JDS

Beyond the Veneer of Competence: How Structured Prompting Fosters Knowledge Transfer in LLM-Augmented Education

I-TING LIN, Department of Data Science, Soochow University, Taiwan
CHIA-HAO CHIU*, Department of Data Science, Soochow University, Taiwan
KO-I LEE, Department of Data Science, Soochow University, Taiwan

Abstract

The integration of Large Language Models into technical education exposes a critical fault line between building genuine conceptual mastery and merely outsourcing cognition. We examined this tension by tracking 123 computer science students across three AI-supported assessments and one unassisted final exam, comparing a standard control group against a cohort trained in structured Prompt Engineering (PE).

Interaction logs and semantic similarity metrics revealed two drastically different approaches to the tool. Untrained students treated the LLM as a shortcut: 76.6% of their submissions heavily mimicked the model's raw output. While this passive extraction maximized short-term grades, it created a severe cognitive dependency. For this group, higher AI usage actively predicted worse performance on the independent final exam.

The PE-trained students, meanwhile, paid an early "transition cost." Their initial scores dropped as they took on the higher cognitive load of breaking down tasks and reformulating AI responses. Yet, this early friction paid off. When AI support was removed, the trained cohort retained significantly more knowledge, achieving a much higher Transfer Index (59.0% vs. 48.7%).

Our data challenges the assumption that educational interfaces should be seamless. To prevent cognitive offloading, educators must treat prompt engineering not as a technical trick, but as a metacognitive skill that injects the "productive friction" necessary for resilient, independent learning.

Review Article

Introduction

Large Language Models (LLMs) have rapidly embedded themselves into higher education as indispensable cognitive scaffolds. Yet, a persistent "usage gap" remains: even as model capabilities advance, many students lack the prompt engineering literacy required to navigate these systems effectively[]. Most interactions are still hampered by imprecise problem formulations and a fundamental failure to decompose tasks, effectively leaving the model's latent reasoning potential untapped.

This proficiency gap raises a critical pedagogical concern: does frequent AI reliance facilitate genuine knowledge internalization, or does it merely produce a "veneer of competence" that vanishes in independent settings? While existing literature focuses heavily on the technical optimization of prompts, we still know very little about how sustained LLM interaction reshapes actual learning behavior.

To untangle these dynamics, we developed a custom AI-mediated testing platform designed to capture every nuance of student–LLM dialogue and transform it into structured behavioral data . Rather than looking solely at task performance, our framework employs semantic evaluation to decouple surface-level completion from deep conceptual mastery. By tracing interaction trajectories across two student cohorts in controlled, high-overlap environments, we expose the direct relationship between specific prompting behaviors and independent learning outcomes. These findings offer a data-driven blueprint for fostering more resilient human–AI collaboration in technical education.

Keywords: Prompt Engineering, Large Language Models (LLMs), Cognitive Dependency, Knowledge Transfer, Learning Analytics, Productive Friction, Technical Education

IT conceived the study, designed the analytical framework, and led the manuscript writing; JH provided the original datasets and oversaw the data curation and visualization; IT and JH co-performed the formal statistical analysis; KI conducted the literature investigation and background research. All authors participated in the revision and approved the final version.

Authors' Contact Information: I-Ting Lin, 11173238@gm.scu.edu.tw, Department of Data Science, Soochow University, Taipei, Taiwan; Chia-Hao Chiu*, starklab2020@scu.edu.tw, Department of Data Science, Soochow University, Taipei, Taiwan; KO-I LEE, 10122144@gm.scu.edu.tw, Department of Data Science, Soochow University, Taipei, Taiwan.

Building on this framework, we examine whether a structured Prompt Engineering (PE) intervention—focusing on strategies such as Chain-of-Thought (CoT), few-shot prompting, and role-based prompting—can influence how students interact with LLMs.

To evaluate this, we test the following hypotheses:

* H1: Prompting Behavior and Interaction Quality Students trained in structured prompting (Group B) are expected to show higher prompting efficiency (η) and more balanced semantic similarity (S_{sem}) compared to the control group. These students are more likely to apply strategies such as CoT to break down complex tasks, leading to more focused and effective interactions.

* H2: Knowledge Transfer and Independent Performance Students in the experimental group (Group B) are expected to achieve a higher Transfer Index (TI) and better performance in the non-AI final exam (P_{final}). This would suggest that structured prompting supports deeper understanding and contributes to improved performance after AI assistance is removed.

The primary objective of this study is to identify the threshold where AI-augmented assistance transitions from a supportive scaffold into a source of cognitive dependency. Through a computational lens—leveraging prompting efficiency (η) and semantic alignment (S_{sem})—we analyze interaction trajectories to understand how specific engagement behaviors correlate with genuine conceptual mastery. Ultimately, this work provides a data-driven roadmap for incorporating LLMs into technical curricula while safeguarding students' capacity for independent reasoning.

Related Work

Prompt Engineering and Interaction Dynamics in LLM Usage

Traditional perspectives often treat prompt engineering as a structured process for extracting outputs from Large Language Models (LLMs). In practice, however, user interaction is rarely this straightforward. Instead, it tends to unfold as an iterative process, where each model response shapes the user's next move. Prompting, in this sense, is less like issuing a one-time instruction and more like an ongoing dialogue between the user and the model .

Early research has largely focused on static settings such as zero-shot or few-shot prompting . While useful, these settings do not fully capture how people use LLMs in more complex tasks. In practice, users engage in multi-turn interactions, gradually refining their inputs based on the model's responses. This back-and-forth process allows them to address hallucinations and clarify unclear outputs. As a result, the sequence of interactions often reveals more about user strategy than any single prompt.

Despite growing interest in prompt engineering, the temporal aspect of these interactions remains relatively underexplored. We have a good understanding of prompt design, but less insight into how users adjust their strategies when things do not go as expected. Looking at interaction trajectories can help uncover whether users are effectively leveraging the model, or simply repeating trial-and-error patterns .

This becomes particularly important in time-constrained settings. Under pressure, users tend to shift from exploratory behavior to more goal-directed interaction. Interestingly, a higher number of interactions does not always indicate better engagement—it may instead reflect a mismatch between the user's needs and the model's responses . On the other hand, more efficient interaction patterns may suggest a better alignment between human and AI. By examining these dynamics, this study seeks to better understand how interaction behavior relates to the quality of human-AI collaboration.

Behavioral Modeling and Learning Analytics in AI-Assisted Environments

Learning analytics(LA) has long utilized behavioral data to model student cognition and predict learning outcomes . Prior research has demonstrated that features such as response time, task sequences, and clickstream interactions can serve as proxies for underlying cognitive processes, including engagement, problem-solving strategies, and knowledge acquisition. These methods are now foundational in digital learning environments to support performance prediction and adaptive learning systems.

Despite their effectiveness, existing behavioral modeling techniques are predominantly based on coarse-grained interaction signals. While such features capture when and how frequently users interact with a system, they provide limited insight into how users reason during the interaction process. As a result, these methods often fail to represent the semantic and strategic dimensions of complex problem-solving tasks.

The emergence of AI-assisted learning environments introduces new challenges for behavioral analysis. Unlike traditional systems with predefined interaction structures, LLMs enable open-ended, natural language interactions, transforming user behavior into unstructured dialogue data. This shift significantly increases the richness of behavioral signals, but also complicates their interpretation. Nevertheless, much of the existing work continues to rely on surface-level metrics, without fully leveraging the semantic information embedded in user-AI interactions.

Recent studies have begun to explore more fine-grained approaches, such as sequence modeling and interaction trajectory analysis, to better capture temporal patterns in learning behavior . However, these methods still largely overlook the semantic structure of user-generated inputs, particularly in prompt-based interactions.

This limitation highlights the need for integrating semantic-level representations into behavioral modeling frameworks. Treating natural language interactions as structured signals offers new opportunities for understanding how users engage with AI systems during complex tasks.

Evaluating Human Performance With and Without LLM Assistance

As LLMs become part of everyday problem-solving, the key question is no longer just whether they help, but how they change the way people approach tasks. Existing research shows that LLMs can improve performance in areas like programming and reasoning, often by providing real-time suggestions or partial solutions that help users move forward.

At the same time, these performance gains do not always mean real progress. When working with LLMs, users may shift from actively solving problems to reviewing and selecting from model-generated outputs . In some cases, this makes the process more efficient; in others, it may point to a growing reliance on the model.

Reliability is another concern. Since LLM outputs are not always consistent, results usually depend on how users verify, revise, and guide the model over multiple turns. This makes the process less about getting a single answer and more about managing an ongoing interaction between user and system.

While prior work has shown clear improvements in AI-assisted settings, we still know relatively little about what happens when that support is no longer available. It remains unclear whether these gains reflect actual skill development or temporary dependence on the model. In particular, there is still limited work comparing performance with and without AI support.

Looking at both interaction behavior and independent performance may help clarify this distinction, and provide a clearer view of how augmentation, dependency, and reliability play out in LLM-based collaboration.

In summary, while LLMs offer undeniable performance gains, the nature of these gains—whether they represent cognitive offloading or genuine skill augmentation—remains ambiguous. To decouple these effects, it is necessary to move beyond aggregate scores and examine the granular, longitudinal patterns of user-AI interaction. The following sections describe our empirical approach to measuring these dynamics through controlled examination environments, specifically focusing on how repetitive task structures and varying levels of AI assistance influence problem-solving trajectories.

Methodology

This study adopts a quasi-experimental design to examine how students' interaction strategies evolve over time and how these changes relate to knowledge internalization following a Prompt Engineering (PE) intervention. Rather than focusing solely on AI usage frequency, we analyze how students interact with LLMs across multiple stages of a course and how these interaction patterns change when AI support is removed.

Participants and Research Context

The study involved 123 first-year undergraduate students enrolled in a foundational Introduction to Computer Science course (group A: ; group B:).

To preserve ecological validity, we maintained the students' original course sections, thereby minimizing potential "experimental awareness" bias.

Group A (Control Group): Followed the standard curriculum. Students were permitted to use Generative AI during assessments but received no formal training on how to optimize interactions with these models.

Group B (PE-Trained Group): At the beginning of the semester, this group participated in a 90-minute Prompt Engineering (PE) intervention. This session extended beyond basic usage and introduced advanced cognitive strategies, including:

Chain-of-Thought (CoT): Guiding the AI to decompose complex logic into manageable steps. Few-shot Prompting: Providing context-rich examples to better align model outputs with technical requirements. Role-based Prompting: Assigning the AI a professional persona to enhance precision and contextual relevance. To investigate the longitudinal impact of AI-assisted learning, we implemented a four-stage assessment sequence following a faded support model. This design gradually transitions students from AI-assisted problem-solving to independent mastery.

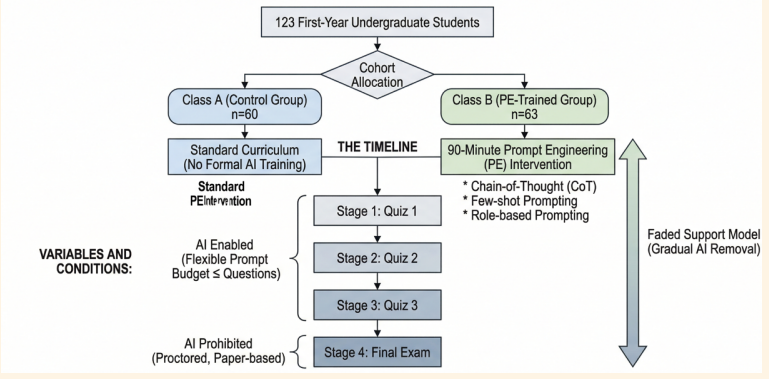

Fig. 1. User Interaction Flow in Comparative Evaluation.

A central component of our methodology is the flexible prompt allocation rule applied during Stages 1–3. Instead of enforcing a rigid one-prompt-per-question constraint,

students were provided with a prompt budget equivalent to the total number of questions, enabling naturalistic interaction behaviors such as iterative refinement of complex problems or balanced distribution across tasks. This design yields a rich dataset for analyzing strategic engagement patterns, including prompting stamina and decision-making processes. The assessment sequence culminates in a paper-based, proctored final exam (Stage 4) without AI assistance, serving as the primary benchmark for evaluating whether prolonged AI use—particularly when supported by Prompt Engineering training—facilitates genuine conceptual internalization or instead leads to cognitive dependency.

Interaction Logging Platform

To capture detailed interaction data, we developed a custom platform that records all student–LLM interactions during the assessments. This avoids variability introduced by external interfaces and ensures consistent data collection.

The platform logs the following information:

Prompt sequences: All user inputs are recorded, including full text and length. This allows us to analyze how prompts evolve over time within a task.

Model responses: All generated outputs are stored, enabling analysis of how different prompting strategies affect response quality.

Answer comparison: Student answers are compared with AI-generated responses using semantic similarity measures. This helps us distinguish between direct copying and more independent responses that incorporate model-generated reasoning.

These logs form the basis for analyzing interaction patterns and their relationship to learning outcomes.

Evaluation Metrics and Computational Framework

A three-dimensional framework is used to characterize student behavior and learning outcomes in AI-assisted settings. The goal is to differentiate between reliance on model outputs and evidence of independent understanding.

(1) Synthesis vs. Mimicry ( )

To examine the extent to which students incorporate AI-generated content, we measure the semantic similarity between model responses ( ) and students' final answers ( ). Sentence embeddings are computed using SBERT, and cosine similarity is applied:

$$S_{sem} = \frac{\mathbf{V_A} \cdot \mathbf{V_S}}{\|\mathbf{V_A}\|\|\mathbf{V_S}\|}$$

Higher similarity values indicate closer alignment with model outputs, which may suggest stronger reliance. Lower values indicate greater divergence, reflecting potential transformation or reinterpretation of the content.

Similarity ranges are used as heuristic references rather than strict thresholds. For example, very high similarity (e.g., $S_{sem} > 0.9$) may indicate direct reuse of model responses, while intermediate ranges (e.g., 0.5–0.7) are interpreted as partial transformation. These interpretations are used cautiously and in combination with other indicators.

(2) Resource Allocation and Prompting Efficiency (η)

To analyze how students manage limited interaction opportunities, we examine prompt usage under a fixed task-to-prompt ratio. Prompting efficiency (η) is defined as the average score gain per prompt:

$$\eta = \frac{\sum_{i=1}^{n} \text{Score}_i}{\sum_{i=1}^{n} \text{Prompt}_i}$$

This metric captures how effectively students use their available prompts during the assessment. Higher values suggest more efficient use of interactions, while lower values may reflect less targeted or exploratory usage.

In addition to overall efficiency, prompt distribution patterns across tasks are analyzed to understand how students allocate resources under time constraints.

(3) Dependency and Knowledge Transfer (TI)

To evaluate whether AI-assisted performance translates into independent ability, we compare performance during AI-supported stages (P_{AI}) with the final, non-AI exam (P_{final}). The Transfer Index (TI) is defined as:

$$TI = \left(\frac{P_{final}}{P_{AI(avg)}} \right) \times 100\%$$

A value close to 100% suggests that performance is maintained without AI support, while lower values indicate a potential gap between assisted and independent performance.

Correlation analysis is conducted between interaction-related measures (e.g., S_{sem}, η) and final exam scores to examine the relationship between interaction behavior and knowledge transfer.

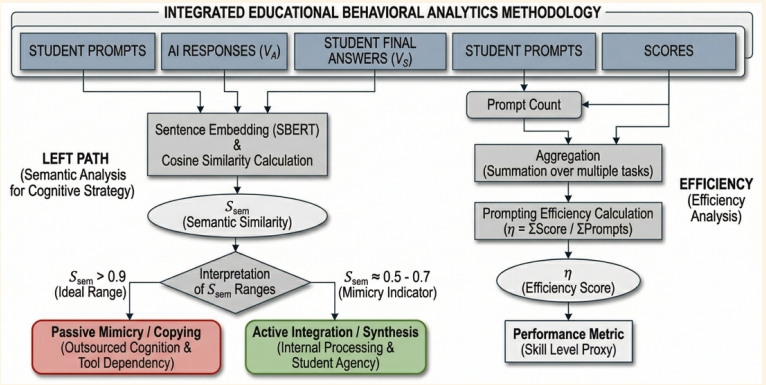

Fig. 2. Density Distribution of Semantic Similarity (S_{sem}) Between Student Answers and LLM Outputs.

Together, these metrics provide a structured way to analyze how students interact with LLMs and how these interactions relate to subsequent independent performance.

Results and Analysis

This section presents a quantitative analysis of the impact of the Prompt Engineering (PE) intervention on students' AI interaction strategies and subsequent knowledge internalization. We examine performance trajectories, longitudinal shifts in semantic dependency, the validity of the semantic similarity metric in open-ended tasks, and the final evaluation of knowledge transfer.

Performance Trajectories and Prompting Efficiency

Longitudinal analysis of performance revealed a significant between-group difference during the second AI-assisted assessment (Quiz 2). The Control Group (group A) achieved a higher average score (49.14) compared to the PE-Trained Group (group B, 30.80) ($p < .001$, $d = 1.53$).To examine whether this difference was attributable to variations in AI usage,

we analyzed system interaction logs. The results indicate that interaction frequencies were comparable across groups. For example, during Quiz 2, the average number of AI interactions in group A (12.41) and group B (11.22) did not differ significantly (p = .202).The performance gap was further reflected in the Prompting Efficiency metric ( ) (see Figure 1). In Quiz 2, group A demonstrated higher efficiency (4.32) than group B (3.08) (p = .005, d = 0.56).

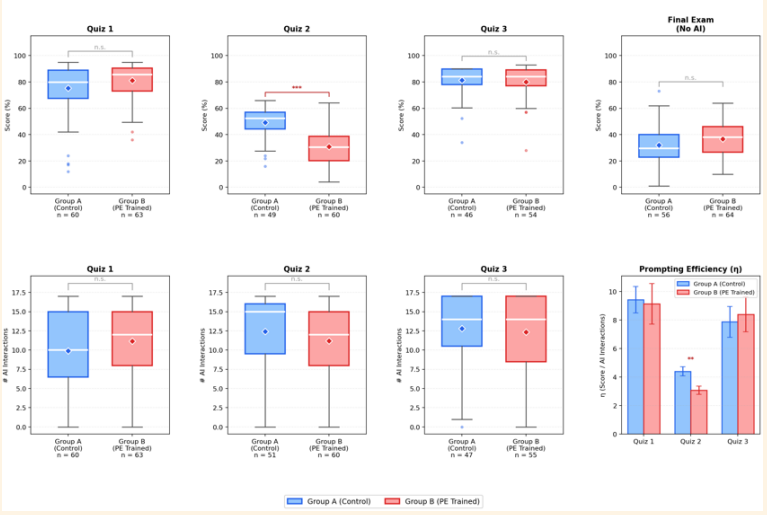

Fig. 3. Score and AI Interaction Comparison Across Stages.

This pattern is consistent with a potential transition cost associated with the adoption of more cognitively demanding prompting strategies. Specifically, the PE intervention may have encouraged a shift from passive extraction to more active forms of engagement, which could temporarily reduce performance efficiency before such strategies are fully stabilized.

Semantic Dependency and Interaction Trajectories

To further examine interaction patterns, we computed the semantic similarity ( ) between students' submitted answers and AI-generated responses.A significant between-group difference was observed in Quiz 2 (p < .001, d = 0.85). In group A, 76.6% of responses fell into a high-similarity range ( ), with no instances of low similarity ( ). In contrast, group B showed a more distributed pattern, with 37.3% in the high-similarity range and 27.5% in the low-similarity range.

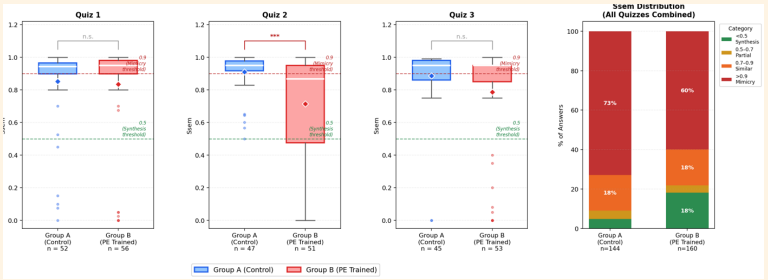

Fig. 4. Semantic Simallarity (Ssem):Synthesls vs. Mimicry Group A (Control) vs Group B (PE Trained).

Within-group analysis further revealed distinct temporal patterns. group B exhibited a significant variation in across the three assessments (Friedman = 9.15, p = .010), with a decrease in Quiz 2 (Wilcoxon p = .003) followed by a partial increase in Quiz 3 (p = .020). In contrast, group A maintained consistently high similarity levels across all stages (Friedman ns).These results suggest that the PE intervention was associated with more dynamic interaction patterns, whereas the control group maintained relatively stable response similarity throughout the assessments.

Validity of Semantic Similarity in Open-Ended Tasks

To assess the interpretability of the semantic similarity metric, we examined its relationship with performance on Short Answer (SA) questions, where generative responses are required.For group A, no significant correlation was observed between S_{sem} and SA scores across all assessments (p > .05). In contrast, for group B, S_{sem} showed a significant negative correlation with SA scores in Quiz 2 (r = -0.492, p < .001) and Quiz 3 (r = -0.564, p < .001).

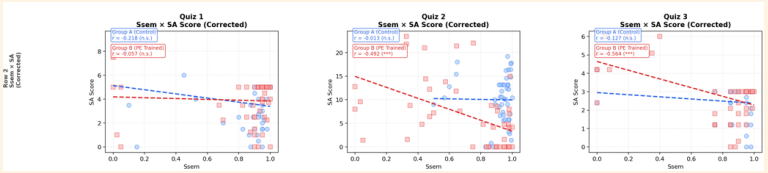

Fig. 5. Divergent Cognitive Pathways: Synthesizing vs. Mimicking.

These findings indicate that, within the PE-trained group, lower semantic similarity—reflecting greater divergence from model-generated responses—was associated with higher performance in open-ended tasks.

Knowledge Transfer and Dependency Patterns

To evaluate knowledge transfer, we analyzed performance in the final exam (Stage 4), where AI assistance was not permitted.

The Transfer Index (TI) was significantly higher in group B (59.04%) compared to group A (48.67%) (t = -2.31, p = .023, d = 0.42). In addition, group B demonstrated a higher mean raw score (36.72) than group A (32.07), although this difference did not reach statistical significance (p = .064).

Further distributional analysis revealed that 60.0% of students in group A retained less than half of their AI-assisted performance (TI < 50%), compared to 34.9% in group B.

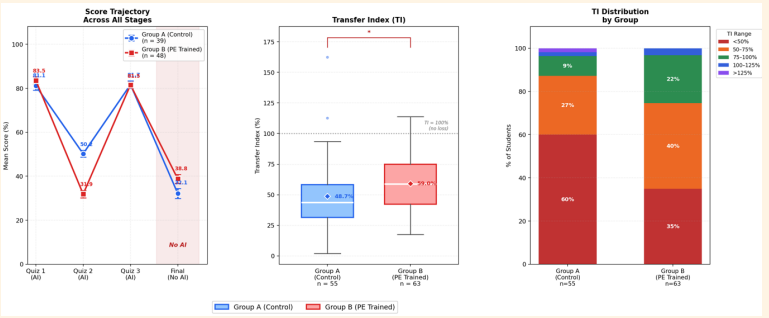

Fig. 6. Distribution of Transfer Index (TI) Across Groups.

We also examined the relationship between AI interaction frequency and knowledge transfer. In group A, total AI interactions were negatively correlated with TI (r = -0.448, p = .004), whereas no significant correlation was observed in group B (r = -0.182, ns).

These results suggest that the relationship between AI usage and learning outcomes differs across groups, with the PE-trained group exhibiting more stable performance under conditions without AI support.

Conclusion

This study examined how structured Prompt Engineering (PE) shapes students' interaction trajectories with Large Language Models and their subsequent capacity for knowledge transfer.

Our findings reveal that while PE training introduces a temporary "transition cost" during early AI-assisted tasks, it ultimately supports deeper cognitive engagement and more resilient knowledge internalization. Instead of defaulting to passive extraction, students trained in structured prompting demonstrated a significant reduction in semantic dependency on AI-generated responses, leading to superior transfer performance in non-AI settings.

These results underscore the necessity of moving beyond surface-level efficiency metrics to examine the granular processes underlying human–AI collaboration. Rather than treating AI as a frictionless shortcut, educational systems must reframe prompt engineering as a foundational metacognitive skill—one that introduces productive pedagogical friction to promote active learning.

Future work should explore how these strategic interaction patterns generalize across diverse academic domains, and how adaptive scaffolding can further optimize the delicate balance between AI assistance and independent reasoning.