Assignment 6

Team name: Lior_Abuhav_Ofri_Kutchinsky

Lior Abuhav ID: 209521400

Ofri Kutchinsky ID: 313611360

Repo link:

https://github.com/ofrik2/prompt-engineering-project

Self-Assessment

Students: Ofri Kutchinsky , Lior Abuhav
Project: Round-Trip Translation Turing Machine with LLM Agents
Self-assigned grade: 94/100

Self-Assessment Statement

What we believe we did well:

We believe the main strength of this project lies in the combination of technical rigor, experimental depth, and reflective use of LLMs as agents, in line with the course objectives.

From a software perspective, we designed the project as a reproducible Python package with a clear modular structure, deterministic execution via a dummy provider, and a command-line interface that allows the entire pipeline to be executed with a single command. Installation, configuration, and usage are documented thoroughly in the README, enabling graders to reproduce results without external dependencies.

From a research perspective, we went beyond a basic comparison of prompt types. Instead of focusing solely on aggregate accuracy, we explored prompt length sensitivity, method disagreement, and failure modes, particularly the phenomenon we term overthinking in Chain-of-Thought prompting. This led to insights that challenge common assumptions about reasoning-based prompts, especially under strict output constraints.

Most importantly, we made extensive and deliberate use of an LLM as an agent throughout the development process. The agent was used to analyze assignment requirements, propose architectural designs, debug evaluation logic, interpret unexpected experimental outcomes, and assist in producing structured academic documentation. These interactions are documented in a curated prompt engineering log, demonstrating iterative, purposeful agent usage rather than ad-hoc querying.

Challenges and what we learned:

A central challenge was handling unexpected experimental behavior, particularly the observation that Chain-of-Thought prompting underperformed the baseline in exact-match accuracy. Rather than treating this as an implementation bug, we learned to analyze it as a methodological issue related to instruction-following versus reasoning. This shift in perspective significantly shaped both our analysis and our conclusions.

Another challenge involved balancing engineering completeness with academic scope. While it was tempting to pursue extensive testing or production-level security features, we

learned to prioritize elements that aligned with the learning objectives of the course: reproducibility, clarity, and thoughtful analysis over exhaustive software hardening.

Finally, documenting agent usage in a meaningful way required reflection. Instead of logging raw conversations, we learned to curate prompts that captured decision-making moments, emphasizing how the agent influenced design and understanding rather than merely recording interactions.

What could be improved:

Despite the overall strength of the project, there are areas where improvement is possible.

Configuration and security practices are adequate but not production-grade. Sensitive information is handled via environment variables and excluded from the repository, but advanced security concerns such as key rotation, permission scoping, or configuration validation are not deeply explored.

Additionally, the experimental scope focuses on constrained classification tasks, which limits the generalizability of conclusions to open-ended generation tasks. Future work could extend the framework to additional domains and models.

These limitations are acknowledged explicitly and are reflected in the self-assigned scores.

Final Justification

Overall, we believe this project demonstrates strong understanding, careful execution, and thoughtful reflection, both technically and academically. The self-assigned grade reflects not only the project's strengths, but also an honest acknowledgment of its limitations. We therefore consider a grade in the low-to-mid 90s to be fair and well-justified.

Category	**Weight**	**My Score**	**Weighted Score**
Project Documentation (PRD & Architecture)	20%	95	19
README & Code Documentation	15%	95	14.25
Project Structure & Code Organization	15%	94	14.1
Configuration & Security	15%	86	12.9
Testing & QA	15%	75	11.25
Research, Analysis & Results	15%	90	13.5
Use of Course Concepts & Terminology	10%	93	9.3
Total	**100%**		**94.3**

3.6 הצהרת יושר אקדמי (Academic Integrity Declaration)

אני מצהיר/ה בזאת ש:

- ☐ ההערכה העצמית שלי היא כנה ואמיתית
- ☐ בדקתי את העבודה מול כל הקריטריונים לפני קביעת הציון
- ☐ אני מודע/ת שציון עצמי גבוה יוביל לבדיקה דקדקנית יותר
- ☐ אני מקבל/ת את העובדה שהציון הסופי עשוי להיות שונה מהציון העצמי
- ☐ העבודה היא פרי עבודתי/נו (של הקבוצה) ואני/ו אחראי/ם לכל תוכנה

חתימה: _____ **תאריך:** 17.12.25

Grade and Comments: