Amarsanaa Amgalanbatar

Below is an **opponent-style review and evaluation** of ***Amarsanaa.pdf*** using the structure and intent of the **Opponent's Evaluation Form** in ***opponensi_lap_en.pdf*** (criteria 1–5, total 0–200; Fail 0–119, Satisfactory 120–139, Average 140–159, Good 160–179, Excellent 180–200).

---

0) Use of AI & suspected plagiarism (integrity check)

What the thesis itself shows

- The thesis **explicitly documents LLM usage** and even includes **"Conversations with LLMs" in the annexes** (8.6), framing AI mainly as **benchmarking / assistance** rather than as the core method.
- The work's **core contribution is deterministic and rule-based** (OAM attributes → rank transform → COCO Y0 → normalized RiskIndex) and is presented as **reproducible engineering + analytical workflow**, which is consistent with independent author work rather than generic AI prose.

Plagiarism indicators (based on document evidence only)

- **No direct red flags** (e.g., abrupt voice changes, citation-free "textbook blocks", inconsistent terminology) stand out from the provided evidence; the thesis is highly self-referential in a structured way (promises mapped to sections, requirements traceability, validation protocol), which is atypical of copy-paste plagiarism.
- However, **absence of evidence ≠ proof of originality**: without a similarity report (Turnitin/iThenticate) I cannot *confirm* plagiarism status—only assess likelihood from internal signals.

Integrity verdict (reasoned)

- **AI use appears disclosed and bounded**, and the thesis demonstrates substantial **author-owned method/implementation content**; **plagiarism suspicion is low** based on internal document structure and transparency sections.

Recommendation (process): run an institutional similarity checker anyway (standard practice), and verify that any AI-assisted text is properly acknowledged per local rules; the thesis already provides transparency material that supports this.

---

1) Topic & Objectives — 38 / 40

Strengths

- The topic is **clear, relevant, and technically meaningful**: improving **transparency, comparability, and reproducibility** in password strength evaluation.
- Objectives are **explicit, measurable, and mapped to delivery locations** ("explicit promises", tasks, scope boundaries), which is unusually strong for a BSc thesis.
- Scope is **properly bounded**: the RiskIndex is stated as **relative structural risk**, **not time-to-crack**, not attacker success probability, and not a production security guarantee.

Minor weaknesses

- The "objective evaluation" claim is defensible only **within the chosen attribute model**; the thesis admits this, but the title can still be read as broader than the dataset-relative normalization allows.

Score rationale: near-excellent topic/objective alignment, only small overreach risk from wording ("objective") beyond the declared boundaries.

---

2) Review of the Literature — 35 / 40

Strengths

- Literature coverage is **broad and well-targeted**: password vulnerabilities, human behavior, password managers, strength meters/metrics, entropy/guessability, and policy guidance.
- The thesis explicitly articulates a **research gap**: lack of a **transparent standardized multi-attribute integration** with auditable scoring logic.
- Strong "bridge" from literature to method: the literature chapter is used to justify **why multi-attribute + deterministic aggregation** is appropriate.

Weaknesses / risks

- Some claims rely heavily on **internal methodological lineage** (COCO taught in-program) rather than peer-reviewed external validation; this is acceptable in context but slightly reduces the "literature-grounded" weight.
- The external benchmarking (zxcvbn) is a good step, but the literature could more explicitly compare against **other academic meters/guessability models** to strengthen positioning beyond one benchmark reference.

Score rationale: clearly above average and close to excellent; slight deduction for limited breadth of *comparative evaluation literature* beyond the chosen references/benchmark.

---

3) Presentation of the Author's Own Work — 52 / 60

Major strengths (method + evidence + artifact)

- **Clear R&D methodology**: controlled dataset (n=100; Weak/Medium/Strong) + deterministic preprocessing and feature extraction + OAM definition with explicit directionality + COCO Y0 aggregation + RiskIndex normalization.
- **Transparency and reproducibility** are operationalized in a **Streamlit + SQLite** prototype (feature extraction, scoring, storage, export) with testing emphasis (determinism, auditability, export integrity, stability).
- **Internal validation protocol** (Delta/Tény sign-consistency via inversion) is described and reported as passing across all objects, which supports computational coherence.
- The thesis includes **benchmark comparison with zxcvbn** showing agreement at extremes and identifying differences in "medium" cases—this is a meaningful sanity check.

Key weaknesses (important, but mostly acknowledged)

1. **Floor effect / loss of discrimination for "Strong" passwords**
   - All Strong-group passwords end up at **RiskIndex = 0**, giving **no within-tier ranking**; the thesis correctly attributes this to min–max normalization and boundary anchoring.
   - This reduces analytical value for the "best" region and suggests the model's feature set + scaling is insufficiently sensitive at the top end.
2. **Constructed labels make "100% alignment" partially circular**
   - The Safe/Risky reference label is defined from the same grouping logic (Safe ↔️ Strong; Risky ↔️ Weak+Medium). Achieving perfect separation is therefore **not independent validation**, and the thesis states it should be interpreted as internal consistency only.
3. **Dataset and external validity limits**
   - A 100-item synthetic dataset is fine for demonstration, but it limits claims about real-world distributions; the thesis explicitly states it is not breach-ground-truth.

Score rationale: strong own contribution (method + implementation + testing). Deductions mainly for: top-tier compression, circularity risk in label alignment, and limited external validity (all of which the thesis largely acknowledges).

---

4) Structure, Style, and Formal Aspects — 37 / 40

Strengths

- **Excellent structure and traceability**: research questions, explicit promises, tasks, scope, and "where delivered" mapping supports defense readiness.
- Clear separation between **method** (Sections 3.1–3.7) and **reference implementation** (3.8+), which improves readability and academic rigor.
- Formal elements are strong: tables/figures, reproducibility notes, artifact inventory, and AI transparency annexing contribute to professional presentation.

Minor weaknesses

- Some passages are **overly verbose / meta-structural** (e.g., repeated reminders of boundaries). This is defensible but slightly impacts concision.

Score rationale: near-excellent formal quality; small deduction for verbosity.

---

5) Overall Impression — 18 / 20

Overall strengths

- The thesis is **methodologically disciplined**: transparent, deterministic, and explicit about what it does *not* claim.
- The implemented artifact + testing + reproducibility posture exceeds the typical baseline for an undergraduate thesis.

Overall weaknesses

- The strongest conceptual limitation is **scale dependence** (dataset-relative RiskIndex) and **limited discrimination among top-tier passwords**; although acknowledged, it meaningfully affects practical interpretability.

Score rationale: strong work with one meaningful methodological limitation that is openly stated and can be improved in future work.

Final scoring (0–200) and grade band

	Criterion	Max	Score
1	Topic & Objectives	40	**38**
2	Literature Review	40	**35**
3	Author's Own Work	60	**52**
4	Structure/Style/Formal	40	**37**
5	Overall Impression	20	**18**
6	**TOTAL**	**200**	**180**

Total: 180 / 200 → "Excellent" (180–200) according to the evaluation bands shown on the form.

Concrete improvement suggestions (high-impact, feasible)

1. **Fix the Strong-tier floor effect**
   - Use a **fixed external normalization range** (not dataset min/max), or a robust scaling (e.g., percentile-based) so "Strong" passwords don't collapse to 0.
2. **Add higher-sensitivity attributes**
   - Introduce n-gram / Markov-style predictability proxies, keyboard-walk detection, or transition-complexity metrics so "good" passwords can still be meaningfully compared.
3. **Strengthen external validity (without changing scope)**
   - Expand evaluation with at least one additional dataset (still synthetic if ethics require), and compare against **multiple** estimators (not only zxcvbn) to triangulate behavior.
4. **AI transparency: keep, but tighten**
   - The annex is a strength; consider adding a short **front-matter AI disclosure statement** summarizing what AI did/did not do, to prevent misinterpretation.

Quick "Opponent summary" (as you'd write on the form)

A well-structured and technically defendable thesis that proposes a transparent, deterministic multi-attribute password evaluation method (OAM + COCO Y0) and implements it as a reproducible Streamlit/SQLite artifact. The work is strong in traceability, testing, and bounded claims. Main limitations are dataset-relative scaling and top-tier score compression, plus the inherently constructed nature of the internal reference labels—both correctly acknowledged. AI usage is documented and appears integrity-compliant; plagiarism suspicion is low based on internal evidence.

If you want, I can also produce a **filled-in Opponent's Evaluation Form narrative** (criterion-by-criterion paragraphs in the exact "opponensi" tone) you can paste into the university template.