Comparing Behavioral Patterns of LLM and Human Tutors: A Population-level Analysis with the CIMA Dataset

Aayush Kucheria
Aalto University
aayush.kucheria@gmail.com

Nitin Sawhney
Aalto University
nitin.sawhney@uniarts.fi

Arto Hellas
Aalto University
arto.hellas@aalto.fi

Abstract

Large Language Models (LLMs) offer exciting potential as educational tutors, and much research explores this potential. Unfortunately, there's little research in understanding the baseline behavioral pattern differences that LLM tutors exhibit, in contrast to human tutors. We conduct a preliminary study of these differences with the CIMA dataset and three state-of-the-art LLMs (GPT-4o, Gemini Pro 1.5, and LLaMA 3.1 450B). Our results reveal systematic deviations in these baseline patterns, particulary in the tutoring actions selected, complexity of responses, and even within different LLMs.

This research brings forward some early results in understanding how LLMs when deployed as tutors exhibit systematic differences, which has implications for educational technology design and deployment. We note that while LLMs enable more powerful and fluid interaction than previous systems, they simultaneously develop characteristic patterns distinct from human teaching. Understanding these differences can inform better integration of AI in educational settings.

1 Introduction

Large Language Models (LLMs) offer unprecedented capabilities for creating educational technologies that can interact with students. Unlike traditional intelligent tutoring systems (ITS), which were often limited by constrained interfaces and rigid interaction patterns (Alkhatlan and Kalita, 2019; Mousavinasab et al., 2021), LLMs provide natural-language interactions that draw on extensive linguistic and contextual training (Brown et al., 2020; Bommasani et al., 2022). This allows LLMs to respond to learner inputs in ways that more closely resemble human tutors, presenting new possibilities for personalized learning experiences.

Despite their potential, important questions remain about how closely LLM tutoring interactions align with human tutoring practices. Existing literature on human tutoring and ITSs emphasize strategies such as scaffolding, immediate feedback, and adaptive questioning to meet the learners' needs (Chi et al., 2001; VanLehn, 2011). However, the conversational and pedagogical behaviors of LLMs in tutoring scenarios remain underexplored.

The current work addresses this research gap. Utilizing the CIMA dataset of language teaching dialogues (Stasaski et al., 2020), which contains multiple responses of human tutors to the same students in an Italian language learning context, we systematically examine and compare the structural pedagogical patterns of human tutors and several state-of-the-art language models, GPT-4o (OpenAI et al., 2024), Gemini Pro 1.5 (Team et al., 2024), and LLaMA 3.1 405B (Grattafiori et al., 2024). We identify and characterize behavioral patterns of LLM tutors and human tutors, focusing on action preferences and response complexity.

Our analysis reveals several key findings:

1. Both human and AI tutors show similar high-level preferences in action selection, with hints comprising approximately 45% of all tutoring actions.

2. Human tutors strongly prefer single-action responses (approximately 72% of interactions), while LLM tutors consistently combine multiple pedagogical actions in their responses.

3. Each LLM exhibits its own characteristic pattern, highlighting the need for LLM-specific tailoring.

As these systems continue to evolve and be deployed in diverse learning contexts, recognizing their distinctive behavioral patterns becomes increasingly important—not to eliminate differences, but to use them more effectively in creating educational experiences that complement human instruction.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 873–881
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

2 Related Work

2.1 Evolution of Intelligent Tutoring Systems

Intelligent Tutoring Systems (ITS) have evolved significantly over decades, from rule-based systems with limited interaction capabilities to increasingly sophisticated architectures. Traditional ITS platforms like Cognitive Tutors (Anderson et al., 1995) and knowledge-based tutors (Akkila et al., 2019) demonstrated effectiveness in specific domains but were constrained by rigid interaction patterns and limited adaptability. These systems typically operated within carefully engineered knowledge frameworks, making them powerful but inflexible (VanLehn, 2011; Ma et al., 2014).

The field has progressively sought more natural and adaptive educational technologies. Dialog-based tutoring systems (Graesser et al., 1999; Rus et al., 2013) attempted to incorporate conversational elements but remained limited by predefined pathways. Recent advances in NLP have enabled more sophisticated systems capable of processing and generating natural language interactions (Rus et al., 2013; Nye et al., 2014), setting the stage for the current generation of LLM-based educational tools.

2.2 Language Models in Educational Applications

Large Language Models represent a fundamental shift in educational technology, offering unprecedented fluidity in natural language interaction coupled with broad knowledge coverage. Recent research has explored various applications of LLMs in education, including personalized learning (Park et al., 2024), assessment (Wang et al., 2024), and tutoring (Kumar et al., 2024).

Studies have demonstrated LLMs' potential to support complex learning processes through adaptive dialogue (Schmucker et al., 2023) and to generate contextually relevant explanations (Naik et al., 2024). LLMs' performance as educational tools has primarily been studied through various metrics such as learning gain (Pardos and Bhandari, 2023) or through assessing the quality or correctness of LLM responses (Kumar et al., 2024).

However, while these systems enable more natural interaction, they simultaneously operate according to statistical patterns learned during training rather than pedagogical principles explicitly encoded by designers (Brown et al., 2020; Bommasani et al., 2022). This tension between fluid interfaces and underlying fixed statistical patterns remains underexplored in educational applications of LLMs.

2.3 Tutoring Patterns and Behaviors

Research on human tutoring has extensively documented the patterns that characterize effective teaching interactions. Chi et al. (2001) identified interactive patterns like scaffolding and feedback loops that support student learning. VanLehn (2011) further explored the balance between different pedagogical moves, noting that expert tutors dynamically adjust their approach based on student needs. Feedback, specifically, has been widely studied, with Hattie and Timperley (2007) emphasizing its critical role in facilitating student learning through targeted interventions.

In comparing AI and human tutoring behaviors, early work by Graesser et al. (1999) examined differences between human tutors and AutoTutor, finding systematic differences in questioning strategies and elaboration patterns. More recent work by Stasaski et al. (2020) with the CIMA dataset highlighted the diversity of valid teaching approaches human tutors employ, noting the low agreement rate (18.1%) between different tutors responding to the same student input. This underscores the complexity of establishing normative patterns for tutoring behavior.

2.4 Interaction Patterns in Language Models

Research on conversational behavior and dialogue generation in LLMs has identified patterns related to turn-taking, conversational coherence, and response complexity (Sandler et al., 2024; Shaikh et al., 2023). These studies highlight that while LLMs produce coherent interactions, the underlying statistical nature can lead to repetitive patterns and superficial dialogues – this behavior has, in part, also led to LLMs being labeled as "stochastic parrots" (Bender et al., 2021).

The few studies that have examined instructional patterns in AI systems have typically focused on direct comparisons of specific responses rather than population-level analysis of behavioral distributions (Puech et al., 2024). These findings emphasize the need to systematically analyze LLM interaction patterns to better understand their educational utility and identify areas for improvement.

2.5 Research Gap

Our research addresses the need to systematically analyze LLM interaction patterns by conducting a detailed comparison of human and LLM tutoring patterns across multiple dimensions of analysis, focusing on action distributions, response complexity, and teaching dynamics. This population-level approach provides a new perspective on how LLMs function in educational contexts compared to human tutors, with implications for both educational technology design and pedagogical theory.

3 Methodology

3.1 Research Questions

This study investigates differences between how language models and humans approach the tutoring task. We examine the underlying patterns in how these systems engage with learners compared to human tutors. This focus can be broken down into specific questions in light 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?

These questions address core theoretical interests about the nature of LLMs as 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.

3.2 Design Principles

Our methodology is shaped by several key principles:

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 in the CIMA dataset.

Reference Distribution Approach: We aggregate human tutor responses to create reference distributions that capture the characteristic patterns of human tutoring behavior. These distributions serve as a baseline for comparative analysis.

Model Comparison: We maintain separate distributions for different LLM configurations, enabling us to distinguish between model-specific behaviors and general LLM characteristics.

This approach 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?".

3.3 Dataset

Our analysis utilizes the CIMA (Conversational Instruction with Multi-responses and Actions) dataset (Stasaski et al., 2020), which provides tutoring dialogues focused on teaching Italian prepositional phrases to English speakers. The dataset is particularly valuable for our study 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.

Key features of the dataset include:

- **Multiple Valid Approaches**: For each student utterance, three different tutors provide responses, showing distinct but equally valid tutoring strategies.

- **Action Labeling**: Each response is annotated with pedagogical actions (Hint, Question, Correction, Confirmation, Other).

- **Progressive Learning**: The dataset captures how concepts build across exercises.

The dataset contains 391 completed exercises across 77 students, with each exercise grounded in both visual and conceptual representations. The mean response lengths (6.82 words for students, 9.99 words for tutors) indicate substantive interactions. This richness, combined with explicit action labeling, provides a strong foundation for analyzing how different tutors structure their teaching interventions.

3.4 Dataset Enhancement with AI Tutors

To enable direct comparison between human and artificial tutoring patterns, we enhanced the CIMA dataset by generating parallel responses from state-of-the-art language models. We selected three advanced instruction-tuned models:

- GPT-4o 2024-08-06 (OpenAI) (OpenAI et al., 2024)

- Gemini Pro 1.5 (Google) (Team et al., 2024).
- LLaMA 3.1 405B instruct nitro (Meta) (Grattafiori et al., 2024)

This selection 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.

For response generation, 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:
>
> - 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 approach ensures consistent action categorization and response formats across all interactions.

3.5 Analysis Framework

Our analysis examines two key dimensions of tutoring behavior:

- **Action Distribution Analysis**: We examine the relative frequency of fundamental tutoring actions across different populations. This analysis compares the baseline distribution derived from human tutors against Language Model behavior, identifying systematic preferences or avoidances in action selection.
- **Action Combination Analysis**: We investigate patterns in how actions are combined within individual responses, including the typical number of actions per response and the balance between single-action and multi-action responses.

3.6 Methodological Limitations

Our analysis framework operates within several important constraints:

- **Dataset Characteristics**: The study utilizes a dataset limited to Italian preposition instruction with crowdsourced rather than professional tutors.
- **Structural Constraints**: The prescribed JSON response format may influence natural interaction patterns, and the restricted action vocabulary limits expressive range.
- **Model Implementation**: Analysis is limited to three model variants with a single prompt template approach and no model fine-tuning.
- **Scope of Conclusions**: While we can identify alignment or deviation from human behavioral patterns, we cannot evaluate the optimality of tutoring choices or assess the quality of specific responses.

Our focus on action distributions represents a deliberate methodological choice, prioritizing the analysis of strategic-level behavioral alignment over response-level quality assessment.

4 Analysis

Our analysis revealed systematic differences in how language models and human tutors approach the educational task, with patterns emerging across multiple dimensions of analysis.

4.1 Action Distributions

Both human and AI tutors demonstrate a strong preference for hints as their primary teaching action, with hints comprising approximately 45% of all actions across both human and LLM sessions (Figure 1). This suggests fundamental alignment in basic tutoring strategy, possibly reflecting the effectiveness of scaffolded guidance over direct instruction.

However, examining the broader action distributions reveals key differences in pedagogical approaches. Human tutors show a more balanced

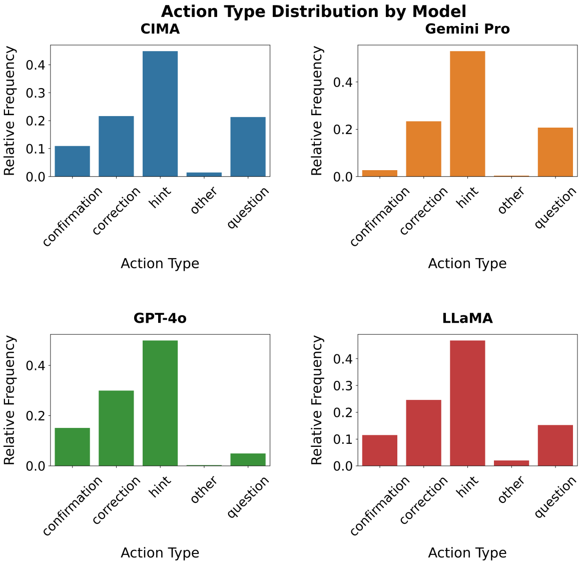

Figure 1: Distribution of actions by different tutors, showing the relative frequency of different pedagogical strategies.

distribution between corrections (20.3%) and questions (21.5%), suggesting a diverse approach. In contrast, AI systems exhibit model-specific patterns - while all maintain the primacy of hints, they differ in secondary strategies. GPT-4o and Gemini Pro 1.5 demonstrate a stronger tendency toward corrections (28.7% and 29.4% respectively) compared to questions (7.3% and 6.8%), while LLaMA 3.1 maintains a more balanced profile closer to human tutors.

Statistical analysis confirmed that the observed differences in action distributions between human and AI tutors were significant (χ^2 = 495.17, p < .001, Cramer's V = 0.124), indicating a weak to moderate effect size. This suggests an interesting pattern: while there is fundamental alignment in primary teaching strategies (the preference for hints), significant differences emerge in how secondary strategies are deployed. This nuanced finding reveals that LLMs have captured core aspects of tutoring behavior while diverging in other dimensions.

4.2 Response Complexity

The most striking difference between human and AI tutors emerges in response complexity (Figure 2). Human tutors demonstrate a strong and consistent pattern for single-action responses, with approximately 71.8% of responses containing just one action, 24.6% containing two actions, and only 3.6% containing three or more. This pattern suggests a teaching strategy focused on clear, targeted interventions.

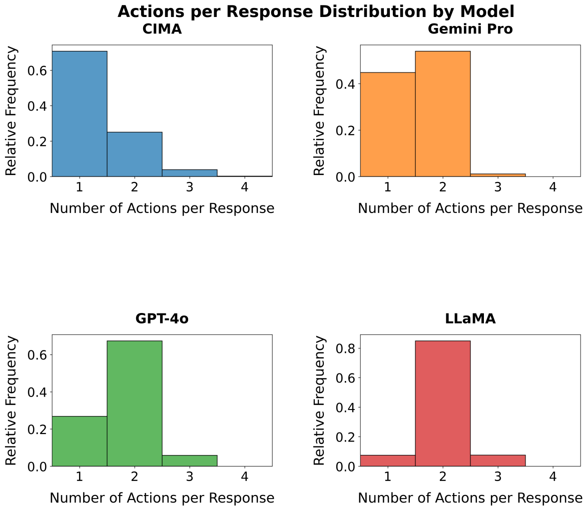

Figure 2: Distribution of the number of pedagogical actions per response in tutoring sessions.

In contrast, AI tutors consistently combine multiple actions in their responses, though with interesting variations between systems. LLaMA shows the strongest preference for dual-action responses (82.3%), while GPT-4o and Gemini Pro display a more balanced distribution. GPT-4o uses single actions in about 31.5% of responses and dual actions in 64.7%, while Gemini Pro shows a more even split between single (42.8%) and dual actions (54.9%).

A Kruskal-Wallis test revealed significant differences in the number of actions per response across the four tutor types (human and three LLMs) (H = 1507.37, p < .001). Post-hoc pairwise comparisons with Bonferroni correction showed significant differences between humans and each LLM (p < .001 for all comparisons), as well as between all pairs of LLMs (p < .001). This confirms that not only do AI tutors differ from human tutors in response complexity, but each AI model exhibits its own statistically distinct pattern in how it structures responses.

5 Discussion

5.1 Summary of Research Findings

Our study provides a comparative population-level view of how LLM tutors and human tutors approach the teaching task.

Our first finding is that both human tutors and LLM tutors share a high-level strategy, where hints are the main tutoring action (approximately 45%

of all actions each). This suggests that LLMs have learned to prioritize guidance much like human experts. The secondary actions show some differences. Human tutors use a somewhat balanced mix of questions and corrections in the interactions (roughly 20% each), indicating an approach that alternates between direct feedback and prompting student thinking. For the secondary actions, LLM tutors show skewed distributions; for example, GPT-4o and Gemini 1.5 rely more heavily on corrections, whereas LLaMa 3.1 maintained a more human-like balance.

These differences in action preference suggest that while LLMs have captured the primary tactic of hinting, they diverge in how they follow up, either by explaining or correcting or by probing the learner.

The second finding is the strong contrast in response complexity between human tutors and LLMs. Human tutors strongly prefer a concise, single-action response (roughly 70% of human tutor responses in the dataset had only one pedagogical action). In comparison, LLM tutors frequently combine multiple actions in a single response; the difference was the strongest for LLaMa 3.1, where over 80% of the responses had two actions). Statistical tests confirmed that these differences in response complexity are significant.

The third finding is that LLMs have unique behavioral signatures. Although the three evaluated LLMs have been trained with large masses of data, each had their distinct tutoring style. This highlights that the way how an LLM interacts reflects the model's design choices or fine-tuning. These results extend the prior observations by Graesser et al. (1999), who noted systematic differences in tutoring style between a classical ITS (AutoTutor) and human tutors. We find that LLM-based tutors likewise deviate from human tutors.

5.2 Pedagogical and Practical Implications

The differences identified in our analysis have implications for educational practice and the design of AI tutoring systems. First, the alignment of primary strategy in terms of heavy use of hints highlights that LLMs have converged on a generally effective tutoring practice. This is encouraging from a pedagogical point of view, as hints are known to facilitate learning by prompting student thinking (Chi et al., 2001).

However, the way how LLM tutors use secondary strategies could affect learning in subtle ways. For example, the LLMs were more likely to provide corrections, and asked prompting questions less frequently than human tutors. Asking questions is often used to encourage active learning – if an LLM tutor predominantly gives corrections, the student might become more passive in the learning process.

On the other hand, providing rapid corrections can be also be beneficial, depending on the scenario. The pedagogical implication is that LLM tutors should be tailored to the contexts and objectives: if the objective is to foster student reasoning, LLMs should be tweaked to ask more open-ended questions rather than providing quick fixes. Furthermore, compound responses might overwhelm the learner, and to avoid this, LLMs should be adapted to match the user competences. That is, there is room for improvement in the pedagogical quality and ability of LLM-driven tutors.

Broadly speaking, our results emphasize that LLM tutors, despite the fluent dialogue, have embedded biases in how they tutor. This resonates with the tension noted by Horvitz between fluid and natural interfaces and the rigid patterns of automated systems (Horvitz, 1999).

5.3 Limitations

Our work comes with a set of limitations, which we acknowledge. Firstly, our study focuses on a single dataset and domain, i.e. the CIMA dataset of Italian language learning dialogues (Stasaski et al., 2020). The tutoring patterns that we focused on (for both humans and LLMs) may be specific to language teaching or even to particular prompts and tasks in CIMA, and it is possible that the balance of actions and complexity would be different to other datasets. This means that the generalizability of the results should be assessed with additional contexts and datasets.

Secondly, our analysis focused on population-level comparisons, but it does not capture how a tutor might adapt over a tutoring session. Human tutors often dynamically adjust their strategy based on students' progress, but we do not know to what extent this holds for LLM tutors, and our current analysis misses these dynamics.

Thirdly, our annotation strategy was automatic, and relied on the existing categories in the CIMA dataset. It is possible that LLM (or human) actions do not always neatly fall into specific categories. We sought to mitigate this by using clear definitions, but acknowledge the presence of noise. Additionally, we relied on LLMs' self-reported action classifications without manual validation. While our population-level patterns are robust to some classification noise, future work should validate the accuracy of these self-classifications. Furthermore, our analysis does not capture subtler nuances in responses; as an example, a human tutor might provide a more encouraging response than an LLM, even if both responses are categorized as hints.

Additionally, as the CIMA dataset was released in 2020 and our tested LLMs were trained on data through 2023-2024, it is possible that the dataset appeared in their training corpora. While this does not invalidate our behavioral analysis—the patterns we observe reflect how these models approach tutoring tasks regardless of prior exposure—it should be considered when interpreting the alignment between LLM and human tutoring strategies.

Finally, we cannot deduce the efficiency of the tutoring. This is a limitation of the practical significance of the work. Despite these issues, our work fills a gap by systematically comparing the baseline behaviors of human and LLM tutors.

6 Conclusion

In this paper, we studied LLM-based tutors differ from human tutors in their interaction patterns, conducting a population-level analysis using the CIMA tutoring dataset. Our focus was on the behavioral structure of the tutoring, composed of what actions the tutors take and how they deliver them.

By generating parallel tutoring responses using three state-of-the-art LLMs and comparing them against human tutor responses, we observe the following: (1) LLM tutors and human tutors have similar high-level tactics with a shared emphasis on giving hints, which indicates that current LLMs have learned or been tuned to adopt some of the existing practices of humans; (2) When going beyond the high-level tactics, there are significant differences in how LLM tutors balance their actions and in how complex the responses are; (3) The differences are not uniform across the LLM tutors, which highlights that each LLM has its own "personal" style of tutoring. These findings were made possible by analyzing the aggregate patterns over many tutoring responses, moving beyond anecdotal or one-to-one comparisons.

Recognizing and understanding these patterns is important when seeking to make informed decisions on how to effectively integrate LLMs into learning environments. The differences that we highlight suggest areas where LLM tutors might benefit from additional tailoring (e.g. tailoring LLMs to the context and objectives, and to match the user competences).

In conclusion, we present a foundation for treating behavioral patterns of LLM tutors as a subject of study by its own right, parallel to how one might study different teaching styles among human tutors. We have also shown how to quantitatively characterize how an LLM "teaches". Such a characterization can help in aligning LLM tutor behavior towards educational best practices, while also benefiting from the existing capacities such as consistency and breadth.

References

Alaa N. Akkila, Abdelbaset Almasri, Adel Ahmed, Naser Masri, Yousef Abu Sultan, Ahmed Y. Mahmoud, Ihab Zaqout, and Samy S. Abu-Naser. 2019. Survey of Intelligent Tutoring Systems Up To the End of 2017. *International Journal of Engineering and Information Systems (IJEAIS)*, 3(4):36–49. Publisher: IJARW.

Ali Alkhatlan and Jugal Kalita. 2019. Intelligent Tutoring Systems: A Comprehensive Historical Survey with Recent Developments. *International Journal of Computer Applications*, 181(43):1–20. 70 citations (Semantic Scholar/DOI) [2024-09-14].

John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Ray. Pelletier. 1995. Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2):167–207. 1991 citations (Semantic Scholar/DOI) [2024-09-14].

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the Opportunities and Risks of Foundation Models. *arXiv preprint*. ArXiv:2108.07258 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Michelene T.H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

Arthur C Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, and Roger Kreuz. 1999. AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. pages 159–166.

Harsh Kumar, Ilya Musabirov, Mohi Reza, Jiakai Shi, Xinyuan Wang, Joseph Jay Williams, Anastasia Kuzminykh, and Michael Liut. 2024. Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–30. ArXiv:2310.13712 [cs].

Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4):901–918. 460 citations (Semantic Scholar/DOI) [2024-09-14].

Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163. 240 citations (Semantic Scholar/DOI) [2024-09-14] Publisher: Routledge _eprint: https://doi.org/10.1080/10494820.2018.1558257.

Atharva Naik, Jessica Ruhan Yin, Anusha Kamath, Qianou Ma, Sherry Tongshuang Wu, Charles Murray, Christopher Bogart, Majd Sakr, and Carolyn P. Rose. 2024. Generating Situated Reflection Triggers about Alternative Solution Paths: A Case Study of Generative AI for Computer-Supported Collaborative Learning. *arXiv preprint*. ArXiv:2404.18262 [cs].

Benjamin D. Nye, Arthur C. Graesser, and Xiangen Hu. 2014. AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469. 207 citations (Semantic Scholar/DOI) [2024-09-14].

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Zachary A. Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint*. ArXiv:2302.06871 [cs].

Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering Personalized Learning through a Conversation-based Tutoring System with Student Modeling.

Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2024. Towards the Pedagogical Steering of Large Language Models for Tutoring: A Case Study with Modeling Productive Failure. *arXiv preprint*. ArXiv:2410.03781 [cs].

Vasile Rus, Sidney D'Mello, Xiangen Hu, and Arthur Graesser. 2013. Recent Advances in Conversational Intelligent Tutoring Systems. *AI Magazine*, 34(3):42–54. Number: 3.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and chatgpt-generated conversations. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 366–380. Springer.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2023. Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems. *arXiv preprint*. ArXiv:2310.01420 [cs].

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2023. Grounding gaps in language model generations. *arXiv preprint arXiv:2311.09144*.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A Large Open Access Dialogue Dataset for Tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*. ArXiv:2403.05530 [cs].

Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4):197–221.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. 2024. Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. *arXiv preprint*. ArXiv:2410.03017 [cs].