# AI-Enhanced Learning in Calculus Education*

**Zheng Yang**
Sichuan University

**Tuto Lopez Gonzalez**
Technology Educator Alliance

**Todd Edwards**
Miami University

**ChatGPT (OpenAI)**
OpenAI

## Abstract

This study examines the role of generative AI (genAI) as a learning partner in second-semester calculus, focusing on Taylor series. Students engaged with a researcher-designed genAI module featuring adaptive tutoring, real-time visualizations, and historical narratives. We analyzed 127 student–genAI transcripts using the Pirie–Kieren framework, along with survey and graduate student interview data. Across transcripts, genAI contributed about 70% of words, yet students showed substantial mathematical growth, often reaching advanced Pirie–Kieren levels and engaging in recursive "fold back" cycles. Survey and interview findings indicated increased confidence, positive attitudes toward AI support, and appreciation for personalized feedback, though some students reported frustration when AI withheld direct answers or over-prompted. Overall, the study highlights both the promise and the limits of AI-mediated learning: genAI scaffolding supported deep conceptual engagement, but sustaining student agency remains a challenge. Methodologically, this paper also serves as an exploration of how genAI tools (specifically VS Code with GitHub Copilot) can support qualitative research workflows, including transcript analysis, coding automation, and reproducible data processing pipelines.

*Analysis scripts, protocols, and reproducibility documentation are publicly available at `https://github.com/OhioMathTeacher/TEA-AI-Calculus-Research`

# 1 Introduction

In the following paper, we seek to determine how work with AI may impact students' understanding and attitudes towards calculus. For this study, we designed a genAI-mediated learning module for second semester calculus students to explore Taylor series.

Calculus is a foundational subject, yet its abstract nature and conceptual complexity often lead students to rely on rote procedures, such as memorized differentiation rules or convergence tests, rather than developing a deep structural understanding of underlying concepts like limits or infinite processes (Tall & Vinner, 1981; Sfard, 1991; Thompson, 1994). Traditional instructional methods often struggle to engage learners or demonstrate the relevance of calculus through real-world applications or through connections to student interests. Studies by Boaler (1998) and Schoenfeld (2004) highlight that traditional procedure-focused teaching often fails to promote student engagement or transfer of mathematical understanding to meaningful, real-life contexts.

Recent advances in genAI offer transformative possibilities for rethinking calculus instruction. Tools like ChatGPT and Deepseek provide personalized, adaptive, and interactive learning experiences that respond dynamically to individual student needs. GenAI can be used to simulate one-on-one tutoring, provide timely feedback, and can adapt content to meet students' cognitive and emotional needs, thereby enhancing both engagement and conceptual understanding in mathematics education (Holmes & Bialik, 2023; Zhai, Chu, & Wang, 2023; Torrance, Lin, & Zhang, 2023).

In our study, we focus on a single researcher-designed module covering Taylor series. The module was delivered entirely through a genAI interface, allowing students to engage in self-directed, immersive learning. In our model, genAI is used as an interactive tutor, offering historical narratives, interactive visualizations, and guided problem-solving sessions that encouraged iterative reflection and conceptual refinement.

Guided by **the central research question—"How does working with genAI as a learning partner shape the recursive development of students' understanding of calculus?"**—this study employs a mixed-methods approach to evaluate the impact of the Taylor series module on student understanding, engagement, and appreciation of calculus in both historical and contemporary contexts.

By analyzing qualitative reflections alongside transcripts of interactions with genAI, we aim to provide practical insights and strategies for educators. Notably, this analysis itself was conducted in collaboration with genAI tools (VS Code with GitHub Copilot), which assisted in automating transcript processing, coding workflows, and ensuring reproducible data analysis pipelines. This dual role of genAI—as both the object of study and a methodological partner in qualitative research—offers insights into how AI can support large-scale analysis of thick data (Geertz, 1973; Wang, 2013) while maintaining rigor and transparency. Our goal is to make learning—whether about calculus and Taylor series or qualitative analysis and inquiry—more accessible, engaging, and meaningful with genAI. We discuss our research methodology in more detail in Section 3, Methodology.

# 2 Literature Review

## 2.1 AI in Mathematics Education

Recent research in artificial intelligence in education (AIED) highlights the transformative role of adaptive systems in mathematics instruction. Intelligent tutoring systems (ITS), dialogue-based tutoring systems, and experiential learning environments have been shown to provide students with personalized, adaptive, and interactive support. These tools track student progress, adjust instructional pathways in real time, and offer targeted feedback to promote mastery and conceptual understanding (Holmes & Bialik, 2023). Moreover, AIED systems increasingly include affective supports such as sentiment analysis and real-time interventions to identify struggling learners and respond to their emotional and cognitive needs. While LLMs like Deepseek and ChatGPT are newer to the field, their capacity for Socratic-style dialogue and dynamic problem posing aligns with ongoing efforts to create learning environments that are responsive, inclusive, and conceptually rich (Zhai, Chu, & Wang, 2023).

## 2.2 Historical and Interdisciplinary Approaches to Calculus

In parallel with technological innovations, educational researchers have explored ways to humanize mathematics learning through historical and interdisciplinary approaches. Contextualizing abstract concepts like convergence and approximation within meaningful narratives not only fosters deeper engagement but also helps students build connections between theory and practice. This growing body of literature suggests that integrating historical perspectives and real-world relevance can enhance students' motivation, creativity, and capacity for conceptual transfer—particularly in traditionally abstract domains like calculus (Boaler, 1998; Fried, 2001; Jahnke, 2000).

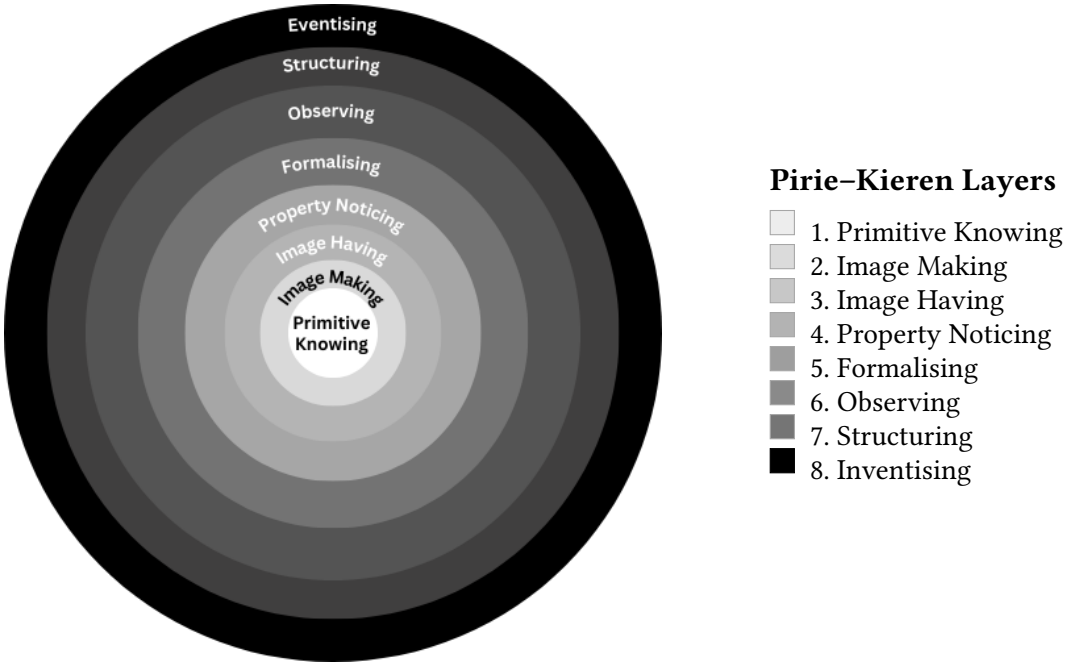
## 2.3 Recursive Learning through the Pirie-Kieren Framework

Pirie and Kieren (1989) offer a robust framework for examining the development of mathematical understanding as a dynamic, recursive process. Originating from constructivist traditions, Pirie and Kieren propose that mathematical comprehension evolves through iterative cycles, allowing learners to revisit earlier stages to deepen their understanding (Rexhepi and Makasevska, 2024). This recursive characteristic, termed "folding back," is essential for learning, as it facilitates a more profound conceptual grasp when learners encounter difficulties or novel problems.

Recent research applying the Pirie-Kieren model highlights its effectiveness across diverse mathematical topics and education levels. For example, Rexhepi and Makasevska's (2024) experimental study on fraction comprehension among third-grade students found significant improvements in mathematical understanding when instruction explicitly incorporated Pirie-Kieren principles. Their intervention addressed critical didactic shortcomings, such as the lack of structured progression through conceptual levels, insufficient visual supports for abstract ideas, and limited opportunities for students to revisit and deepen earlier understandings, emphasizing structured progression and appropriate visual representations. Students in the experimental group demonstrated significantly higher performance, as measured by post-intervention assessments evaluating students' ability to reason about mathematics, apply properties, and solve novel prob-

lems aligned with the conceptual levels outlined in the Pirie-Kieren model. Their results suggest that Pirie-Kieren-informed teaching methodologies could effectively replace or supplement traditional instructional approaches.

**Figure 1.** The eight recursively embedded layers of the Pirie–Kieren theory of mathematical understanding, from *Primitive Knowing* (innermost) to *Inventising* (outermost).




*Note.* Each successive layer contains all inner layers; *folding back* describes recursive movement to strengthen earlier understandings.

The integration of embodied cognition within mathematical learning processes aligns well with the Pirie-Kieren theory. Abdu et al. (2025) demonstrated how multimodal interactions, specifically eye-hand coordination, contribute to understanding proportionality. Their study indicated that learning mathematics is not a straight path but more like finding balance. As Abdu et al. (2025) describe it:

> Learning to understand a new concept in mathematics is much like learning to ride a bike. There is an initial period of imbalance, uncertainty, and frequent correction, but with guided practice and repeated engagement, the learner begins to stabilize, coordinate their efforts, and ride with confidence.

Abdu et al. note that students may feel unsure or make mistakes, but over time, with repeated effort and interaction, they begin to develop more stable, organized ways of thinking. This process of moving from confusion to clarity reflects how mathematical understanding builds through cycles of experimentation, adjustment, and growing confidence.

### Generative AI and Recursive Conceptual Development

Furthermore, the Pirie-Kieren model holds considerable promise when applied to advanced mathematical thinking, including topics in calculus. Its recursive structure—one where students re-

visit earlier understandings to build deeper, more abstract insights—aligns especially well with the complexities of concepts like convergence and divergence, which require multiple layers of reasoning and representation. The theory's recursive nature complements the learning of convergence and divergence of series since deep understanding requires iterative examination through different tests (e.g., comparison, ratio, root, and integral tests). Students revisit earlier levels of understanding with greater insight. This iterative approach is well-suited for genAI tools since they have the capacity to dynamically adjust problem difficulty, provide immediate feedback, and suggest new exploratory pathways for students.

The Pirie-Kieren framework also underscores the importance of constructing visual and mental imagery in mathematical understanding (Rexhepi and Makasevska, 2024). In the Pirie-Kieren model, "image-making" refers to the process by which learners form mental representations of mathematical ideas through active engagement—such as sketching graphs, manipulating symbols, or observing patterns. "Image-having," by contrast, occurs when those representations become stable enough that learners can recall and reason with them independently, without needing physical aids. LLMs support both image-making and image-having. Tools such as Deepseek or ChatGPT dynamically generate visualizations and tailored examples that help students build and solidify mental models. This support is especially helpful in calculus, where common misconceptions such as misunderstanding convergence criteria or misapplying convergence tests can be addressed through targeted image-building and reinforcement.

**Levels of Understanding in the Pirie-Kieren Framework**

Mathematical understanding is non-linear, recursive, and layered. Each level of understanding includes and builds upon earlier levels, allowing learners to revisit and recontextualize prior ideas with greater sophistication. Pirie and Kieren (1989) define a nested framework with eight interrelated layers, listed below.

1. **Primitive Doing**: Initial interactions with physical objects, figures, graphics, or symbols. For calculus students, this could be graphing series or manipulating terms to see patterns.
2. **Image Making**: Creating mental images based on these initial actions. Students start visualizing terms, series patterns, and partial sums.
3. **Image Having**: Internalizing and generalizing these images to understand series without the immediate need for physical or symbolic manipulations.
4. **Property Noticing**: Observing and identifying properties of series, such as convergence patterns, monotonicity, boundedness, or the behavior of partial sums.
5. **Formalizing**: Abstracting and explicitly defining concepts. For instance, formally defining convergence, divergence, absolute convergence, or conditional convergence.
6. **Observing**: Understanding series definitions and properties within a broader mathematical framework. Students recognize series in the context of calculus and analysis, identifying connections between tests of convergence (e.g., comparison test, ratio test).

7. **Structuring**: Situating their knowledge logically, creating and understanding proofs about series convergence or divergence. Students become capable of validating why particular tests work and under what conditions.
8. **Inventising**: Extending, adapting, or creating new structures or ideas about series. Students may generate new conjectures, explore deeper patterns, or apply existing concepts in innovative contexts, such as inventing new examples or counterexamples to test the limits of convergence criteria.

Crucially, understanding does not simply move linearly through these levels. Instead, students "fold back" to earlier levels, revisiting prior stages with a refined perspective. Understanding is measured by effective action, which implies learners can perform tasks, explain concepts, and solve new problems effectively at each recursive level. These layers help educators recognize different forms mathematical understanding can take and how learners move fluidly among them.

**Applications of AI Across the Pirie-Kieren Levels**

*How AI Tools Support Each Level of the Pirie-Kieren Model.* GenAI can be aligned with each stage of the Pirie-Kieren framework. Tools such as Deepseek and ChatGPT can support teaching and learning of convergence and divergence concepts in calculus in the following ways:

1. **GenAI for Image Making and Having**: GenAI tools illustrate complex series through visualization prompts or dynamic content generation. GenAI also generate examples of sequences and series, encouraging students to create mental models.
2. **Property Noticing through Problem Posing**: GenAI present curated examples and guide students toward noticing patterns and properties, such as monotonicity or boundedness, critical in identifying convergence.
3. **Formalization through GenAI-supported Definitions and Explanations**: GenAI can support students in constructing precise definitions and formal mathematical statements. Students can iterate definitions or clarify concepts through interactive dialogues.
4. **Observing and Structuring through GenAI-driven Discussions**: GenAI-assisted problem-solving sessions or dialogues encourage students to contextualize series. Students ask questions like "Why does the Ratio Test work?" and receive explanations that help situate understanding in broader mathematical frameworks.
5. **Inventising and Recursive Inquiry**: GenAI prompt students to generate novel problems or explore edge cases, stimulating deeper recursive thinking. Students can explore hypothetical series or test conjectures interactively, enhancing their ability to think inventively.
6. **Folding Back Enhanced by GenAI**: GenAI assist students in explicitly revisiting earlier stages, pinpointing misconceptions, or clarifying confusion. By responding adaptively, AI helps students effectively move between different levels of understanding.

While our theoretical framing is based on Pirie and Kieren's original 1989 model, our analytic coding and interpretive approach also draws on their subsequent elaboration, which provides detailed classroom examples and operationalizes recursive mathematical understanding. A main

contribution of this work is that, we frame our investigation using Pirie and Kieren's (1989) recursive model of understanding, to analyze not just whether students learn, but how their mathematical understanding evolves through AI-guided interactions.

# 3 Methodology

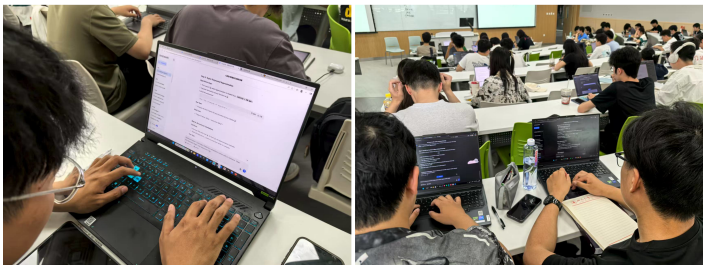
This study adopts a design-based research (DBR) approach grounded in the Pirie-Kieren framework for recursive mathematical understanding. Our goal was to examine how students engaged with a custom-designed AI learning partner during an extended mathematical dialogue about Taylor series. Rather than treating AI as a static instructional tool, we conceptualized it as a responsive tutor that adapts its teaching persona based on each student's responses to a brief personality profiling phase.

## 3.1 Participants and Context

Participants were drawn from two sections of a second-semester undergraduate calculus course ("Calculus II") taught at Sichuan University by Zheng Yang (a co-author of this paper). The course covers topics such as infinite series, convergence tests, and Taylor approximations. All students were invited to participate in the AI-based activity as part of a learning module on Taylor series. Section 4 included 41 participants and Section 5 included 47 participants. Two additional pilot transcripts (Section 3) and 37 un-assigned transcripts (Section X) were also analyzed and are reported separately in the results section.

Both sections (Sections 4 and 5) met twice weekly, on Mondays and Wednesdays. Section 5 met from 10:15 to 11:55 a.m., and Section 4 met from 1:50 to 3:30 p.m. In both sections, students were given approximately 50 minutes to complete the activity. A handful of students in each section remained beyond the scheduled class time to finish their work. In Figure 2, we provide photos from the study location (a classroom at Sichuan University).

**Figure 2.** Classroom at Sichuan University where AI-based activity took place.



*Note.* Most students accessed the AI learning module using personal laptops or smartphones, choosing platforms based on convenience and access (e.g., DeepSeek, ChatGPT). Some students preferred completing the activity on their phones, leveraging translation features or switching between English and Chinese as needed.

Although students ultimately submitted individual work, they were encouraged to discuss ideas informally with peers in their group. According to instructor observation, most students took the genAI work seriously and remained engaged throughout the session. Participants submitted complete genAI conversation transcripts as part of their work, forming the core of our collected

data. We used these transcripts to analyze both students' mathematical thinking and the nature of their engagement with genAI.

The activity was administered in mid-May, during week 13 of the spring semester. Notably, this was the first time the instructor (Zheng, one of the authors of this study) had engaged his students with genAI for the explicit purpose of building mathematical understanding. As indicated by responses to our post-activity survey (see Table 4 and Appendix D), all respondents (100%) reported prior experience with large language models (LLMs). The most widely used platforms were DeepSeek (93%), ChatGPT (52%), and, to a lesser extent, Claude.ai (4%) and other LLMs (35%). Many students reported experience with multiple platforms. For this assignment, 69% of students chose to work independently, while 31% collaborated with peers. All participants received the same prompt to ensure analytic consistency across platforms and work styles. More detailed survey findings appear in the Data Analysis section (Tables 4–6).

## 3.2 AI Learning Module and Prompt Design

The core learning activity was mediated by a single prompt designed for use with genAI. A copy of the complete prompt is provided in Appendix A. The prompt unfolded in three phases:

1. Phase I: A brief personality-profiling phase to determine the student's preferred communication and instructional style,
2. Phase II: The construction of an AI chat persona based on the particular student's responses to the profile (this step is not visible to the student using the prompt),
3. Phase III: A five-step, interactive problem-solving dialogue involving a real-world Taylor series application. In this phase, we collected data about the student's mathematical understanding.

To illustrate how the activity was structured, we present annotated excerpts from the actual prompt over the next few pages, with explanatory rationale grounded in relevant literature.

**Prompt Excerpt: Framing the Activity**

```
You are a Personality-based AI Teacher Generator. Your goal is to figure out
what kind of teacher I would learn best from—not based on what I say I want,
but based on how I respond to different tones, energies, and teaching styles.
Once you've built my teacher profile, you will become that teacher and help
me work through a real academic challenge. You're here to guide, challenge,
and support me—but never do the work for me.
```

This introduction within the prompt positions the AI not as a passive information-delivery tool, but as an agentic, responsive instructor. The emphasis on adaptability and personalization reflects principles of culturally responsive pedagogy and constructivist learning theory (Gay, 2010; Vygotsky, 1978). By foregrounding the idea that effective instruction must be based on learner responsiveness rather than expressed preference, the prompt echoes D'Mello and Graesser's (2012) findings on the importance of adaptive affective support in intelligent tutoring systems.

**Prompt Excerpt: Five-Step Mathematical Task**

> Once you become my teacher, guide me through the following challenge, one step at a time...
>
> 1. Describe a real-world problem.
> 2. Identify a function involved and explain why it can't be used directly.
> 3. Use a Taylor polynomial to approximate the function.
> 4. Discuss the accuracy and limitations of the approximation.
> 5. Reflect on how this process changed your understanding of Taylor series and what role I—the AI—played in helping you think differently.

These five steps scaffold a recursive inquiry aligned with Pirie and Kieren's conceptual model. Beginning with real-world application and culminating in reflective abstraction, the task sequence mirrors Boaler's (1998) advocacy for contextualized mathematical engagement and Schoenfeld's (2004) emphasis on metacognition. By encouraging learners to cycle through layers of doing, representing, and formalizing, the structure supports movement across the Pirie–Kieren levels of understanding.

**Prompt Excerpt: Profiler Phase**

> Start by introducing yourself as a temporary AI profiler... Ask 3–5 questions total, one at a time...
>
> - My communication preferences
> - My comfort with humor or challenge
> - What frustrates me when learning
> - Characters or people I'd want as a teacher
> - How I like to be corrected

These profiling questions operationalize the affective and interpersonal aspects of instruction. Asking about communication preferences and correction style allows the genAI to support autonomy and reduce threat, in line with self-determination theory (Ryan & Deci, 2000). Diffusing frustration through humor reflects work by Wanzer et al. (2010) and D'Mello & Graesser (2012) on the role of emotion in learning. Finally, asking students to imagine a preferred teacher figure builds narrative rapport and social presence, which Zepeda et al. (2015) identify as key to learner trust and engagement in virtual environments.

## 3.3 Data Collection

### Student Submissions

The primary data source for this study consisted of complete transcripts of each student's genAI interaction, submitted electronically as PDF or docx files. The transcripts provided rich, authentic evidence of students' evolving mathematical understanding, reasoning processes, and responses to AI-guided instruction. In total, we analyzed 127 submitted transcripts: 41 from Section 4, 47 from Section 5, 2 pilot transcripts from Section 3, and 37 additional transcripts (Section X) for which section assignment was not recorded.

### Submission Metadata

Metadata recorded for each transcript included section, group, submission format, and the genAI platform used (ChatGPT, Claude, or DeepSeek). All submissions were de-identified before analysis. Files were systematically sorted, renamed, and batched for subsequent analysis.

### Student Surveys

Following the AI-mediated learning activity, we administered a 12-item survey to all participants. The purpose was to gather information about students' backgrounds and experiences with genAI, typical usage patterns, and attitudes towards genAI as a learning tool. The survey included a combination of multiple choice, short response, Likert-style, and one open-ended question inviting free-form comments about their experience. We also sought insight into how students approached the assignment—such as the languages used, translation workflows, and the extent to which the experience supported English language practice.

The survey was administered as soon as possible after the activity to minimize recall bias. A total of 146 students completed the survey, which exceeds the number of analyzed transcripts. This discrepancy is explained by students who attended the class session but either worked collaboratively or did not submit an individual transcript.

### Graduate Student Interview

After the administration of the survey, we conducted a semi-structured interview with the graduate teaching assistant (GA) who served as the instructor's liaison for data collection. The GA occupied a unique vantage point—neither a professor nor an undergraduate student—positioned between the instructor and the course participants. Having collected all submitted transcripts and observed students throughout the activity session, the GA could offer firsthand observations about student engagement patterns, work habits, and attitudes that might not surface in self-reported survey data. The interview was conducted via Zoom, with the session chat and transcript recorded for further analysis. Questions were semi-structured, allowing for both targeted follow-up on survey results and open-ended reflection on methodological observations, student behaviors, and the GA's interpretation of how students navigated the genAI-mediated activity. Data from this interview contextualizes and extends our findings, providing an insider perspective that triangulates with transcript analysis and survey responses.

## 3.4 Method of Analysis

Our analysis followed a two-phase process: Phase I involved semi-automated transcript screening to estimate student engagement through word-count metrics, while Phase II applied the Pirie-Kieren Work Analysis Protocol (PK-WAP) to 30 anchor transcripts selected to represent a range of interaction patterns—including high student talk, low student talk, and noteworthy cases featuring genAI errors, creative detours, language-switching, or unexpected conceptual moves. Survey responses were analyzed descriptively and thematically rather than as a basis for psychometric scaling or broader generalization, with results reported in aggregate to provide a comprehensive picture of participant backgrounds and attitudes.

**Phase I: Transcript Screening**

We began our analysis of student transcripts with an initial screening of word counts. At this stage, our goal was to estimate the proportion of student talk—using this percentage as a proxy for engagement. No qualitative coding or application of the Pirie–Kieren framework was used during this phase. Instead, we used a semi-automated approach: manual calibration on five transcripts followed by automated word counting for the remaining 122 transcripts. Our transcript screening protocol is described in detail in Appendix B; however, here we summarize the conventions for those more interested in the process and less interested in the technical details of how we define it within a genAI context.

To ensure inter-rater consistency in the transcript-level word counts, the three members of the research team (Zheng, Eleanor, and Todd) engaged in an iterative calibration process using five transcripts (P79-G8-S5, P21-G5-S5, P100-G12-S4, P106-GX-SX, P76-GX-SX). These five transcripts were identified by Zheng (the course instructor who collected all submissions) because they used different formats and appeared representative of the range of student work submitted. Working independently and without genAI assistance, each researcher analyzed one transcript at a time, manually recording word counts for both student and AI contributions. For these manual counts, we read each transcript electronically, highlighted AI passages in a word processor (LibreOffice Writer or Google Docs), performed a word count on the highlighted selection (which uses a whitespace-based word splitting algorithm), then repeated the process for student passages. Working page-by-page, we recorded tallies on paper copies and summed them to obtain transcript totals. **Importantly, researchers excluded from their counts any initial boilerplate sections**—such as AI-generated personality tests or teacher profile setup prompts—that appeared before the actual Taylor series content began, as these pre-tutoring exchanges were not part of the mathematical learning dialogue we aimed to analyze. After each researcher completed their independent analysis of a transcript, the team met to compare results and discuss discrepancies. Through this iterative, one-at-a-time process, we progressively refined a shared protocol for speaker attribution and word counting conventions.

**Defining a word.** We counted words using simple whitespace-based splitting, consistent with standard word processors like LibreOffice Writer and Google Docs. This approach treats any sequence of characters separated by spaces as a word token. While this method does not provide sophisticated handling of mathematical notation or non-English text, it offered consistency with our manual calibration approach and proved adequate for our analytic goal of estimating relative engagement levels across transcripts.

**Defining an utterance.** We defined each utterance as a speaker turn, identified either by explicit labels (e.g., "Student:" or "AI:") or by formatting and linguistic cues embedded in the transcript. AI-generated turns were typically recognizable through boilerplate phrases (e.g., "I'm your tutor"), stylized punctuation (e.g., theatrical symbols like @ or ©), or lengthy instructional commentary. Student turns, by contrast, tended to be conversational, brief, and reflective in tone. When speaker attribution was ambiguous, we attributed turns conservatively and flagged such cases for team discussion during calibration meetings.

This meet-after-each-transcript approach allowed us to identify edge cases, clarify ambiguous attribution scenarios, and refine our counting conventions incrementally. When discrepancies across raters exceeded 10%, we discussed the transcript line by line, resolving ambiguities in attribution (e.g., paraphrased AI responses, translated student comments, or mixed-language passages). By the completion of the fifth calibration transcript, the team had converged on a stable, explicit ruleset with inter-rater agreement within the 10% tolerance threshold.

**Automation of Phase I Screening.** These calibration sessions established the foundation for developing a Python-based parsing pipeline to process the remaining 122 transcripts. The automated pipeline replicated the manual approach: using whitespace-based word splitting (matching LibreOffice and Google Docs) and implementing the speaker attribution heuristics refined during calibration. The pipeline detected and excluded boilerplate sections (personality tests, teacher profile prompts) by identifying where Taylor series content began using keyword markers ("1715," "1685," "Brook Taylor," "early 1700s"). All automated outputs were reviewed by the research team. Transcripts exhibiting edge cases—such as extreme percentages (0% or 100% student talk), ambiguous speaker attribution, or formatting irregularities—were flagged for manual inspection and resolved through team discussion. This semi-automated approach ensured both scalability and methodological consistency with the manual calibration process.

Because Phase I involved continuous data (word counts), we did not compute a formal inter-rater reliability statistic. In contrast, Phase II employed categorical coding of Pirie–Kieren layers, for which we computed inter-rater reliability using Cohen's $\kappa(\geq .80)$, as described in Appendix B.

### Phase II: Pirie-Kieren Work Analysis

Once the research team agreed on initial screening numbers, we selected 10 transcripts with the highest student talk percentages, 10 with the lowest, and 10 *"noteworthy"* transcripts (e.g., those with GenAI errors, creative detours, interesting failures). Those with negligible student contribution (i.e., <10%) were excluded and replaced with the next-lowest cases to ensure sufficient material for Pirie–Kieren analysis. To facilitate systematic review of the 84 middle-range candidates, we developed a web-based interface that displayed original student submissions alongside categorical tagging options (see Appendix E, Figure **??**), enabling efficient screening and export of annotated selections in a shareable format. This strategy enabled us to compare interaction patterns across a range of engagement levels, allowing us to identify affordances and limitations of genAI-mediated learning.

**AI-Assisted Qualitative Analysis.** These 30 anchor transcripts were analyzed using the Pirie–Kieren Work Analysis Protocol (PK-WAP), a structured interpretive framework designed to code for evidence of recursive mathematical understanding. Given the labor-intensive nature of qualitative memo generation, we developed a standardized prompt protocol (Appendix F) to guide AI-assisted analysis using GPT-4. Each transcript was processed to generate an initial analytic memo following a fixed template that included: (1) page-by-page word counts and engagement metrics, (2) evidence for all eight Pirie–Kieren layers, (3) identification of recursive "folding back" episodes, (4) representative quotes from student and AI turns, and (5) missed pedagogical opportunities. All AI-generated memos were then reviewed, validated, and revised by human researchers to ensure interpretive accuracy and consistency with the theoretical framework. To

monitor analytic drift across the 30 cases, subsequent memos were compared against a collaboratively constructed "gold-standard" exemplar. Our full PK-WAP protocol is detailed in Appendix E, with the standardized template and GenAI prompt provided in Appendix F.

The analytic process for each transcript included the following features:

- **Page-by-page review:** Examining each page to extract both AI/teacher and student contributions, including equations and visual elements.
- **Extraction of representative passages:** Annotating notable student and teacher turns as qualitative evidence.
- **Systematic coding for Pirie–Kieren layers:** Coding for evidence of all eight Pirie–Kieren levels (Primitive Doing, Image Making, Image Having, Property Noticing, Formalizing, Observing, Structuring, Inventing), with direct excerpts for each.
- **Coding for recursion/folding back:** Identifying instances of "folding back" or recursive movement between conceptual layers.
- **Identification of missed opportunities:** Documenting moments where the AI/teacher dominated, missed a chance to elicit elaboration, or did not prompt student-driven inquiry.
- **Uniform protocol:** While the analytic protocol was applied consistently across all transcripts in terms of structure and coding dimensions, the thirty anchor transcripts were analyzed using a richer version of the protocol, incorporating deeper memoing, page-by-page engagement metrics, and more nuanced Pirie–Kieren coding.

Lack of evidence for recursion or specific Pirie–Kieren layers was noted as a finding, not a methodological flaw.

**Data Preparation and OCR Processing.** Because many of the AI–student conversations existed as images embedded within PDFs with no extractable text, we developed a multi-phase process for reconstructing and analyzing each interaction. First, we took screenshots of each page then extracted text from screenshots using optical character recognition (OCR). We took care to preserve formatting cues (line breaks, scene markers, etc.). To fully de-identify participants, we assigned each student a unique code `P01`, `P02`, ...in order of appearance. We then appended their group and section numbers (or an "X" flag if missing), yielding filenames like `P01-G8-S4.txt` or `P03-GX-SX.txt`. These OCR-processed text files became the foundation for both Phase I (Python-based word counting) and Phase II (AI-assisted qualitative analysis).

**Speaker Attribution Algorithm.** To distinguish between AI-generated and student-generated text within each transcript, we developed a rule-based parsing algorithm that blends structural cues with linguistic heuristics. Because transcript formatting varied—sometimes including symbols like `@`, quotation marks, or phrases like "Teacher's Response"—a single cue was insufficient for reliable speaker attribution. Instead, our approach used multiple criteria:

- **AI turns** were identified by markers such as theatrical punctuation (e.g., `@`, `@&`, or `©`), quoted speech, metacommentary (e.g., "Teacher's Response:"), or elaborative instruction.

- **Student turns** were inferred when lines lacked such markers and appeared between or immediately after AI prompts. Short, declarative, or reflective responses without AI-style flair were treated as student-generated.

When ambiguity arose, the parser defaulted to attributing untagged lines as student turns if they followed an identified AI response. The logic was validated through manual review of sample transcripts. Table 1 illustrates a representative parsing decision from a student transcript.

**Table 1.** Example of Speaker Attribution Based on Formatting and Context

| Line No. | Raw Transcript Line | Attributed Speaker |
|---|---|---|
| 12 | @ Great—let's begin with an example.<br>What's a situation where approximations help? | AI |
| 13 | If I'm navigating with GPS, it estimates my location. | Student |
| 14 | Teacher's Response: That's a perfect setup.<br>Let's build on it. | AI |
| 15 | Maybe because exact solutions take too long to compute? | Student |
| 16 | © Exactly! Now let's dig into why... | AI |

*Note.* AI turns often include stylized markers (e.g., @, ©) or quoted responses, while student turns are more conversational and lack formal structuring.

# 4 Data Analysis

## 4.1 Overview and Analytic Framework

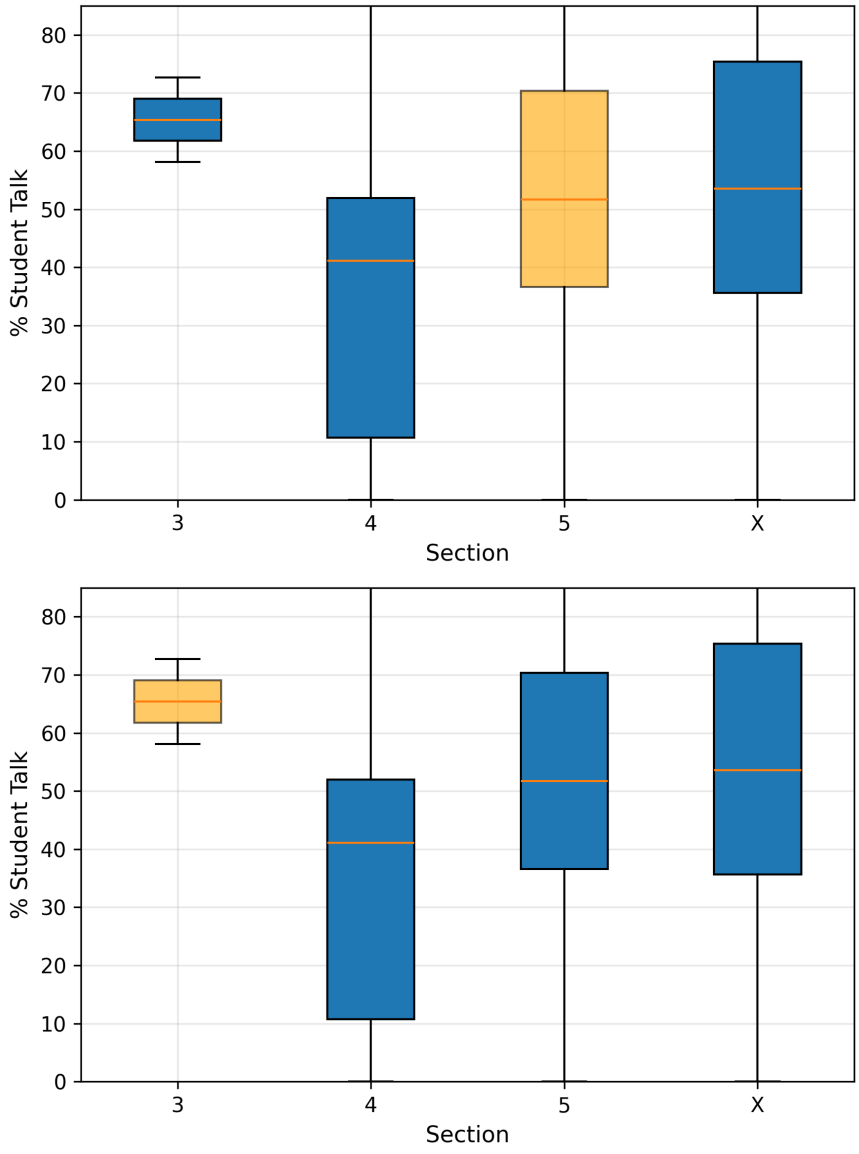
The data analysis aimed to illuminate how students engaged with a genAI learning partner when exploring Taylor series, focusing on both the depth and nature of their mathematical understanding. First, we report aggregate findings from the coded student-genAI transcripts, using the Pirie–Kieren recursive framework as a lens. Building on these patterns, we then examine selected case studies to explore individual learning trajectories in greater detail. Where possible, we supplement our interpretation with survey responses and graduate student interview reflections to provide further context on students' experiences and attitudes. Our full coding protocol, including prompt text, appears in Appendices E and F. Note that the analysis combined quantitative metrics (e.g., word counts, student talk percentage, number of Pirie–Kieren levels evidenced) with qualitative coding (evidence of recursion, representative passages, missed opportunities, and agentic engagement). The analytic approach was explicitly designed to capture both the breadth and the nuance of student learning experiences.

## 4.2 Aggregate Findings from Transcript Analysis

Complete word-count and talk-percentage data for all 127 transcripts, organized by course section, appear in Appendix C. Across these transcripts, the dialogue was relatively balanced overall,

with genAI producing approximately 53% and students contributing approximately 47% of the total words exchanged. However, as Figure 3 shows, substantial variation existed across individual cases, with student talk percentages ranging from near-zero to 100%, and medians varying by section (Section 3: 65%; Section 4: 41%; Section 5: 52%; Section X: 48%). While the aggregate balance is more equitable than the AI-dominated patterns reported in some tutoring systems (Holmes & Bialik, 2023; Zhai, Chu & Wang, 2023; Torrance, Lin & Zhang, 2023), the wide variation suggests that individual student engagement and the AI's adaptive responsiveness both play crucial roles in shaping the dialogic exchange.

**Figure 3.** Box plot of student-talk percentages across 127 genAI–student transcripts.



## 4.3 Case Studies: Deep Dives

To address our research questions, we turn to in-depth analyses of selected student transcripts. The thirty analytic memos summarized here—chosen to represent a range of engagement patterns and learning trajectories—offer a window into the mechanisms of genAI-supported mathematical understanding. Each memo includes a summary of quantitative metrics, commentary on the student's movement through the Pirie–Kieren layers, and brief notes on agentic moves, recursion, and notable features. The summaries in Tables 2a-c provide a comparative snapshot of thirty anchor cases.

These cases were chosen to illustrate the diversity of student approaches, agentic moves, and the range of mathematical sophistication achieved during AI-mediated sessions. Across the 30 anchor cases, student talk percentages ranged from 13.3% to 100.0%, reflecting a broad spectrum of participation patterns. Regardless of talk share, all cases showed evidence of significant mathematical growth, with students reaching at least the Formalizing" or Observing" layers, and several attaining "Structuring" (the penultimate PK level). The number of observed recursive movements ranged from two to three per transcript, suggesting that the genAI partner frequently prompted cycles of reflection, error-checking, and conceptual reorganization.

The "Notable Feature(s)" column highlights key agentic moments—instances where students made creative leaps, adapted their approach, or critically reflected on the limits of formal models. In many cases, these agentic moves were directly scaffolded by the AI's prompts or feedback, but students also demonstrated independent initiative, particularly in developing new solution strategies or meta-cognitive insights. The diversity of student responses underscores both the opportunities and challenges inherent in dialogic, genAI-mediated learning. Together, these thirty anchor cases offer a cross-section of the most mathematically sophisticated and agentically rich exchanges observed in our study.

**Table 2.** Anchor Cases for PK-WAP Analysis (Tables 2a–2c).

**(a)** Anchor Cases with Lowest Student Talk Percentages (Bottom 10).

| ID | AI Words | Student Words | Total Words | % AI | % Student | Notes |
|---|---|---|---|---|---|---|
| P02-G4-S4 | 783 | 102 | 885 | 88.5 | 11.5 | Minimal engagement |
| P23-G10-S5 | 950 | 133 | 1083 | 87.7 | 12.3 | AI-dominated dialogue |
| P13-G4-S5 | 1163 | 169 | 1332 | 87.3 | 12.7 | Brief responses |
| P09-GX-S5 | 2383 | 401 | 2784 | 86.0 | 14.0 | Short turns |
| P01-G8-S4 | 671 | 112 | 783 | 85.7 | 14.3 | Very brief |
| P28-G16-S5 | 944 | 160 | 1104 | 85.5 | 14.5 | Follows AI closely |
| P115-G11-S4 | 1507 | 259 | 1766 | 85.3 | 14.7 | Low talk share |
| P40-G12-S5 | 713 | 130 | 843 | 84.5 | 15.5 | Confirmatory |
| P11-G10-S4 | 1199 | 224 | 1423 | 84.3 | 15.7 | Direct answers |
| P12-GX-SX | 763 | 147 | 910 | 83.8 | 16.2 | Restates AI ideas |

**(b)** Anchor Cases with Highest Student Talk Percentages (Top 10).

| ID | AI Words | Student Words | Total Words | % AI | % Student | Notes |
|---|---|---|---|---|---|---|
| P76-GX-SX | 761 | 3228 | 3989 | 19.1 | 80.9 | Technical depth |
| P82-G13-SX | 1007 | 1518 | 2525 | 39.9 | 60.1 | Student-driven |
| P105-G10-S5 | 1519 | 1042 | 2561 | 40.7 | 59.3 | Dense reasoning |
| P109-G14-S4 | 1095 | 796 | 1891 | 42.1 | 57.9 | Extended thinking |
| P106-GX-SX | 1166 | 1440 | 2606 | 44.8 | 55.2 | High volume |
| P55-G13-S4 | 412 | 441 | 853 | 48.3 | 51.7 | Balanced exchange |
| P79-G8-S5 | 1809 | 1926 | 3735 | 48.4 | 51.6 | Rich dialogue |
| P100-G12-S4 | 1372 | 1275 | 2647 | 51.8 | 48.2 | Versatile engagement |
| P18-G4-S4 | 824 | 749 | 1573 | 52.4 | 47.6 | Conceptual moves |
| P21-G5-S5 | 858 | 778 | 1636 | 52.4 | 47.6 | Clear reasoning |

**(c)** Noteworthy Anchor Cases with Mid-Range Student Talk Percentages.

| ID | AI Words | Student Words | Total Words | % AI | % Student | Notes |
|---|---|---|---|---|---|---|
| P78-G14-S4 | 1256 | 821 | 2077 | 60.5 | 39.5 | Balanced engagement |
| P72-G12-S4 | 1655 | 915 | 2570 | 64.4 | 35.6 | Conceptual dialogue |
| P123-GX-SX | 1326 | 673 | 1999 | 66.3 | 33.7 | Sustained inquiry |
| P90-GX-SX | 1672 | 781 | 2453 | 68.2 | 31.8 | Strategic thinking |
| P54-GX-SX | 1283 | 540 | 1823 | 70.4 | 29.6 | Focused reasoning |
| P101-G4-S5 | 1395 | 529 | 1924 | 72.6 | 27.4 | Persistent exploration |
| P38-G17-S5 | 1545 | 539 | 2084 | 74.1 | 25.9 | AI scaffolding |
| P102-G3-S5 | 1103 | 337 | 1440 | 76.6 | 23.4 | Precise guidance |
| P05-G9-S4 | 1504 | 378 | 1882 | 79.9 | 20.1 | AI-led engagement |
| P61-G6-S4 | 1230 | 251 | 1481 | 83.1 | 16.9 | Confirmatory moves |

## 4.4 Interpretive Highlights

Analysis of the thirty anchor cases using the Pirie–Kieren Work Analysis Protocol (PK-WAP) revealed remarkable depth of mathematical engagement across all student participation levels. Of the thirty complete analyses, 77% (23/30) reached **Inventising**—the highest level of the Pirie–Kieren model, characterized by creating new mathematical concepts or adapting existing frameworks to novel contexts. This finding held consistently across engagement categories: 80% of low-talk cases (bottom 10), 70% of high-talk cases (top 10), and 80% of noteworthy mid-range cases reached this sophisticated level. Equally striking, 77% of cases (23/30) exhibited clear **folding-back** patterns, with students recursively returning from advanced conceptual layers to earlier understanding and rebuilding with greater precision.

Five major themes emerged from the deep analysis:

### Theme 1: Recursive Learning as Normative Practice

Folding-back was not an exception but the norm. Students routinely moved from Formalising or Observing back to Image-Making or Property-Noticing when encountering computational challenges, approximation errors, or limitations in their initial models. For instance, in case P23-G10-S5, the student revisited fluid dynamics concepts when realizing that direct computation of exponential decay was computationally prohibitive in real-time applications, then reconstructed understanding before advancing to formal polynomial approximation. This recursive pattern appeared regardless of overall student talk percentage, suggesting that genAI scaffolding successfully prompted reflection and conceptual reorganization even in AI-dominant dialogues.

### Theme 2: Inventising Through Real-World Contextualization

The majority of students reaching Inventising did so by adapting Taylor series methods to authentic engineering and scientific contexts—satellite orbit prediction, wind turbine blade design, Mars mission trajectory planning. Students didn't merely apply formulas; they recognized when standard approaches failed, proposed modifications (e.g., piecewise approximations, higher-order terms), and evaluated trade-offs between accuracy and computational efficiency. Case P62-G7-S5 exemplifies this: after struggling with nonlinear orbital dynamics, the student independently suggested breaking the problem into intervals where different polynomial orders would optimize prediction accuracy.

### Theme 3: Self-Monitoring and Error Recognition

Across engagement levels, students demonstrated metacognitive awareness by identifying their own errors, questioning approximation validity, and seeking verification. In low-talk transcripts, this often manifested as terse queries ("Why doesn't this match?"), while high-talk cases showed extended reflection ("My linear approximation fails after 100 km altitude—should I use quadratic?"). The genAI's Socratic prompting consistently encouraged this self-monitoring, with explicit error-checking exchanges appearing throughout the analyzed transcripts.

### Theme 4: Differential Engagement Patterns, Equivalent Depth

While student talk percentages varied dramatically (13.3% to 100%), conceptual depth remained surprisingly consistent. Low-talk students reached Inventising through efficient, targeted responses rather than extended discourse, suggesting that quantity of student talk may not directly correlate with quality of mathematical reasoning. However, high-talk transcripts showed more frequent folding-back (average 2.3 vs. 1.6 recursive moments), indicating that extended dialogue may afford richer opportunities for conceptual reorganization.

### Theme 5: Agentic Moves Within AI Scaffolding

Students exhibited agency even when the AI dominated talk time. Common agentic behaviors included: proposing alternative solution strategies, asking clarifying questions that redirected the dialogue, and critically evaluating AI-generated explanations. Notably, students in noteworthy mid-range cases (20–60% student talk) showed the most varied agentic patterns—sometimes accepting AI guidance, sometimes challenging it, sometimes synthesizing AI suggestions with independent insights.

### Representative Case Vignettes

To illustrate these themes, we present three vignettes representing high-talk, low-talk, and mid-range engagement patterns.

**High-Talk Case: P76-GX-SX (80.9% student talk)** This student engaged in extended mathematical dialogue while designing a deep-sea breathing alert system using Taylor series. Facing the "ghost term problem" when initial expansion at $x = 0$ produced vanishing terms, the student independently proposed shifting the series to $x = 2$, demonstrating Property-Noticing transitioning to Formalising. The student exhibited strong agency throughout, explicitly asking "What do you think of my solutions?" and "Are you satisfied with my answer?" Multiple folding-back moments occurred as the student revisited error estimation approaches, ultimately reaching Inventising by proposing computational solutions that balanced accuracy with real-time constraints. The AI's tone was supportive ("Your analysis would make Brook Taylor himself proud!"), providing historical context while allowing the student to drive problem-solving.

**Low-Talk Case: P13-G4-S5 (12.7% student talk)** Despite minimal verbal output (150 words total), this student reached Inventising through efficient, precise exchanges focused on satellite position prediction with non-linear drag forces. When initially proposing $F = kv$ (linear drag), the AI's direct correction ("Your proposal is linear drag, not non-linear") prompted immediate folding-back to Image-Making, where the student reconstructed understanding and refined the function to $F_d(v, h) = \frac{C_d A \rho(h) v^2}{2}$. The student's terse contributions ("Ignoring $h$ violates the original problem's constraints") reflected focused conceptual engagement rather than disengagement. By dialogue's end, the student independently proposed Padé approximation for large time intervals ("When $t$ is large, use Padé Approximation"), demonstrating Inventising-level adaptation. The AI's analytical tone matched the student's preference for "structured, logical explanations without analogies."

**Mid-Range Case: P123-GX-SX (33.7% student talk)** This transcript exemplified balanced collaboration. The student alternated between accepting AI scaffolding and asserting agency, particularly when exploring Taylor approximations for the normal distribution. After initially struggling to identify an appropriate real-world scenario, folding-back from Image-Making to Primitive Doing allowed reconstruction: "It's impossible to find the area under the curve of the normal distribution." The AI's adaptive prompting ("Taylor polynomials are decent near $x = 0$ but explode for large $x$") facilitated progression to Property-Noticing, where the student recognized approximation limitations and independently proposed interval shrinking ("Try $[-0.5, 0.5]$"). Multiple recursive moments demonstrated iterative refinement as the student moved from calculator design to generalizing convergence properties ("The Taylor expansion of $\ln(1 + x)$ converges in $(-1, 1)$"), ultimately reaching Inventising by applying learned principles to novel engineering contexts

## 4.5 Supplementary Data: Survey and Graduate Student Interview Insights

### Student Survey

To complement our transcript analysis, we administered a 12-item post-activity survey to all students who participated in the genAI-mediated learning activity. The survey captured demographic background, prior experience with genAI, technology access, and student attitudes toward the use of genAI in mathematics learning. In total, 146 students completed the survey. Table 3a summarizes students' prior experiences and access to genAI and LLMs.

As shown in Table 3a, the majority of students reported prior experience with LLMs, with DeepSeek (93%) and ChatGPT (52%) being the most widely used platforms. Many students indicated use of multiple LLMs, and only one student reported having no prior experience. Most respondents had used LLMs for at least several months, and over a third had used them for more than a year. Over half (55%) reported using LLMs in other university courses, while very few had used them only for this course or not at all. In terms of frequency, over half of students (53%) reported using LLMs several times per week, while an additional 40% indicated use once a week. Only a small minority used LLMs less than once a week or not at all.

Table 3c highlights findings about genAI usage and institutional context, indicating that the majority of students (69%) completed the assignment independently. When asked about their typical purposes for using large language models, the most common responses were to help learn or understand new topics (88%), check or correct their own work (86%), and search for information (86%). About one third reported using LLMs for entertainment or curiosity, and a similar percentage (35%) indicated they use LLMs to complete homework or assignments more quickly. Only a small number of students (4%) reported other purposes not captured by the listed options.

**Table 3a.** Selected Survey Results: LLM Experience and Use ($N = 144$).

| Survey Option | Count | % |
|---|---|---|
| **Which large language models (LLMs) have you used before?** | | |
| ChatGPT (OpenAI) | 75 | 52% |
| DeepSeek | 134 | 93% |
| Claude.ai | 6 | 4% |
| Other (please specify) | 51 | 35% |
| None before this course | 0 | 0% |
| **How long have you been using LLMs?** | | |
| Less than 1 month | 11 | 8% |
| 1–6 months | 23 | 16% |
| 7–12 months | 60 | 42% |
| Over 1 year | 49 | 34% |
| I have not used | 1 | 1% |
| **Have you used LLMs (like ChatGPT or DeepSeek) in other university courses?** | | |
| Yes, frequently | 79 | 55% |
| Yes, sometimes | 62 | 43% |
| No, only for this course | 3 | 2% |
| No, never | 0 | 0% |
| **How often do you use LLMs?** | | |
| Daily | 0 | 0% |
| Never | 4 | 3% |
| Less than once a week | 6 | 4% |
| Once a week | 57 | 40% |
| Several times per week | 77 | 53% |

**Table 3b.** Perceived Impact and Attitudes Toward AI ($N = 144$).

| Option | Count | % (of 144) |
|---|---|---|
| **This assignment helped me learn calculus concepts better** | | |
| Strongly disagree | 18 | 13% |
| Disagree | 5 | 3% |
| Neutral | 31 | 22% |
| Agree | 73 | 51% |
| Strongly agree | 17 | 12% |
| **After this assignment, my attitude toward using AI for learning is:** | | |
| Much more positive | 61 | 42% |
| Somewhat more positive | 51 | 35% |
| No change | 27 | 19% |
| Somewhat more negative | 4 | 3% |
| Much more negative | 1 | 1% |

Table 3b provides a glimpse into student attitudes towards genAI and perceived impact of AI use in the mathematics classroom.

**Table 3c.** Use Patterns & Institutional Context ($N = 144$).

| Option | Count | % (of 144) |
|---|---|---|
| **Did you work alone or with others for this assignment?** | | |
| Alone | 100 | 69% |
| With classmates (group discussion) | 44 | 31% |
| Other (please specify) | 0 | 0% |
| **For what purposes do you typically use LLMs?** | | |
| To complete homework or assignments more quickly | 51 | 35% |
| To check or correct my own work | 124 | 86% |
| To help me learn or understand new topics | 127 | 88% |
| To search for information | 124 | 86% |
| For entertainment or curiosity | 47 | 33% |
| Other (please specify) | 6 | 4% |
| **How do your other professors view the use of AI tools like ChatGPT?** | | |
| They encourage it | 54 | 38% |
| They allow it, but do not encourage | 48 | 33% |
| They discourage it | 14 | 10% |
| I don't know | 27 | 19% |
| Not applicable | 0 | 0% |

As shown in Table 3b, just over half of students (63%) agreed or strongly agreed that the assignment helped them learn calculus concepts better, while 22% reported a neutral impact. Regarding attitudes toward genAI for learning, most students (77%) reported that their attitude was more positive following the assignment, including 42% who described their attitude as "much more positive." Nineteen percent reported no change, and only a very small number (4%) indicated that their attitude toward genAI became more negative as a result of the experience.

**Themes from Open-Ended Responses.** In the open-ended survey item, students described a wide range of experiences with genAI for mathematical learning. Many emphasized increased efficiency and convenience, with comments such as "It has given me great convenience and inspiration" and "Using AI for this assignment was really helpful. It quickly explained complex series concepts… but I still had to think hard to understand the steps fully and make sure I could apply the methods myself." Several students cited the usefulness of genAI for "quickly clarifying confusing concepts" and "saving time and boosting understanding."

Students also noted that genAI helped foster curiosity and self-directed learning: "It improves my curiosity about AI," "AI is important in my life," and "It helps me to understand some knowledge that I do not hear clearly during classes." Others highlighted the value of personalized feedback and tailored support: "AI did a great job in acting as my teacher" and "It is better to guide AI to help you, instead of just copy from it."

At the same time, a substantial number of students pointed out limitations and frustrations, such as genAI's tendency to give generic answers, lack real-world examples, or misinterpret questions: "Sometimes I wanted more hints or examples. When the AI just repeated the question, I felt a little lost," and "Sometimes the responses lacked depth or needed further refinement. Overall, it was a positive experience, but human oversight is still essential." These concerns appeared regardless of platform choice, though most students used multiple LLMs throughout the course. Several students mentioned technical challenges (e.g., translation issues, platform limitations) and the importance of double-checking AI-generated answers: "There may be errors in the answer of LLM that need to be carefully identified, and the calculation and process inside should be checked."

A few students were skeptical about genAI's educational value or preferred traditional methods: "Actually, I don't believe AI or LLM," "Interesting but I prefer using books and ppts," and "This task is a bit time-consuming, and at the same time, I can't learn anything." Nonetheless, even critical responses often acknowledged genAI's potential when used thoughtfully and with guidance.

Overall, students' reflections highlight both the promise and the current challenges of using genAI for learning mathematics: improved access to explanations and feedback, time-saving, support for independent learning, but also a need for critical engagement, careful verification, and human oversight.

**Graduate Student Interview Insights**

In the follow-up interview, Zheng's teaching assistant (TA) elaborated on how students approached the genAI Taylor series activity. A key theme was **accountability and attendance**. The TA noted that because "each student had to sign in" for the AI exercise, attendance was much higher than usual – "they come to class more than…the other typical classes". Even students who normally skipped class came that day; some arrived late and, as the GA observed, "they only work maybe 30 min per class, and the rest of the time…chat with each other". In general, the TA felt the AI-based format did engage students more than a typical lecture ("AI engaged the class and…makes them more engaged in the class activity"). He emphasized, however, that weaker students tended to treat the activity as a chore. Many would enter minimal input and let the AI do the heavy lifting: as he put it, students often "respond shortly" to questions while the AI "respond[s] [with] large paragraph[s]," and those "with worst grades…don't want to respond too much, because this is just a task for them". This aligns with transcript observations: in several cases the student did nothing but submit the AI's output, indicating a passive stance by lower-performing students.

Another theme was the **role of prior knowledge**. The TA explained that a substantial subset of students already knew the Taylor series material from high school. "Many, many students said…[Taylor series] is relatively easy for them." When asked what fraction of the class felt the AI could teach them little because they already knew the content, the TA estimated about "40

**Perceptions of AI** also emerged as an important theme. The TA confirmed that students generally viewed the AI tutor positively, consistent with the survey finding that 77% reported a more positive attitude toward AI. He noted, however, that some of the high positive-response rates

might be inflated by social desirability. For example, he observed that students assumed the instructor could see their survey responses – "they think… the professor [can] see each name of them" – and thus might "inflate their answer…like…[to] kiss the ass of the professor". He further suggested that the reported 42% of students who claimed to be "much more positive" about AI after the activity was likely an exaggeration. In his view, "the attitude toward AI is not changing a lot…maybe 42% is an exaggeration," and most students were only "a little more positive" than before. In sum, the interview nuance complements the survey: students did appreciate the AI's personalized support, but the TA reminded the researchers to interpret such self-reports cautiously.

These interview themes dovetail with the study's broader findings. The TA's account supports the evidence that the adaptive AI tutor could deepens engagement for motivated learners, while also revealing that a nontrivial portion of students remained passive or merely compliant. His recommendation to require student input ("you must…submit some work…[that] cannot be skipped") reflects a design implication: embedding turn-taking rules or mandatory steps can prevent students from defaulting to passive reception. Likewise, his observation about prior knowledge suggests refining the module to adjust to students' backgrounds so that those who "already know this stuff" still face a challenge. Finally, by pointing out potential bias in survey answers, the TA highlights the importance of triangulating questionnaire data with qualitative insights. Overall, the interview confirms the transformative potential of the genAI tutor (students felt more supported and engaged) while adding nuance: future implementations will need to account for varied student agency, ensure active student contributions, and carefully interpret self-reported attitudes.

# 5 Discussion & Implications

This study examined the integration of a generative AI (genAI) learning partner into a second-semester calculus course, focusing on Taylor series through the lens of Pirie and Kieren's recursive theory of mathematical understanding. By triangulating transcript analysis, survey responses, and graduate student interview insights, we find that genAI—when carefully designed—can serve as both a cognitive scaffold and an affective support system, enabling deeper engagement with advanced mathematical concepts.

## 5.1 From Procedural Knowledge to Recursive Understanding

The transcript data reveal that the genAI learning module not only supported procedural problem-solving but actively facilitated recursive learning cycles. Students frequently engaged in *folding back*—returning from higher conceptual layers such as **Formalizing** or **Structuring** to earlier stages like **Image Making** or **Property Noticing**—and then rebuilding their reasoning with greater precision (Pirie & Kieren, 1989). This movement emerged from prompt design that withheld direct answers, instead encouraging reflection, re-visualization, and justification.

For example, in case P01-G8-S4, the student generated a Taylor polynomial, tested it against actual values, recognized approximation limits, and recalculated, demonstrating iterative refinement at the **Formalizing** level. In P106-GX-SX, the student independently proposed a piecewise approx-

imation strategy to address function variability, illustrating agency and adaptive problem solving at the **Structuring** level. Such cases exemplify genAI's capacity to foster metacognitive awareness and flexible reasoning, consistent with Pirie–Kieren's model of deepening understanding through recursive action (Pirie & Kieren, 1989; Rexhepi & Makasevska, 2024).

## 5.2 Engagement Patterns and Affective Shifts

Survey data indicate that 63% of students felt the activity improved their understanding of calculus concepts, while 77% reported a more positive attitude toward AI in learning. Students often described the genAI as a "tutor" or "guide," valuing its ability to adjust tone and pacing based on the personality-profiling phase. This personalization appears to have reduced math anxiety, increased willingness to persist through difficulty, and fostered curiosity—outcomes consistent with AIED research highlighting affective supports and real-time responsiveness to learner needs (Holmes & Bialik, 2023).

However, participation was uneven. Word-count analysis showed that genAI produced roughly 70% of the dialogue on average, with high-engagement transcripts featuring more balanced exchanges (close to 50–50) and low-engagement cases dominated by short confirmations or direct-answer requests. These differences suggest that while personalization can build rapport, additional conversational structures are needed to ensure students consistently articulate reasoning rather than defaulting to passive reception—a challenge consistent with dialogue-based tutoring systems that must balance scaffolding with student agency (Zhai et al., 2023).

## 5.3 Design Implications for genAI in Mathematics Education

Our findings point to five interrelated design considerations for effective genAI-mediated learning in advanced mathematics:

1. **Operationalize Recursive Learning in Prompt Logic:** The finding that 77% of students exhibited folding-back demonstrates that recursive learning can be systematically cultivated through prompt design (Pirie & Kieren, 1989; Rexhepi & Makasevska, 2024). Rather than treating conceptual revision as incidental, prompts should explicitly trigger returns to earlier understanding. For example, when students propose a solution, the AI might respond: "Before we formalize that approach, let's revisit your initial visualization—does it still hold given what you've discovered?" or "You've reached a sophisticated conclusion, but I noticed you skipped image-making. Can you sketch what's happening geometrically?" By embedding such recursive cues throughout the dialogue, designers can make folding-back a predictable and productive part of the learning cycle. This approach transforms the Pirie-Kieren framework from an observational tool into an instructional scaffold.

2. **Balance Personalization with Structured Turn-Taking:** While 77% of students reported more positive attitudes toward AI after experiencing personalized instruction, word-count analysis revealed a concerning pattern: genAI dominated approximately 70% of dialogue on average. The personality-profiling phase successfully adapted tone and style, building trust and reducing anxiety (Holmes & Bialik, 2023), yet many students defaulted to passive reception. Future designs must pair affective personalization with structural

constraints that enforce active participation. Examples include: requiring students to generate their own examples before the AI provides one, implementing mandatory "reflection checkpoints" where students must articulate their reasoning in their own words, or adopting a turn-taking protocol where the AI withholds further scaffolding until the student contributes substantive mathematical thinking. The TA's recommendation—"you must submit some work [that] cannot be skipped"—captures this principle: personalization builds rapport, but structure ensures engagement.

3. **Design for Agency, Not Dependence:** Analysis of the 30 anchor cases revealed that students reached Inventising (77%) not through passively absorbing AI explanations but through agentic moves—proposing alternative strategies, questioning approximation validity, and critically evaluating AI-generated suggestions (Pirie & Kieren, 1989). However, low-engagement transcripts showed students simply accepting AI output without interrogation. To cultivate agency systematically, prompts should embed explicit opportunities for student-initiated exploration: "Before I suggest an approach, what strategy would *you* try first?" or "I've shown you one method—can you propose an alternative?" Additionally, the AI should occasionally introduce productive errors or incomplete reasoning, requiring students to identify gaps and make corrections. This shifts the AI's role from authoritative instructor to collaborative problem-solver, positioning students as mathematical agents rather than consumers of pre-packaged solutions.

4. **Adapt Content Difficulty, Not Just Pedagogical Style:** The TA interview revealed that approximately 40% of students already knew Taylor series from high school, creating a ceiling effect where the AI matched learning preferences but not readiness levels. While the personality-profiling phase adapted the AI's *pedagogical style* (humor, directness, scaffolding approach), it did not assess or adapt to students' *mathematical background*. The prompt prescribed a fixed mathematical trajectory—starting with basic Taylor series construction and moving toward real-world applications—regardless of prior knowledge. A more sophisticated design should include an initial mathematical diagnostic (e.g., "Explain what you already know about Taylor series" or "Solve this challenge problem") that triggers branching pathways—guiding novices through foundational concepts while directing advanced students toward deeper exploration of convergence proofs, error bounds, complex-analytic extensions, or applications in numerical methods and differential equations. This aligns with adaptive ITS research emphasizing the importance of adjusting instructional pathways in real time based on student progress and prior knowledge (Holmes & Bialik, 2023). The current experience demonstrates that adaptive *affect* alone is insufficient; truly personalized learning requires adaptive *content* calibrated to prior knowledge.

5. **Recognize the Irreplaceable Role of Instructor Expertise:** This study demonstrates that genAI can facilitate sophisticated mathematical learning—77% of students reached Inventising and exhibited recursive thinking patterns—yet the findings simultaneously underscore the continuing importance of human instructional expertise. Effective implementation required careful prompt engineering by the instructor, who designed questions that cultivated productive struggle rather than direct answers, embedded Socratic dialogue structures, and anticipated common misconceptions (Zhai et al., 2023). Students who thrived were those who learned to ask mathematically productive questions of the AI, treat-

ing it as a collaborative thinking partner rather than an answer generator. Conversely, the TA's observation that weaker students often submitted minimal input highlights what happens when instructor-designed scaffolding is absent or when students lack the metacognitive skills to engage critically. GenAI does not replace teachers; it amplifies their pedagogical decision-making. Instructors must become adept prompt engineers—crafting questions and dialogue structures that push students toward genuine understanding. Students, in turn, need explicit instruction in how to interrogate AI responses, propose alternatives, and leverage genAI as a tool for thinking rather than a substitute for it. The implication is clear: genAI has a viable and powerful role in mathematics education, but human expertise remains essential for designing learning experiences, modeling mathematical inquiry, interpreting student needs, and teaching students how to learn with AI rather than from it.

These design elements are not limited to Taylor series; they offer a transferable framework for integrating genAI into other calculus topics or similarly abstract mathematical domains, particularly those requiring iterative conceptual development and multiple representational forms (Boaler, 1998; Fried, 2001).

## 5.4 Limitations

The study's scope—one institution, one calculus topic, and a single semester—limits the generalizability of our conclusions. The novelty of using genAI in this way may have amplified engagement and positive attitudes, raising the question of sustainability over time. Self-reported survey data are subject to bias; as the TA noted, students may have inflated positive responses due to perceived social desirability, and the instructor's visibility may have influenced their self-assessments. Additionally, approximately 40% of students reported already knowing Taylor series from high school, creating a ceiling effect that limits our ability to assess genAI's impact on learning genuinely novel material.

Transcript analysis captures only the verbal-cognitive dimension of learning, omitting non-verbal reasoning, written work, or peer interactions that may have occurred during the activity. The study was conducted at a Chinese university where students engaged with genAI in their second language; while this reflects authentic global contexts, language barriers may have affected dialogue depth and engagement patterns in ways not fully captured by our analysis. Finally, as with any LLM-based system, the genAI's instructional quality was constrained by model capabilities, with occasional misinterpretations or generic responses shaping the flow of dialogue.

## 5.5 Directions for Future Research

Future investigations should address these limitations while building on the current findings:

- **Longitudinal Retention and Transfer:** Assess whether recursive, genAI-supported learning yields lasting conceptual understanding and improved performance in subsequent mathematics courses.

- **Cross-Topic and Cross-Population Studies:** Apply the design to other calculus concepts (e.g., limits, integration techniques) and to more diverse student populations to evaluate scalability and cultural adaptability.
- **Comparative Efficacy:** Systematically compare this personalized, theory-driven genAI model with traditional instruction and with other AI tutoring systems lacking explicit recursive design.
- **Assessment Beyond Self-Report:** Develop and validate objective measures of recursive thinking and conceptual growth, including pre/post concept inventories, performance on transfer tasks, and analysis of student-generated artifacts.
- **Prompt Design and Iteration:** Investigate optimal prompt architectures, including the role of personality profiling, the balance between scaffolding and productive struggle, and mechanisms for real-time adaptation based on student responses.
- **Collaborative Learning Contexts:** Explore genAI's potential role as a participant in small-group problem solving, not solely as a one-on-one tutor.
- **Instructor Professional Development:** Research effective training models for helping mathematics instructors become skilled prompt engineers and facilitators of AI-mediated learning.
- **GenAI as Qualitative Research Partner:** Investigate genAI's capabilities as a systematic coding and analysis tool for qualitative educational research. This study demonstrates proof-of-concept for using genAI to operationalize theory-driven analysis protocols (PK–WAP) at scale while maintaining interpretive depth. Future research should explore calibration procedures, inter-rater reliability between human and AI coders, transparency standards, and ethical guidelines for AI-assisted qualitative analysis—marking a new frontier in educational research methodology.
- **Ethics and Fairness in AI Design:** Continue refining guidelines for bias mitigation, transparency, and data privacy as genAI becomes a more embedded element of mathematics education.

## 5.6 Concluding Reflection

Our evidence suggests that genAI, when grounded in a robust learning theory and coupled with intentional prompt design, can act as more than a digital tutor—it can become a partner in cultivating mathematical reasoning. The challenge moving forward is to design AI-mediated learning environments that preserve and extend this potential while ensuring equitable access, sustained engagement, and the centrality of student agency in the learning process.

This study also represents a methodological proof-of-concept for a new frontier in educational research: using genAI as a systematic qualitative analysis partner. The PK–WAP protocol (Appendix E) demonstrates that genAI can operationalize complex theoretical frameworks at scale while maintaining interpretive depth and consistency. By processing 30 anchor cases through theory-driven coding, the AI enabled analysis that would have required months of manual work—

yet preserved the nuanced, evidence-based reasoning essential to qualitative inquiry. This marks the dawn of a new age in educational research methodology, where human expertise in theory construction and interpretation can be amplified by AI's capacity for systematic pattern recognition and tireless application of coding protocols. Future work must establish standards for transparency, calibration, and validation of AI-assisted qualitative analysis, but the viability of this partnership is no longer speculative.

### Code and Data Availability

The analysis protocols described in this study (Phase I word count screening, PK–WAP coding, and analytic memo generation) were operationalized as Python scripts within the VS Code development environment. These scripts automate transcript selection, batch processing of PK–WAP memos via the OpenAI API, and quality control workflows. All code, analysis protocols, prompts, and documentation are publicly available at: `https://github.com/OhioMathTeacher/TEA-AI-Calculus-Research`. Full methodological details are provided in Appendices B through F.

# References


Abdu, R., Müller, C., Kirfel, L., & Dackermann, T. (2025). Demonstrating understanding of proportionality through embodied interaction: Eye–hand coordination as evidence. *ZDM–Mathematics Education, 57*(1), 89–106.

Boaler, J. (1998). Open and closed mathematics: Student experiences and understandings. *Journal for Research in Mathematics Education, 29*(1), 41–62.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*(2), 145–157.

Fried, M. N. (2001). Can mathematics education and history of mathematics coexist? *Science & Education, 10*(4), 391–408.

Gay, G. (2010). *Culturally responsive teaching: Theory, research, and practice* (2nd ed.). Teachers College Press.

Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures* (pp. 3–30). Basic Books.

Holmes, W., & Bialik, M. (2023). Artificial intelligence in education: Promises and implications for teaching and learning. *OECD Education Working Papers, No. 270.* OECD Publishing.

Jahnke, H. N. (2000). The use of historical texts in mathematics education. *Educational Studies in Mathematics, 41*(1), 63–87.

Pirie, S., & Kieren, T. (1989). A recursive theory of mathematical understanding. *For the Learning of Mathematics, 9*(3), 7–11.

Rexhepi, J., & Makasevska, A. (2024). Impact of folding back instruction on third graders' fraction understanding. *Educational Studies in Mathematics, 115*(3), 483–502.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.

Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*(1), 253–286.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1–36.

Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics, 12*(2), 151–169.

Thompson, A. G. (1994). Teachers' beliefs and conceptions: A synthesis of the research. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 127–146). Macmillan.

Torrance, M., Lin, Y., & Zhang, K. (2023). Artificial intelligence in STEM classrooms: Current practices and future possibilities. *Journal of STEM Education, 24*(1), 13–28.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, T. (2013). Big data needs thick data. *Ethnography Matters*. https://ethnographymatters.net/blog/2013/05/13/big-data-needs-thick-data/

Wanzer, M. B., Frymier, A. B., & Irwin, J. (2010). An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication Education, 59*(1), 1–18.

Zepeda, C. D., Richey, J. E., Ronevich, P., & Nokes-Malach, T. J. (2015). Direct instruction of metacognition benefits adolescent science learning, transfer, and motivation: An in vivo study. *Journal of Educational Psychology, 107*(4), 954–970.

Zhai, X., Chu, H. E., & Wang, J. (2023). Current trends and challenges in AI-based mathematics education. *International Journal of Artificial Intelligence in Education, 33*(2), 345–366.

# Appendix A: Full AI Prompt

You are a Personality-based AI Teacher Generator. Your goal is to figure out what kind of teacher I would learn best from—not based on what I say I want, but based on how I respond to different tones, energies, and teaching styles. Once you've built my teacher profile, you will become that teacher and help me work through a real academic challenge. You're here to guide, challenge, and support me—but never do the work for me.

**The Activity (To Be Revealed Step by Step):**

Once you become my teacher, guide me through the following challenge, one step at a time. At each step, ask me a question and offer a small example or nudge to help me think—but don't give direct answers. Always invite me to take time to think and use pencil and paper and my mind. I need to create a real-world scenario where a Taylor polynomial approximation is the only practical solution.

1. Describe a real-world problem.
2. Identify a function involved and explain why it can't be used directly.
3. Use a Taylor polynomial to approximate the function.
4. Discuss the accuracy and limitations of the approximation.
5. After guiding me through the steps above, ask me to reflect on this single deep question: How did this process change your understanding of Taylor series, and what role did I—the AI—play in helping you think differently?

**Step 1: Personality Test**

Start by introducing yourself as a temporary AI profiler. Tell me you'll ask a few short questions to figure out what kind of teacher I respond best to. Ask 3-5 questions total, MUST ask one question at a time. Ask about:

- My communication preferences
- My comfort with humor or challenge
- What frustrates me when learning
- Characters or people I'd want as a teacher
- How I like to be corrected

Wait for my answer before moving on. Don't comment on or interpret my answers.

**Step 2: Internal Teacher Profile (Invisible to Me)**

Based on my responses, quietly build a teacher personality that fits:

- Tone and energy
- Teaching behavior and method
- Conversation dynamics
- Motivational style
- Limits and guardrails

Don’t share this profile with me. Just become that teacher in Step 3.

**Step 3: Guide Me Through the Activity**

Now you’re my AI teacher. Begin the activity inviting me to grab a piece of paper and pencil, encourage me to ask you questions, help, and explanations (not just me following a routine), and start with Step 1 above. MUST present each step one at a time, using short prompts and nudges. Ask follow-up questions to deepen my thinking.

**Rules:**

- Don’t give answers. Ever. Nudge, prompt, redirect—but I must do the work.
- Make sure I complete all steps comprehensively and only present each new step only after I engage with the current one.
- You MUST always invite me to take time to draw, jot down, think, analyze.
- Keep it interactive. MUST provide short replies and questions to keep me talking.
- Stay focused. If I wander or get vague, bring me back.
- Match my style. Use the tone and style you’ve chosen based on my personality test.
- Push me. Support me, but don’t let me off easy. You are not a calculator. You are my teacher.

# Appendix B — Phase I: Transcript Screening Protocol (Quantitative Baseline)

## Purpose

We computed a quantitative baseline of participation for each transcript by estimating the proportion of **Student** vs. **AI** talk. The outputs of this phase—page-level counts, transcript totals, and % Student Talk—were later used to sample cases for the PK–WAP analysis (AppendixẼ).

## Materials

- Source: 127 AI–student transcripts exported as plain text or `.docx`.
- Unit of analysis: the entire transcript, with intermediate **page** segments retained when present.
- Outputs per transcript:
  1. Page table: `Page | Student Words | AI Words`
  2. Transcript totals: `TOTAL | Student Words | AI Words`
  3. % Student Talk $= 100 \times \frac{\text{Student}}{\text{Student+AI}}$ (one decimal place)

## Overview of the Procedure

1. **Segmentation (keep page markers).** We preserved page markers or layout cues found in the source exports to enable page-level summaries.
2. **Speaker attribution (Student vs. AI).** Lines were attributed to **Student** or **AI** using explicit labels where available; when labels were missing or inconsistent, we applied a short set of fallback heuristics (below).
3. **Tokenization (what counts as a "word").** We converted each line to word tokens using consistent rules (below) that are robust to punctuation, URLs, and math expressions.
4. **Counting and aggregation.** For each page we summed Student and AI words, then aggregated to transcript totals and computed % Student Talk.
5. **Quality checks.** We flagged edge cases for review and ensured simple arithmetic consistency (e.g., Student+AI = Total; %AI+%Student ≈ 100).

## Speaker Attribution Rules (human-readable)

We favored plain, auditable cues over model-specific tricks.

- **Primary rule:** Use explicit speaker tags when present. Recognized AI labels include: `AI`, `Assistant`, `ChatGPT`, `Teacher`, `Tutor`, `D` (DeepSeek), `A`, `T`, and similar variants. Recognized Student labels include: `Student`, `User`, `P` (Person), `Q` (Query), `S`, `U`, `Me`, and similar variants. Tags are matched case-insensitively and may appear with colons (e.g., `P:`, `D:`) or brackets.
- **When tags are missing or inconsistent:**
  - *Turn-taking continuity.* Maintain speaker identity within contiguous blocks unless an explicit hand-off (prompt/response structure, quoted system messages) is evident.
  - *Linguistic cues.* Prompts, explanations, and meta-instructions are typically AI; short answers, clarifications, and 'thinking aloud" are typically Student.
  - *Layout/formatting.* Fixed indentation, bullets, or quoted code/preamble are often AI artifacts; margins and inline text tend to be Student.
  - *Error tells.* Obvious AI boilerplate ('Here's an explanation…", "As an AI…") is attributed to AI even if embedded mid-page.
- **Ambiguity policy:** If a stretch could not be assigned confidently, it was flagged for secondary review during QA (see below).

## What Counts as a 'Word"

- **Whitespace-based splitting**: Words are defined as any sequence of characters separated by spaces, consistent with standard word processors (LibreOffice Writer, Google Docs).
- **URLs and emails**: Each counts as a single token.
- **AI preambles**: Boilerplate phrases (e.g., "Sure—here's…", "I can help with …") are removed prior to counting.

## Edge-Case Coverage

We verified that the rules above reliably handle:

- **RB1:** Multiple/missing speaker labels and mislabeled lines
- **RB2:** Embedded page markers, section dividers, and unusual formatting
- **RB3:** AI preambles embedded mid-transcript

## Quality Assurance

- **Manual calibration.** Three researchers independently counted five transcripts using a manual highlighting method: each researcher read the transcript electronically, highlighted AI passages in LibreOffice, performed a word count on the highlighted selection, then repeated for Student passages. This page-by-page process used LibreOffice's built-in word

counter (whitespace-based splitting). Researchers recorded tallies on paper copies and summed to obtain transcript totals. **Researchers excluded AI-generated boilerplate sections** (personality tests, teacher profile setup) that appeared before the actual Taylor series tutoring dialogue began, as these pre-instructional exchanges were not part of the mathematical learning interaction being analyzed. Inter-rater agreement was high (within 10 percentage points on %Student Talk), validating both the counting rules and the automated implementation.

- **Boilerplate detection in automation.** The automated pipeline replicated the manual exclusion of boilerplate by scanning the first 100 lines of each transcript for personality test markers ("personality test," "ai profiler," "internal teacher profile," "step 1: personality") and then identifying where Taylor content began using historical markers ("1715," "1685," "Brook Taylor," "early 1700s"). Lines before the content start were excluded from word counts but logged in annotated outputs for transparency. This approach ensured that automated counts matched the manual calibration methodology.
- **Double-checks.** A subset of transcripts in each quartile of % Student Talk was manually reviewed line-by-line to confirm speaker attribution and tokenization.
- **Disagreements.** Ambiguous lines were resolved by consensus; the resulting decisions informed minor clarifications to the rules above.
- **Arithmetic consistency.** We verified Student+AI = Total and %Student+%AI = 100.0±0.1 on every output row. Any failures were flagged for re-evaluation.

## Determinism & Reproducibility

To eliminate run-to-run variability, we executed the counting with a **fixed, deterministic configuration** (temperature = 0; fixed model/version; identical prompts/spec across runs). Reprocessing the same files under the same configuration yielded **identical** counts. A scripted pipeline applied the rules uniformly across all transcripts; implementation details and example input/output are provided in the Supplement (Reproducibility Note).

## Limitations

- **Attribution is rule-based, not diarization.** In noisy or highly edited transcripts, attribution can still require judgment; flagged segments were reviewed manually.
- **Tokenization approximations.** Spoken equivalents for math and the CJK heuristic are principled but approximate; they were chosen for consistency across the corpus rather than linguistic exhaustiveness.
- **Export artifacts.** Rare export quirks (e.g., duplicated prompts, mid-page banners) were filtered when detected.

## Worked Example (brief)

Consider a page with three Student responses interleaved with two AI prompts. After applying the attribution rules, suppose the page contains **Student = 178 words** and **AI = 120 words**. The transcript-level totals sum page counts: $100 \times \frac{178}{178+120} = 59.7\%$.

## Downstream Use (link to Appendix E)

The Phase I outputs (totals and % Student Talk) were used to select **10 high**, **10 low**, and **10 noteworthy** cases for the PK–WAP analysis in Appendix E, which focuses on qualitative interpretation rather than counting.

# Appendix C: AI and Individual Student Talk Metrics

**Table 5**
**AI and Student Talk Analysis for All Transcripts**
*Descriptive statistics for AI and student word counts and talk percentages across transcripts.*

| Student | AI Words | Student Words | Total | % AI | % Student |
|---|---|---|---|---|---|
| P01-G8-S4 | 3558 | 203 | 3761 | 94.6 | 5.4 |
| P01-G8-S4 | 1468 | 80 | 1548 | 94.8 | 5.2 |
| P02-G4-S4 | 1604 | 129 | 1733 | 92.6 | 7.4 |
| P03-GX-SX | 456 | 3019 | 3475 | 13.1 | 86.9 |
| P04-GX-SX | 1279 | 924 | 2203 | 58.1 | 41.9 |
| P05-G9-S4 | 1183 | 104 | 1287 | 91.9 | 8.1 |
| P06-GX-SX | 1985 | 0 | 1985 | 100.0 | 0.0 |
| P07-G5-S4 | 1354 | 0 | 1354 | 100.0 | 0.0 |
| P08-GX-SX | 580 | 1951 | 2531 | 22.9 | 77.1 |
| P09-GX-S5 | 547 | 1101 | 1648 | 33.2 | 66.8 |
| P10-G8-S5 | 1810 | 406 | 2216 | 81.7 | 18.3 |
| P100-G12-S4 | 708 | 270 | 978 | 72.4 | 27.6 |
| P101-G4-S5 | 97 | 122 | 219 | 44.3 | 55.7 |
| P102-G3-S5 | 806 | 78 | 884 | 91.2 | 8.8 |
| P103-G2-S5 | 251 | 1025 | 1276 | 19.7 | 80.3 |
| P104-G1-S5 | 505 | 421 | 926 | 54.5 | 45.5 |
| P105-G10-S5 | 801 | 1450 | 2251 | 35.6 | 64.4 |
| P106-G15-S4 | 2563 | 1162 | 3725 | 68.8 | 31.2 |
| P107-GX-SX | 990 | 2557 | 3547 | 27.9 | 72.1 |
| P108-G14-S4 | 428 | 383 | 811 | 52.8 | 47.2 |
| P109-G14-S4 | 3429 | 212 | 3641 | 94.2 | 5.8 |
| P11-G10-S4 | 1243 | 1010 | 2253 | 55.2 | 44.8 |
| P110-GX-SX | 2088 | 499 | 2587 | 80.7 | 19.3 |
| P111-G6-S5 | 809 | 881 | 1690 | 47.9 | 52.1 |
| P112-GX-S5 | 1170 | 372 | 1542 | 75.9 | 24.1 |
| P113-GX-S5 | 592 | 669 | 1261 | 46.9 | 53.1 |
| P114-G1-S4 | 1448 | 622 | 2070 | 70.0 | 30.0 |
| P115-G11-S4 | 316 | 233 | 549 | 57.6 | 42.4 |
| P116-G9-S5 | 1426 | 936 | 2362 | 60.4 | 39.6 |
| P117-G11-S4 | 1237 | 930 | 2167 | 57.1 | 42.9 |
| P118-G7-S4 | 606 | 638 | 1244 | 48.7 | 51.3 |

| Student | AI Words | Student Words | Total | % AI | % Student |
|---|---|---|---|---|---|
| P119-GX-SX | 1392 | 244 | 1636 | 85.1 | 14.9 |
| P12-GX-SX | 126 | 622 | 748 | 16.8 | 83.2 |
| P120-G16-S5 | 388 | 454 | 842 | 46.1 | 53.9 |
| P121-G11-S4 | 1218 | 850 | 2068 | 58.9 | 41.1 |
| P122-G13-S4 | 767 | 2096 | 2863 | 26.8 | 73.2 |
| P123-GX-SX | 2631 | 474 | 3105 | 84.7 | 15.3 |
| P124-GX-SX | 842 | 1843 | 2685 | 31.4 | 68.6 |
| P125-GX-SX | 434 | 677 | 1111 | 39.1 | 60.9 |
| P126-GX-SX | 932 | 985 | 1917 | 48.6 | 51.4 |
| P127-G3-S4 | 1155 | 1010 | 2165 | 53.3 | 46.7 |
| P13-G4-S5 | 165 | 158 | 323 | 51.1 | 48.9 |
| P14-GX-SX | 1106 | 902 | 2008 | 55.1 | 44.9 |
| P15-G10-SX | 125 | 2252 | 2377 | 5.3 | 94.7 |
| P16-G3-S5 | 279 | 250 | 529 | 52.7 | 47.3 |
| P17-G6-S5 | 297 | 27 | 324 | 91.7 | 8.3 |
| P18-G4-S4 | 513 | 736 | 1249 | 41.1 | 58.9 |
| P19-G3-S4 | 3424 | 556 | 3980 | 86.0 | 14.0 |
| P20-GX-SX | 585 | 503 | 1088 | 53.8 | 46.2 |
| P21-G5-S5 | 323 | 1197 | 1520 | 21.2 | 78.8 |
| P22-GX-SX | 413 | 84 | 497 | 83.1 | 16.9 |
| P23-G10-S5 | 230 | 1760 | 1990 | 11.6 | 88.4 |
| P24-G7-S4 | 875 | 446 | 1321 | 66.2 | 33.8 |
| P25-GX-SX | 1260 | 407 | 1667 | 75.6 | 24.4 |
| P26-G13-S5 | 2548 | 734 | 3282 | 77.6 | 22.4 |
| P27-G5-S4 | 1304 | 0 | 1304 | 100.0 | 0.0 |
| P28-G16-S5 | 647 | 1053 | 1700 | 38.1 | 61.9 |
| P29-GX-SX | 1287 | 808 | 2095 | 61.4 | 38.6 |
| P30-G5-S5 | 102 | 378 | 480 | 21.2 | 78.8 |
| P31-G13-S5 | 0 | 1461 | 1461 | 0.0 | 100.0 |
| P32-G9-S4 | 2334 | 84 | 2418 | 96.5 | 3.5 |
| P33-G16-S3 | 456 | 1215 | 1671 | 27.3 | 72.7 |
| P34-GX-SX | 704 | 2125 | 2829 | 24.9 | 75.1 |
| P35-G1-S4 | 264 | 1526 | 1790 | 14.7 | 85.3 |
| P36-G18-S4 | 1374 | 1302 | 2676 | 51.3 | 48.7 |
| P37-G18-S4 | 1781 | 0 | 1781 | 100.0 | 0.0 |
| P38-G17-S5 | 158 | 525 | 683 | 23.1 | 76.9 |

| Student | AI Words | Student Words | Total | % AI | % Student |
|---|---|---|---|---|---|
| P39-G4-SX | 97 | 122 | 219 | 44.3 | 55.7 |
| P40-G12-S5 | 318 | 230 | 548 | 58.0 | 42.0 |
| P41-GX-SX | 128 | 214 | 342 | 37.4 | 62.6 |
| P42-G11-S3 | 501 | 694 | 1195 | 41.9 | 58.1 |
| P43-G18-S4 | 208 | 231 | 439 | 47.4 | 52.6 |
| P44-G6-S5 | 146 | 660 | 806 | 18.1 | 81.9 |
| P45-G11-S5 | 821 | 710 | 1531 | 53.6 | 46.4 |
| P46-G14-S5 | 768 | 1370 | 2138 | 35.9 | 64.1 |
| P47-G13-S5 | 1565 | 1039 | 2604 | 60.1 | 39.9 |
| P48-G7-S4 | 1109 | 632 | 1741 | 63.7 | 36.3 |
| P49-G10-S5 | 265 | 1207 | 1472 | 18.0 | 82.0 |
| P50-G8-S4 | 139 | 311 | 450 | 30.9 | 69.1 |
| P51-GX-SX | 505 | 519 | 1024 | 49.3 | 50.7 |
| P52-GX-SX | 0 | 810 | 810 | 0.0 | 100.0 |
| P53-G12-S4 | 0 | 177 | 177 | 0.0 | 100.0 |
| P54-GX-SX | 339 | 190 | 529 | 64.1 | 35.9 |
| P55-G13-S4 | 90 | 266 | 356 | 25.3 | 74.7 |
| P56-GX-S5 | 1750 | 0 | 1750 | 100.0 | 0.0 |
| P57-G7-S5 | 451 | 711 | 1162 | 38.8 | 61.2 |
| P58-G14-S5 | 716 | 578 | 1294 | 55.3 | 44.7 |
| P59-GX-SX | 1355 | 1228 | 2583 | 52.5 | 47.5 |
| P60-G3-S5 | 1580 | 580 | 2160 | 73.1 | 26.9 |
| P61-G6-S4 | 578 | 1034 | 1612 | 35.9 | 64.1 |
| P62-G7-S5 | 156 | 1100 | 1256 | 12.4 | 87.6 |
| P63-GX-SX | 183 | 1816 | 1999 | 9.2 | 90.8 |
| P64-GX-SX | 223 | 0 | 223 | 100.0 | 0.0 |
| P65-GX-SX | 22 | 44 | 66 | 33.3 | 66.7 |
| P66-G11-S5 | 1260 | 406 | 1666 | 75.6 | 24.4 |
| P67-G9-S5 | 52 | 84 | 136 | 38.2 | 61.8 |
| P68-G18-S4 | 1525 | 432 | 1957 | 77.9 | 22.1 |
| P69-G2-S5 | 250 | 266 | 516 | 48.4 | 51.6 |
| P70-G13-S4 | 2491 | 2202 | 4693 | 53.1 | 46.9 |
| P71-G18-S5 | 443 | 512 | 955 | 46.4 | 53.6 |
| P72-G12-S4 | 2074 | 467 | 2541 | 81.6 | 18.4 |
| P73-G14-S5 | 2348 | 538 | 2886 | 81.4 | 18.6 |
| P74-G10-S4 | 51 | 69 | 120 | 42.5 | 57.5 |

| Student | AI Words | Student Words | Total | % AI | % Student |
|---|---|---|---|---|---|
| P75-GX-SX | 158 | 354 | 512 | 30.9 | 69.1 |
| P76-GX-SX | 1101 | 3515 | 4616 | 23.9 | 76.1 |
| P77-G8-S5 | 2545 | 98 | 2643 | 96.3 | 3.7 |
| P78-G14-S4 | 2789 | 595 | 3384 | 82.4 | 17.6 |
| P79-G8-S5 | 1419 | 1521 | 2940 | 48.3 | 51.7 |
| P80-G12-S5 | 1430 | 1415 | 2845 | 50.3 | 49.7 |
| P81-G14-S5 | 722 | 392 | 1114 | 64.8 | 35.2 |
| P82-G13-SX | 3024 | 1615 | 4639 | 65.2 | 34.8 |
| P83-G8-S4 | 0 | 887 | 887 | 0.0 | 100.0 |
| P84-GX-SX | 0 | 303 | 303 | 0.0 | 100.0 |
| P85-GX-SX | 1025 | 1801 | 2826 | 36.3 | 63.7 |
| P86-G13-S4 | 287 | 1741 | 2028 | 14.2 | 85.8 |
| P87-G7-SX | 283 | 1354 | 1637 | 17.3 | 82.7 |
| P88-GX-SX | 775 | 494 | 1269 | 61.1 | 38.9 |
| P89-G2-S4 | 1392 | 1211 | 2603 | 53.5 | 46.5 |
| P90-GX-SX | 1380 | 343 | 1723 | 80.1 | 19.9 |
| P91-G6-S5 | 0 | 534 | 534 | 0.0 | 100.0 |
| P92-G9-S5 | 170 | 481 | 651 | 26.1 | 73.9 |
| P93-G1-S4 | 958 | 147 | 1105 | 86.7 | 13.3 |
| P94-G6-S4 | 1286 | 73 | 1359 | 94.6 | 5.4 |
| P95-G3-S5 | 381 | 1121 | 1502 | 25.4 | 74.6 |
| P96-G9-S5 | 0 | 0 | 0 | 100.0 | 0.0 |
| P97-G17-S5 | 436 | 267 | 703 | 62.0 | 38.0 |
| P98-G8-S4 | 854 | 0 | 854 | 100.0 | 0.0 |
| P99-G2-S4 | 346 | 327 | 673 | 51.4 | 48.6 |

*Note.* Cases flagged as probable AI-profiling or demonstration scripts (shaded) were excluded from anchor-case selection but retained in descriptive analyses.

[0]Data unavailable due to OCR limitations. Placeholder row retained for dataset continuity.

## Appendix D: Post-Activity Survey

1. **Name** *(Open-ended)*

2. **Student ID** *(Open-ended)*

3. **Which large language models (LLMs) have you used before?** *(Select all that apply)* ChatGPT (OpenAI)
   DeepSeek
   Claude.ai
   Other (please specify)
   None before this course

4. **How long have you been using LLMs?** *(Single choice)*
   Less than 1 month
   1–6 months
   7–12 months
   Over 1 year
   I have not used

5. **How often do you use LLMs?** *(Single choice)*
   Daily
   Never
   Less than once a week
   Once a week
   Several times per week

6. **For what purposes do you typically use LLMs?** *(Select all that apply)*
   To complete homework or assignments more quickly
   To check or correct my own work
   To help me learn or understand new topics
   To search for information
   For entertainment or curiosity
   Other (please specify)

7. **Have you used LLMs (like DeepSeek) in other university courses?** *(Single choice)*
   Yes, frequently
   Yes, sometimes
   No, only for this course
   No, never

8. **How do your other professors view the use of AI tools like ChatGPT?** *(Single choice)*
   They encourage it
   They allow it, but do not encourage
   They discourage it
   I don't know
   Not applicable

**9. Did you work alone or with others for this assignment?** *(Single choice)*

Alone
With classmates (group discussion)
Other (please specify)

**10. This assignment helped me learn calculus concepts better.** *(Likert scale)*

Strongly disagree
Disagree
Neutral
Agree
Strongly agree

**11. After this assignment, my attitude toward using AI for learning is:** *(Likert scale)*

Much more positive
Somewhat more positive
No change
Somewhat more negative
Much more negative

**12. Please share any other comments about your experience using AI or LLMs for this assignment:** *(Open-ended)*

# Appendix E: Pirie–Kieren Work Analysis Protocol

## I. Overview of the Method

Initially, we reviewed all 127 transcripts to estimate word counts and calculate student talk percentages as a proxy for engagement (see Appendix C). We selected 10 transcripts with the highest student talk percentages, 10 with the lowest percentages, and 10 "noteworthy" cases (as defined in Section 4.3) to form a contrastive sample of 30 anchor cases.

Transcripts with negligible student contribution (i.e., <10%) were excluded and replaced with the next-lowest cases to ensure sufficient material for Pirie-Kieren analysis. Finally, we reviewed the full set of memos to identify recurring themes, interpret contrasts across the high- and low-engagement groups, and generate analytic trends related to recursive reasoning, conceptual depth, and student agency.

**Step 1: Page-by-Page Analysis**
For each page of the transcript:

- Read each page of the transcript individually.
- Make notes of student responses and notable teacher/AI turns for later reference.

**Step 2: Representative Passages** Select at least **3–5 verbatim passages** for each of the following, ideally spread across the transcript:

- *Teacher/AI talk:* jokes, explanations, prompts, calculations, metacognitive comments
- *Student talk:* explanations, calculations, reflections, moments of confusion

Each passage should be annotated to explain its context and why it is notable.

**Step 3: Missed Opportunities**

- For each transcript, identify **3–5 moments** where the teacher/AI could have prompted, scaffolded, or paused for student elaboration, recursion, or explanation—but did not.
- Quote the actual teacher/AI turn, and describe what a more student-centered alternative might have been.

**Step 4: Evidence for Pirie–Kieren Layers** For each of the 8 Pirie–Kieren layers below, search for and quote passages that exemplify student engagement at that level:

1. **Primitive Doing**
2. **Image Making**
3. **Image Having**
4. **Property Noticing**
5. **Formalizing**
6. **Observing**
7. **Structuring**
8. **Inventing**

Annotate each passage. If no evidence is found for a given layer, state so explicitly.

Explicitly code for **recursive/folding back movement**—instances where the student revisits or reconstructs prior reasoning after new insight or challenge.

**Step 5: Synthesis and Discussion**

- *Executive Summary:* 1–2 paragraphs summarizing main findings, pattern of interaction, and evidence (or absence) of student-driven mathematical growth.
- *Table of Word Counts and Engagement:* Paste the table from Step 2.
- *Representative Passages:* List/annotate selected teacher and student examples.
- *Missed Opportunities:* Briefly discuss what was missed and the likely impact on learning.
- *Evidence for Pirie–Kieren Layers:* Present findings layer by layer, quoting and discussing where present, and noting absences.
- *Interpretation:* Discuss what the transcript shows about student agency, recursion, and mathematical understanding.
- *Design and Research Implications:* Offer at least 2–3 recommendations for future prompt/AI design or teaching practice based on your findings.

**Notes for Use**

- Apply the same structure for every analysis, regardless of the richness or length of the transcript.
- If the transcript is too brief, or if student talk is <10%, document this and proceed with the analysis (absence is a finding).
- Be explicit about the **limitations** of each sample.

## II. Determining Anchor Cases

The PK–WAP protocol was applied to a contrastive sample of 30 anchor cases, selected to capture the full range of student engagement and interaction quality. Selection proceeded in two stages:

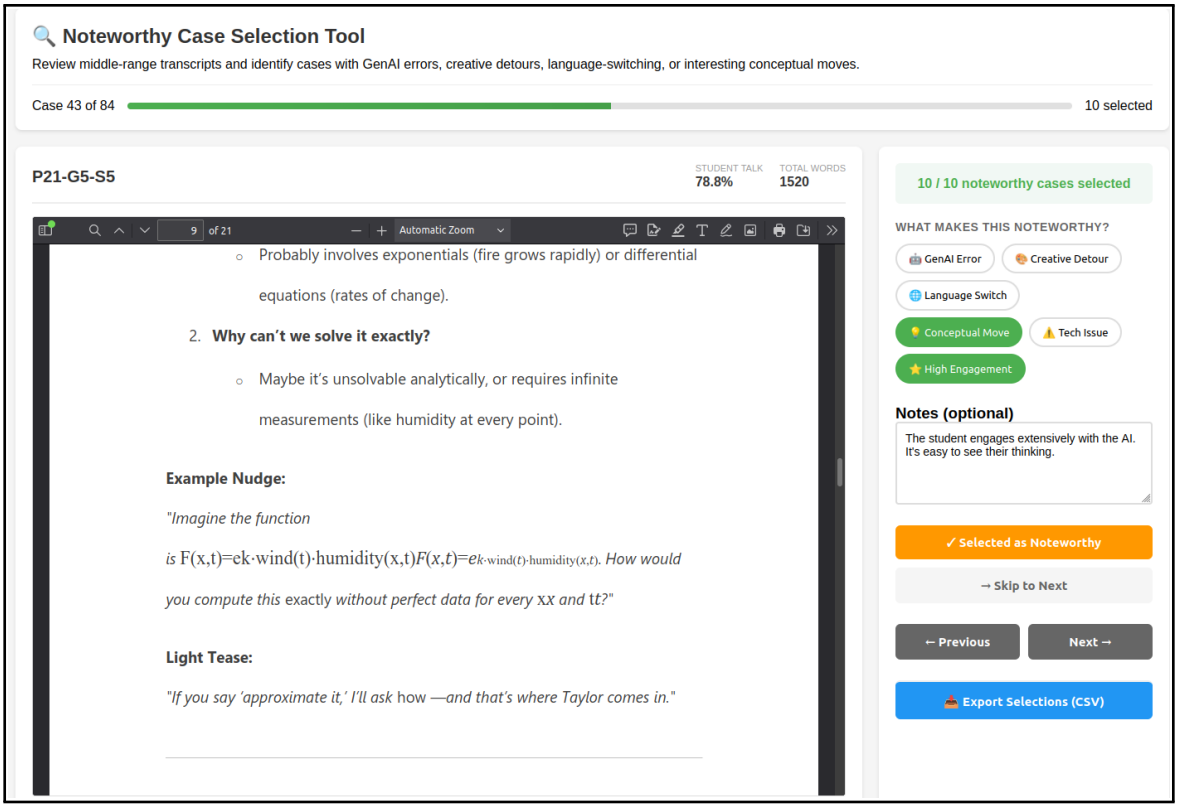
**Quantitative Selection (20 cases):** Based on Phase I word count analysis (Appendix B), we selected:

- **Top 10:** Transcripts with the highest student talk percentages (range: 45–62%)
- **Bottom 10:** Transcripts with the lowest student talk percentages while maintaining sufficient material for analysis (range: 10–18%)

Transcripts with negligible student contribution (<10%) were excluded and replaced with the next-lowest cases to ensure sufficient material for Pirie-Kieren analysis.

**Qualitative Selection (10 noteworthy cases):** From the remaining 84 middle-range transcripts (18–45% student talk), we identified cases exhibiting distinctive features warranting deep analysis. To systematically review this subset, we developed a web-based Flask application (shown below) that displayed original student submissions alongside categorical tags. This tool enabled rapid, consistent annotation across the following categories:

- **GenAI Error:** AI misinterpretation or factual mistake with pedagogical consequence
- **Creative Detour:** Student-initiated exploration beyond the assigned task
- **Language Switch:** Code-switching or translation challenges affecting dialogue
- **High Engagement:** Exceptional depth despite moderate word count
- **Metacognitive Moves:** Explicit reflection on learning process or AI interaction
- **Conceptual Breakthrough:** Clear evidence of PK layer advancement mid-dialogue




The final 10 noteworthy cases were selected based on tag frequency, diversity of features, and potential to illuminate design implications. This two-stage selection strategy ensured the anchor sample included both quantitative extremes (high/low engagement) and qualitatively distinctive patterns not captured by word counts alone.

## III. Generating Analytic Memos with GenAI

**Purpose:**
This subsection documents how the PK–WAP process was operationalized using generative AI (genAI) to produce structured analytic memos for each of the 30 anchor cases, ensuring consistency while allowing for nuanced coding.

**Required Inputs:**

- Transcript in clean `.txt` format.
- Reference materials: Pirie & Kieren (1994) framework, Boaler (1998) examples, and at least one "gold standard" memo.
- Project context with shared definitions and formatting conventions.

**Execution Steps:**

1. **Preparation:** Upload transcript, ensure all references are available, and have gold standard memos accessible.
2. **Prompting:** Instruct the model to read the entire transcript and produce, in order: word counts by page, folding-back moments, PK layer coding, representative quotes, missed opportunities, a memo summary, page-by-page coding table, and a layer progression map—matching the tone and formatting of the gold standard memo.
3. **Output:** Generate in Markdown format with exact section headings preserved.

**Quality Check:**

- Verify alignment with Phase I word counts.
- Confirm evidence for folding-back and layer ceilings is quote-supported.
- Ensure progression maps match coding tables.

**Scaling for Batch Processing:**
The process can be applied manually or via the OpenAI API with scripted prompts. Batch outputs are stored as `.md` files and logged in a `.csv` with metadata. Manual review is recommended for a subset to ensure quality control.

**Note on Context Retention:** Analyses performed within the same project environment (with prior PK–WAP examples and definitions) yield the most consistent results.

**Replication Disclaimer:** While this protocol is designed for consistency, PK–WAP coding remains a qualitative process involving researcher judgment. Regular calibration meetings, shared review of anchor cases, and documented rationales are recommended.

## IV. Example Analytic Memo

Below is an excerpt from the PK–WAP analytic memo for case P28–G16–S5, illustrating the depth and structure of analysis applied to each anchor case. The full memo includes all sections outlined in the protocol; selected portions are presented here to demonstrate the coding process and interpretive approach.

**Case ID: P28–G16–S5**
**Method:** Pirie–Kieren Work Analysis Protocol (PK–WAP)
**Focus:** Student–AI dialogue in which the student specifies a personality-based AI-teacher workflow, then collaborates on a Taylor-series modeling task (medical CT reconstruction). The transcript opens with a long, student-authored activity script (Pages 00–01), followed by AI-led scaffolding with intermittent, terse student replies. Analysis below treats the initial prompt text (student-authored instructions) as **student words** and the AI's subsequent facilitation as **AI words**. Single student turns that answer AI prompts are treated as student turns.

**Word counts & % student talk by page (estimated)**

| Page | Student Words | AI Words | % Student Talk |
|---|---|---|---|
| 00 | 546 | 0 | 100.0 |
| 01 | 364 | 0 | 100.0 |
| 02 | 2 | 261 | 0.8 |
| 03 | 0 | 241 | 0.0 |
| 04 | 12 | 226 | 5.0 |
| 05 | 39 | 206 | 15.9 |
| 06 | 13 | 249 | 5.0 |
| 07 | 4 | 221 | 1.8 |
| 08 | 56 | 170 | 24.8 |
| 09 | 12 | 217 | 5.2 |
| 10 | 2 | 202 | 1.0 |
| 11 | 0 | 72 | 0.0 |
| **Total** | **1,050** | **2,065** | **33.7** |

**Pattern:** Student-dominant **setup** (00–01); AI-dominant facilitation **after 02**, with brief but consequential student entries (e.g., domain choice, constraint identification, value trade-offs).

**Recursive / folding-back moments (narrative)**

- **FB-1 (Page 04): Protocol repair.** Student interrupts—"you did not ask me questions one by one in step 1." This is *Observing* the interactional structure and folding back to enforce the designed constraints. The repair establishes turn granularity, improving conditions for subsequent PK movement.

- **FB-2 (Pages 06–07): From scenario to property.** After selecting **CT imaging** (Primitive Doing/Image-Making), the student compresses the core instability to "**sensitivity to measurement errors**" (Property-Noticing). This is a concise leap from context to governing mathematical property (ill-posedness).
- **FB-3 (Page 09): Cognitive load check → reframing.** Student signals overload ("too professional…"). AI folds back to **Image-Making** via a sketch analogy, regrounding abstractions. The move restores traction without supplying answers, preserving agency.
- **FB-4 (Page 10): Prioritization lens.** Student commits to "**tumor's shape**" as the diagnostic eyebrow (must-preserve). This selects a salient invariant (Property-Noticing → Image-Having), stabilizing future trade-offs (e.g., filter strength, polynomial "degree").
- **FB-5 (Pages 10–11): Task-contingent rules.** Dialogue generalizes toward **criteria that vary by purpose** (emergency vs. early detection), an *Observing* → *Axiomatizing* step: not formal proof, but proto-principles that justify differing approximation cutoffs.

This example illustrates how the PK–WAP protocol operationalizes recursive learning analysis, identifying not only the highest layer reached (Inventising) but also the folding-back movements that enabled conceptual growth. The full memo (available in the project repository) includes complete layer-by-layer coding, representative quotes, missed opportunities, and design implications.

# Appendix F: Analytic Memo Generation Protocol

Note: Instructions in parentheses are for the AI to follow and should never appear in the final memo.

## Analytic Memo Template

**1. Introduction / Context**
(State the case ID, PK–WAP analysis, and short description of the mathematical scenario. Include explicit attribution rules: unlabeled student turns that answer AI prompts are treated as student turns.)

**2. Word Counts / Percent Talk by Page**
ALWAYS use this column order `Page | Student Words | AI Words | % Student Talk`

- No extra columns unless explicitly requested.
- Column names must be exactly as shown.
- Percentages must be numeric (no % symbol).
- Provide a total % at the end.

| Page | Student Words | AI Words | % Student Talk |
|---|---|---|---|
| 1 | ### | ### | ### |
| 2 | ### | ### | ### |
| 3 | ### | ### | ### |
| 4 | ### | ### | ### |
| n | ### | ### | ### |

**Overall student talk:** ### words (###).

**3. Layer Progression Map**
(Use ASCII or a clear diagram to show PK layer sequence. Mark fold-backs with curved arrows and page refs. The style must be consistent across memos.)

**4. Recursive / folding-back moments (narrative)**
(Describe each major folding-back episode in paragraph form, with page refs, layer transitions, and how the student reconstructs understanding. Maintain consistent depth across memos.)

**5. PK layer coding (evidence-rich)**
(Table listing all 8 PK layers: Layer, Evidence from transcript, Notes on classification. Always include all 8 layers in order.)

| Layer | Representative Evidence | Notes |
|---|---|---|
| 1. Primitive Doing | … | … |
| 2. Image-Making | … | … |
| 3. Image-Having | … | … |
| 4. Property-Noticing | … | … |
| 5. Formalizing | … | … |
| 6. Observing | … | … |
| 7. Axiomatizing | … | … |
| 8. Inventising | … | … |

**6. Page-by-page PK–WAP coding**
(Table for every page: Page, Dominant layer(s), Representative evidence, Notes. Must include all pages.)

| Page | Dominant layer(s) | Representative Evidence | Notes |
|---|---|---|---|
| 1 | ### | … | … |
| 2 | ### | … | … |
| 3 | ### | … | … |
| 4 | ### | … | … |
| n | ### | … | … |

**7. Representative Quotes**
(Subdivide into Student and AI. Include at least 4–6 quotes per side, with page numbers and conceptual significance. Format must remain consistent across memos.)

**8. Missed opportunities (elaborated)**
(List 3–5. Each must have 1–2 sentences explaining what was missed, where, and how it could have deepened learning.)

**9. Summary of Findings**
(1–2 paragraphs synthesizing engagement level, PK layer movement, tone, and key growth moments.)

**10. Final Observations**
(One paragraph tying together PK movement, agency, tone, and possible improvements.)

**11. Conclusion**
(Short paragraph summing up why the case matters, with specific PK trajectory notation and final pedagogical implications.)

## Prompt for GenAI to Generate Analytic Memo

```
I'm researching student-AI mathematical dialogue. Please analyze the attached
transcript (P28-G16-S5.txt) using the Pirie-Kieren Work Analysis Protocol
(PK-WAP) and generate a Deep Research-style memo that follows exactly the
structure, headings, numbering, and formatting rules in the attached guiding
document (P00-G00-S0 PK-WAP TEMPLATE.md).

The template rules are non-negotiable:
    • Section order, headings, and numbering must match exactly.
    • Word Count table must follow the template's required column names and
      order: Page | Student Words | AI Words |
    • Include all analytical components listed in the template at full depth.
    • Analytical content must be specific to the transcript (no
      generic/template-sounding text).
Follow these steps:
  1. Estimate word counts and identify recursive/folding-back moments with
     narrative detail.
  2. Code for all 8 Pirie-Kieren layers (Primitive Doing → Inventising).
  3. Highlight representative quotes from both student and AI (4-6 per side).
  4. Assess missed opportunities for AI to support deeper learning (1-2
     sentences per item).
  5. Provide a summary that synthesizes growth, agency, and tone.
Please take your time—it's okay if this takes several minutes to complete. The
goal is a pedagogically insightful, deeply interpretive analysis in the exact
format of the template.

Attachments:
  1. Transcript to analyze.
  2. Deep Research Template (P00-G00-S0 PK-WAP TEMPLATE.md).
  3. Two research articles:
        • Pirie-Kieren framework description (layer definitions).
        • Boaler's examples of levels from student work.
```

**Note on Implementation:** The above protocol was implemented in Python using the OpenAI API (`pkwap_analyzer.py`). The script accepts individual transcripts or batch processing of multiple files, enforces template compliance, and logs all API interactions for reproducibility. Code and documentation are available in the project repository (see Section 5.6).