Grade Appeal - HW3: Team *Promptopia*

Team: Guy Raz (302632740), Gil Chen (316019975) | **Repository:** GitHub Link

Self-Evaluation Verdict: 95/100 (Level 4 — Excellent)

We stand by a self-grade of 95. We believe the AI evaluator's score of 82 unfairly penalizes the project for missing standard artifacts irrelevant to this specific LLM orchestration architecture. We prioritized experimental validity over redundant unit testing.

Evaluator	Grade	Verdict
AI Assessment	82/100	Level 3 — Very Good
Self-Evaluation	**95/100**	**Level 4 — Excellent**

AI Evaluation Summary

Acknowledged Strengths:

- **Documentation (19/20):** Comprehensive PRD with clear objectives.
- **Research (14/15):** Systematic experiments across 3 error levels.
- **Code Quality (14/15):** Modular design with single-responsibility agents.
- **Orchestration:** Effective chaining via `translation-round-trip`.

Disputed Gaps:

- Missing automated testing (unit/integration).
- Missing architecture artifacts (diagrams/ADRs).
- Limited README & No CI/CD.
- Missing Jupyter notebooks.

Defense & Justification

We dispute the specific deductions regarding testing and artifacts as architectural misunderstandings.

1. Automated Testing Strategy

- **Rebuttal:** Standard unit tests are unnecessary as the system relies on agent interactions, not complex internal algorithms. The complexity lies in LLM calls, not code logic.
- **Validation:** The experiment itself, conducted across three error levels, serves as the rigorous system validation.

2. Reproducibility & Artifacts

- **Rebuttal:** Reproducibility is achieved via a robust CLI rather than static notebooks.
- **Execution:** Running `python experiment.py` fully replicates the study. We chose a clean, executable pipeline over redundant artifacts.