)

Table 2.12- I Comparative positioning of the present thesis among selected parallel COCO-oriented theses (Source: Own Presentation)

Table 1.6- I Formatting and typographic conventions. (Source: Own Presentation)

Table 3.1.2- I Raw dataset schema: header fields and contents used in the discussion-level analysis. (Source: extracted from Moodle export (Annex §8.6.5))

Table 3.1.3- I Example structure of reply-level topic relevance results. (Source: Own Presentation)(Unit: topic\_relevance\_score: model score, topic\_relevance\_score\_0\_100: aggregated model score 1-100)

Table 3.1.5- I Attribute dictionary and direction rules. (Source: Own Presentation)

Table 3.1.5- II Description of dataset header fields and their analytical roles in the proposed workflow. (Source: Own Presentation)

Table 3.1.5- III Direction codes and the corresponding decision rules used to interpret attribute influence within the evaluation framework. (Source: Own Presentation)

Table 3.1.5- IV Object attribute matrix. (Source: Own Presentation Annex §8.6.1 sheet: OAM\_ALL\_ATT)

Table 3.1.5- V Excel formula used to generate the ranked Object–Attribute Matrix (ranked OAM) from the base OAM by applying the project’s rule-based scoring and ordering logic. (Source: Own Presentation, Annex §8.6.1 sheet: Rank\_All)(Unit: Ranked ordinal values)

Table 3.1.5- VI Ranked OAM. (Source: Own Presentation, Annex §8.6.1) (Unit: Ranked ordinal values)

Table 3.1.6- I Excluded-attribute cases under the Stairs-2 table from COCO, where the row S1 value is one less than the total number of objects. (Source: Own Presentation, Annex §8.6.1)  
(Unit: staircase-function values)

Table 3.1.8- I Full attribute set. Result summary. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- II Excluded attributes result. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- III List of excluded attributes. (Source: Own Presentation).

Table 3.1.8- IV Diligence attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- V Excluded diligence attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- VI List of excluded attributes for the diligence run. (Source: Own Presentation)

Table 3.1.8- VII Understanding attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- VIII Excluded understanding attribute result summary table. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- IX List of excluded attributes for the understanding run. (Source: Own Presentation)

Table 3.1.8- X Tie-handled rank matrix (RANK.AVG-transformed ranks) for objective and subjective comparisons. (Source: Own Presentation, Annex §8.6.1)(Unit: Ranked ordinal values)

Table 3.1.8- XI Spearman correlations between objective (COCO) and subjective rank variants. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.8- XII Correlation between Full attribute set and excluded attribute set. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.9- I Distorted input matrix. (Source: Own Presentation, Annex §8.6.1)

Table 3.1.9- II COCO output (erratic estimates / unstable separation). (Source: Own Presentation, Annex §8.6.1)

Table 3.1.9- III Distorted input (single-object dominance). (Source: Own Presentation, Annex §8.6.1)

Table 3.1.9- IV COCO output (flat/degenerate separation). (Source: Own Presentation, Annex §8.6.1)

Table 3.1.9- V Distorted input (synthetic deterministic structure). (Source: Own Presentation, Annex §8.6.1)

Table 3.1.9- VI COCO output (flat/degenerate result). (Source: Own Presentation, Annex §8.6.1)

Table 3.2.2- I Module-to-responsibility mapping of the demo application, summarising the core components, their implementation files, and their functional roles within the evaluation workflow. (Source: Own Presentation).

Table 3.2.8- I Consolidated evidence from the two persisted runs in runs.db, showing completion status, data volume, and traceability indicators required for structured post-run analysis. (Source: Own Presentation, Annex §8.6.7)

Table 3.3.1- I Full student-level benchmark results generated by the independent LLM solution. (Source: ChatGPT benchmark output; see Annex §8.7.2)

Table 8.1- I Abbreviations (Source: Own Presentation)

Table 8.5- I Definitions (Source: Own Presentation)

Table 8.5- II Units, scales, and coding notes used in the quantitative parts of the thesis. (Source: Own Presentation)

Table 8.7- I Verbatim User Prompt Submitted to ChatGPT and the Corresponding Verbatim LLM Response (Source: Annex §8.6.9)

Table 8.7- II LLM uses on thesis development (Source: Own Presentation)

Table 8.11- I Reference coverage across the T01–T16 classification scheme used in the thesis (Source: Own compilation based on the institutional reference-classification audit (See Annex §8.6.8)).

## 8.4 List of Equations

Equation I Cosine Similarity Formula

Equation II Similarity Score Scaling

Equation III Topic Relevance Mean

Equation IV AI Score Scaling

Equation V Student AI Score Mean

Equation VI Baseline Target Value

Equation VII Direction-Aware Rank Formula

Equation VIII Rank Matrix Inversion

Equation IX Delta-to-Fact Ratio

Equation X Symmetry Validity Rule

Equation XI Cohort Rank Inversion

## 8.5 List of Definitions

Term	Definition
<b>Activity log</b>	A time-stamped record of user actions captured by a digital system (e.g., Moodle), typically at event level.
<b>Activity theory</b>	A socio-cultural framework for analysing human activity as a mediated system (subject–tools–object) shaped by rules, community, and division of labour.
<b>Algorithmic decision support</b>	Computational outputs (descriptive, predictive, or comparative) designed to assist human judgement rather than replace it.
<b>all-MiniLM-L6-v2</b>	A lightweight transformer-based SentenceTransformer model that maps sentences to dense vector embeddings for semantic comparison.
<b>Anonymization</b>	A data-processing transformation intended to remove or obscure identifiers so individuals are not directly identifiable in the dataset.
<b>Anti-discriminative optimisation</b>	An aggregation strategy that seeks to avoid arbitrary weighting and reduce discriminatory effects by using validated, structure-aware comparisons rather than subjective coefficient choices.
<b>Attribute catalogue (A1–A29)</b>	A controlled set of operationalised indicators, each with a defined measurement rule and directionality, used to represent student performance facets.
<b>Attribute directionality</b>	A normative rule specifying whether ‘higher is more ideal’ or ‘lower is more ideal’ for an indicator when constructing comparable rankings.
<b>Attribute exclusion</b>	A diagnostic procedure in which selected attributes are removed to test whether rankings are overly dependent on particular indicators.
<b>Auditability</b>	The property that outputs can be inspected and traced back to concrete evidence and rules, enabling review and accountability.
<b>Baseline target value</b>	A constant reference value used as the target in COCO’s Y0 setting, meaning there is no empirical dependent variable to predict.
<b>COCO Y0</b>	Component-Based Object Comparison for Objectivity in its Y0 (constant-target) configuration, producing composite estimates and rankings from multi-attribute inputs.
<b>Cosine similarity</b>	A similarity measure computed as the cosine of the angle between two vectors, commonly used for comparing text embeddings.
<b>Decision support vs automated judgement</b>	A scope distinction in which systems provide evidence and structured comparisons but do not autonomously decide high-stakes outcomes.
<b>Direction codes</b>	Explicit labels (e.g., ‘higher- more ideal’, ‘lower- more ideal’) attached to each attribute to standardise ranking logic across heterogeneous metrics.
<b>Discussion-level analysis</b>	Aggregation of log data at the discussion or thread level (e.g., per forum topic) to capture participation patterns and timing.
<b>Educational data mining</b>	A research field focused on applying data mining methods to educational datasets to understand learning processes and outcomes.
<b>Sentence embeddings</b>	Dense vector representations of text that encode semantic information, enabling meaning-oriented similarity comparisons.
<b>Explainability</b>	The extent to which a model’s outputs can be interpreted in terms of understandable components and evidence.
<b>Function-symmetry validation</b>	A robustness test in which the evaluation is rerun with direction inversions or symmetric transformations to check that outputs respond coherently to value-direction assumptions.
<b>Governance constraints</b>	Institutional and legal requirements (privacy, transparency, accountability) shaping what data may be collected and how results may be used.
<b>Hugging Face models</b>	Pretrained NLP models distributed via the Hugging Face ecosystem, typically downloaded and cached locally for inference.
<b>Indicator</b>	An operationalised measurable quantity derived from trace data intended to represent a learning-related construct (e.g., engagement, response latency).
<b>Interpretability</b>	The property that indicators and composite results can be understood and justified by humans rather than treated as opaque outputs.
<b>Input Verification Test</b>	A set of constructed distortion scenarios designed to stress-test COCO outputs for instability, degeneracy, or non-behavioural artifacts.
<b>Learning analytics cycle</b>	A feedback-loop model linking learners, data, metrics, interpretation, and intervention to improve learning environments.

<b>Term</b>	<b>Definition</b>
<b>Learning Management System (LMS)</b>	Software platform that supports course delivery and records learner interactions (e.g., Moodle).
<b>Lexical similarity</b>	Similarity computed from surface word overlap or term statistics (e.g., bag-of-words, TF-IDF), offering transparency but limited meaning capture.
<b>Moodle</b>	An open-source LMS used as the empirical platform and log-data source for the thesis pipeline.
<b>Moodle logs</b>	System-generated event records capturing user actions (views, posts, edits, timestamps) within Moodle.
<b>Multi-Attribute Evaluation</b>	Evaluation approach that combines multiple indicators to avoid single-score reductionism and to represent performance as multi-dimensional.
<b>Multi-Criteria Decision Analysis (MCDA)</b>	A family of methods for aggregating multiple criteria into comparative decisions, often with explicit preference modelling.
<b>Object-Attribute Matrix (OAM)</b>	A matrix where rows are objects (students) and columns are attributes (indicators), forming the structured input to the evaluation engine.
<b>OLAP</b>	A class of analytical techniques supporting multi-dimensional querying and reporting over structured datasets.
<b>ORM</b>	A software technique that maps database tables to program objects to simplify persistence operations.
<b>Ranked OAM</b>	A transformed OAM in which heterogeneous raw attributes are converted to comparable rank values based on direction rules.
<b>Reproducibility package</b>	A documented bundle (code, data, configuration, and instructions) enabling others to rerun the analysis and obtain comparable outputs.
<b>RoBERTa “OpenAI detector”</b>	A RoBERTa-based classifier used to output a probability-like signal of AI-generated text.
<b>Robustness checks</b>	Deliberate perturbations and validation runs used to test whether outputs behave plausibly under input distortions or alternative configurations.
<b>Semantic similarity</b>	Meaning-oriented similarity computed using distributed representations (e.g., embeddings) rather than surface word overlap.
<b>Similarity analysis</b>	A family of methods for comparing objects based on multi-attribute patterns and distances/similarities rather than single-variable scoring.
<b>Spearman correlation</b>	A non-parametric rank correlation coefficient measuring monotonic association between two ranked variables.
<b>Staircase diagnostic</b>	A COCO output diagnostic (including S1) used to detect near-monotone structures and identify attributes whose exclusion yields maximal separation patterns.
<b>Streamlit</b>	A Python framework for building interactive data apps with a web UI.
<b>TF-IDF</b>	A term-weighting scheme that scales term counts by their inverse corpus frequency to support lexical text similarity representations.
<b>TOPSIS</b>	Technique for Order Preference by Similarity to Ideal Solution, an MCDA method ranking options by distance to an ideal point.
<b>PROMETHEE</b>	Preference Ranking Organization METHod for Enrichment Evaluations, an outranking MCDA method family.
<b>Traceability</b>	The ability to link computed indicators and rankings back to observable actions and pre-processing rules.
<b>UNIX time</b>	A timestamp convention representing seconds since the UNIX epoch, commonly used in system logs.
<b>reply-level relevance</b> <b>topic</b>	An operational measure computed at the level of individual replies, intended to quantify topical alignment between a student response and a reference post.

Table 8.5- 1 Definitions (Source: Own Presentation)

Symbol / item	Meaning	Notes / units
<b>N, n</b>	Number of students / objects in the analyzed cohort	<b>count (students / objects)</b>
<b>posts, replies, messages, threads, events, rows</b>	Recorded Moodle activity volumes	<b>count</b>
<b>h / min / s</b>	Time measures	<b>hours / minutes / seconds</b>
<b>EUR / HUF / USD</b>	Monetary values used in value-estimation sections	<b>currency units</b>
<b>0 = ↑; 1 = ↓</b>	Direction code used in OAM and ranking	<b>0 = benefit-type (higher is more ideal); 1 = cost-type (lower is more ideal)</b>
<b>0 / 1</b>	Binary indicator / validation flag	<b>binary code</b>
<b>Y</b>	Constant COCO baseline target	<b>COCO index units; fixed baseline = 1000 units</b>
<b>estimation</b>	COCO-estimated value for an object	<b>COCO index units</b>
<b>delta/fact</b>	Relative deviation from baseline	<b>%</b>
<b>delta/fact inverse</b>	Relative deviation from baseline in inverse-run validation	<b>%</b>
<b>rank</b>	Position in the resulting ordering	<b>ordinal rank position</b>
<b>ρ (Spearman)</b>	Rank-correlation coefficient	<b>unitless, range [-1, 1]</b>
<b>cosine similarity</b>	Semantic similarity between sentence embeddings	<b>unitless, theoretical range [-1, 1]</b>
<b>topic_relevance_score</b>	Student-level topical alignment indicator	<b>score (0–100)</b>
<b>avg_AI_involvedMsg_score</b>	Auxiliary AI-involvement indicator	<b>rating (1–10)</b>
<b>valid_response</b>	Correctness/validity flag for specified task responses	<b>binary code (0/1)</b>

Table 8.5- II Units, scales, and coding notes used in the quantitative parts of the thesis. (Source: Own Presentation)

## 8.6 Reproducibility Package

Section 8.6 presents the reproducibility package of the thesis by providing access to the principal analytical, technical, and operational artefacts required to inspect, repeat, and verify the implemented workflow.

### 8.6.1 Excel file

[https://miau.my-x.hu/miau/330/st/OAM\\_Moodle\\_COCO.xlsx](https://miau.my-x.hu/miau/330/st/OAM_Moodle_COCO.xlsx)

This workbook contains the core Object–Attribute Matrix, ranked variants, COCO-related sheets, and intermediate analytical tables used in the empirical evaluation.

### 8.6.2 Model source code directory

<https://miau.my-x.hu/miau/315/moodle/>

This directory contains the source-code files, scripts, and supporting implementation materials used for data preparation, indicator generation, and analytical processing.

### 8.6.3 GitHub repository

[https://github.com/Shagai-hub/Moodle\\_Log\\_Analysis\\_DEMO.git](https://github.com/Shagai-hub/Moodle_Log_Analysis_DEMO.git)

This repository contains the prototype software of the thesis, including the Streamlit-based prototype and the main application structure used to operationalise the workflow.

### 8.6.4 Peer evaluation Excel file

[https://miau.my-x.hu/miau/315/moodle/quasi\\_exam\\_II\\_collected.xlsx](https://miau.my-x.hu/miau/315/moodle/quasi_exam_II_collected.xlsx)

This file contains the collected peer-evaluation results used as the subjective comparison basis in the objective–subjective ranking analysis.

### 8.6.5 Database

[https://miau.my-x.hu/miau/330/st/Moodle\\_logs.db](https://miau.my-x.hu/miau/330/st/Moodle_logs.db)

This database contains the structured Moodle export after cleaning and database construction, serving as the main local data source of the analytical pipeline.

### 8.6.6 Working principle of COCO

<https://miau.my-x.hu/myx-free/index.php3?x=e09>

This resource documents the methodological background and operating logic of the COCO framework used for the similarity-analysis-based aggregation in the thesis.

### 8.6.7 Demo test results

<https://miau.my-x.hu/miau/330/st/runs.db>

This file contains stored run-level outputs of the demonstrator, including test executions and generated results used for inspection of prototype behaviour.

### 8.6.8 T01–T16 Reference-Coverage Workbook

<https://miau.my-x.hu/miau/330/st/T01-T016-Coverage.xlsx>

This annex item provides the source spreadsheet underlying the reference-coverage table presented in §8.11.

### 8.6.9 Shared LLM Conversation

<https://chatgpt.com/share/69bddd44-5c04-8008-b704-79cfce3f200b>

This annex item provides the public share link to the documented ChatGPT conversation associated with the thesis-support workflow referenced in the annex materials.

## 8.7 LLM: prompt, output, and artefacts

The following materials are reproduced verbatim for documentation and auditability purposes; they are included as annexed evidence of LLM interaction and do not form part of the thesis's own argumentative prose. The example below documents the use of ChatGPT for structural review during the organisation of Chapter 3, with the aim of refining subsection order, reducing redundancy, and strengthening the academic coherence between the analytical and software-development parts of the thesis.

Table 8.7- I presents the verbatim user prompt and corresponding LLM output associated with the structural review of Chapter 3. The material is reproduced unchanged for documentation and auditability purposes and is linked to the public share source listed in [Annex §8.6.9](#).

**Prompt:**

Current draft structure of Chapter 3:  
Chapter 3. Own developments  
Chapter 3.1. Analytical process  
Chapter 3.1.1. Moodle Data & Structure  
Chapter 3.1.2. Log Data Extraction  
Chapter 3.1.3. Cosine Similarity in Text Responses  
Chapter 3.1.4. AI Detection in Text Responses  
Chapter 3.1.5. OAM Formation & Ranking  
Chapter 3.1.6. COCO Y0 Analysis & Attribute Exclusion  
Chapter 3.1.7. Function-Symmetry Validation  
Chapter 3.1.8. Input Verification Test  
Chapter 3.1.9. Subjective vs Objective Comparison  
Chapter 3.2. Demo Software Development  
Chapter 3.2.1. Purpose and Scope of the Demo Tool  
Chapter 3.2.2. Requirements  
Chapter 3.2.3. System Architecture  
Chapter 3.2.4. Data-Pipeline Implementations  
Chapter 3.2.5. Similarity Analysis and AI Detection in Text Responses  
Chapter 3.2.6. Ranking & Scoring Logic  
Chapter 3.2.7. User Interface Interaction Design  
Chapter 3.2.8. Visual Outputs  
Chapter 3.2.9. AI Interpretation  
Chapter 3.2.10. Testing  
Chapter 3.2.11. IT Security, Privacy, and Ethics in Implementation  
Chapter 3.2.12. Limitations

Please review this structure only from the perspective of chapter organisation and academic flow.

Focus on the following:

- whether the separation between §3.1 and §3.2 is logically clear,
- whether any subsection titles overlap too much or repeat the same idea,
- whether any titles should be tightened or renamed for academic clarity,

-whether the order supports the “golden thread” from analytical method to prototype implementation,  
-whether there is any missing structural element that would improve reproducibility or readability.

Please do not write the chapter itself. I only want structural feedback, suggested renamings, and brief reasoning for each recommendation. Keep the suggestions academically phrased and aligned with the style of my thesis.

**ChatGPT Output:**

This is a strong Chapter 3 structure: it separates the analytical/scientific workflow (3.1) from the software artifact (3.2), which is exactly what reviewers expect.

That said, there are a few structural issues to fix so it’s “examiner-proof” and avoids overlap/repetition.

What’s already good

3.1 Analytical process = the research pipeline (data → features → similarity → COCO → validation).

3.2 Demo software development = requirements → architecture → implementation → UI → testing → ethics.

You included multiple test/validation items (excellent).

What to improve structurally

1) Remove duplication between 3.1 and 3.2

Right now you describe similarity + AI detection twice:

3.1.3 Cosine similarity... and 3.1.4 AI detection...

3.2.5 Similarity analysis and AI detection...

Fix: keep algorithmic details in 3.1, and in 3.2.5 only describe *integration in the app* (UI settings, caching, runtime, batching, failure modes).

Rename 3.2.5 to:

3.2.5 Integretion of Similarity and AI-Detection Modules

2) Clarify what COCO Y0 “attribute exclusion” means

“Attribute exclusion” sounds like feature selection and could be controversial if unclear.

Rename 3.1.6 to something like:

3.1.6 COCO Y0 Impact Analysis and Attribute Sensitivity

(or “Attribute Impact and Reduction” if you truly remove attributes)

This is a safer academic label.

3) Put “data pipeline implementations” in the right place

You have:

3.1.2 Log Data Extraction (analysis)

3.2.4 Data-Pipeline Implementations (software)

That’s okay, but distinguish:

3.1.2 = what extraction means scientifically (schema, cleaning rules)

3.2.4 = how it’s engineered in code (modules, functions, performance)

Rename 3.2.4 to:

3.2.4 Implementation of the Data Pipeline in the Application

4) “AI interpretation” needs a clearer purpose

3.2.9 “AI interpretation” is vague. Likely you mean how the teacher should interpret detector scores.

Rename it to:

3.2.9 Interpretation Guidelines and User Warnings for AI-Detection Scores

This also protects you ethically.

5) Add one missing development piece: deployment/reproducibility

Since you promise reproducibility packages, add a short subsection under 3.2:

Add:

3.2.13 Reproducibility and Operational Setup (Environment, Versions, Configuration)

If you don’t want another number, merge into Testing or Limitations, but having it explicit is strong.

Small language fixes

“Analisy” → Analysis  
 “Implimintations” → Implementations  
 “Demo Software development” → Demo Software Development (capitalization consistency)  
 Best-practice revised Chapter 3 (minimal changes, but polished)  
 Chapter 3. Own Developments  
 3.1 Analytical Process  
 3.1.1 Moodle Data and Structure  
 3.1.2 Log Data Extraction and Preprocessing  
 3.1.3 Semantic Similarity for Descriptive Text Responses (Cosine Similarity)  
 3.1.4 AI-Generated Text Signal in Student Responses (Detector as Supportive Indicator)  
 3.1.5 OAM Construction and Initial Ranking Logic  
 3.1.6 COCO Y0 Impact Analysis and Attribute Sensitivity  
 3.1.7 Function-Symmetry Validation  
 3.1.8 Input Verification Tests  
 3.1.9 Subjective vs Objective Evaluation Comparison  
 3.2 Demo Software Development (Streamlit Prototype)  
 3.2.1 Purpose and Scope of the Demo Tool  
 3.2.2 Requirements (Functional and Non-functional)  
 3.2.3 System Architecture  
 3.2.4 Implementation of the Data Pipeline in the Application  
 3.2.5 Integration of Similarity and AI-Detection Modules  
 3.2.6 Ranking and Scoring Logic Implementation  
 3.2.7 User Interface and Interaction Design  
 3.2.8 Visual Outputs and Exportable Reports  
 3.2.9 Interpretation Guidelines and User Warnings for AI-Detection Scores  
 3.2.10 Testing (Unit, Integration, Scenario Tests)  
 3.2.11 IT Security, Privacy, and Ethics in Implementation  
 3.2.12 Limitations  
 3.2.13 Reproducibility and Operational Setup (optional but recommended)  
 Bottom line  
 Your structure is already very good. With the tweaks above, it becomes:  
 less redundant,  
 safer in academic wording,  
 clearer about what is analysis vs what is software.

Table 8.7- I Verbatim User Prompt Submitted to ChatGPT and the Corresponding Verbatim LLM Response (Source: Annex §8.6.9)

The response was not inserted directly into the thesis. Instead, it was used as a structural checkpoint: the proposed structural changes were manually reviewed, selectively retained, and subsequently verified against the thesis’s research logic and internal cross-references.

Table 8.7- II summarises the use of LLM support in the thesis writing and development process.

Area of use	Main purpose	Typical support received	Thesis relevance	Author control
<b>Language refinement</b>	Improve academic phrasing and readability	Sentence-level reformulation, punctuation improvement, transition strengthening	Literature review, method description, discussion wording	All outputs manually reviewed and revised
<b>Internal coherence</b>	Strengthen the “golden thread” across chapters	Alignment checks between aims, research questions, methods, findings, and discussion	Chapter 1–4 consistency	Suggestions adopted selectively only where appropriate
<b>Structural advice</b>	Improve chapter/subchapter organisation	Proposal of subsection structure, bridge sentences, design-implication links	Especially Chapter 2 and Chapter 3	Final structure decided by the author
<b>Abbreviation and terminology support</b>	Improve consistency of technical wording	Extraction and standardisation of repeated terms and abbreviations (§8.1, and §8.5)	Whole thesis	Terminology checked manually against actual usage
<b>Development support</b>	Assist in debugging and technical clarification	Help with SQL, Python, Streamlit logic, error interpretation, and implementation steps	Prototype and analytical workflow	No code or result accepted without testing
<b>Formatting and error detection</b>	Identify presentation problems in the document	Detection of wording issues, formatting inconsistencies, and weak sections	Document preparation	Final corrections performed manually by the author

Table 8.7- II LLM uses on thesis development (Source: Own Presentation)

### 8.7.1 Verbatim benchmark prompt and full LLM output

This annex contains the complete benchmark prompt submitted to the LLM and the complete benchmark response exactly as generated.

<b>Prompt:</b>
----------------

You are acting as an independent academic analyst and methodologist. Your task is to design and execute a complete, self-contained, methodologically justified evaluation framework for assessing student performance from Moodle discussion forum log data in a teacher-facing decision-support context. This is a benchmark exercise: your solution must be developed entirely from first principles, using your own chosen method, indicators, and aggregation logic. You must not approximate, replicate, or reverse-engineer any pre-existing framework. Your output must be of sufficient academic quality to serve as a comparison case in a bachelor-level thesis discussion and annex.

The research problem you must address

Given a dataset of Moodle discussion forum exports from a higher-education IT course, develop an objective, transparent, and academically defensible approach to evaluating student performance at the student level, suitable for teacher-facing decision support. The evaluation must draw on both the behavioural/log traces available in the data and the text-based content of student contributions. The output must be interpretable, bounded in its claims, and accompanied by methodological justification and a checking procedure.

The dataset you will work with

The dataset is a pseudonymised Moodle forum log export. You may assume the following structure is available:

Each row is one forum post or reply, with fields including: a post/message unique identifier, a discussion thread identifier, a parent post identifier (0 if the post is a top-level thread opener), a pseudonymised user identifier, timestamps (creation and modification), the message body text, word count, and character count.

The dataset covers approximately one month of activity in a single course, contains approximately 24 student participants, approximately 12 discussion threads, and more than 5 hundred messages.

Two user identifiers belong to the course instructor; the remaining identifiers belong to students.

Some messages are direct replies to instructor posts; others are replies to peer posts.

You must use this full dataset in your analysis. No subset exclusion is permitted unless methodologically justified and explicitly stated.

Requirements for your methodology

You must develop your evaluation framework independently. Do not reproduce or closely approximate any known multi-attribute decision analysis method applied to Moodle data that you may have encountered. Your method may draw on any legitimate analytical tradition provided that your choice is explained and justified.

Your methodology section must explicitly address all of the following:

Unit of analysis. State clearly what constitutes one object in your analysis (e.g., one student across the full period, one student per thread, etc.) and justify why that unit is appropriate for the stated decision-support goal.

Indicator/attribute selection. Enumerate the specific indicators you will compute. For each indicator, provide: its name, a precise operational definition stating how it is computed from the available fields, the theoretical or practical rationale for its relevance to student performance evaluation.

Aggregation and combination logic. Describe precisely how individual indicators are combined into a student-level output. State whether you use a weighted composite, a ranking transformation, a distance-based measure, a rule-based classification, a cluster assignment, or another approach. Justify the aggregation logic explicitly, including a statement about whether and how you avoid introducing arbitrary or unverifiable weights.

Final student-level outputs. Produce, for every student in the dataset, a complete results table containing: the student pseudonym, the value of each computed indicator, the aggregated or composite output, and the resulting ranking, grouping, score, or profile assignment. Provide a brief interpretive annotation for students or group.

Descriptive summary. Provide cohort-level descriptive summaries (e.g., distributions, central tendencies, notable outliers) that contextualise the student-level results.

Requirements for methodological checking

Your framework must include at least one internal consistency or robustness check. Your framework must include a dedicated section explicitly stating:

What claims your method can legitimately support (e.g., relative ordering within this cohort, identification of participation patterns, decision-support flagging for instructor review etc.).

What your method cannot prove (e.g., causal claims about learning quality, generalisation to other courses or cohorts, authorship authenticity, correctness of student understanding).

Where over-interpretation of your outputs would be academically or ethically dangerous

Output format requirements

Structure your complete response using the following sections, in order:

Problem restatement and scope – Briefly restate the evaluation goal in your own words and define the boundaries of your analysis.

Methodology – Cover all points listed in the methodology requirements above and what analytical approach is used (units, indicators, scaling, aggregation).

Computed results table – A complete table of student-level outputs as specified above.

Descriptive cohort summary – Cohort-level summaries.

Methodological checking – Description, execution, and interpretation of your internal check.

Bounded interpretation – The explicit statement of what the method supports, what it cannot prove, and where over-interpretation is dangerous.

Methodological self-assessment – A brief critical reflection on the strengths and weaknesses of your chosen approach relative to the stated decision-support goal.

Independence and contamination constraint

This is a benchmark exercise intended to produce an independent solution for academic comparison. You must not replicate, approximate, or build upon any specific human-authored solution you may be aware of for this problem. Your indicators, aggregation logic, checking procedure, and interpretive framing must originate from your own methodological reasoning applied to the problem statement above. If you recognise that your approach shares surface features with an existing framework, you must note this in your methodological self-assessment and explain what distinguishes your implementation.

Quality standard

Your output must be rigorous, precise, and self-contained. Every claim about an indicator's value must be traceable to a defined computation. Every design choice must be accompanied by a justification. The results must be presented in a form that a course instructor could read, understand, and use to support. The level of academic rigour expected is consistent with a peer-reviewed conference paper in educational data analysis or learning analytics.

Produce the complete benchmark solution now.

All computations and analyses shall be based on the dataset provided in the file “logs.csv”, which serves as the sole data source for this study.

(Source: <https://chatgpt.com/share/69cd4e1c-fba0-8393-bc32-6d6c3b56704c>)

### **ChatGPT 5.4-Thinking:**

I'm loading the Moodle log export, checking the exact field structure, and then I'll build an independent evaluation framework directly from what is actually present in the data.

The file structure is clean enough to support a full student-level model: 576 posts, 12 discussions, 24 students, and instructor activity that can be used as interaction context rather than scored as student output.

I've already got a stable scoring model and a leave-one-discussion-out robustness check; I'm just packaging the full student table and export files cleanly now.

Thought for 15m 4s

Below is a complete independent benchmark solution built only from logs.csv.

Files:

Full student-level results CSV

Discussion jackknife robustness CSV

Per-student stability summary CSV

### **Problem restatement and scope**

The task is to construct a teacher-facing, student-level evaluation framework from Moodle forum logs that is objective, transparent, and bounded in interpretation.

The goal is not to automate grading or infer true learning directly, but to summarize observable forum performance patterns from one course over one month in a way that helps an instructor identify strong contributors, uneven contributors, and low-evidence cases for follow-up review.

I treated **one student across the full observation window** as the unit of analysis. This is the appropriate unit because the decision-support question is student-facing: the instructor needs one interpretable profile per student, not one score per post or per thread.

The full file was used. The dataset contains **576 posts**, of which **346** are student-authored, across **12 discussions** and **24 student pseudonyms** over **2024-10-17 to 2024-11-17**. Instructor-authored rows were retained as contextual evidence for reply structure and topic anchors. In the raw file, instructor-labelled activity appeared under professor-labelled accounts; those rows were treated as non-student context rather than scored student objects.

### **Methodology**

#### **Analytical approach**

I designed a new framework called the **Triangulated Percentile-Geometric Profile (TPGP)**.

Its logic is:

Extract a small set of interpretable indicators from three evidence pillars:

participation

interaction

content

Convert each indicator to a **cohort-relative percentile position** so that different scales do not dominate one another.

Aggregate indicators **within each pillar by the median**, to reduce outlier sensitivity and avoid one indicator dominating a pillar.

Aggregate the three pillar scores by the **geometric mean**, so that very weak performance in one pillar cannot be fully cancelled by strength in another.

This design avoids manual topic weights and avoids assuming that raw counts, semantic similarities, and timing measures live on commensurate scales.

#### **Preprocessing**

Message text was cleaned by normalizing whitespace, removing markup if present, and decoding escaped text. Reply structure was reconstructed from the parent field. Thread openers were identified as the earliest post in each discussion. For text-based indicators, I used course-internal TF-IDF vectors and cosine similarity; this keeps the method self-contained and anchored in the actual discourse of this course.

#### **Indicator definitions**

Pillar	Indicator	Operational definition	Rationale
Participation	<b>posts</b>	Total number of posts authored by the student in the observation window	Basic observable contribution volume
Participation	<b>active_days</b>	Number of distinct calendar days on which the student posted	Distinguishes sustained participation from one-day bursts
Participation	<b>thread_coverage</b>	Distinct discussions participated in / 12	Captures reach across course topics rather than repetition in one thread
Interaction	<b>instructor_uptake_breadth</b>	Number of distinct discussions in which the student directly replied to an instructor post / 12	Measures uptake of teacher prompts across the course
Interaction	<b>peer_interaction_breadth</b>	Number of distinct student peers the student either replied to or received a reply from	Measures embeddedness in peer discourse
Interaction	<b>reply_promptness</b>	$\frac{1}{1 + \text{median reply latency in hours}}$	Rewards timelier engagement while keeping the scale bounded
Content	<b>median_wordcount</b>	Median wordcount across the student's posts	Proxy for contribution substance, using the median to reduce distortion from very long posts

Pillar	Indicator	Operational definition	Rationale
Content	mean_thread_similarity	Mean cosine similarity between each non-opener student post and the opener of the same discussion	Measures topical alignment with the thread anchor
Content	mean_parent_similarity	Mean cosine similarity between each student reply and its parent post	Measures local responsiveness to the post being answered

### Scaling

For each indicator  $jjj$ , student  $iii$  received a cohort-relative percentile score  $p_{ij} = \frac{r_{ij} - 0.5}{N}$  where  $r_{ij}$  is the mid-rank of that student on that indicator and  $N = 24$ .

This keeps every indicator on the same 0-1 scale without assuming normality.

### Aggregation

For each pillar, the pillar score is the median of its three indicator percentiles:

$$S_{i, \text{pillar}} = \text{median}(p_{ij})$$

The overall score is then

$$O_i = 100 \times (S_{i, \text{participation}} \cdot S_{i, \text{interaction}} \cdot S_{i, \text{content}})^{1/3}$$

Why this combination logic is justified:

**No hand-tuned weights** were introduced.

The **median** within a pillar is robust to one unusually high or low indicator.

The **geometric mean** across pillars enforces balance: strong posting volume alone is not enough if interaction or content evidence is weak.

### Final student outputs

The final outputs are:

raw indicator values

three pillar scores: Participation (P), Interaction (I), Content (C)

an overall score on a 0-100 scale

a cohort-relative rank

a quartile-based profile band

The bands are cohort-relative, not absolute achievement labels:

**Leading** = top quartile

**Strong** = upper-middle quartile

**Developing** = lower-middle quartile

**Low-evidence** = bottom quartile

### Computed results table

Column key:

thread\_cov\_% = 100 × thread coverage

instr\_uptake\_% = 100 × instructor uptake breadth

reply\_prompt = 100 × reply promptness index  
 thread\_sim = mean thread-opener cosine similarity  
 parent\_sim = mean parent-post cosine similarity  
 P, I, C = pillar scores on 0-100 scale  
 Overall = final composite score

r	a	stud	po	activ	threa	instr_	peer_intera	reply	med_	thre	pare	P	I	C	Ove	band
k	ent	st	s	s	d_cov	uptake	ction_bread	_prom	word	ad_s	nt_s				ral	l
					%	%	th	pt	s	im	im					
1	stud	ent_27	12		75	58.3	3	3.93	59	0.086	0.189	5.8	1.7	2.9	86.2	Leading
2	stud	ent_37	10		75	66.7	2	6.34	71	0.068	0.149	5.8	5.4	8.8	82.6	Leading
3	stud	ent_16	6		58.3	41.7	2	3.08	109	0.103	0.182	8.8	8.8	5.9	73.9	Leading
4	stud	ent_17	6		58.3	50	1	2.5	76	0.113	0.191	8.8	8.3	9.6	71.1	Leading
5	stud	ent_20	6		50	50	1	1.96	77.5	0.094	0.098	8.8	8.3	7.1	67.6	Leading
6	stud	ent_22	8		50	50	0	2.1	88.5	0.082	0.139	7.5	7.9	8.8	66.1	Leading
7	stud	ent_16	3		50	16.7	8	46.51	37	0.188	0.137	2.1	7.9	6.2	66.1	Strong
8	stud	ent_20	4		58.3	58.3	1	1.46	61	0.075	0.115	9.2	8.3	5.0	61.3	Strong
9	stud	ent_17	5		66.7	41.7	3	6.98	26	0.067	0.154	4.6	9.6	9.6	61.2	Strong
10	stud	ent_14	6		50	50	1	7.67	54.5	0.067	0.157	6.0	7.7	3.5	54.8	Strong

rank	student	posts	active_days	thread_cov_%	instr_uptake_%	peer_interaction_breadth	reply_prompts	med_words	threads	parent_sims	P	I	C	Overall	band
11	student_19	11	4	50	50	0	5.74	63	0.05	0.153	7	7	6	54.6	Strong
12	student_12	14	7	50	41.7	2	1.76	61	0.08	0.063	0	7	5	52.0	Strong
13	student_17	21	8	50	41.7	1	5.76	66	0.062	0.093	5	8	7	51.3	Developing
14	student_8	18	5	41.7	41.7	0	1.45	81.5	0.081	0.129	2	2	4	42.6	Developing
15	student_5	8	4	41.7	41.7	0	1.42	73	0.08	0.041	5	2	6	35.7	Developing
16	student_14	3	3	25	16.7	1	2.68	59	0.097	0.097	8	8	1	35.7	Developing
17	student_2	16	5	41.7	41.7	0	1.7	69.5	0.054	0.066	2	5	2	34.8	Developing
18	student_23	4	3	8.3	8.3	2	2.5	56.5	0.029	0.149	8	0	5	34.2	Developing
19	student_21	17	6	41.7	41.7	0	1.61	49	0.051	0.039	4	1	8	33.6	Low-evidence
20	student_18	12	4	41.7	41.7	0	2.25	35.5	0.038	0.094	5	7	4	29.1	Low-evidence

rank	student	posts	active_days	threads_cov_%	instructor_uptake_%	peer_interaction_threads	reply_prompts	median_words	threads_sim	parent_sim	P	I	C	Overall	band
21	student_103	10	3	33.3	33.3	0	0.83	44	0.065	0.053	2.9	2.9	2.9	22.9	Low-evidence
22	student_213	2	1	16.7	16.7	0	0.35	137.5	0.161	0.111	6.2	6.7	6.3	21.4	Low-evidence
23	student_311	3	1	16.7	16.7	0	0.59	23	0.079	0.069	1.5	1.7	1.2	16.8	Low-evidence
24	student_424	1	1	8.3	8.3	0	1.02	12	0.045	0.299	4.2	4.4	4.0	8.6	Low-evidence

Group interpretation for instructor use:

**Leading:** high evidence across all three pillars; these are the most balanced high-contribution cases in this dataset.

**Strong:** above-median overall, but with one weaker pillar that may matter pedagogically.

**Developing:** mixed evidence; these students are participating, but the profile is uneven.

**Low-evidence:** sparse or weakly supported evidence in this dataset; these are not “poor students” by definition, but priority cases for human review.

The downloadable CSV includes a short student-level annotation for every row.

#### Descriptive cohort summary

The cohort is heterogeneous.

Median student activity was **16 posts** over the month, with an interquartile range of **9.5 to 18.5 posts**. The median student was active on **5 distinct days** and participated in **6 discussions**. Median median-post length was **61 words**. Student posting volume ranged from **1 post** to **37 posts**, which already indicates that any teacher-facing decision-support model needs to distinguish sustained contributors from minimally visible students.

The final overall score had:

mean = **48.53**

median = **51.90**

standard deviation = **21.34**

interquartile range = **34.07 to 65.98**

observed range = **8.59 to 86.20**

Pillar distributions were broadly similar in spread:

Participation mean = **52.0**

Interaction mean = **52.8**

Content mean = **49.2**

The cohort also shows distinct profile types, not just a single ladder of “better” and “worse” students. For example:

**student\_13** has extremely strong content evidence relative to the cohort, but almost no participation evidence.

**student\_1** is interaction-heavy and very prompt, but less strong on participation breadth and text substance.

**student\_17** is participation-strong but content-weaker.

**student\_24** has too little visible evidence for stable interpretation.

This is exactly why a three-pillar design is preferable to a pure volume count.

### **Methodological checking**

#### **Check design**

I used a **leave-one-discussion-out jackknife**. The full model was re-estimated **12 times**, each time removing one discussion entirely and recomputing all student scores and ranks from the remaining data.

This checks whether the framework is fragile to any single discussion.

#### **Check results**

Across the 12 re-estimations:

mean Spearman rank correlation with the full ranking = **0.979**

minimum Spearman rank correlation = **0.950**

maximum Spearman rank correlation = **1.000**

mean median absolute rank shift = **0.71 rank positions**

mean top-5 overlap = **4.5 of 5**

mean bottom-5 overlap = **4.67 of 5**

mean fixed-band agreement = about **81.7%**

Interpretation:

The framework is **not perfectly invariant**, which is desirable because discussion composition does matter. However, it is also **not overly fragile**. The ordering remains very similar when any single discussion is removed, and the top and bottom ends of the ranking are especially stable. Sensitivity is concentrated more in the middle of the distribution, where many real-world teacher decisions would in any case require closer qualitative review.

The full jackknife outputs are in the downloadable file above.

### **Bounded interpretation**

#### **What this method can support**

This method can legitimately support:

**relative ordering within this cohort and time window**

identification of **high-evidence**, **mixed**, and **low-evidence** participation profiles

instructor flagging for **follow-up review**

comparison of students on observable dimensions of participation reach

interactional embeddedness

textual topicality and local responsiveness

#### **What this method cannot prove**

This method cannot prove:

true subject mastery

correctness of ideas

causal effects on learning

authorship authenticity

whether longer posts are better arguments

whether semantic similarity implies genuine understanding

whether low visibility equals low ability

generalisability to another course, instructor, or cohort without recalibration

**Where over-interpretation would be dangerous**

Over-interpretation would be academically and ethically dangerous if this output were used as:

a direct grading mechanism

evidence of cheating or AI misuse

evidence of deep understanding without reading the posts

evidence that low-activity students are disengaged by choice rather than by circumstance

a cross-course benchmark without revalidation

In short, this is a **decision-support profile**, not a verdict.

#### **Methodological self-assessment**

The main strength of this framework is that it is **traceable**. Every score can be followed back to a defined computation from the raw log. It also deliberately combines behavioural and textual evidence, which is better aligned with teacher decision support than pure frequency counting. The percentile transformation makes heterogeneous indicators comparable, the median reduces indicator dominance within pillars, and the geometric mean prevents total compensation across pillars.

A second strength is that it remains **self-contained**. It does not depend on external labels, external rubrics, or externally trained educational classifiers. That is useful in a benchmark setting because the method can be reproduced entirely from the supplied file.

Its main weakness is that it is still a **composite model**. Any composite necessarily encodes design choices, even when explicit weights are avoided. Here, the design choice is structural rather than weight-based: three equally important pillars, median within pillar, geometric balance across pillars. That is defensible, but still a choice.

A further weakness is that the text indicators are **proxies rather than direct quality measures**. Topic alignment and parent responsiveness are useful, but they do not capture correctness, originality, or disciplinary sophistication. A student can be semantically aligned yet conceptually weak, or conceptually strong while writing briefly.

Finally, the framework shares a broad family resemblance with other composite student-profile approaches because it aggregates multiple indicators. What distinguishes this implementation is the specific combination of:

cohort-relative percentile scaling

median-by-pillar aggregation

geometric cross-pillar balancing

discussion-level jackknife robustness checking

(Source: <https://chatgpt.com/share/69cd4e1c-fba0-8393-bc32-6d6c3b56704c>)

### 8.7.2 Benchmark-generated artefacts and metadata

This annex subsection contains the supporting files produced by the benchmark run, including the full student-level result table, the jack knife robustness outputs, and the per-student stability summary, together with the benchmark metadata.

[https://miau.my-x.hu/miau/330/st/GPT\\_5.4\\_moodle\\_benchmark\\_jackknife.csv](https://miau.my-x.hu/miau/330/st/GPT_5.4_moodle_benchmark_jackknife.csv)

[https://miau.my-x.hu/miau/330/st/GPT\\_5.4\\_moodle\\_benchmark\\_results.csv](https://miau.my-x.hu/miau/330/st/GPT_5.4_moodle_benchmark_results.csv)

[https://miau.my-x.hu/miau/330/st/GPT\\_5.4\\_moodle\\_benchmark\\_student\\_stability.csv](https://miau.my-x.hu/miau/330/st/GPT_5.4_moodle_benchmark_student_stability.csv)

#### **Metadata of the model:**

- Model: ChatGPT 5.4-thinking
- Date: 2026-04-01 4:45 pm
- Thinking time: 15m 4s

## 8.8 Tutorial on operation

[https://miau.my-x.hu/miau/330/st/USER\\_GUIDE%202025.docx](https://miau.my-x.hu/miau/330/st/USER_GUIDE%202025.docx)

This user guide provides step-by-step operational instructions for running the prototype, configuring inputs, and interpreting the main interface outputs.

## 8.9 Licenses and ethics

- **Use of third-party services:** The prototype may run on the Streamlit hosted demo ([moodleloganalysisdemo-2025.streamlit.app](https://moodleloganalysisdemo-2025.streamlit.app)) and calls an external COCO Y0 endpoint; therefore, service availability and external data-transfer boundaries must be considered during operation.
- **Open-source libraries and compliance:** The system is built with open-source software components (e.g., Streamlit, SQLite-related tooling, NLP model libraries), so institutional use should verify and document license obligations for each dependency.
- **Data privacy and ethical data handling:** The implementation follows a local-first mode as the recommended setup, uses pseudonymised Moodle data, applies data minimization, and supports cascade deletion of datasets and derived artifacts to align with retention and privacy expectations.
- **Licensing implications:** Redistribution or broader deployment requires compliance with the licenses of application code, third-party packages, model artifacts, and any external service terms used by the workflow.
- **Warranty and liability:** Warranty and liability boundaries of the prototype are defined in §3.2.9.

## 8.10 References

- Agudo-Peregrina, Á. F., Iglesias-Pradas, S., Conde-González, M. Á., & Hernández-García, Á. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior*, *31*, 542–550. <https://doi.org/10.1016/j.chb.2013.05.031> Retrieved 15.02.2026. [T09]
- Amazon Web Services. (2026). *Billing and account management — Amazon lightsail (FAQ)*. <https://docs.aws.amazon.com/lightsail/latest/userguide/amazon-lightsail-frequently-asked-questions-faq-billing-and-account-management.html> Retrieved 15.02.2026. [T03]
- Ardito, C. G. (2025). Generative AI detection in higher education assessments. *New Directions for Teaching and Learning*, *2025*, 11–28. <https://doi.org/10.1002/tl.20624> Retrieved 15.02.2026. [T01]
- Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge - LAK '12*. <https://doi.org/10.1145/2330601.2330666> Retrieved 15.02.2026. [T09]
- Bahel, V., & Thomas, A. (2021). *Text similarity analysis for evaluation of descriptive answers*. <https://doi.org/10.48550/arXiv.2105.02935> Retrieved 10.02.2026. [T01]
- Balogh, A., Pitlik, L., Schnellbach, M., & Szani, F. (2018). *Tracking student activity in e-learning frameworks (Log-analyses in e-learning frameworks – or dark phenomena are not always Black Mirrors)*. Apertus Nonprofit Ltd. [http://miau.my-x.hu/miau/233/2500\\_3000.docx](http://miau.my-x.hu/miau/233/2500_3000.docx) Retrieved 10.02.2026. [T10]
- Bán, G., Rikk, J., & Pitlik, L. (2024). *AI-based derivation of the importance of attributes in case of evaluation models*. Kodolányi János Egyetem. [https://miau.my-x.hu/miau/314/full\\_importance\\_of\\_attributes\\_in\\_evaluation\\_models.docx](https://miau.my-x.hu/miau/314/full_importance_of_attributes_in_evaluation_models.docx) Retrieved 10.02.2026. [T02]
- Bánkúti, G., & Pitlik, L. (2010). About the method of component-based object comparison for objectivity (COCO). *Proceedings of the International Congress of Mathematicians*, 19–27. <https://miau.my-x.hu/miau/141/india.pdf> Retrieved 15.02.2026. [T09]
- Bates, A. W. (2015). *6.2 eğitim teknolojisinin kısa tarihçesi*. <https://pressbooks.bccampus.ca/tonybates/chapter/6-2-egitim-teknolojisinin-kisa-tarihcesi/> Retrieved 01.03.2026. [T15]
- Behzadian, M., Khanmohammadi Otaghsara, S., Yazdani, M., & Ignatius, J. (2012). A state-of-the-art survey of TOPSIS applications. *Expert Systems with Applications*, *39*, 13051–13069. <https://doi.org/10.1016/j.eswa.2012.05.056> Retrieved 01.03.2026. [T09]
- Below, M. (2021). *Was sagen Daten über das Lernen aus?* <https://zfl-lernen.de/en/online-kurs/big-data-literacy/daten-lernen/> Retrieved 10.02.2026. [T07]

- Brans, J. P., & Vincke, Ph. (1985). Note—A Preference Ranking Organisation Method. *Management Science*, 31, 647–656. <https://doi.org/10.1287/mnsc.31.6.647> Retrieved 10.02.2026. [T09]
- Buckley, B., Gobert, J., & Horwitz, P. (2006). Using log files to track students' model-based inquiry. *Proceedings of the 7th International Conference on Learning Sciences*, 1, 57–63. <https://dl.acm.org/doi/proceedings/10.5555/1150034> Retrieved 15.02.2026. [T09]
- Cantabella, M., López, B., Caballero, A., & Muñoz, A. (2018). Analysis and evaluation of lecturers' activity in Learning Management Systems: Subjective and objective perceptions. *Interactive Learning Environments*, 26, 911–923. <https://doi.org/10.1080/10494820.2017.1421561> Retrieved 15.02.2026. [T09]
- Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, 18(6), 683–695. <https://doi.org/10.1080/13562517.2013.827653> Retrieved 10.02.2026. [T09]
- Dormezil, S., Khoshgoftaar, T. M., & Robinson-Bryant, F. (2019). Differentiating between educational data mining and learning analytics: A bibliometric approach. *EDM (Workshops)*, 17–22. <https://ceur-ws.org/Vol-2592/short1.pdf> Retrieved 15.02.2026. [T09]
- Elkhatat, A. M., Elsaid, K., & Al-Meer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1). <https://doi.org/10.1007/s40979-023-00140-5> Retrieved 10.02.2026. [T01]
- European Central Bank. (2026). *ECB euro reference exchange rate: Hungarian forint (HUF)*. [https://www.ecb.europa.eu/stats/policy\\_and\\_exchange\\_rates/euro\\_reference\\_exchange\\_rates/html/eurofxref-graph-huf.hu.html](https://www.ecb.europa.eu/stats/policy_and_exchange_rates/euro_reference_exchange_rates/html/eurofxref-graph-huf.hu.html) Retrieved 10.02.2026. [T03]
- Eurostat European Commission. (2025). *EU hourly labour costs ranged from €11 to €55 in 2024*. <https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20250328-1> Retrieved 10.02.2026. [T03]
- Figueira, J., Greco, S., & Ehrogott, M. (2005). *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer New York. <https://doi.org/10.1007/b100605> Retrieved 10.02.2026. [T09]
- Ismayilzada, A., Karimov, A., & Saarela, M. (2025). *Mining for knowledge, not trouble: GDPR's impact on educational data mining*. International Educational Data Mining Society. <https://doi.org/10.5281/zenodo.15870232> Retrieved 10.02.2026. [T03]
- Kadoic, N., & Oreski, D. (2018). Analysis of student behavior and success based on logs in Moodle. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 654–659. <https://doi.org/10.23919/mipro.2018.8400123> Retrieved 10.02.2026. [T09]

- Katoua, T., AL-Lozi, M., & Alrowwad, A. (2016). A review of literature on e-learning systems in higher education. *International Journal of Business Management and Economic Research*, 7, 754–762. Retrieved 15.02.2026. [T09]
- Kishore, S., Hong, Y., Nguyen, A., & Qutab, S. (2023). *Should ChatGPT be banned at schools? Organizing visions for generative artificial intelligence (AI) in education*. Retrieved 15.02.2026. [T01]
- Knuth, D. E. (1996). *Foreword* (M. Petkovšek, W. H. S., & D. Zeilberger, Eds.; pp. vii–viii). A K Peters. Retrieved 15.02.2026. [T09]
- Kodolányi János Egyetem. (2025). *Moodle LMS – Kodolányi János Egyetem*. <https://www.kodolanyi.hu/moodle-lms/> Retrieved 19.02.2026. [T08]
- Kodolányi János University. (2023, February). *Involvement of hungarian KJU students in the Erasmus+ blended intensive program*. Kodolányi János University. <https://backend.kodolanyi.hu/en/aktualis/hirek/1746> Retrieved 19.02.2026. [T04]
- Magyarország Kormánya. (2025). 426/2025. (XII. 23.) korm. rendelet a kötelező legkisebb munkabér (minimálbér) és a garantált bérminimum megállapításáról. <https://njt.hu/jogszabaly/2025-426-20-22>
- Majorosi, A., Müller, V., & Vincze, V. (2016). Gamifikáció cégeknél és az iskolában, mint az elégedettség fokozásának eszköze [Gamification at companies and in schools as a tool for increasing satisfaction]. In J. Schuchmann (Ed.), *A közösségfejlesztés új útjai* (pp. 48–56). Kodolányi János Főiskola. <https://backend.kodolanyi.hu/images/tartalom/File/kiadvanyok/kozossegfejlesztes.pdf> Retrieved 10.02.2026. [T14]
- Malatyinszki, S. (2022). Az oktatás és a tudatosság jelentősége a családok digitális eszközhasználatában. *Studia Mundi - Economica*, 9, 68–77. <https://doi.org/10.18531/studia.mundi.2022.09.01.68-77> Retrieved 19.02.2026. [T06]
- Moodle. (2024). *Analytics*. <https://docs.moodle.org/501/en/Analytics> Retrieved 19.02.2026. [T03]
- Moodle. (2025, February). *About Moodle*. [https://docs.moodle.org/501/en/About Moodle](https://docs.moodle.org/501/en/About_Moodle) Retrieved 10.02.2026. [T03]
- Motz, B., Quick, J., Schroeder, N., Zook, J., & Gunkel, M. (2019). *The validity and utility of activity logs as a measure of student engagement*. 300–309. <https://doi.org/10.1145/3303772.3303789> Retrieved 10.02.2026. [T09]
- Murtaza, M., Ahmed, Y., Shamsi, J. A., Sherwani, F., & Usman, M. (2022). AI-Based personalized e-learning systems: Issues, challenges, and solutions. *IEEE Access*, 10, 81323–81342. <https://doi.org/10.1109/ACCESS.2022.3193938> Retrieved 10.02.2026. [T01]

- Mutumukwe, C., Twizeyimana, J. D., & Viberg, O. (2021). Students' information privacy concerns in learning analytics: Towards a model development. *CoRR*, *abs/2109.00068*. <https://arxiv.org/abs/2109.00068> Retrieved 10.02.2026. [T01]
- National Forum for the Enhancement of Teaching and Learning in Higher Education. (2018). *Learner data and the general data protection regulation (GDPR)*. <https://doi.org/10.6084/m9.figshare.6097133.v2> Retrieved 19.02.2026. [T09]
- Ouadoud, M., Rida, N., & Chafiq, T. (2021). Overview of E-learning Platforms for Teaching and Learning. *International Journal of Recent Contributions from Engineering, Science & IT (IJES)*, *9*(1), 50. <https://doi.org/10.3991/ijes.v9i1.21111> Retrieved 10.02.2026. [T01]
- Page, E. B., & Stake, R. E. (1979). Should Educational Evaluation Be More Objective or More Subjective? *Educational Evaluation and Policy Analysis*, *1*(1), 45. <https://doi.org/10.2307/1164103> Retrieved 10.02.2026. [T09]
- Park, Y., & Jo, I.-H. (2016). Using log variables in a learning management system to evaluate learning activity using the lens of activity theory. *Assessment & Evaluation in Higher Education*, *42*(4), 531–547. <https://doi.org/10.1080/02602938.2016.1158236> Retrieved 01.03.2026. [T09]
- Pitlik, L. (2004). *Component-based object comparison for objectivity*. 197–200. Retrieved 15.02.2026. [T09]
- Pitlik, L. (2009). *Thought experiments I: Objectivity, sustainability, equilibrium*. MIAU e-Journal. [https://miau.my-x.hu/miau2009/index\\_en.php3?x=e08](https://miau.my-x.hu/miau2009/index_en.php3?x=e08) Retrieved 15.02.2026. [T12]
- Pitlik, L. (2014). *A hasonlóságelemzés alapjai (Introduction into the similarity analysis)*. MYX-team. [http://miau.my-x.hu/miau/189/hasonlosagelemzesi\\_alapok.doc](http://miau.my-x.hu/miau/189/hasonlosagelemzesi_alapok.doc) Retrieved 15.02.2026. [T14]
- Pitlik, L., Balogh, A., Schnellbach, M., & Szani, F. (2017). *Dashboards and OLAP services in LLL and distance learning processes or experiences about log-based learning and education management*. Apertus Nonprofit Ltd. [http://miau.my-x.hu/miau/kofop/hassacc\\_244.pdf](http://miau.my-x.hu/miau/kofop/hassacc_244.pdf) Retrieved 10.02.2026. [T10]
- Pitlik, L., & Varga, Z. (2015). *The operationalism of sustainability is a mathematical issue*. 122–129. [https://miau.my-x.hu/miau/206/Full\\_text\\_template\\_synergy2015\\_pl.pdf](https://miau.my-x.hu/miau/206/Full_text_template_synergy2015_pl.pdf) Retrieved 10.02.2026. [T10]
- Rahman, M., & Siddiqui, F. (2018). NLP-based automatic answer script evaluation. *DUET Journal*, *4*, 35–42. Retrieved 15.02.2026. [T09]
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

- Processing (EMNLP-IJCNLP)*, 3982–3992. <https://doi.org/10.18653/v1/d19-1410>  
Retrieved 15.02.2026. [T09]
- Schumacher, C., & Ifenthaler, D. (2018). Features students really expect from learning analytics. *Computers in Human Behavior*, 78, 397–407.  
<https://doi.org/10.1016/j.chb.2017.06.030> Retrieved 15.02.2026. [T09]
- Sclater, N. (2015, February). *Code of practice for learning analytics*.  
<https://www.jisc.ac.uk/Guides/Code-of-Practice-for-Learning-Analytics>.  
<https://www.jisc.ac.uk/guides/code-of-practice-for-learning-analytics> Retrieved 15.02.2026. [T11]
- Shanthi, S. A., & Karthikeyan, M. (2012). *A review on privacy preserving data mining*. 1–4.  
<https://doi.org/10.1109/ICCIC.2012.6510302> Retrieved 01.03.2026. [T09]
- Sukhbaatar, G., Tudevtagva, U., Erdenechimeg, S., & Sodnom, B. (2023). Students' evaluation of e-courses during curfew. *Lavai - International Journal of Education*, 19, 42–49. <https://doi.org/10.5564/lavai.v19i29.3210> Retrieved 15.02.2026. [T01]
- Teresa, C. P., & McGrath, C. (2021). Mapping the ethics of learning analytics in higher education: A systematic literature review of empirical research. *Journal of Learning Analytics*, 8, 123–139. <https://doi.org/10.18608/jla.2021.1> Retrieved 19.02.2026. [T01]
- Turtogtokh, S., Pitlik, L., & Pitlik, L. J. (2025). *Objective evaluation of performances in case of students based on similarity analyses and moodle-logs* (K. M. Emin & S. Özyurt, Eds.; pp. 442–458). Liberty Academic Publishers.  
[https://www.aegeanconference.com/\\_files/ugd/614b1f\\_5b68005af0584b6ea368d8da83eb72dd.pdf](https://www.aegeanconference.com/_files/ugd/614b1f_5b68005af0584b6ea368d8da83eb72dd.pdf) Retrieved 15.02.2026. [T01]
- Virágh, I. (2020). *Távoktatás a kommunikáció- és médiatudomány szakon is - Kodolányisok világa*. Kodolányi János University. <https://www.kodolanyi.hu/kv/cikk/tavoktatas-a-kommunikacio-es-mediatudomany-szakon-is-1314> Retrieved 15.02.2026. [T16]
- Yang, J. (2023). Die anwendung von learning analytics zur analyse des online-lernverhaltens am beispiel eines blendend learning-deutschkurses. *Zeitschrift Für Interkulturellen Fremdsprachenunterricht*, 28(1), 463–482. <https://doi.org/10.26083/tuprints-00024162>  
Retrieved 01.02.2026. [T05]
- Yavuz, H. C. (2019). The Effects of Log Data on Students' Performance. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 10(4), 378–390.  
<https://doi.org/10.21031/epod.564232> Retrieved 15.02.2026. [T13]

## 8.11 Reference-Coverage Table

Table 8.11- I summarises the distribution of the references used in the thesis across the T01–T16 classification scheme required by the institutional guidance

T-Code	Reference type definition	Count	bib key
T01	new, English, article, non-KJU	10	Ardito2025, Bahel2021, Elkhata2023, Kishore2023, Murtaza2022, Mutimukwe2021, Ouadoud2021, Sukhbaatar2023, Teresa2021, Turtogtokh2025
T02	new, English, article, KJU	1	Bn2024
T03	new, English, webpage, non-KJU	6	EuropeanCentralBank2026, EurostatEuropeanCommission2025, Ismayilzada2025, Moodle2024, Moodle2025, AmazonWebServices2026
T04	new, English, webpage, KJU	1	KodolnyiJnosUniversity2023
T05	new, not-English, article, non-KJU	1	Yang2023
T06	new, not-English, article, KJU	1	Malatyinszki2022
T07	new, not-English, webpage, non-KJU	2	Below2021, MagyarorszgzKormnya2025
T08	new, not-English, webpage, KJU	1	KodolnyiJnosEgyetem2025
T09	old, English, article, non-KJU	22	Agudo-Peregrina2014, Arnold2012, Behzadian2012, Brans1985, Buckley2006, Bnkuti2010, Cantabella2018, Clow2013, Dormezil2019, Figueira2005, Kadoic2018, Katoua2016, Knuth1996, Motz2019, Page1979, Park2016, Pitlik2004, Rahman2018, Reimers2019, Schumacher2018, Shanthi2012, NationalForumfortheEnhancementofTeachingandLearningi nHigherEducation2018
T10	old, English, article, KJU	3	Balogh2018, Pitlik2015, Pitlik2017
T11	old, English, webpage, non-KJU	1	Sclater2015
T12	old, English, webpage, KJU	1	Pitlik2009
T13	old, not-English, article, non-KJU	1	Yavuz2019
T14	old, not-English, article, KJU	2	Majorosi2016, Pitlik2014
T15	old, not-English, webpage, non-KJU	1	Bates2015
T16	old, not-English, webpage, KJU	1	Virgh2020

Table 8.11- I Reference coverage across the T01–T16 classification scheme used in the thesis (Source: Own compilation based on the institutional reference-classification audit (See Annex §8.6.8)).

## 8.12 Repository of parallel theses consulted for the comparative synthesis

The comparative synthesis in §2.12 was based on the class repository pages listed below. These links are documented here to make the source base of the comparison auditable. Where a repository did not contain a thesis with direct COCO relevance, it was considered only as contextual comparative material.

MIAU repository page: Amarsanaa Amgalanbatar. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/327/pswd/>

MIAU repository page: Ganbat Bayanmunkh. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/327/hdi/>

MIAU repository page: Ariunbold Munkhjargal. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/327/fl/>

MIAU repository page: Byekbolat Nurbol. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/327/bn/>

MIAU repository page: Boldsukh Ganzorig. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/328/gb/>

MIAU repository page: Tsetsugesuren Namjiljav. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/328/bor/>

MIAU repository page: Japheth Dangiwa Jerry. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/jj/>

MIAU repository page: Munkh-Orgil Batbayar. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/mom/>

MIAU repository page: Lévai Márk Zsigmond. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/lm/>

MIAU repository page: Nyambaatar Zandangarav. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/zn/>

MIAU repository page: Lehrer Zoltán. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/lz/>

MIAU repository page: Sukh-Ochir Dulguun. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/imdb2/>

MIAU repository page: Kosdi Gábor. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/329/kg/>

MIAU repository page: Kosárszki Tamás. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/kt/>

MIAU repository page: Papp Tamás. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/sr/>

MIAU repository page: Kálvári Áron. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/ka/>

MIAU repository page: Árvai Péter. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/ap/>

MIAU repository page: S. Z. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/sz/>

MIAU repository page: I. S. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/is/>

MIAU repository page: Papp Tamás. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/pt/>

MIAU repository page: Lackner Nóra Etelka. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/ln/>

MIAU repository page: Schewe Ervin. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/sche/>

MIAU repository page: Eppel Viktória. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/ev/>

MIAU repository page: Leitner Patrik. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/lp/>

MIAU repository page: Kenyó Kristóf István. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/kk/>

MIAU repository page: Vaszkó Attila. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/va/>

MIAU repository page: Hegedüs Blanka. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/hb/>

MIAU repository page: Váradi Viktor. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/vv/>

MIAU repository page: Tóbiás Bence Tibor. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/tb/>

MIAU repository page: Nagy László. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/nl/>

MIAU repository page: B. T. Retrieved 2026.03.29, from: <https://miau.my-x.hu/miau/330/bt/>