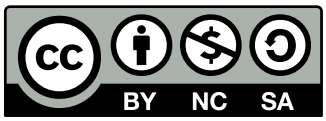
Aalto University
School of Science

Master's programme in Computer, Communication, and Information Sciences

Fluid Interfaces and Fixed Patterns: Understanding LLM Behavior in Educational Contexts

Aayush Kucheria

Master's Thesis
2024



Aalto University
School of Science

Author Aayush Kucheria

Title Fluid Interfaces and Fixed Patterns: Understanding LLM Behavior in Educational Contexts

Degree programme Computer, Communication, and Information Sciences

Major Data Science, Machine Learning, and Artificial Intelligence (MACADEMIA)

Supervisor Prof. Nitin Sawhney

Advisor Prof. Tomi Kauppinen

Date 31 Dec, 2024 **Number of pages** 57 **Language** English

Abstract

As Large Language Models (LLMs) emerge as potential tutoring agents, they promise more fluid, adaptive educational interactions than traditional intelligent tutoring systems. However, the extent to which LLM behavior actually aligns with human tutoring patterns remains poorly understood. This thesis examines this tension between fluid interfaces and fixed behavioral patterns in ITS. Drawing on constructivist learning theory and analysis of historical constraints in educational technology, we investigate how LLMs process and respond in the tutoring task compared to human teachers. Through systematic analysis of the CIMA dataset, we compare action distributions and response patterns between human tutors and three state-of-the-art LLMs (GPT-4o, Gemini Pro 1.5, and LLaMA 3.1 405B) in language teaching dialogues. Rather than evaluating performance or effectiveness, we focus on understanding fundamental differences in how artificial and human tutors structure their teaching interactions. Our results reveal systematic deviations in LLM behavior from human tutoring patterns, particularly in action selection and response adaptation to student behavior. These findings suggest that while LLMs enable more fluid interaction, they may develop fixed behavioral patterns distinct from human teaching strategies. This research contributes to both theoretical understanding of ITS behavior and practical development of more effective educational technologies, while raising important questions about the nature of machine teaching and learning.

Keywords Large Language Models, Tutoring Behavior, Educational Technology, Human-AI Interaction, Action Pattern Analysis, Intelligent Tutoring Systems

Preface

This thesis was a result of my curiosities around how we learn and grow as humans and how technology might support—or hinder—that process. The intellectual giants that I strive to stand on the shoulders of include Bret Victor, Alan Kay, Seymour Papert, J.C.R. Licklider, and Douglas Engelbart, among other pioneers who dreamed of augmenting humanity through technology.

Along with my interest in investigating research questions about intelligent tutoring and human learning in the age of LLMs, this thesis was also a way for me to learn and reflect on my own interests, values, and ways of sensemaking. To engage with both conceptual and empirical research in an interdiscliplinary context. I am grateful to my supervisor, Nitin Sawhney, for his steady guidance and patience throughout this difficult path. I also appreciate my advisor Tomi Kauppinen, Kaisla Kajava, and Fedor Vitiugin, who were together on this long journey with me. Special thanks to Sofi, who helped me reframe my relationship with work and opened me up to more expansive ways of engaging with it.

Thesis writing is difficult, as I found out during the process. There were two books that were influential in helping me sense-make this craft: "How to write a Thesis" by Umberto Eco, and "Writing your Dissertation in Fifteen Minutes a Day" by Joan Bolker.

While this work presents some interesting findings about how AI systems function in tutoring, its deeper value to me has been a foundation for future exploration. It represents a personal milestone in how I think, work, and learn.

Lauttasaari, 31 Dec 2024

Aayush Kucheria

Contents

Abbreviations

AI	Artificial Intelligence
CAI	Computer Assisted Instruction
CIMA	Conversational Instruction with Multi-responses and Actions (dataset)
ITS	Intelligent Tutoring Systems
LLM	Large Language Model
RLHF	Reinforcement Learning with Human Feedback

1 Introduction

1.1 The Promise and Challenges of Personalized Learning

The field of education often talks about personalisation, but so does Science Fiction. Neal Stephenson, in his book "The Diamond Age" [1] introduces an AI-powered interactive book that personalizes to the reader. Orson Scott Card's "Ender's Game" [2] features a personal AI tutor named Jane. Even Isaac Asimov, one of the "Big Three" science fiction writers, envisions in "The Fun They Had" [3] a future school with personalized instruction and robot teachers. These visions capture something fundamental about our aspirations for education: the idea that learning could be perfectly tailored to each individual.

When we move from science fiction to educational research, the study that often gets cited is the Bloom 2-Sigma Effect [4]. In 1984, Benjamin Bloom conducted a study that seemed to validate these aspirations for personalized learning. Students who received one-on-one tutoring combined with mastery learning performed two standard deviations better than those in traditional classrooms. To put this in context, a 2 standard deviation improvement shifts a student from the middle of the class to the top 2% - a dramatic change that suggests the transformative potential of personalized instruction.

But Bloom's dramatic results raise some interesting questions when we examine later research. Many researchers have tried to replicate these findings, but their effects have typically been around 0.8 sigma, still impressive, but notably different from Bloom's 2 sigma effect [5]. VanLehn's 2011 meta-analysis [6], examining studies from 1975 to 2010, found that human tutoring improved test scores by an effect size of 0.79 compared to no tutoring. Similarly, a 2014 meta-analysis by Ma and colleagues [7], looking at 107 effect sizes across 73 studies, found meaningful but more modest improvements.

What explains this gap between Bloom's dramatic results and later findings? Was there something unique about Bloom's implementation of tutoring and mastery learning? Or does this point to a deeper challenge in how we measure and replicate educational interventions? Or was Bloom's effect an anomaly of its experimental details rather than a statement about human learning? These remain open questions.

Regardless of this problem of replicability of the strong results Bloom generated, there is still a strong case for optimism. Even if we can't hit the two sigma mark, the evidence points to a robust trend showing that personalized instruction, mastery-based learning, and feedback-oriented approaches lead to significant improvements in student performance. This is evidenced by a variety of intelligent tutoring systems [8, 9] that are achieving positive results, along with meta-analyses [6, 7] showing significant effects when comparing tutoring with more traditional learning environments. Importantly, the trend is also well-supported by theories of learning and cognitive development, like constructivism [10] that making learning more individualized, feedback-oriented, and mastery-based does lead to noticeable and significant improvements in student performance. For example, AutoTutor [8] is a popular Intelligent Tutoring System (ITS) that has been producing learning gains about 0.8 standard deviations (SD) above

controls who read static materials for the same amount of time. In simpler terms, it boosts students at the 50% percentile to the 79% percentile. Another example is Salman Khan, of Khan Academy, who in his book "Brave New Words" [11] writes about his aspiration of creating a tutor for every learner in the world.

This pattern, where later studies show significant but smaller effects, points to something important about the challenge of personalized learning. It's not just about whether personalization works (it clearly does) [6], but about understanding the complex factors that make it effective. A theoretical framework that helps us understand this complexity comes from developmental psychology, particularly the constructivist perspective championed by Jean Piaget [10] and later expanded by Seymour Papert [12] among others. Constructivism approaches learning as creating environments where learners can build their own understanding and models of the world [13], rather than as pouring knowledge into empty vessels [14].

This constructivist view aligns naturally with the field of Intelligent Tutoring Systems (ITS), which takes this philosophical foundation and attempts to build systems that can customize learning based on individual needs. These systems are already making an impact: Cognitive Tutor [9], for example, has grown from a small 1993 study to reach over half a million students across 3000 schools [15].

Yet despite these successes, personalized education hasn't revolutionized learning in the way early advocates predicted [16]. The core challenge isn't theoretical, we have strong evidence that personalization works, but practical. Current approaches to user modeling, crucial for personalized learning, don't scale well [17]. The same applies to generating datasets or employing experts to train these systems. This creates a crucial tension: we know personalized learning can be transformative, but our traditional approaches to implementing it face serious scalability limitations.

This scalability challenge becomes particularly interesting when we consider one of the grand challenges set by the field of AI in Education (AIED): "mentors for every learner" [17]. It's a vision that echoes those science fiction stories we started with, but now with the technical capabilities to potentially make it real. Recent advances in AI, particularly in Large Language Models (LLMs), suggest new approaches to this long-standing challenge of scaling personalized education.

1.2 Why Now Could Be Different

What is particularly interesting about our current moment in educational technology is how it differs from previous attempts to solve the scalability problem. Earlier approaches to automated tutoring faced an important limitation: they could only work with explicitly programmed knowledge and behaviors [18, 19]. This created a "scaling paradox": The more we tried to make systems adaptable to individual learners, the more explicit programming they required, making them increasingly difficult to scale [20].

Large Language Models (LLMs) [21] take a fundamentally different approach. Unlike traditional systems that rely on carefully engineered rules and structures, LLMs are trained to exhibit capabilities through exposure [22] to vast amounts of training data. For example, this architectural difference leads to what we might call "emergent

adaptation" [21, 23]: the ability to calibrate responses based on subtle cues in how learners express themselves.

Consider what happens when a physics concept comes up in conversation with an LLM. The system can generate adjusted responses in its language and explanations based on whether it is speaking with a graduate student or a middle school learner [24]. This adjustment happens not through predefined categories or levels, but through the model's ability to recognize and respond to the distribution of words, ideas, and ways of framing concepts that the learner uses [25]. It's a kind of personalization that emerges naturally from the interaction rather than being explicitly programmed.

What makes this adaptation particularly powerful is that it emerges from the inherently conversational nature of these systems [24]. Traditional tutoring systems needed separate modules for understanding student input, deciding how to respond, and generating appropriate feedback [26]. LLMs integrate all these capabilities into a single system that can maintain context, adjust in real-time to new information, and improvise responses based on the unique path each conversation takes [27–29].

This capability for fluid, context-sensitive interaction [30, 31] addresses one of the core challenges in scaling personalized education: the ability to maintain natural, pedagogically sound conversations without requiring explicit programming for every possible interaction. When a student asks about the relationship between acceleration and velocity, for instance, an LLM can generate explanations that connect to the student's current understanding, draw relevant analogies, and adjust its approach based on the student's responses - all without needing these variations to be pre-programmed.

However, this potential comes with important caveats. The same flexibility that allows LLMs to adapt to individual learners also raises questions about consistency and pedagogical soundness [32, 33]. Just because a system can generate plausible-sounding explanations doesn't necessarily mean those explanations support effective learning. This tension between flexibility and pedagogical structure represents one of the key challenges in leveraging LLMs for education.

Moreover, while LLMs show impressive capabilities for natural interaction, we're still in the early stages of understanding how to harness these capabilities for effective teaching [24, 34]. We know LLMs can engage in educational conversations, but whether they can do so in ways that consistently support learning objectives while adapting to individual needs is still being worked out.

This brings us to the focus of this research: understanding how we might bridge the gap between LLMs' potential for natural, adaptive interaction and the structured, pedagogically sound approach needed for effective tutoring. By examining how LLMs can simulate expert tutor behavior while maintaining pedagogical effectiveness, we aim to explore whether these systems might finally offer a path to scaling personalized education without sacrificing quality.

1.3 Thesis Statement and Objectives

As we touched on in the previous section, the emergence of Large Language Models represents a pivotal moment in educational technology, suggesting new possibilities for scaling personalized learning through more natural, adaptive interactions. However, as

we've discussed, improved technical capability for fluid interaction doesn't necessarily translate to more effective or human-like teaching behavior. This tension between technological capability and pedagogical alignment raises crucial questions about integrating artificial teaching behavior into human-centered systems.

This investigation examines the interplay between fluid interaction capabilities [31] and fixed behavioral patterns in how LLMs approach the tutoring task. While these systems enable more natural educational interactions than previous approaches [24], we hypothesize that they exhibit characteristic patterns distinct from human teaching strategies. Our central thesis is that even as LLMs transcend previous technical limitations in educational technology, they may introduce new patterns of interaction that neither mirror human teaching nor simply extend traditional computer-aided instruction.

Through systematic analysis of tutoring interactions, we aim to understand whether the ability to engage in fluid conversation necessarily leads to human-like teaching behavior. This question becomes particularly crucial as we consider the deployment of these systems at scale. If LLMs exhibit their own characteristic teaching patterns, we need to understand these patterns to effectively harness their capabilities while ensuring pedagogical soundness.

Our research objectives are threefold:

1. To identify and characterize systematic differences between how language models and human tutors structure their teaching interactions, focusing on patterns of action selection and response adaptation rather than performance metrics. This objective moves beyond simple effectiveness measures to understand fundamental differences in how artificial and human tutors approach the teaching task.

2. To examine whether different language models, despite their shared capability for fluid interaction, develop distinct "tutoring patterns": consistent behavior that differs both from human tutors and from each other. This investigation helps us understand whether observed patterns represent general characteristics of LLMs in educational contexts or specific artifacts of particular models and training approaches.

3. To investigate how these artificial tutoring patterns emerge and persist across different student behaviors and learning contexts, providing insight into whether they represent fundamental characteristics of how current language models function in the teaching task. This includes examining how LLMs adapt their teaching strategies in response to different student behaviors and learning situations.

These objectives reflect a deliberate focus on understanding behavioral patterns rather than evaluating performance or effectiveness. While much research examines how well LLMs perform as tutors [34, 35], we focus instead on understanding fundamental differences in how they approach the teaching task. This emphasis allows

us to explore questions about the nature of artificial teaching behavior without making assumptions about what constitutes "correct" or "effective" tutoring.

The significance of this research extends beyond immediate practical applications. By examining how systems capable of fluid interaction develop fixed behavioral patterns, we contribute to broader discussions about the nature of intelligent tutoring and learning using LLMs. These insights have implications for both theoretical understanding of ITS behavior and practical development of educational technologies that can effectively scale personalized learning.

This investigation builds on recent work examining LLMs' capabilities for natural interaction [24, 34], while addressing a crucial gap in our understanding: how closely can artificial tutoring behavior align with human teaching patterns, and what systematic differences emerge even when technical limitations on fluid interaction are removed? As we will explore in the following sections, this question connects to fundamental theories about the nature of learning, teaching, and human-computer interaction.

2 The Evolution of Educational Technology

2.1 Intelligent Tutoring Systems: A Historical Overview

When we examine the history of intelligent tutoring systems, a key question emerges: how have our attempts to automate tutoring evolved alongside our understanding of what makes teaching effective? Importantly, this question brings to light an understanding of how different technical capabilities shaped our assumptions about what automated teaching could look like.

The story begins in the 1970s, when researchers started exploring how computers might help with education through what they called Computer Assisted Instruction (CAI). Carbonell [36] saw that by incorporating artificial intelligence techniques into CAI, we could create systems that did more than just present information - they could adapt to individual students. This marked one of the first shifts from seeing computers as dead delivery mechanisms to imagining them as interactive, alive, teaching assistants.

The field of ITS emerged at an interesting intersection of three different intellectual traditions: computer science, cognitive psychology, and educational research [26]. Computer scientists brought their understanding of how to structure and process knowledge. Cognitive psychologists contributed insights about how people learn and build mental models. Education researchers brought their practical experience with what makes teaching effective in real classrooms.

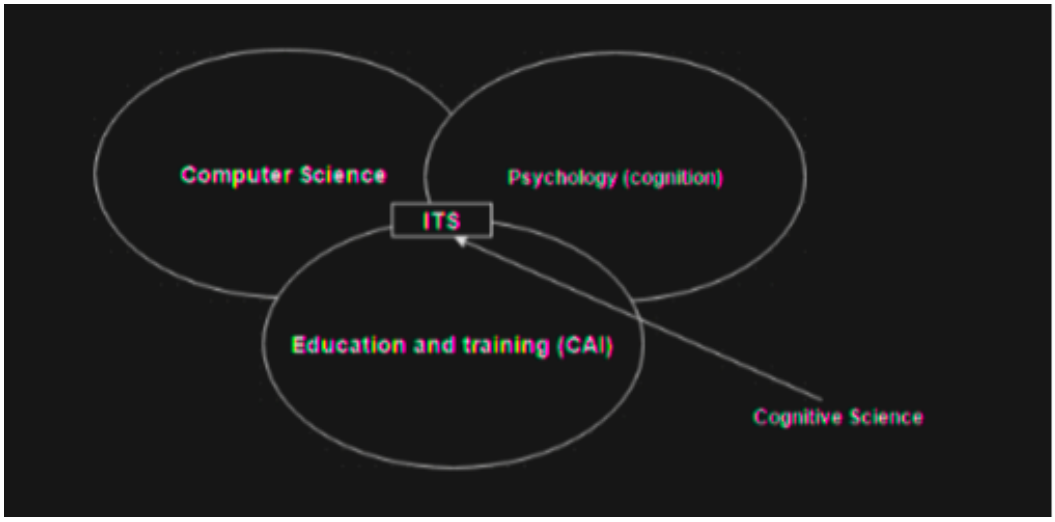


Figure 1: The interdisciplinary origins of Intelligent Tutoring Systems (ITS), adapted from [26].

Self [37] captured this complexity well when they argued that we should view ITS as an engineering design field. It meant approaching the challenge of automated teaching as something that required both technical sophistication and deep understanding of human learning. Twenty years after Self made this argument, the field had developed its own identity, combining insights from all three of its parent disciplines while

developing new approaches specific to the challenges of intelligent tutoring [18].

The early systems approached this challenge in a way that made sense given the technology of the time - they tried to break down teaching into explicit rules and structures [18]. Take the overlay model [18], for example. It worked by comparing student knowledge directly against expert knowledge, treating learning as a process of filling in missing pieces. This seems reasonable at first, after all, isn't learning about acquiring expert knowledge? But this approach built-in some assumptions about learning that were not necessarily true.

Subsequently developed systems tried to get more sophisticated by adding "bug libraries" - collections of common student mistakes and misconceptions [38]. This was a step forward in recognizing that learning is more than simply acquiring correct knowledge, but it still worked within the same basic framework of explicitly programmed states and responses.

These technical approaches created particular patterns in how systems could interact with students. Most early platforms could only evaluate final answers [39, 40], not engage with the problem-solving process itself. This wasn't because process-level engagement wasn't important, any teacher will tell you it's crucial, but because the technical architecture made it impractical to implement.

Koedinger [15] recounts his early struggles trying to integrate ANGLE, a geometry proof tutor, into existing textbook curricula. These practical challenges pushed developers to think about ITS not as standalone tools but as components of a complete learning environment. The question wasn't just "How do we make the system work?" but "How do we make it work within the complex reality of actual classrooms?"

Following an adjacent historical thread, the development of systems like Andes Physics Tutor [41] marked an important shift. Instead of just checking answers, Andes could work with students step-by-step through problems. This might seem like a small change, but it represented a different way of thinking about how computers could support learning. The focus shifted from evaluating final answers to engaging with the problem-solving process itself: from results-centered to process-centered.

AutoTutor [42] pushed this evolution further by introducing natural language dialogue and animated agents. The results were positive: students using AutoTutor typically improved by nearly a letter grade [8]. They made the interaction feel more natural, but more importantly, showed that breaking free from rigid interaction patterns could actually lead to better learning outcomes.

This push toward more natural interaction took many forms. Researchers explored free-form explanation feedback [43], collaborative peer tutoring [44], and systems that could support student metacognition [45, 46]. What's particularly interesting is that many of these innovations weren't limited by their pedagogical value, they showed genuine promise in research settings, but by the technical infrastructure available to implement them [15]. Today's language models make many of these design interventions more technically feasible than ever before. But this improved technical capability raises a question: just because we can make these interactions more fluid and natural, should we always do so?

This question connects to an insight from early ITS developers. Koedinger and his colleagues [15] viewed their systems "not as a substitute for teachers but as part

of the toolkit for teachers that would enhance their ability to do the things they can uniquely do." This perspective helps us think more carefully about which aspects of tutoring we should make more fluid and which structures or constraints serve valuable pedagogical purposes.

This foundational architecture of ITS (Figure 2) [26] reflects these multidisciplinary influences in its core components. The Expert Knowledge Module contains the domain expertise that the system aims to teach - the structured representation of knowledge that computer scientists excel at creating. The Student Model Module, informed by cognitive psychology, tracks and analyzes how learners understand and misunderstand the material. The Tutoring Module, drawing from educational research, implements teaching strategies to bridge the gap between expert knowledge and student understanding. All of this is mediated through a User Interface Module that facilitates interaction with the student.

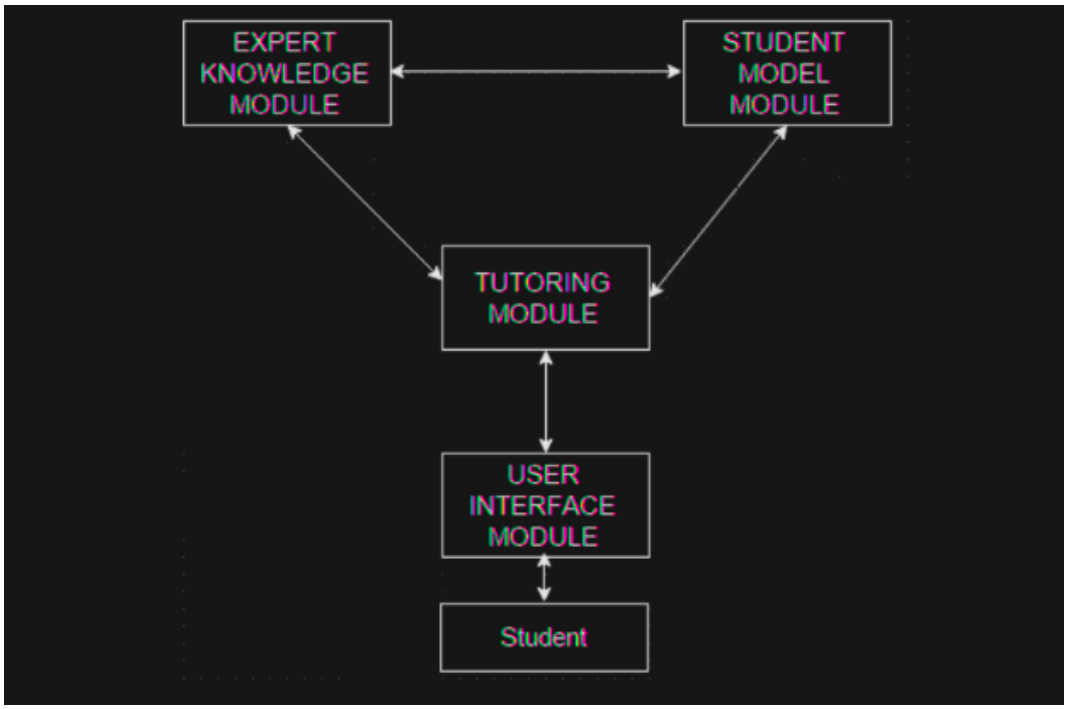


Figure 2: Core components of a traditional Intelligent Tutoring System (ITS) architecture, showing the interaction between Expert Knowledge, Student Model, and Tutoring modules, mediated through the User Interface to support student learning, adapted from [26]

Consider Koedinger's open question [15]: "How much of the more open-ended problem solving and reasoning that we want students to learn can intelligent tutoring systems support?" It brings to light questions about understanding which parts of the learning process benefit from structure and which parts need more fluid, open-ended exploration. Some constraints in traditional ITS existed purely because of technical limitations [47] - the inability to process natural language responses, the need for pre-programmed problem-solving paths. But others, like the structured decomposition

of complex problems or the careful sequencing of concepts, might serve valuable pedagogical purposes.

As the field matured, researchers began systematically reviewing and categorizing developments in ITS. Nwana's 1990 survey [26] provided one of the first comprehensive frameworks for understanding ITS components. This was followed by deeper analyses from Shute [48] and others [18].

What we want to highlight is how the field's understanding of effective tutoring evolved. Early systems often assumed a transmission model of learning [44], where the computer's role was to transfer knowledge to the student. But research into human tutoring showed something different: good tutors actually let students do as much work as possible, providing guidance to keep them on track rather than direct instruction [49]. This insight began influencing ITS design, leading to systems that focused more on supporting student exploration and self-directed learning.

The shift from rule-based to neural approaches [50] in ITS also suggests something interesting about how we think about artificial intelligence in education. Instead of trying to program every possible interaction and knowledge state explicitly, newer systems can develop their understanding through training. But this capability for more fluid, adaptive interaction doesn't mean we should eliminate all structure. The challenge becomes identifying which constraints were truly just technical limitations and which ones scaffold meaningful learning experiences.

Looking at these historical systems helps us understand which constraints were actually necessary for effective teaching, and which were just limitations of the technology we had at the time. Modern approaches, especially those using large language models, are showing us that many of these historical constraints weren't pedagogically necessary at all. But they're also forcing us to think more carefully about the role of structure in learning. We're no longer bound by the same technical limitations as early ITS developers, but their fundamental insights about supporting rather than replacing human teaching remain relevant. The field continues to evolve, with new technologies enabling more natural, adaptive, and student-centered approaches to automated tutoring.

2.2 The QWERTY Phenomenon in Educational Technology

The persistence of technically-driven constraints in tutoring systems reflects what Papert [12] calls the "QWERTY phenomenon": where historical technical limitations become calcified into standard practice, persisting long after the original technical constraints have been removed. Just as the QWERTY keyboard layout was designed to prevent mechanical typewriter keys from jamming but persists in our digital age, many structures in educational technology reflect past technical limitations rather than optimal learning approaches.

This phenomenon manifests particularly strongly in intelligent tutoring systems, where early technical constraints have shaped not just implementation, but our very conception of how computer-aided learning should work. Consider three levels where this calcification appears:

2.2.1 Interface Level

The most obvious manifestation is in how we expect learners to interact with tutoring systems. Early systems could only handle limited, structured inputs such as multiple choice questions, fill-in-the-blank responses, or highly constrained natural language [8]. While modern language models can engage in much more natural dialogue [51, 52], many systems still default to these restricted interaction patterns. We have institutionalized these limitations to the point where they feel "normal" both to designers and learners.

2.2.2 Pedagogical Level

More subtly, technical constraints have shaped our pedagogical assumptions. The need to pre-program explicit rules and responses led to highly structured, sequential approaches to teaching. This mirrors the "computer-as-tool" paradigm [12] we discussed earlier where the computer's role is to deliver pre-structured content rather than engage in genuine dialogue. Even as technology has evolved to enable more dynamic, adaptive approaches, this structured paradigm often persists.

2.2.3 Conceptual Level

Perhaps most profoundly, these technical limitations have influenced how we conceptualize the basic nature of tutoring itself. The need to make tutoring "computer-readable" has pushed us toward models of teaching and learning that emphasize explicit rules, clear sequences, and unambiguous outcomes. This represents a kind of reverse adaptation where human teaching concepts have been simplified and structured to match what early computers could process.

2.3 Limitations of Traditional Approaches

The recognition of these QWERTY-like phenomena in tutoring systems suggests a question: How do we distinguish between constraints that serve genuine pedagogical purposes and those that merely reflect historical technical limitations?

Some key considerations:

- **Natural vs. Engineered Interaction**: Does a given structure exist because it supports learning, or because it was easier to implement in early systems?
- **Flexibility vs. Control**: Are rigid structures maintained for pedagogical reasons, or have we simply grown accustomed to the control they provide?
- **Explicit vs. Emergent Structure**: Do we require pre-defined structures because they're pedagogically optimal, or because we haven't trusted systems to handle more organic approaches?

These considerations nudge oneself towards thinking of fluid interfaces [31], and to build interfaces that offer an opportunity to challenge and potentially liberate ourselves from long-standing assumptions about the nature of computer-aided learning.

This connects to an important dichotomy in Human-Computer Interaction: the distinction between computers as tools versus mediums of expression [12]. It becomes particularly relevant when we examine the evolution of intelligent tutoring systems. Traditional ITS approaches have been shaped not just by pedagogical requirements but by technical limitations, restrictions that have often been accepted as necessary constraints on how tutoring must work.

Consider the typical constraints in traditional tutoring systems:

- Fixed dialogue trees for tutor responses
- Predetermined paths through subject matter
- Limited ability to recognize and adapt to student context
- Rigid assessment formats
- Standardized presentation of concepts

Many of these constraints reflect technological limitations rather than pedagogical necessities. Just as early computer interfaces required humans to adapt to machine limitations (like command-line interfaces), early tutoring systems required both tutors and students to adapt to technical constraints in how knowledge and interaction could be represented.

The emergence of large language models, viewed through the lens of fluid interfaces [31], suggests an opportunity to re-examine which constraints are truly necessary for effective tutoring and which are artifacts of previous technical limitations. This is particularly relevant for tutor simulation, where we have historically had to make significant compromises in how we model tutor behavior [53].

Traditional approaches to tutor simulation often try to explicitly model tutor decision-making [44] - creating rule systems or decision trees that attempt to capture tutoring expertise. This reflects the tool-based paradigm [12], where we try to reduce tutoring to a set of programmable rules. But just as fluid interfaces suggest moving beyond rigid structures in interface design [31], we might similarly re-imagine tutor simulation as something more organic and adaptive.

What might this look like in practice? Instead of trying to create a perfect model of tutor behavior, we might focus on creating systems that can:

- Maintain natural, flowing conversations while preserving pedagogical purpose
- Adapt their communication style based on subtle cues in student interaction
- Generate contextually appropriate examples and explanations

- Switch between different representations of concepts based on student understanding
- Allow structure to emerge from the interaction rather than being predetermined

This approach to tutor simulation prompts several questions regarding its implementation and efficacy. We must consider how to balance pedagogical effectiveness with more natural, fluid interactions between tutors and students. This leads us to examine which structural elements are genuinely essential for effective tutoring, rather than merely conventional. Furthermore, we need to determine reliable methods for evaluating the success of these more dynamic tutoring approaches, particularly when they deviate from traditional, highly structured formats.

2.3.1 Examining Traditional Constraints in Tutoring Systems

Let's examine specific constraints that have historically shaped intelligent tutoring systems and explore which might be relaxed through fluid interfaces and modern language models:

1. Linear Conversation Flow

Traditional constraint: Tutoring interactions follow a predetermined turn-taking pattern: the student responds, then the tutor responds, following a rigid back-and-forth structure. This constraint emerged from early dialogue systems' [36] limitations rather than pedagogical necessity. Human tutors naturally maintain multiple parallel threads of conversation [54], circle back to previous points [55], and weave concepts together organically [6].

With LLMs' ability to maintain complex conversation state and context, we might imagine tutoring systems that can:

- Maintain multiple parallel discussion threads
- Revisit and connect concepts organically as opportunities arise
- Adjust the pace and flow of conversation based on student engagement
- Allow for more natural, overlapping dialogue patterns

2. Fixed Knowledge Representation

Traditional constraint: Concepts must be pre-encoded in structured formats that the system can process, limiting the flexibility of explanation and exploration. This constraint arose from the need to make knowledge machine-readable, not from how humans best understand concepts. Human tutors naturally shift between different representations of the same concept based on student understanding [56, 57].

Modern language models suggest possibilities for:

- Dynamically generating multiple representations of concepts
- Creating custom examples that connect to student interests
- Adapting explanations based on student background
- Allowing concepts to emerge through exploration rather than being pre-defined

3. Rigid Assessment Structure

Traditional constraint: Student understanding must be evaluated through structured questions with predetermined correct answers [58]. This limitation reflects historical technical constraints in natural language understanding, not optimal assessment practices [59]. Human tutors evaluate understanding through natural conversation and open-ended exploration [55].

Fluid interfaces could enable:

- Recognition of understanding demonstrated through natural dialogue
- Dynamic generation of contextual assessment opportunities
- Multiple paths to demonstrate mastery
- Continuous evaluation integrated into normal conversation

4. Context Isolation

Traditional constraint: Each tutoring session starts fresh, with limited ability to build on previous interactions or broader student context. This constraint emerged from memory and processing limitations [16], not pedagogical best practices [60]. Human tutors naturally build on previous interactions and understand broader student context [61].

Modern systems could:

- Maintain long-term memory of student interests and challenges
- Connect concepts across different tutoring sessions
- Adapt to student's evolving understanding over time
- Integrate awareness of broader learning context

These constraints suggest an experimental approach: we could select one constraint, perhaps the linear conversation flow, and compare learning outcomes between a traditional system with rigid turn-taking, a fluid system that allows for more natural conversation patterns, and a human tutor as baseline. This could help demonstrate how relaxing technically-driven constraints might actually enhance rather than compromise pedagogical effectiveness.

The broader implication is that many "necessary" structures in tutoring systems reflect what Papert called the "QWERTY phenomena" [12]: patterns that persist due to historical technical limitations rather than current needs or optimal learning approaches. By identifying and questioning these constraints, we can begin to imagine tutoring systems that better align with natural learning processes.

3 Theoretical Foundations

A fundamental divide in educational technology is the tension between computers as tools versus computers as mediums of expression [62]. One that becomes particularly relevant as we consider the role of Large Language Models in learning. When Seymour Papert wrote Mindstorms in 1980 [12], personal computers weren't even really a thing yet. But what's interesting is how he was already pushing back against the obvious way people wanted to use them in education, as sort of like digital teachers, delivering instruction to students. Instead, Papert saw computers as something quite different: as a medium that children could use to express themselves.

This shift in perspective from computers-as-teachers to computers-as-medium might seem subtle, but it points to something pretty fundamental about how we think about learning. Traditional intelligent tutoring systems exemplify the "computer as tool" paradigm. They approach learning from the top down, beginning with predetermined content and curricula [20] rather than with the learner themselves. This approach makes an implicit assumption: that we can know what a student needs to learn before we understand who they are and what they know.

This assumption creates a fundamental mismatch with how learning actually works. As Piaget articulates through his theory of genetic epistemology [10], learning must begin from what one already knows. Learning is inherently personal and individual. Each learner brings their own context, interests, and existing knowledge to any new concept. When we treat computers purely as tools for delivering predetermined content, we ignore this fundamental nature of learning.

Traditional intelligent tutoring systems face a limitation in how they approach personalization itself. These systems typically attempt to adapt to learners through carefully engineered features such as predefined levels of expertise, structured pathways, and explicitly programmed responses. This approach reflects a broader pattern in artificial intelligence that Sutton describes in "The Bitter Lesson" [63]: we often assume we can encode our understanding of learning and teaching directly into our systems. Yet historically, this assumption has consistently led to suboptimal solutions. Just as we cannot fully codify the complexities of chess strategy [64] or image recognition [65] into explicit rules, we cannot completely capture the nuances of how humans learn and adapt through predetermined frameworks [66].

Modern language models suggest a possibility for transcending these limitations. Unlike the rigid expert systems of the past, which operate from a limited set of predetermined responses, LLMs work with a "latent space" [67]: an incredibly intricate and complex representation of language, knowledge, and human interaction [21]. This space functions as a kind of tiny representation of the world [68], vastly broader and more holistic than what traditional systems could achieve. Rather than relying on explicitly engineered features, LLMs are trained for their capacity for personalization through exposure to vast amounts of training data.

What's particularly fascinating about this shift is how it might finally let us realize Papert's vision of computers as a medium for learning [12]. When Papert talked about computers as a medium, he wasn't just suggesting a different way to deliver content. He was proposing a different relationship between learner and machine. He envisioned

environments where children could explore mathematical ideas through programming, where the computer became an "object to think with" rather than a tool to learn from.

This vision aligns remarkably well with how LLMs function. LLMs are built from the ground up to engage in natural dialogue at a complexity that approaches or even exceeds human capability [21]. They can maintain context, adjust in real-time to new information, and improvise responses based on the unique path each conversation takes [69]. This is fundamentally different from traditional systems where each possible interaction must be explicitly programmed in advance.

Let's come back to the example we touched on earlier. Consider what happens in a conversation with an LLM about a physics concept. The system doesn't just deliver pre-programmed explanations, it engages in a kind of collaborative meaning-making, adjusting its language and examples based on how the learner expresses their understanding. This isn't just personalization in the traditional sense of selecting from predetermined options. It's more like what happens when you use language itself as a medium for thinking - the conversation shapes and reshapes itself based on the interaction between participants.

But this potential raises questions about educational technology. If we treat LLMs as a medium rather than a tool, how do we ensure they support effective learning? How do we maintain pedagogical purpose while allowing for the kind of open-ended exploration that characterizes genuine learning? These questions become particularly crucial when we consider the scalability challenge in education. We are not just trying to create effective learning experiences, but to do so at a scale that could meaningfully impact educational access.

This problem and the potential for flexibility suggests creating what we might call "alive" interfaces [70, 71]: learning environments that can adapt organically to each learner's needs and interests. But what exactly makes an interface "alive"? The answer lies in a shift in how we approach the relationship between humans and computational systems.

3.1 Fluid Interfaces and Learning

Traditional interfaces, even "intelligent" ones, operate on the principle of structured adaptation [72]. They adjust their behavior based on predefined parameters and pathways. Even when these pathways are sophisticated, they still represent a fundamentally tool-based approach where the computer's behavior is constrained by pre-engineered responses to anticipated user needs.

What's emerging now is something quite different: what we might call "fluid interfaces" [70, 71]. Rather than trying to anticipate and program for every possible interaction, these interfaces function as a medium that naturally conforms to and amplifies human patterns of thought and expression. This is a philosophical reorientation, enabled by technological evolution, in how we think about human-computer interaction [73].

Consider how a human tutor adapts their teaching style. They don't select from a predefined set of strategies. Instead, they continuously reshape their approach based on subtle cues in the student's expression, engagement, and understanding [6, 61].

They maintain interactions that evolve organically rather than following prescribed paths.

The shift to fluid interfaces in learning environments represents an attempt to capture this organic quality of human interaction. Rather than forcing learners to adapt to the computer's way of structuring knowledge and interaction, these systems aim to meet learners where they are, adapting their form and function to match the natural flow of human thought and learning.

This approach aligns remarkably well with constructivist principles [10] not just in theory but in actual operation. While traditional systems attempt to implement constructivism through carefully engineered features and pathways, fluid interfaces embody it at a more fundamental level. They create environments where learning can emerge naturally from the interaction between learner and system, much as Piaget [10] described knowledge emerging from the interaction between child and environment.

To make this shift from structured to fluid interfaces more concrete, consider how we've evolved in handling user authentication [70]. We've moved from requiring users to remember passwords (adapting to computer constraints) to using face detection (honoring natural human states). This represents a fundamental principle of fluid interfaces. Rather than forcing humans to translate their natural modes of being into machine-readable formats, we expand our systems to work with human patterns.

This principle becomes particularly intriguing when we consider how fluid interfaces might handle more abstract domains like learning and understanding. Traditional systems often require learners to demonstrate their knowledge through structured formats [8] like multiple choice questions, formal proofs, standardized essays. But what if we could create interfaces that recognize and work with more natural expressions of understanding?

Several questions naturally arise when we propose such a radical shift in interface design:

- **Reliability:** How do we ensure these more fluid interfaces maintain educational effectiveness?
- **Scalability:** Can these interfaces adapt meaningfully across different subjects and learning styles while maintaining their fluid nature?
- **Balance**: How do we maintain enough structure to guide learning while allowing for organic exploration?

One might worry that such fluid interfaces could lead to chaos [74] - learning experiences that meander without purpose. However, this concern reflects a false dichotomy between structure and fluidity [75]. Just as a river is both fluid and channeled by its banks, fluid interfaces can maintain direction while adapting to individual learners' paths.

The power of fluid interfaces resides in shifting the source of structure rather than eliminating it entirely. Instead of enforcing top-down structure through predefined pathways, these interfaces facilitate the organic emergence of structure through learner-system interactions. This paradigm shift transforms educational technology

from restrictive tools to adaptable mediums, reshaping how we perceive and utilize technology in education.

Looking ahead, we might imagine learning environments that:

- Adapt their representation of concepts based on the learner's emerging understanding
- Allow for multiple simultaneous views of the same concept, each tailored to different aspects of the learner's comprehension
- Create dynamic spaces for exploration that maintain pedagogical effectiveness while honoring individual learning paths

To illustrate how fluid interfaces might manifest in practice, consider a "Dynamic Idea Playground" where instead of just discussing concepts through text, the interface creates interactive, playable representations of what is being discussed. When exploring game strategy, a playable board might emerge. When discussing mathematical concepts, an interactive graph appears. For therapy or complex decision-making, a dynamic mind map evolves with the conversation. Rather than forcing ideas into fixed formats, the interface morphs to create the most natural representation for each concept being explored. In this, we are seeing Papert's [12] vision of computers as "objects to think with" realized in a fluid form. Instead of merely presenting content, the interface generates a dynamic environment that adapts and adjusts to reflect the natural progression of ideas and conversations.

This example particularly illustrates two key theoretical principles we discussed:

1. The shift from structured adaptation to organic response: the system doesn't respond from predetermined visualizations but generates appropriate representations based on the emerging conversation
2. The emphasis on meeting learners where they are by creating multiple simultaneous views of concepts, each tailored to different aspects of understanding

Or consider how we might transform the process of thought development. Traditional systems require us to pre-structure our thoughts into rigid formats like bullet points, formal arguments, structured notes. But what if our interfaces could adapt to our natural thought patterns instead? Picture a system that lets thoughts flow naturally through voice or text, then helps discover and highlight emerging patterns and connections, while preserving the original voice and thinking style. When we allow thoughts to flow naturally and let structure emerge through interaction, we are seeing constructivist principles in action. Just as Piaget [10] described knowledge emerging from interaction with the environment, here we see understanding emerging from the interaction between learner and system.

The examples presented illustrate a significant change in our understanding of structure within learning environments. Instead of imposing an external structure, fluid

interfaces offer the opportunity for structure to develop naturally through interactions, similar to the fluid yet directed flow of a river.

The power of fluid interfaces lies in their ability to resolve what previously seemed like irreconcilable tensions in educational technology. They suggest ways to maintain direction without imposing rigid structure, to provide guidance without constraining exploration, and to assess understanding without forcing standardized expressions of knowledge [70].

4 LLMs in Education

4.1 The Promise of Natural Interaction

The emergence of Large Language Models represents a potential breakthrough in educational technology, particularly in addressing the longstanding challenge of creating natural, adaptive learning interactions at scale [67]. Unlike traditional intelligent tutoring systems that rely on explicitly programmed rules and structures [66], LLMs are trained in capabilities through exposure to vast amounts of training data [21]. This difference suggests new possibilities for realizing the fluid interfaces discussed in previous sections.

Early research into LLMs' educational capabilities reveals both promise and limitation [23]. While models like GPT-4 can engage in sophisticated dialogue and generate contextually appropriate responses [76], they also face unique challenges in educational contexts. Studies of math explanations generated by GPT-3 found that teachers rejected approximately 50% of the outputs [77]. Even with more recent models like ChatGPT, while the quality of generated hints improved (with 70% deemed acceptable by experts) [35], these hints may still produce lower learning gains than human-generated alternatives [35].

4.2 From Tool to Medium: LLMs and Fluid Interaction

The integration of LLMs into education, to reiterate, represents more than a technical advancement. It suggests a shift in how we think about educational technology. This shift aligns closely with our earlier discussion of computers as mediums rather than tools [12]. Traditional intelligent tutoring systems exemplified the tool paradigm, requiring explicit programming for every possible interaction [66].

In contrast, LLM's capability for fluid, context-sensitive interaction [23] addresses one of the core challenges in scaling personalized education: the ability to maintain natural, pedagogically sound conversations without requiring explicit programming for every possible interaction. When a student asks about a concept, an LLM can generate explanations that connect to their current understanding, draw relevant analogies, and adjust its approach based on their responses - all without these variations being pre-programmed.

4.3 Simulation and Modeling: Understanding LLM Capabilities

Recent work on LLMs' simulation capabilities provides crucial context for our experimental focus on tutor behavior simulation. Studies by Argyle et al. [78] have demonstrated that LLMs can effectively simulate multiple human perspectives in complex dialogues, while Aher et al. [79] showed promising results in replicating human subject studies. Additional work [80] demonstrated LLMs' ability to maintain consistent interaction patterns across extended conversations, suggesting potential for maintaining stable tutoring styles.

The simulation capabilities of LLMs become particularly interesting in the context of expert behavior modeling. Kim et al. [81] identified key patterns in expert tutor behavior, including strategic question selection, adaptive feedback provision, and dynamic scaffolding of complex concepts. Recent work by Nye et al. [82] suggests that LLMs might be capable of learning these patterns through exposure to expert demonstrations, rather than requiring explicit programming of tutoring rules.

4.3.1 Evolution of Student Modeling

Another interesting implementation of LLMs has been their role in student modeling. Traditional approaches to student modeling, as exemplified by Corbett and Anderson's knowledge tracing [20, 83], relied heavily on explicit feature engineering. This approach required careful decomposition of domain knowledge into discrete components and explicit programming of assessment rules.

Later work by Baker et al. [84] introduced more sophisticated approaches that could capture affective states and engagement levels, but still relied on predetermined features and structured data collection. The emergence of deep knowledge tracing [85] suggested possibilities for more flexible student modeling, but remained constrained by the need for large amounts of structured training data.

LLMs suggest a different approach to student modeling. One where student understanding emerges naturally through conversation rather than being explicitly mapped to predefined states. Recent work demonstrates that LLMs can maintain sophisticated models of student knowledge through natural dialogue, adjusting their responses based on subtle cues in student expression and engagement [86].

4.3.2 Advances in Tutoring Simulation

Research on LLM-based tutoring has revealed several promising capabilities:

- **Adaptive Response Generation**: Studies [86–88] show that LLMs can generate contextually appropriate explanations that adapt to student prior knowledge and learning style.
- **Strategic Interaction Planning**: Recent work [89] also demonstrates LLMs' ability to maintain longer-term interaction strategies, planning sequences of explanations and questions that build toward learning objectives.
- **Multimodal Integration**: Recent advances by Chi et al. [55] show promising results in combining language understanding with visual and interactive elements, suggesting possibilities for richer tutoring interactions.
- **Metacognitive Support**: Research [45, 46] indicates that LLMs can effectively prompt and support student self-reflection and metacognitive development.

These capabilities suggest potential for "organic tutoring" where the interaction naturally adapts to student needs without relying on rigid scripts or predetermined paths. This aligns with constructivist learning principles and supports the kind of fluid interaction discussed in previous sections.

4.3.3 Challenges in Tutor Simulation

Several key challenges emerge when using LLMs for tutor simulation, as identified through both empirical studies and theoretical analysis:

- **Pedagogical Fidelity**: While LLMs can generate plausible responses, ensuring these responses align with effective teaching practices remains challenging. A particularly concerning limitation emerges from LLMs' fundamental architecture as predictive text generators: their ability to produce coherent, well-reasoned text can sometimes result in what can be quite compelling but incorrect explanations.
- **Strategic Decision Making**: Expert tutors don't just respond to students, they maintain complex mental models of student understanding, calibrate their responses based on subtle cues, and strategically select between different pedagogical moves. Studies demonstrate that expert tutors engage in sophisticated multi-level decision making that current LLMs struggle to replicate consistently [90, 91].
- **Consistency and Reliability**: LLMs can produce inconsistent or contradictory responses across multiple interactions, potentially undermining the trust necessary for effective learning relationships [92, 93]. This becomes particularly problematic in sequential learning tasks where consistency is crucial for building understanding.
- **Cultural and Contextual Awareness**: Work by Garcia et al. [94] highlights challenges in ensuring LLM tutors maintain appropriate cultural sensitivity and contextual awareness across diverse student populations.
- **Assessment Accuracy**: Studies [95–97] indicate that while LLMs can engage in sophisticated dialogue, they sometimes struggle to accurately assess student understanding, particularly for complex conceptual knowledge.

These challenges point to a broader question about the nature of tutoring. As Kim [81] argue, effective tutoring involves more than just generating appropriate responses - it requires maintaining a coherent pedagogical strategy while adapting to individual student needs. This tension between consistency and adaptability represents a core challenge in LLM-based tutor simulation.

4.4 Looking Forward: Research Directions

Our experimental work builds on these findings while addressing a crucial gap in the literature: understanding how closely LLM behavior can align with expert tutor action patterns. While previous work has shown LLMs can simulate human behavior in general contexts [79], tutoring presents unique challenges that require careful investigation.

The methodological challenge here mirrors a broader question in educational AI: how do we evaluate systems that might achieve the same pedagogical goals

through different means? Our approach, detailed in the following sections, focuses on understanding the distribution of tutor actions and strategic choices rather than surface-level response matching.

This investigation has implications beyond just technical capability. If LLMs can effectively simulate expert tutor behavior while maintaining pedagogical effectiveness, it suggests new possibilities for scaling personalized education without sacrificing quality. However, realizing this potential requires careful attention to both the technical capabilities and limitations of LLMs and the pedagogical principles that guide effective teaching and learning.

5 Methodology

5.1 Research Questions

This study investigates differences between how language models and humans approach the tutoring task. While much research focuses on evaluating the performance or effectiveness of AI tutors [34, 35], we take a different approach: examining the underlying patterns in how these systems engage with learners compared to human tutors. This focus stems from several broader questions about the nature of ITS and AI:

1. How do artificial tutoring systems function when given the same context as human tutors?
2. What systematic differences emerge in how AI and human tutors structure their teaching interactions?
3. How do AI tutors adapt their teaching strategies in response to different student behaviors?

These questions address core theoretical interests about the nature of ITS while avoiding assumptions about what constitutes "correct" or "effective" tutoring. By focusing on behavioral patterns rather than performance metrics, we aim to understand fundamental differences in how artificial and human tutors approach the teaching task.

5.2 Design Principles

Our methodology is shaped by several key principles that inform how we approach comparing LLM and human tutoring behavior:

5.2.1 Population-Level Analysis

Rather than attempting direct turn-by-turn comparisons between human and LLM responses, we focus on analyzing aggregate behavioral patterns across the entire dataset. This approach is particularly important given the low agreement rate (18.1%) observed between human tutors, which suggests that seeking exact matches between responses would be methodologically unsound. By examining population-level patterns, we can better understand the overall alignment of tutoring strategies while accommodating natural variations in individual approaches.

5.2.2 Reference Distribution Approach

We aggregate human tutor responses to create reference distributions that capture the characteristic patterns of human tutoring behavior. While this approach necessarily sacrifices some granular information about inter-tutor variance, it provides a more stable baseline for comparative analysis. These reference distributions serve as a robust framework for evaluating whether LLM behavior falls within the expected range of human tutoring actions.

5.2.3 Model Comparison Strategy

Our approach maintains separate distributions for different LLM configurations, enabling us to:

- Distinguish between model-specific behaviors and general LLM characteristics
- Assess the consistency of LLM responses across different implementations

This framework fundamentally reorients our research question from "Does this LLM respond like a human tutor would?" to "Does this LLM's pattern of action choices align with the patterns we observe in human tutors?" This shift enables a more nuanced and meaningful evaluation of LLM tutoring capabilities.

5.3 Dataset

5.3.1 Overview

Our analysis utilizes the CIMA (Conversational Instruction with Multi-responses and Actions) dataset [98], which provides a rich collection of tutoring dialogues focused on teaching Italian prepositional phrases to English speakers. The dataset is particularly valuable for studying tutoring behavior patterns as it captures multiple valid tutoring responses for each student interaction, reflecting the reality that there is rarely one "correct" way to respond in a tutoring context, as illustrated by the following interaction:

Student: How do you say "behind" in Italian?
Tutor A: "Behind" is "dietro". Would you like to try making a complete phrase?
Student: The cat is dietro the purple bed?
Tutor A: You're on the right track! Remember to use "è" for "is".
Also, we need "al" before "letto" when saying "to the bed".
Tutor B: Close! The complete phrase would be "il gatto è dietro al letto viola"
Tutor C: Good attempt! Let's break this down:
- "The cat" = "il gatto"
- "is" = "è"
- "behind" = "dietro"
- "the purple bed" = "al letto viola"
Would you like to try again with these pieces?

This example illustrates several key features of the dataset:

1. **Multiple Valid Approaches:** For each student utterance, three different tutors provide responses, showing distinct but equally valid tutoring strategies. Tutor A offers guided scaffolding, Tutor B provides direct correction, while Tutor C breaks down the components systematically.

2. **Action Labeling:** Each response is annotated with pedagogical actions. In the example above:
    - Tutor A: [Question, Hint]
    - Tutor B: [Correction]
    - Tutor C: [Hint, Question]
3. **Progressive Learning:** The dataset captures how concepts build across exercises. Consider this sequence showing knowledge transfer:

> Exercise 1:
> Student: Is "il gatto è vicino al letto" correct for "the cat is next to the bed"?
> Tutor: Yes, perfect! "vicino al" is correct for "next to the".
>
> Exercise 2:
> Student: Then would "il cane è vicino alla scatola" work for "the dog is next to the box"?
> Tutor: Excellent thinking! Yes, you're applying the pattern correctly.
> Notice how we use "alla" instead of "al" because "scatola" (box) is feminine.

This progression shows how students build on previously learned concepts, and how tutors recognize and reinforce these connections.

The dataset [98] contains 391 completed exercises across 77 students, with each exercise grounded in both visual and conceptual representations. What makes CIMA particularly suitable for studying tutoring behavior is its capture of natural variation in teaching strategies. Consider these parallel responses to a student error:

> Student: il gatto è in cima il letto viola [attempting: "the cat is on top of the purple bed"]
>
> Tutor A: Good try! We need to say "al" instead of "il" after "in cima" - this is because we're saying "on top of THE bed". Want to try again?
>
> Tutor B: Close! Remember that "in cima" needs to be followed by "al" when we're talking about a masculine noun like "letto". The correct phrase is "il gatto è in cima al letto viola".
>
> Tutor C: Let's think about this - when we say "on top OF the bed", we need a special word to show that connection. In Italian, we use "al" after "in cima". Can you spot where this needs to go in your sentence?

These responses, while addressing the same error, demonstrate different tutoring philosophies:

- Tutor A provides minimal correction with encouragement
- Tutor B offers comprehensive explanation with the correct answer
- Tutor C engages the student in error analysis

The mean response lengths (6.82 words for students, 9.99 words for tutors) indicate substantive rather than minimal interactions. This richness of interaction, combined with explicit action labeling, provides a strong foundation for analyzing how different tutors (both human and AI) structure their teaching interventions.

5.3.2 Dataset Creation and Implications

The CIMA dataset [98] was created through an asynchronous crowdsourcing approach, where Amazon Mechanical Turk workers role-played as both students and tutors. To ensure quality, workers were required to have 95% approval rates over at least 1,000 previous tasks, and a subset of responses from each worker was manually checked. Importantly, tutor-workers were not required to know Italian. Instead, they were provided with relevant vocabulary and grammar rules to simulate the knowledge a tutor would have.

A distinctive feature of the CIMA dataset is also its collection of multiple valid tutoring responses for each student interaction. Specifically, each student turn receives responses from three different tutors, allowing us to analyze variation in teaching strategies and establish a more comprehensive understanding of possible pedagogically sound responses to any given student input.

This collection method has several implications for our analysis of tutoring behavior patterns:

- **Natural Variation:** Because responses come from a large pool of crowdworkers (209 distinct tutors) rather than a small group of professional tutors, the dataset captures a broader range of natural tutoring instincts and strategies. This variety is particularly valuable for our analysis, as it helps establish a more representative baseline of human tutoring behavior against which to compare LLM patterns.
- **Controlled Knowledge:** By providing tutors with explicit grammar rules and vocabulary rather than requiring Italian proficiency, the dataset isolates tutoring behavior from language expertise. This aligns well with our focus on analyzing tutoring patterns rather than pedagogical effectiveness, as it captures how people naturally structure teaching interactions when given a fixed knowledge base, similar to how LLMs work with their training data.
- **Task-Focused Interactions:** The asynchronous nature of data collection, where tutors respond to individual student turns rather than maintaining extended conversations, emphasizes immediate tutoring decisions over long-term strategy. This granular focus helps us identify specific patterns in how tutors choose actions and structure responses.

For our investigation of LLM tutoring behavior, these characteristics make CIMA particularly suitable. The dataset's emphasis on capturing multiple valid responses to the same student input allows us to study how both human tutors and LLMs handle the inherent flexibility of teaching interactions. Moreover, the controlled knowledge aspect creates a more direct comparison between human and ITs patterns, as both are working from explicit rather than internalized language knowledge.

Creating an Ensemble View

The multiple-response structure of CIMA enables us to create an "ensemble view" of human tutoring behavior. Rather than treating any single response as the ground truth, we can observe patterns across multiple tutors' responses to the same situation. This approach aligns well with our research questions - we're not trying to identify optimal tutoring strategies, but rather understand patterns in how different tutors (human and AI) approach the same teaching moments.

5.3.3 Additional Details

The dataset captures and categorizes both student and tutor actions through a systematic labeling system:

Role	Action Type	Example
Student	Guess	"Is it 'il gatto e vicino alla scatola rosa'?"
	Question	"How would I say 'pink' in Italian?"
	Affirmation	"Oh, I understand now!"
	Other	"Which I just said."
Tutor	Hint	"Remember that 'l" is used before words starting with vowels."
	Question	"Are you sure you have all the words in the right order?"
	Correction	"Very close, but 'viola' comes after 'letto'."
	Confirmation	"Exactly! Well done using the correct article."
	Other	"Let's break this down step by step."

Table 1: Categorization of student and tutor actions with representative examples

Each interaction in the dataset is labeled with one or more of these action types, enabling analysis of teaching patterns and response strategies. This labeling system

allows us to track how different LLMs and human tutors vary in their choice of actions in response to specific student behaviors.

The system employs a deliberate strategy in selecting and sequencing exercises to optimize learning opportunities. When generating a sequence of exercises for a student, the system prioritizes questions that share at least one concept with previously encountered exercises. This approach creates natural opportunities for knowledge transfer and reinforcement learning. For example, a student who has just learned about color terms in one context might encounter the same colors in a different grammatical construction, allowing them to build upon existing knowledge while acquiring new skills.

5.4 Dataset Enhancement (addition of AI Tutors)

To enable direct comparison between human and artificial tutoring patterns, we enhanced the CIMA dataset [98] by generating parallel responses from state-of-the-art language models. This enhancement maintains the dataset's fundamental structure while introducing a new dimension for analysis.

5.4.1 Model Selection and Implementation

Our implementation strategy focused on controlling for basic conversational capability to isolate tutoring-specific behavioral patterns. We selected three advanced instruction-tuned language models, each representing different approaches to large-scale language modeling:

- GPT-4o 2024-08-06 (OpenAI) [99]
- Gemini Pro 1.5 (Google) [100]
- LLaMA 3.1 405B instruct nitro (Meta) [101]

This selection of models from different providers, each with distinct architectural choices and training approaches, allows us to distinguish between behaviors fundamental to language models in general versus those specific to particular implementations.

5.4.2 Response Generation Process

To ensure consistent comparison with human responses, we developed a structured prompting system that provides each model with equivalent context to what human tutors received in the original dataset. Each interaction uses a prompt template that specifies:

```
You are a language tutor teaching Italian. Available actions
(one response can correspond to multiple action types):
- Question: Ask student for clarification or to elaborate
- Hint: Provide indirect guidance
```

```
- Correction: Point out and fix errors
- Confirmation: Acknowledge correct responses
- Other: Any other type of response

Context:
- Target phrase (IT): {target_phrase['it']}
- Target phrase (EN): {target_phrase['en']}
- Grammar rules: {grammar_rules}
Conversation history: {conversation_history}

Please provide a response as a tutor to the student's last message.
Respond in JSON format with:
{
    "response": "your response text",
    "actions": ["your action types"]
}
```

This structured approach ensures that model responses can be directly compared with human responses in the original dataset, maintaining consistent action categorization and response formats across all interactions. Our complete experiment code is visible at [102].

5.5 Experimental Structure

5.5.1 Analysis Framework

Action Distribution Analysis

We examine the relative frequency of fundamental tutoring actions (Question, Hint, Correction, Confirmation, Other) across different populations. This analysis compares the baseline distribution derived from human tutors in the CIMA dataset [98] against Language Model behavior. By analyzing these distributions, we can identify systematic preferences or avoidances in action selection, revealing whether LMs exhibit biases toward particular tutoring strategies.

Action Combination Analysis

To understand the complexity and sophistication of tutoring responses, we investigate patterns in how actions are combined within individual responses. This includes analyzing:

- The typical number of actions per response
- The balance between single-action and multi-action responses

These patterns provide insight into how tutors structure their interventions and whether LLMs mirror the natural complexity of human tutoring interactions.

Conditional Response Analysis

We study how tutoring strategies adapt to different types of student input by examining response patterns conditioned on student action types. This analysis focuses on tutor responses to:

- Student questions and requests for help
- Student attempts and guesses at solutions
- Student expressions of understanding or confusion

Through this analysis, we can evaluate how effectively LLMs mirror the dynamic, responsive nature of human tutoring, particularly in their ability to provide context-appropriate support.

5.6 Evaluation Framework

5.6.1 Observable Dimensions

Our analysis framework enables measurement of several key dimensions of tutoring behavior:

Primary Metrics

We focus on these aspects of tutoring interactions:

1. Distribution of fundamental tutoring actions (hints, questions, corrections, confirmations)
2. Conditional response patterns based on student behavior
3. Significant deviations from established human behavioral patterns
4. Systematic biases in action selection (both overuse and avoidance)

Given the natural variation in human tutoring strategies [98], we expect to observe broad distributions in the baseline data. This inherent variance shapes our analytical approach: rather than seeking exact behavioral matches, we focus on identifying meaningful departures from human patterns. For instance, while human tutors might employ hints with a frequency ranging from 40-80% in response to student questions, an LLM utilizing hints only 10% of the time would constitute a significant deviation from typical human behavior.

5.6.2 Methodological Limitations

Our analysis framework operates within several important constraints:

Dataset Characteristics

The study utilizes the CIMA dataset [98], which has specific properties that influence generalizability:

- Domain specificity: Limited to Italian preposition instruction
- Tutor population: Relies on crowdsourced rather than professional tutors
- Environmental constraints: Interactions occur in a controlled, structured setting

Structural Constraints

The experimental design imposes certain limitations:

- Prescribed JSON response format may influence natural interaction patterns
- Restricted action vocabulary limits expressive range
- No direct measurement of response quality or effectiveness

Model Implementation

The current implementation has specific boundaries:

- Analysis limited to three model variants
- Single prompt template approach
- No model fine-tuning, adaptation, or usage of base models

Scope of Conclusions

It is crucial to acknowledge what our analysis cannot determine. While we can identify alignment or deviation from human behavioral patterns, we cannot:

- Evaluate the optimality of tutoring choices
- Assess the quality or appropriateness of specific responses
- Make claims about general tutoring capability beyond this specific context

Our focus on action distributions represents a deliberate methodological choice, prioritizing the analysis of strategic-level behavioral alignment over response-level quality assessment. While this approach necessarily sacrifices some depth of evaluation, it enables robust comparative analysis at the population level.

Despite using crowdsourced tutors rather than professionals, prior research suggests that crowdworker behavior can effectively approximate general tutoring patterns, lending validity to our comparative framework.

6 Results

The analysis of tutoring interactions between language models and human tutors reveals systematic differences in how these systems approach the teaching task. These patterns emerge across multiple dimensions of analysis, from basic action selection to complex interaction flows.

6.1 Action Distribution Analysis

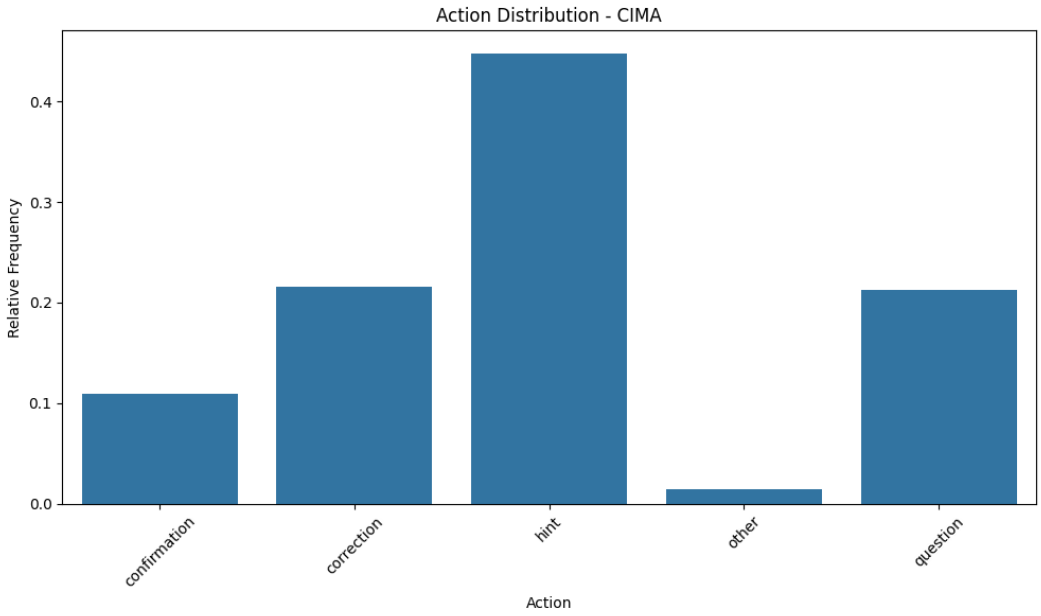


Figure 3: Distribution of tutoring actions in human tutoring sessions (CIMA dataset), showing the relative frequency of different pedagogical strategies.

The distribution of tutoring actions shows both similarities and systematic differences between human and AI tutors. As shown in Figures 3 and 4, both groups demonstrate a strong preference for hints as their primary teaching action, with hints comprising approximately 45% of all actions across both human and ITS sessions. This suggests some fundamental alignment in basic tutoring strategy, possibly reflecting the effectiveness of scaffolded guidance over direct instruction.

However, examining the broader patterns reveals key differences in how these actions are deployed. Human tutors (Figure 3) show a more balanced distribution between corrections (around 20%) and questions (about 20%), suggesting a more diverse pedagogical approach. In contrast, AI systems (Figure 4) show some variation among themselves - while all maintain the primacy of hints, they differ in their secondary strategies. GPT-4o and Gemini Pro 1.5 demonstrate a stronger tendency toward corrections (around 30%) compared to questions (about 5%), while LLaMA 3.1 maintains a more balanced profile closer to human tutors.

The relatively low frequency of confirmations across all systems (10-15%) is noteworthy, as is the minimal use of "other" actions, suggesting that both human

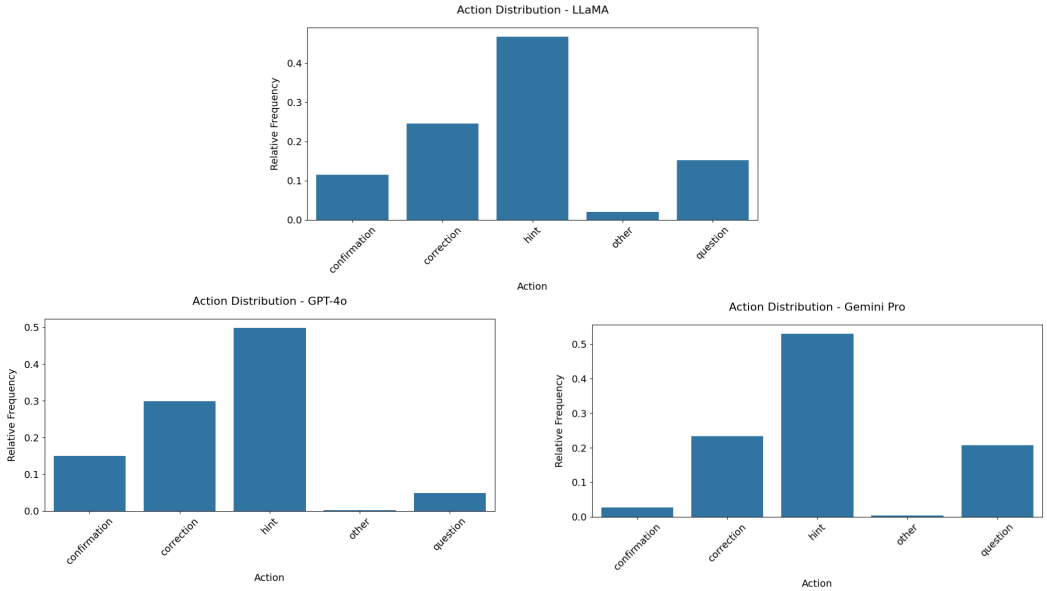


Figure 4: Distribution of tutoring actions across different ITS (LLaMA 3.1 405B, GPT-4o, and Gemini Pro 1.5

and AI tutors primarily rely on a core set of pedagogical moves. This consistency might indicate either basic constraints of the tutoring task or shared understanding of effective teaching strategies.

These patterns raise interesting questions about whether the differences in action distribution reflect meaningful pedagogical choices or artifacts of the underlying architectures. The higher correction rates in AI systems might suggest a more directive interaction pattern, while humans' more frequent use of questions could indicate a more Socratic approach to guidance.

6.2 No. of Actions per Response

The most striking difference between human and AI tutors emerges in response complexity. As shown in Figure 5, human tutors demonstrate a strong and consistent pattern for single-action responses - choosing to either provide a hint, make a correction, or ask a question, but rarely combining these actions. Approximately 70% of human responses contain just one action, with only about 25% containing two actions and fewer than 5% containing three. This pattern suggests a teaching strategy focused on clear, targeted interventions.

In contrast, AI tutors across all models (Figure 6) consistently seem to combine multiple actions in their responses, though with interesting variations between systems. LLaMA shows the strongest preference for dual-action responses (over 80%), while GPT-4o and Gemini Pro display a more balanced distribution between single and dual actions. GPT-4o uses single actions in about 30% of responses and dual actions in about 65%, while Gemini Pro shows an almost even split between single (45%) and dual actions (55%). All systems rarely use three or more actions in a single response.

This systematic tendency toward more complex responses appears to be a consistent

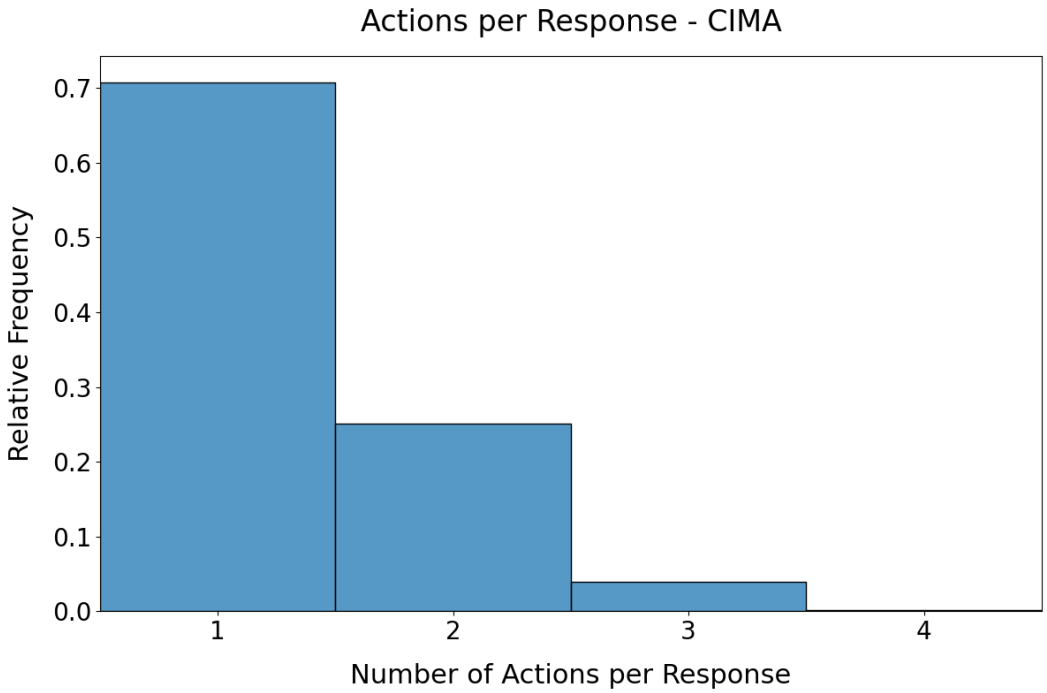


Figure 5: Distribution of the number of pedagogical actions per response in human tutoring sessions (CIMA dataset), demonstrating humans' preference for single-action responses.

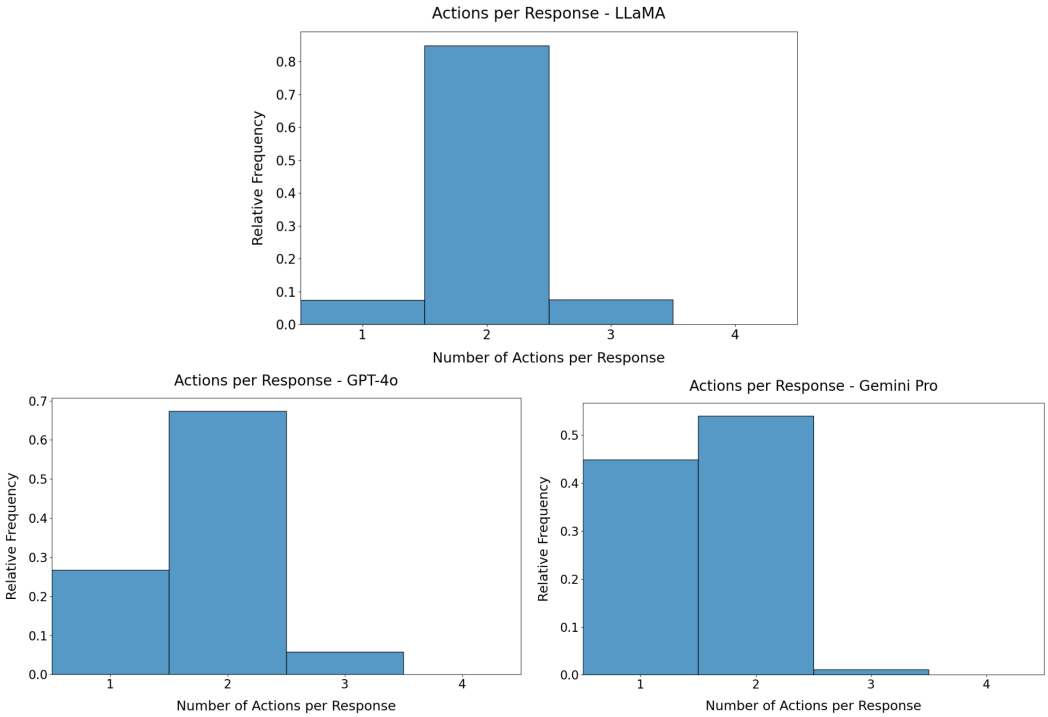


Figure 6: Distribution of the number of pedagogical actions per response for ITS (LLaMA 3.1, GPT-4o, and Gemini Pro), showing varying levels of response complexity.

feature of how language models function in the tutoring task, rather than a quirk of

any particular architecture. This pattern may reflect these systems' training as helpful assistants, where elaboration and comprehensiveness are typically rewarded. However, it raises questions about whether such complexity actually serves pedagogical goals effectively, given that experienced human tutors tend to favor simpler, more focused interventions.

The contrast between human and AI approaches to response complexity might also reflect different underlying assumptions about effective teaching. Human tutors' preference for single actions could indicate an intuitive understanding of cognitive load theory, while AI systems' more complex responses might suggest an implicit model of teaching as information delivery rather than guided discovery.

6.3 Response Patterns and Teaching Dynamics

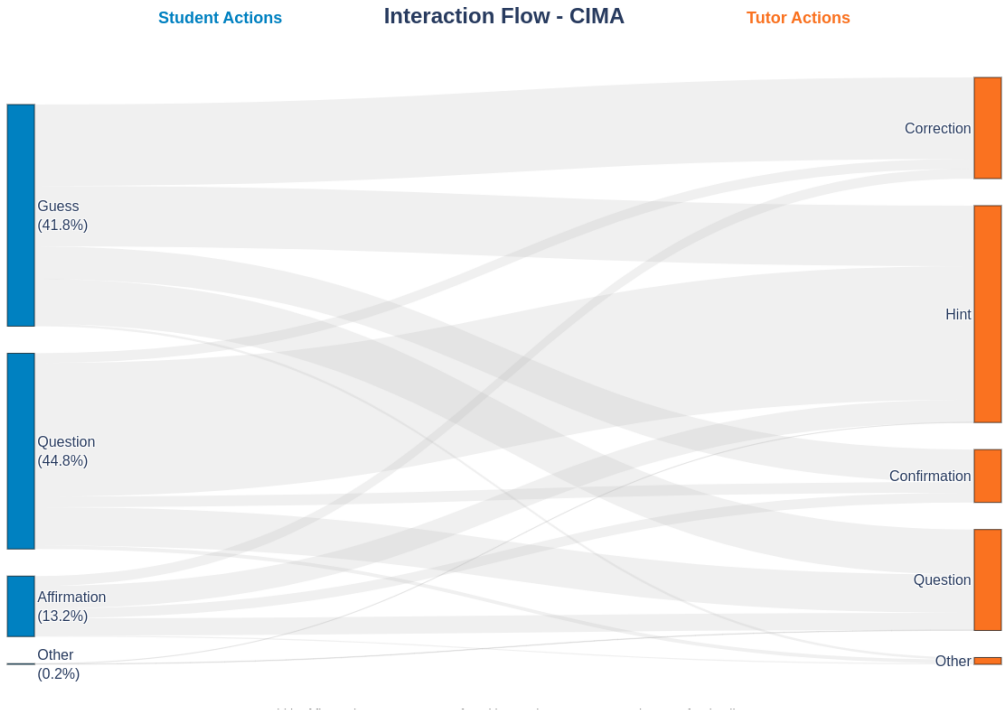


Figure 7: Sankey diagram showing the flow between student actions (left, blue) and tutor responses (right, orange) in human tutoring sessions (CIMA dataset). The width of each flow indicates the relative frequency of that particular student-tutor action pair.

The interaction flow diagrams (Figures 7 and 8) reveal the dynamic patterns of student-tutor exchanges, with several notable insights. The response patterns (shown in orange on the right) reveal distinct teaching strategies. Human tutors (Figure 7) show clear, consistent patterns in how they respond to specific student actions. When students make guesses, human tutors predominantly respond with hints or corrections, suggesting a focused strategy of either guiding students toward the correct answer or directly addressing misconceptions. When students ask questions, human tutors strongly favor providing hints, indicating a preference for scaffolded guidance over direct answers, as seen in the previous section as well.

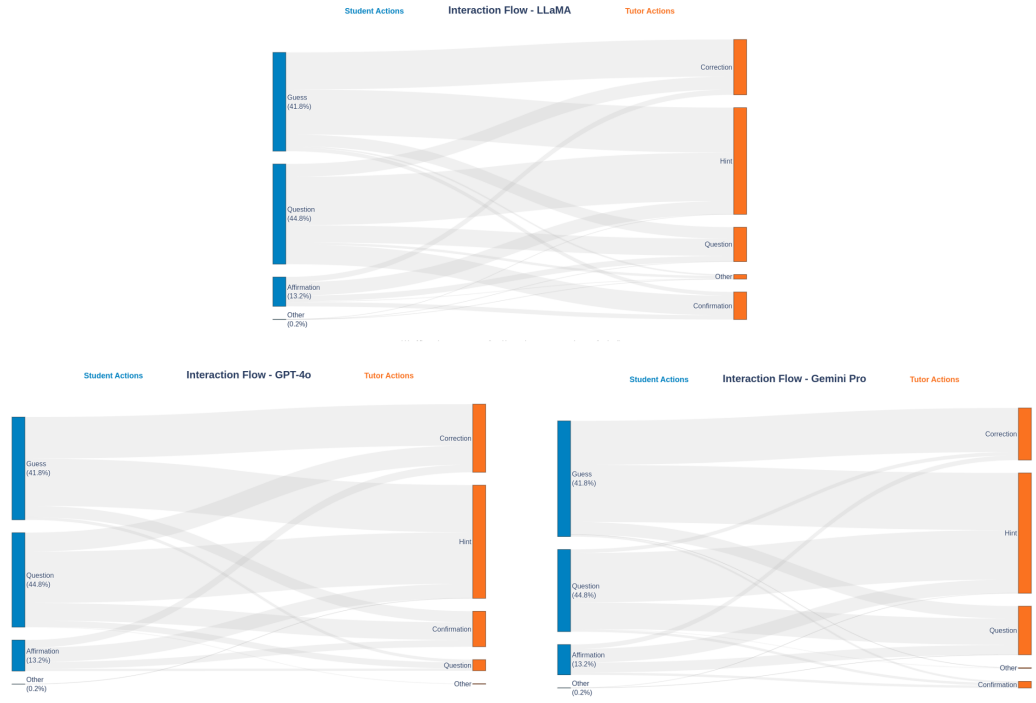


Figure 8: Sankey diagrams showing interaction flows between student actions and tutor responses across three ITS (LLaMA 3.1, GPT-4o, and Gemini Pro). Flow widths represent the frequency of specific student-tutor action pairs.

In contrast, AI tutors (Figure 8) demonstrate more distributed response patterns. While they maintain some basic alignment with human strategies such as frequently using hints in response to questions, they show less distinct mappings between student actions and tutor responses. For example, LLaMA appears to use corrections more uniformly across different student actions, while GPT-4o and Gemini Pro 1.5 show stronger tendencies to respond to questions with hints but maintain more varied responses to guesses.

Particularly interesting is how each AI model exhibits distinct systematic patterns that differ both from human tutors and from other models. This suggests that while language models can engage in fluid, natural conversation, they exhibit characteristic response patterns distinct from human teaching strategies. These differences raise important questions about whether such variations reflect meaningful pedagogical choices or are artifacts of the models' underlying architectures and training approaches.

These differences in tutor response patterns reflect genuine variations in teaching strategy rather than differences in the types of situations being handled. This makes these patterns particularly relevant for understanding how different approaches to automated tutoring might affect the learning experience.

7 Discussion

7.1 Theoretical Implications

Our investigation into how language models approach tutoring reveals a tension that connects directly to our theoretical framework. While LLMs enable more fluid, natural interaction than any previous educational technology, seemingly realizing the vision of computers as a dynamic medium, they simultaneously exhibit their own fixed patterns that neither mirror human teaching nor simply extend traditional computer-aided instruction. Each research question we posed at the start, yielded insights that contribute to our understanding of this artificial teaching behavior.

7.1.1 Natural Behavior Patterns in LLM-based ITS

Addressing our first research question about how AI tutors naturally behave, our findings reveal both alignment with and deviation from human tutoring patterns. The consistent preference for hints (45% of actions) across both human and AI tutors suggests some fundamental alignment in basic teaching strategy. However, this surface-level similarity masks deeper differences in how these hints are deployed and combined with other actions.

7.1.2 Systematic Differences in Teaching Structure

Our second research question sought to identify systematic differences in how AI and human tutors structure their teaching interactions. The most striking difference emerged in response complexity - while human tutors strongly prefer single-action responses (70%), AI tutors consistently combine multiple actions in their responses. This pattern persists across different models and contexts, suggesting it may be an inherent characteristic of how current language models approach teaching rather than a quirk of any particular implementation. Human tutors' preference for simpler responses might reflect an intuitive understanding of cognitive load theory [103]: the idea that learning is optimized when information is presented in manageable chunks. AI tutors, despite their sophisticated language capabilities, exhibit patterns consistent with different information processing approaches.

Second, this complexity pattern might reflect deeper assumptions built into these systems during training. The tendency toward comprehensive responses could stem from training data where more detailed, complete answers are typically rewarded. This suggests that even as LLMs enable more natural interaction, they maintain certain biases from their training that manifest in systematic behavioral patterns.

Thirdly, this complexity difference might represent different approaches to knowledge construction. While human tutors tend to guide learning through focused, incremental steps - aligning with constructivist principles of building understanding progressively - AI tutors seem to favor a more comprehensive, multi-faceted approach to each teaching moment.

Each AI model also exhibits distinct patterns of responses, as evidenced by distinct patterns in their action distributions and response structures. This emergence of model-specific patterns raises interesting questions about the nature of artificial teaching behavior. While these systems can engage in fluid, natural conversation, they exhibit fixed patterns distinct from both human teachers and each other.

7.1.3 Adaptation to Student Behavior

Our third research question examined how AI tutors adapt their teaching strategies to different student behaviors. The interaction flow analysis reveals that while AI tutors maintain some basic alignment with human response patterns (such as using hints in response to questions), they show less distinct mappings between student actions and tutor responses. This suggests a more distributed, less context-sensitive approach to adaptation compared to human tutors.

7.2 Practical Implications

These findings have several important implications for the development and deployment of ITS:

7.2.1 Design Considerations

The tendency toward complex, multi-action responses might need to be explicitly addressed in system design if we want AI tutors to more closely match human teaching patterns. However, this raises a question: should we modify these observed patterns to match human behavior, or might these differences represent valid alternative approaches to teaching?

Rather than viewing divergence from human patterns as necessarily problematic, we might instead focus on understanding when different approaches are most effective. For example, multi-action responses might be particularly valuable for review sessions or when synthesizing multiple concepts, while simpler, focused responses might better serve initial concept introduction.

7.2.2 Model Selection Impact

The emergence of distinct patterns across different models has significant implications for educational deployment. Each model's characteristic patterns, from GPT-4o's balanced approach to LLaMA 3.1's stronger preference for dual-action responses, suggest that model selection might significantly impact the learning experience. This extends beyond simple performance metrics to more basic questions about teaching style and approach.

7.2.3 Interaction Design

The less distinct mapping between student actions and tutor responses in AI systems suggests a need for more structured guidance in how these systems respond to different

types of student behavior.

7.3 Limitations and Future Directions

Several limitations of our current study suggest directions for future research:

- **Dataset Specificity:** Our analysis focused on Italian preposition learning, and patterns might differ in other educational contexts.
- **Model Selection:** While we examined three prominent language models, future work could explore how these patterns manifest across a broader range of models and architectures.
- **Base Model Behavior:** An interesting direction for future research would be examining how these patterns differ in base models before instruction tuning and RLHF.
- **Learning Outcomes:** Our study focused on behavioral patterns rather than educational effectiveness. Future work could examine how these different interaction patterns affect student learning outcomes.

Our findings contribute to ongoing discussions about the role of AI in education, particularly regarding the scalability of personalized learning [17]. While language models show promise in enabling more natural educational interactions than previous systems, their tendency to exhibit fixed patterns distinct from human teaching raises important questions about the nature of artificial teaching.

Rather than viewing these differences as limitations, they might represent an opportunity to develop new hybrid approaches to tutoring that combine the consistency and scalability of AI systems with the adaptive sophistication of human teachers. The challenge becomes identifying which AI patterns to preserve and which to modify for optimal learning outcomes.

Our findings suggest the need to reconsider some assumptions about educational technology. The emergence of fixed patterns in systems capable of fluid interaction challenges the simple dichotomy between rigid and adaptive systems. Instead, we might need new theoretical frameworks that account for how artificial teaching behavior develops its own characteristics even while exhibiting adaptive responses.

This connects back to Papert's vision [12] of computers as a medium for learning, but suggests a more complex relationship than originally envisioned. Rather than simply serving as a neutral medium, these systems appear to develop their own consistent patterns that color how they facilitate learning interactions.

8 Conclusion

This research reveals insights about the nature of intelligent tutoring using LLMs and the complex relationship between fluid interfaces and fixed patterns in educational

AI. Through a preliminary analysis of tutoring interactions, we discovered that even as language models enable more natural educational interactions than any previous system, they simultaneously exhibit characteristic patterns distinct from both human teaching and traditional computer-aided instruction.

Our findings illuminate several key dynamics:

- First, while both human and AI tutors show similar high-level preferences (like the predominant use of hints), they differ markedly in how they structure these interventions. The striking contrast in response complexity with AI tutors consistently preferring multi-action responses while human tutors favor focused, single-action interventions suggests fundamental differences in how these systems handle the teaching task.
- Second, each AI model develops its own distinctive interaction pattern: consistent patterns of behavior that differ both from human tutors and from other models. This emergence of model-specific patterns challenges simplistic notions about AI tutors simply mimicking human behavior, suggesting instead that these systems exhibit distinct approaches to teaching.
- Third, AI tutors show less distinct mappings between student actions and tutor responses compared to the clear patterns exhibited by human teachers. This more distributed response pattern suggests a different approach to contextual adaptation, raising questions about how these systems process and generate responses according to student needs.

These findings contribute to both theoretical understanding and practical development of educational technology. Theoretically, they suggest we need new frameworks for understanding artificial teaching behavior. Frameworks that can account for how systems maintain adaptivity while developing fixed patterns. Practically, they indicate that effective ITS might not come from perfectly mimicking human teachers, but from understanding and appropriately leveraging the characteristic patterns these systems naturally develop.

Looking forward, this work opens new questions about artificial teaching and learning. As these systems continue to evolve, the challenge become more than making them more human-like. They include understanding how to harness their unique patterns of behavior in service of effective learning. This suggests a possible future where AI tutors don't simply replicate human teaching strategies but complement them with their own distinctive interaction patterns to tutoring.

The tension between fluid interfaces and fixed patterns emerges not as a limitation to be overcome, but as a characteristic of ITS that must be understood and thoughtfully integrated into educational design. This understanding becomes crucial as we work toward developing educational technologies that can truly scale personalized learning while maintaining pedagogical effectiveness.

This research thus marks a step in understanding how AI systems process and generate responses in teaching contexts, contributing to both the theoretical foundations of

AI in education and the practical development of more effective learning technologies. As we continue to develop and deploy these systems, maintaining this dual focus on understanding behavioral patterns while ensuring pedagogical effectiveness will be crucial for realizing the potential of AI in education.

References

[1] Neal Stephenson. *The Diamond Age: Or, a Young Lady's Illustrated Primer*. en. Google-Books-ID: aAV6wV4Rn00C. Random House Worlds, Aug. 2003. ISBN: 978-0-553-89820-0.

[2] Orson Scott Card. *Ender's Game*. en. Google-Books-ID: Ojqi8KbWuLwC. Macmillan, Apr. 2010. ISBN: 978-1-4299-6393-0.

[3] Isaac Asimov. "The Fun They Had". en. In: (1951). Page Version ID: 1255384084.

[4] Benjamin S. Bloom. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring". In: *Educational Researcher* 13.6 (1984). Publisher: [American Educational Research Association, Sage Publications, Inc.], pp. 4–16. ISSN: 0013-189X. DOI: 10.2307/1175554.

[5] Chen-Lin C. Kulik, James A. Kulik, and Robert L. Bangert-Drowns. "Effectiveness of Mastery Learning Programs: A Meta-Analysis". In: *Review of Educational Research* 60.2 (1990). Publisher: [Sage Publications, Inc., American Educational Research Association], pp. 265–299. ISSN: 0034-6543. DOI: 10.2307/1170612.

[6] Kurt VanLEHN. "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems". en. In: *Educational Psychologist* 46.4 (Oct. 2011), pp. 197–221. ISSN: 0046-1520, 1532-6985. DOI: 10.1080/00461520.2011.611369.

[7] Wenting Ma et al. "Intelligent tutoring systems and learning outcomes: A meta-analysis." en. In: *Journal of Educational Psychology* 106.4 (Nov. 2014). 460 citations (Semantic Scholar/DOI) [2024-09-14], pp. 901–918. ISSN: 1939-2176, 0022-0663. DOI: 10.1037/a0037123.

[8] Benjamin D. Nye, Arthur C. Graesser, and Xiangen Hu. "AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring". en. In: *International Journal of Artificial Intelligence in Education* 24.4 (Dec. 2014). 207 citations (Semantic Scholar/DOI) [2024-09-14], pp. 427–469. ISSN: 1560-4292, 1560-4306. DOI: 10.1007/s40593-014-0029-5.

[9] Kenneth R Koedinger et al. "Intelligent Tutoring Goes To School in the Big City". In: 1036 citations (Semantic Scholar/DOI) [2024-09-14] Artwork Size: 165322 Bytes. Carnegie Mellon University, 1997, 165322 Bytes. DOI: 10.1184/R1/6470153.V1.

[10] Jean Piaget. *The Construction Of Reality In The Child*. London: Routledge, July 2013. ISBN: 978-1-315-00965-0. DOI: 10.4324/9781315009650.

[11] Salman Khan. *Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing)*. en-US. Viking, 2024. ISBN: 978-0-593-65695-2.

[12] Seymour Papert. *Mindstorms: children, computers, and powerful ideas*. USA: Basic Books, Inc., 1980. ISBN: 978-0-465-04627-0.

[13] Ernst von Glasersfeld. *RADICAL CONSTRUCTIVISM*. London: Routledge, Aug. 2013. ISBN: 978-0-203-45422-0. DOI: 10.4324/9780203454220.

[14] Paulo Freire. "Pedagogy of the Oppressed*". In: *Toward a Sociology of Education*. Num Pages: 13. Routledge, 1978. ISBN: 978-0-429-33953-0.

[15] Kenneth R. Koedinger and Vincent Aleven. "An Interview Reflection on "Intelligent Tutoring Goes to School in the Big City"". en. In: *International Journal of Artificial Intelligence in Education* 26.1 (Mar. 2016). 160 citations (Semantic Scholar/DOI) [2024-09-14], pp. 13–24. ISSN: 1560-4306. DOI: 10.1007/s40593-015-0082-8.

[16] Beverly Woolf. *Building Intelligent Interactive Tutors, Student-Centered Strategies for Revolutionizing E-Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN: 978-0-08-092004-7.

[17] Beverly Park Woolf et al. "AI Grand Challenges for Education". en. In: *AI Magazine* 34.4 (Dec. 2013). 124 citations (Semantic Scholar/DOI) [2024-09-14], pp. 66–84. ISSN: 0738-4602, 2371-9621. DOI: 10.1609/aimag.v34i4.2490.

[18] Ali Alkhatlan and Jugal Kalita. "Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments". In: *International Journal of Computer Applications* 181.43 (Mar. 2019). 70 citations (Semantic Scholar/DOI) [2024-09-14], pp. 1–20. ISSN: 09758887. DOI: 10.5120/ijca2019918451.

[19] Elham Mousavinasab et al. "Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods". In: *Interactive Learning Environments* 29.1 (Jan. 2021). 240 citations (Semantic Scholar/DOI) [2024-09-14] Publisher: Routledge _eprint: https://doi.org/10.1080/10494820.2018.1558257, pp. 142–163. ISSN: 1049-4820. DOI: 10.1080/10494820.2018.1558257.

[20] John R. Anderson et al. "Cognitive Tutors: Lessons Learned". en. In: *Journal of the Learning Sciences* 4.2 (Apr. 1995). 1991 citations (Semantic Scholar/DOI) [2024-09-14], pp. 167–207. ISSN: 1050-8406, 1532-7809. DOI: 10.1207/s15327809jls0402_2.

[21] Tom B. Brown et al. *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. DOI: 10.48550/arXiv.2005.14165.

[22] Jared Kaplan et al. *Scaling Laws for Neural Language Models*. arXiv:2001.08361 [cs]. Jan. 2020. DOI: 10.48550/arXiv.2001.08361.

[23] Jason Wei et al. *Emergent Abilities of Large Language Models*. arXiv:2206.07682 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2206.07682.

[24] Minju Park et al. *Empowering Personalized Learning through a Conversation-based Tutoring System with Student Modeling*. en. Mar. 2024. DOI: 10.1145/3613905.3651122.

[25] Leonard Salewski et al. *In-Context Impersonation Reveals Large Language Models' Strengths and Biases*. arXiv:2305.14930 [cs]. Nov. 2023. DOI: 10.48550/arXiv.2305.14930.

[26] Hyacinth S. Nwana. "Intelligent tutoring systems: an overview". en. In: *Artificial Intelligence Review* 4.4 (Dec. 1990), pp. 251–277. ISSN: 1573-7462. DOI: 10.1007/BF00168958.

[27] Robin Schmucker et al. *Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems*. arXiv:2310.01420 [cs]. Nov. 2023. DOI: 10.48550/arXiv.2310.01420.

[28] Atharva Naik et al. *Generating Situated Reflection Triggers about Alternative Solution Paths: A Case Study of Generative AI for Computer-Supported Collaborative Learning*. arXiv:2404.18262 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2404.18262.

[29] Soya Park, Hari Subramonyam, and Chinmay Kulkarni. *Thinking Assistants: LLM-Based Conversational Assistants that Help Users Think By Asking rather than Answering*. arXiv:2312.06024 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2312.06024.

[30] Michel Beaudouin-Lafon. "Instrumental interaction: an interaction model for designing post-WIMP user interfaces". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. CHI '00. New York, NY, USA: Association for Computing Machinery, Apr. 2000, pp. 446–453. ISBN: 978-1-58113-216-8. DOI: 10.1145/332040.332473.

[31] Eric Horvitz. "Principles of Mixed-Initiative User Interfaces". en-US. In: May 1999, pp. 159–166.

[32] Shashank Sonkar, Naiming Liu, and Richard G. Baraniuk. *Student Data Paradox and Curious Case of Single Student-Tutor Model: Regressive Side Effects of Training LLMs for Personalized Learning*. arXiv:2404.15156 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2404.15156.

[33] Hasan Abu-Rasheed et al. *Supporting Student Decisions on Learning Recommendations: An LLM-Based Chatbot with Knowledge Graph Contextualization for Conversational Explainability and Mentoring*. arXiv:2401.08517 [cs]. Jan. 2024. DOI: 10.48550/arXiv.2401.08517.

[34] Harsh Kumar et al. "Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW2 (Nov. 2024). arXiv:2310.13712 [cs], pp. 1–30. ISSN: 2573-0142. DOI: 10.1145/3687038.

[35] Zachary A. Pardos and Shreya Bhandari. *Learning gain differences between ChatGPT and human tutor generated algebra hints*. arXiv:2302.06871 [cs]. Feb. 2023. DOI: 10.48550/arXiv.2302.06871.

[36] Jaime R. Carbonell. *Mixed-Initiative Man-Computer Instructional Dialogues. Final Report*. en. Tech. rep. ERIC Number: ED040585. May 1970.

[37] John Self. "Theoretical foundations for intelligent tutoring systems | Journal of Artificial Intelligence in Education". In: *Journal of Artificial Intelligence in Education* (1990).

[38] Alaa N. Akkila et al. "Survey of Intelligent Tutoring Systems Up To the End of 2017". en. In: *International Journal of Engineering and Information Systems (IJEAIS)* 3.4 (Apr. 2019). Publisher: IJARW, pp. 36–49.

[39] Kenneth R. Koedinger and Vincent Aleven. "Exploring the Assistance Dilemma in Experiments with Cognitive Tutors". en. In: *Educational Psychology Review* 19.3 (Sept. 2007). 540 citations (Semantic Scholar/DOI) [2024-09-14], pp. 239–264. ISSN: 1040-726X, 1573-336X. DOI: 10.1007/s10648-007-9049-0.

[40] Grace Guo et al. "Visualizing Intelligent Tutor Interactions for Responsive Pedagogy". In: *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*. arXiv:2404.12944 [cs]. June 2024, pp. 1–9. DOI: 10.1145/3656650.3656667.

[41] Kay G. Schulze et al. "Andes: An Intelligent Tutor for Classical Physics". en. In: *Journal of Electronic Publishing* 6.1 (Sept. 2000). Publisher: Michigan Publishing, University of Michigan Library. ISSN: 1080-2711. DOI: 10.3998/3336451.0006.110.

[42] Arthur C Graesser et al. "AutoTutor: A simulation of a human tutor". In: *Cognitive Systems Research* 1.1 (Dec. 1999), pp. 35–51. ISSN: 1389-0417. DOI: 10.1016/S1389-0417(99)00005-4.

[43] Nigel Bosch et al. "Students' Verbalized Metacognition During Computerized Learning". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–12. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445809.

[44] Noboru Matsuda et al. "Cognitive anatomy of tutor learning: Lessons learned with SimStudent". In: *Journal of Educational Psychology* 105.4 (2013). Place: US Publisher: American Psychological Association, pp. 1152–1163. ISSN: 1939-2176. DOI: 10.1037/a0031955.

[45] Kenneth Holstein and Vincent Aleven. *Designing for human-AI complementarity in K-12 education*. arXiv:2104.01266 [cs]. July 2021. DOI: 10.48550/arXiv.2104.01266.

[46] Ido Roll et al. “Metacognitive Practice Makes Perfect: Improving Students’ Self-Assessment Skills with an Intelligent Tutoring System”. en. In: *Artificial Intelligence in Education*. Ed. by Gautam Biswas et al. Berlin, Heidelberg: Springer, 2011, pp. 288–295. ISBN: 978-3-642-21869-9. DOI: 10.1007/978-3-642-21869-9_38.

[47] Alan L. Tharp and Woodrow E. Robbins. “Using computers in a natural language mode for elementary education”. In: *International Journal of Man-Machine Studies* 7.6 (Nov. 1975), pp. 703–725. ISSN: 0020-7373. DOI: 10.1016/S0020-7373(75)80034-4.

[48] Valerie J. Shute and Joseph Psotka. “Intelligent Tutoring Systems: Past, Present, and Future.” In: Fort Belvoir, VA: Defense Technical Information Center, May 1994. DOI: 10.21236/ADA280011.

[49] Douglas C. Merrill et al. “Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems”. en. In: *Journal of the Learning Sciences* 2.3 (July 1992). 311 citations (Semantic Scholar/DOI) [2024-09-14], pp. 277–305. ISSN: 1050-8406, 1532-7809. DOI: 10.1207/s15327809jls0203_2.

[50] Benedict du Boulay. “Artificial Intelligence as an Effective Classroom Assistant”. In: *IEEE Intelligent Systems* 31.6 (Nov. 2016). Conference Name: IEEE Intelligent Systems, pp. 76–81. ISSN: 1941-1294. DOI: 10.1109/MIS.2016.93.

[51] Rose E. Wang et al. *Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise*. arXiv:2410.03017 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.03017.

[52] Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. “AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails”. en. In: *Proceedings of the Eleventh ACM Conference on Learning @ Scale*. 2 citations (Semantic Scholar/DOI) [2024-09-14]. Atlanta GA USA: ACM, July 2024, pp. 5–15. ISBN: 9798400706332. DOI: 10.1145/3657604.3662041.

[53] Romain Puech et al. *Towards the Pedagogical Steering of Large Language Models for Tutoring: A Case Study with Modeling Productive Failure*. arXiv:2410.03781 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.03781.

[54] Vasile Rus et al. “Recent Advances in Conversational Intelligent Tutoring Systems”. en. In: *AI Magazine* 34.3 (Sept. 2013). Number: 3, pp. 42–54. ISSN: 2371-9621. DOI: 10.1609/aimag.v34i3.2485.

[55] Michelene T.H. Chi et al. “Learning from human tutoring”. en. In: *Cognitive Science* 25.4 (2001), pp. 471–533. ISSN: 1551-6709. DOI: 10.1207/s15516709cog2504_1.

[56] Shaaron Ainsworth. “DeFT: A conceptual framework for considering learning with multiple representations”. en. In: *Learning and Instruction* 16.3 (June 2006), pp. 183–198. ISSN: 09594752. DOI: 10.1016/j.learninstruc.2006.03.001.

[57] Mitchell J. Nathan and Anthony Petrosino. “Expert Blind Spot Among Preservice Teachers”. en. In: *American Educational Research Journal* 40.4 (Jan. 2003), pp. 905–928. ISSN: 0002-8312, 1935-1011. DOI: 10.3102/00028312040004905.

[58] Albert T. Corbett and John R. Anderson. “Knowledge tracing: Modeling the acquisition of procedural knowledge”. en. In: *User Modeling and User-Adapted Interaction* 4.4 (Dec. 1994), pp. 253–278. ISSN: 1573-1391. DOI: 10.1007/BF01099821.

[59] Ido Roll and Ruth Wylie. “Evolution and Revolution in Artificial Intelligence in Education”. en. In: *International Journal of Artificial Intelligence in Education* 26.2 (June 2016), pp. 582–599. ISSN: 1560-4306. DOI: 10.1007/s40593-016-0110-3.

[60] John R. Anderson. “The Architecture of Cognition”. en. In: Edition: 0. Psychology Press, Nov. 2013. ISBN: 978-1-317-75953-9. DOI: 10.4324/9781315799438.

[61] Mark R. Lepper and Maria Woolverton. “Chapter 7 - The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors”. In: *Improving Academic Achievement*. Ed. by Joshua Aronson. Educational Psychology. San Diego: Academic Press, Jan. 2002, pp. 135–158. DOI: 10.1016/B978-012064455-1/50010-5.

[62] A. Kay and A. Goldberg. “Personal Dynamic Media”. In: *Computer* 10.3 (Mar. 1977). Conference Name: Computer, pp. 31–41. ISSN: 1558-0814. DOI: 10.1109/C-M.1977.217672.

[63] Richard Sutton. “The bitter lesson”. In: *Incomplete Ideas (blog)* 13.1 (2019), p. 38.

[64] Monty Newborn. *Deep Blue: An Artificial Intelligence Milestone | SpringerLink*. English. 1st ed. Springer New York, NY. ISBN: 978-1-4684-9568-3.

[65] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. en. In: *Nature* 521.7553 (May 2015). Publisher: Nature Publishing Group, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539.

[66] Kurt Vanlehn. “The Behavior of Tutoring Systems”. In: *Int. J. Artif. Intell. Ed.* 16.3 (Aug. 2006), pp. 227–265. ISSN: 1560-4292.

[67] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. arXiv:2108.07258 [cs]. July 2022. DOI: 10.48550/arXiv.2108.07258.

[68] Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. “Deep learning for AI”. In: *Association for Computing Machinery* 64.7 (2021). ISSN: 0001-0782. DOI: 10.1145/3448250.

[69] Aakanksha Chowdhery et al. *PaLM: Scaling Language Modeling with Pathways*. arXiv:2204.02311 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2204.02311.

[70] Sahil. "Live Machinery: An Interface Design Philosophy for Wholesome AI Futures". en. In: (Nov. 2024).

[71] Brenda Laurel. *Computers as Theatre*. 75 Arlington Street, Suite 300 Boston, MA, United States: Addison-Wesley Longman Publishing Co., Inc., Jan. 1991. ISBN: 978-0-201-51048-5.

[72] Liang Zeng. "Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Edition) by B. Shneiderman and C. Plaisant". In: *International Journal of Human–Computer Interaction* 25.7 (Sept. 2009). Publisher: Taylor & Francis, pp. 707–708. ISSN: 1044-7318. DOI: 10.1080/10447310903187949.

[73] Douglas C Engelbart. "Augmenting human intellect: A conceptual framework". In: *Augmented Education in the Global Age*. Routledge, 2023, pp. 13–29.

[74] Pierre Dillenbourg. "Design for classroom orchestration". In: *Computers & Education* 69 (2013), pp. 485–492. ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2013.04.013.

[75] Christopher Alexander. *The Timeless Way of Building*. en. Google-Books-ID: H6CE9hlbO8sC. Oxford University Press, 1979. ISBN: 978-0-19-502402-9.

[76] OpenAI et al. *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. Mar. 2024. DOI: 10.48550/arXiv.2303.08774.

[77] Ethan Prihar and Morgan Lee. "Comparing different approaches to generating mathematics explanations using large language models". In: *International Conference on Artificial Intelligence in Education* (2023), pp. 290–295.

[78] Lisa P. Argyle et al. "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3 (July 2023). 323 citations (Semantic Scholar/arXiv) [2024-09-14] 323 citations (Semantic Scholar/DOI) [2024-09-14] arXiv:2209.06899 [cs], pp. 337–351. ISSN: 1047-1987, 1476-4989. DOI: 10.1017/pan.2023.2.

[79] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. 206 citations (Semantic Scholar/arXiv) [2024-09-14] 37 citations (Semantic Scholar/DOI) [2024-09-14] arXiv:2208.10264 [cs]. July 2023. DOI: 10.48550/arXiv.2208.10264.

[80] Saffron Huang et al. "Collective Constitutional AI: Aligning a Language Model with Public Input". en. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. Rio de Janeiro Brazil: ACM, June 2024, pp. 1395–1417. ISBN: 9798400704505. DOI: 10.1145/3630106.3658979.

[81] Byungsoo Kim et al. “AI-Driven Interface Design for Intelligent Tutoring System Improves Student Engagement”. In: *ArXiv* (Sept. 2020). DOI: arXiv:2009.08976.

[82] Maxwell Nye et al. *Show Your Work: Scratchpads for Intermediate Computation with Language Models*. 533 citations (Semantic Scholar/arXiv) [2024-09-14] arXiv:2112.00114 [cs]. Nov. 2021. DOI: 10.48550/arXiv.2112.00114.

[83] Albert T. Corbett and John R. Anderson. “Knowledge tracing: Modeling the acquisition of procedural knowledge”. en. In: *User Modelling and User-Adapted Interaction* 4.4 (1995). 1932 citations (Semantic Scholar/DOI) [2024-09-14], pp. 253–278. ISSN: 0924-1868, 1573-1391. DOI: 10.1007/BF01099821.

[84] Ryan S. Baker. “AI and self-regulated learning theory: What could be on the horizon?” en. In: *Computers in Human Behavior* 147 (Oct. 2023). 0 citations (Semantic Scholar/DOI) [2024-09-14], p. 107849. ISSN: 07475632. DOI: 10.1016/j.chb.2023.107849.

[85] Chris Piech et al. “Deep Knowledge Tracing”. In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc., 2015.

[86] Swarnadeep Saha, Peter Hase, and Mohit Bansal. *Can Language Models Teach Weaker Agents? Teacher Explanations Improve Students via Personalization*. arXiv:2306.09299 [cs]. Nov. 2023. DOI: 10.48550/arXiv.2306.09299.

[87] Gautam Yadav, Ying-Jui Tseng, and Xiaolin Ni. *Contextualizing Problems to Student Interests at Scale in Intelligent Tutoring System Using Large Language Models*. arXiv:2306.00190 [cs]. May 2023. DOI: 10.48550/arXiv.2306.00190.

[88] Annabel Marie Latham. “Personalising Learning with Dynamic Prediction and Adaptation to Learning Styles in a Conversational Intelligent Tutoring System”. en. doctoral. Manchester Metropolitan University, Dec. 2011.

[89] Martin Klissarov et al. *On the Modeling Capabilities of Large Language Models for Sequential Decision Making*. arXiv:2410.05656 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.05656.

[90] Jionghao Lin et al. *Using Large Language Models to Provide Explanatory Feedback to Human Tutors*. arXiv:2306.15498 [cs]. June 2023. DOI: 10.48550/arXiv.2306.15498.

[91] Karl Cobbe et al. *Training Verifiers to Solve Math Word Problems*. arXiv:2110.14168 [cs]. Nov. 2021. DOI: 10.48550/arXiv.2110.14168.

[92] Angelica Chen et al. *Two Failures of Self-Consistency in the Multi-Step Reasoning of LLMs*. arXiv:2305.14279 [cs]. Feb. 2024. DOI: 10.48550/arXiv.2305.14279.

[93] Manas Gaur and Amit Sheth. *Building Trustworthy NeuroSymbolic AI Systems: Consistency, Reliability, Explainability, and Safety*. arXiv:2312.06798 [cs]. Dec. 2023. DOI: 10.48550/arXiv.2312.06798.

[94] Silvia García-Méndez, Francisco De Arriba-Pérez, and María Del Carmen Somoza-López. "A Review on the Use of Large Language Models as Virtual Tutors". en. In: *Science & Education* (May 2024). 2 citations (Semantic Scholar/DOI) [2024-09-14]. ISSN: 0926-7220, 1573-1901. DOI: 10.1007/s11191-024-00530-2.

[95] Anastasia Olga et al. *Generative AI: Implications and Applications for Education*. arXiv:2305.07605 [cs]. May 2023. DOI: 10.48550/arXiv.2305.07605.

[96] Lixiang Yan et al. "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review". In: *British Journal of Educational Technology* 55.1 (Jan. 2024). arXiv:2303.13379 [cs], pp. 90–112. ISSN: 0007-1013, 1467-8535. DOI: 10.1111/bjet.13370.

[97] Changrong Xiao and Wenxing Ma. "Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs". In: (2024). DOI: arXiv:2401.06431.

[98] Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. "CIMA: A Large Open Access Dialogue Dataset for Tutoring". en. In: *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Seattle, WA, USA → Online: Association for Computational Linguistics, 2020, pp. 52–64. DOI: 10.18653/v1/2020.bea-1.5.

[99] OpenAI et al. *GPT-4o System Card*. arXiv:2410.21276 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2410.21276.

[100] Gemini Team et al. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. arXiv:2403.05530 [cs]. Dec. 2024. DOI: 10.48550/arXiv.2403.05530.

[101] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. arXiv:2407.21783 [cs]. Nov. 2024. DOI: 10.48550/arXiv.2407.21783.

[102] AayushKucheria. *tutorActionSimulation*. 2024.

[103] John Sweller, Jeroen J. G. van Merrienboer, and Fred G. W. C. Paas. "Cognitive Architecture and Instructional Design". en. In: *Educational Psychology Review* 10.3 (Sept. 1998), pp. 251–296. ISSN: 1573-336X. DOI: 10.1023/A:1022193728205.