Abstract

Timely, individualized, and technology appropriate feedback is very important for effective learning in university-level computer science education, yet it presents a significant logistical challenge in large foundational courses like Computer Architecture. This thesis investigates the feasibility of leveraging modern Large Language Models (LLMs) to automate the feedback process for technical assignments.

This thesis comparatively evaluates four state-of-the-art models—ChatGPT-4o, Gemini 2.5 Pro, Claude Sonnet 4, and DeepSeek-V3—on a dataset of 380 real, anonymized student submissions from a Computer Architecture course. The assignments include low-level tasks such as RISC-V assembly programming, microprogramming, and pipeline analysis. To assess the impact of instructional design, each submission was processed using four distinct prompt configurations, ranging from a basic rubric to a hybrid prompt containing both a reference solution and a feedback exemplar. The quality of the 5,960 generated feedback responses was evaluated across six criteria: Correctness, Clarity, Depth, Consistency, Usefulness, and Strictness.

The empirical results indicate that current LLMs are capable of producing technically accurate, coherent, and pedagogically valuable feedback, especially when guided by well-designed prompts. Among the tested models, Gemini 2.5 Pro achieved the highest overall performance, while ChatGPT-4o exhibited the most balanced and consistent results. Claude Sonnet 4 demonstrated superior clarity and tone, and DeepSeek-V3 provided a cost-efficient open-source alternative with moderate accuracy. A key insight of this study is that prompt engineering exerts an influence on feedback quality comparable to, and in some cases exceeding, model selection. The hybrid prompt combining exemplars and reference solutions consistently yielded the highest rubric-aligned performance across all systems.

Despite these promising findings, the study is limited by its single-rater evaluation design, limited prompt variety, and dependency on specific model versions available in 2025. Nevertheless, the results provide empirical evidence that LLM-based sys-