From Dialogue to Mastery: Investigating Question-Asking and Interactive Learning with Large Language Models

Anonymous ACL submission

Abstract

This paper investigates the potential of large language models (LLMs) to shift from passive data absorption to active, interactive learning through simulated student-teacher dialogues. We introduce a dataset of 1,322 contexts spanning domains like song lyrics, news articles, movie plots, academic papers, and images, and analyze conversational interactions to assess the ability of LLMs to gain knowledge about these contexts. Our findings show that interactive learning significantly boosts performance, with interactive student models surpassing static learning approaches in just four dialogue turns on average. However, student models still trail behind teacher models equipped with full context knowledge. To further assess learning dynamics, we introduce the Cumulative Information Coverage (CIC) metric, revealing that more insightful questions drive better outcomes, although rigid questioning patterns remain a limitation. These findings suggest that advancing interactive learning methods and extending machine learning theories could better capture the dynamics of conversational learning, paving the way for effective machine intelligence and educational technologies.

1 Introduction

The future of machine intelligence depends on creating systems that not only learn passively from data but also engage in dynamic, interactive learning processes akin to human cognition. Language, crucial in human learning and pedagogy (Vygotsky and Cole, 1978), facilitates the active construction of knowledge through dialogues. Whether learning about new movie plots or complex academic theories, students often refine their understanding through conversational interactions with teachers, resolving ambiguities and deepening their comprehension (see Figure 1). In contrast, machine learning has predominantly followed an inductive learning approach, focusing on static datasets of labeled examples. Consequently, the role of machine learning models in dynamic settings or personalized applications has been limited.

Although some earlier efforts integrated conversational capabilities (Eric and Manning, 2017; Liu et al., 2018), they were constrained by the limitations of the NLP and generative capabilities of the time. Despite advances in large language models (LLMs) (Achiam et al., 2023; AI@Meta, 2024), their potential to learn from conversations remains underexplored. In this paper, we ask the question: *How effectively can LLMs learn new concepts through conversational interactions?*

Interactive learning marks a shift from traditional inductive learning and paradigms like active learning (Lewis and Gale, 1994; Cohn et al., 2004), enabling models to refine their understanding through dialogue. This can lead to models that better capture the complexities of adaptive learning environments. In educational technologies, LLMs simulating student-teacher interactions can provide personalized and adaptive learning experiences. Interactive learning can also enhance human-AI collaboration in fields like healthcare and research. It could also drive innovation in multimodal learning, integrating diverse data types and allowing clearer alignment with human values.

In this work, we investigate how LLMs can acquire knowledge about new concepts that were not part of their pre-training data, simulating real-world situations where AI must learn new information. These include concepts across diverse domains and modalities. We introduce a dataset comprising 1,322 contexts spanning multiple domains, including song lyrics, news articles, movie plots, academic papers, and images; all unseen by the LLMs during pretraining. This diverse collection allows for a rigorous assessment of performance in an eclectic range of scenarios and complexity levels.

We compare two modes of teaching: *static lessons*, where a teacher model provides a con-

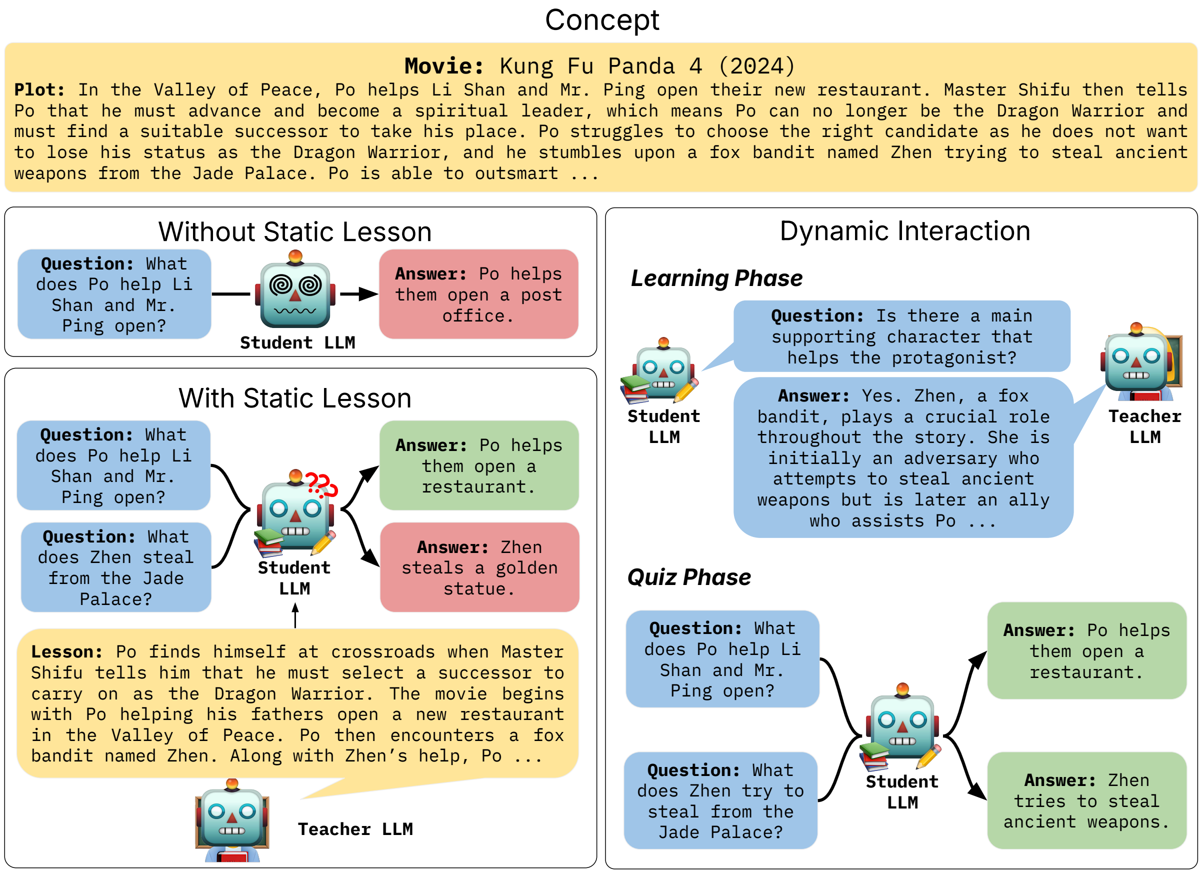

Figure 1: Given the concept of a movie plot (*top*) from a time-period outside of LLM training data, non-interactive approaches such as zero-shot prompting (*left-top*) and static lessons (*left-bottom*) fail due to lack of information or intricacies in the concept. Through dynamic interaction with a teacher (*right*), a student can learn about a concept more comprehensively to perform well on tasks.

densed summary of key content, and *dynamic interactions*, where a student LLM actively engages by posing questions. We evaluate the effectiveness of these approaches by simulating interactions between student and teacher models. To measure student learning, we compute the LLM's performance after receiving the static lesson or at the end of each dialogue turn with the teacher in the dynamic interaction setting. The latter enables a study of the ability of LLMs to ask questions that most help understanding new concepts.

Our experiments indicate that conversational interactions enable LLMs to acquire concepts with substantially greater effectiveness, outperforming static learning in 4 dialogue turns on average. However, we also observe that the performance of LLMs saturates and falls short of the performance of teacher models. Our contributions are:

- We develop a framework and datasets to assess the learning capabilities of LLMs in static and dynamic interaction setups.
- We show that LLMs consistently learn more effectively from dynamic interactions, suggesting promise for interactive learning approaches.

These findings have significant implications for advancing machine learning theory, educational technologies, and human-AI collaboration.

2 Related Work

Conversational Machine Learning. Previous works have utilized language instructions to guide machine learning tasks (Srivastava et al., 2017; Hancock et al., 2018; Arabshahi et al., 2020), typically focusing on single-turn dialogues where a student model is taught using instructions and limited labeled examples. These approaches often suffer from incomplete task understanding due to instruction complexity. Prior research addresses this with two strategies: (i) active learning through teacher annotations (Collins et al., 2008; Tamkin et al., 2022), and (ii) language-based advice or clarifications from teachers. Our work aligns with the latter, enhancing student comprehension through teacher guidance. Unlike studies that rely on external modules to generate questions based on statistical measures (Rao and Daumé III, 2018; Srivastava et al., 2019), we employ LLMs to dynamically generate contextually relevant questions, tailored to address gaps in student knowledge.

Interactive Learning with LLMs. Recent research shows that LLMs can improve task performance with human-provided explanations (Wei et al., 2022; Lampinen et al., 2022) and self-generated feedback (Madaan et al., 2023; Chen et al., 2024). Smaller LLMs also benefit from fine-tuning on explanations from larger models (Ho et al., 2023). While these studies focus on enhancing student performance through teacher-provided information (Saha et al., 2023), our research shifts the focus to the student's ability to ask informative questions, enabling more comprehensive teacher explanations. Related work includes using LLMs to learn human preferences through dialogue (Li et al., 2023) and evaluating LLMs in conversational question-answering (Abbasiantaeb et al., 2024). Our study uniquely examines LLMs' ability to engage in conversations with teachers to learn concepts across various domains, extending beyond accuracy metrics to assess the novelty and effectiveness of student questions.

3 Experiment Setup

In this section, we delineate the problem setup (§ 3.1), outline the creations of datasets (§ 3.2), describe the different interaction scenarios (§ 3.3) and models used (§ 3.4), and define our evaluation metrics (§ 3.5).

3.1 Problem Setup

In this work, a *concept* refers to a distinct unit of knowledge that captures abstract ideas or information embedded in documents across various domains such as literature, sciences, and current world events. Practically, our concepts are expressed through contexts, which comprise of detailed information pertinent to the concepts being taught. For instance, the concept of a specific movie is defined by the context of its corresponding Wikipedia article, which contains the plot, while the concept of a specific image is defined by the context provided by the image itself. For our study, we explore how a student LLM, denoted by $\mathcal{S}$, can learn concepts by interacting with a teacher, denoted by $\mathcal{T}$. The student-teacher dynamic forms a central part of our experimental design.

The student, $\mathcal{S}$ is an LLM capable of following instructions and asking open-ended, information-seeking questions. The teacher, $\mathcal{T}$, on the other hand, can be either a human expert or another language model. For the purpose of this study, the teacher is also an LLM but with one critical difference from the student LLM: the teacher has direct access to a context that allows it to respond accurately and effectively to the student's questions. For example, in the task where we teach $\mathcal{S}$ about new movies, we provide $\mathcal{T}$ access to movie plots available on Wikipedia (§3.2). By adopting this configuration of the student and teacher, we aim to isolate the effects of learning from interactions on concept acquisition in LLMs.

3.2 Datasets

LLMs possess extensive world knowledge as a result of large-scale pre-training on open web-text (Roberts et al., 2020). Evaluating their learning abilities on concepts within their pre-training data can thus lead to ambiguous and misleading interpretations. To ensure a robust analysis of concept acquisition, we compiled datasets comprising a range of previously unseen concepts.

We sourced new and unseen concepts by both automated scraping and manual compilation from a broad spectrum of sources to gather diverse materials, including song lyrics, movie plots, news articles, and academic papers, all published after July 2023 (since we tested LLMs trained on data obtained before this period). These documents were collected from platforms such as Genius, Wikipedia, AP News, and arXiv. This heterogeneous dataset, carefully curated, spans a broad spectrum of complexity and information types, enabling a comprehensive evaluation of LLM's interactive learning performance in various scenarios.

Dataset Composition Our evaluation dataset comprises a diverse collection of 1,322 contexts spanning multiple domains, as detailed in Table 4 in Appendix B. This comprehensive compilation includes images for visual interpretation tasks, movie plots for narrative analysis, and song lyrics for assessing comprehension of artistic and poetic language. Additionally, it features academic papers from various disciplines and a wide range of news articles covering different topics. The diversity in content types and the substantial number of contexts in each domain ensure a robust evaluation across a wide spectrum of complexity levels and subject matters. This carefully curated dataset allows us to thoroughly assess the concept acquisition and teaching capabilities of large language models across different types of information and communication styles. Next we describe the different subsets of the dataset.

Song Lyrics (417 contexts) Sourced from Genius, this subset challenges the interpretation of poetic and artistic language, often rich with metaphor and emotional expression. The brevity and ambiguity of lyrics test the models' ability to extract meaning from concise, creative texts.

News Articles (412 contexts) Gathered from AP News, this subset spans various categories: World News (72), Sports (67), Science (55), Politics (54), Entertainment (48), US News (51), Business (41), and Oddities (24). This domain evaluates the accurate transmission of factual, often timely information and the ability to distinguish between objective reporting and subjective commentary.

Movie Plots (179 contexts) Compiled from Wikipedia, this subset tests the models' ability to comprehend and convey complex story elements such as characters, settings, and events. The complexity of the plots varies, allowing for evaluation across different difficulty levels.

Academic Papers (164 contexts) Sourced from arXiv, this category spans various disciplines: Computer Science (25), Economics (13), Electrical Engineering & Systems Science (25), Mathematics (25), Physics (25), Quantitative Biology (18), Quantitative Finance (8), and Statistics (25). This domain examines the communication of specialized and technical language, complex logical structures, and the handling of citations and references. It tests the models' capacity to understand and teach detailed, scholarly content, engaging with in-depth analysis and evidence-based arguments.

Images (150 contexts) Drawn from the COCO dataset (Lin et al., 2014)[1], this subset assesses visual interpretation skills and the conversion of visual information into textual explanations. This multimodal aspect challenges the models to integrate visual data into coherent educational content.

[1]Images, unlike text, do not carry direct semantic content that could be memorized or specifically encoded in a language model's training data. Therefore, the age of the images is inconsequential to the model's ability to analyze and interpret visual information.

Quiz Generation for Evaluating Learning Performance To assess the concept-learning abilities of the student LLMs, as shown in the *quiz phase* in Figure 1, we generated a set of 10 questions and their respective answers for each context. For textual contexts, we utilized `gpt-3.5-turbo`, while for the image domain, we employed `gpt-4-turbo` due to its multimodal capabilities. This approach ensured that each question was directly relevant to its source material, simulating realistic scenarios. Figure 1 provides examples of quiz questions from the movie plots domain. Appendix B (Table 3) shows examples of quiz questions for each domain.

3.3 Student-Teacher Interaction Scenarios

In this work, we explore three student-teacher interaction scenarios to assess the conversational learning capabilities of LLMs, categorized as *static* and *dynamic* interactions:

1. **Static Student with Lesson:** The student is presented with a *static lesson* generated by the teacher.
2. **Dynamic Student without Lesson:** The student asks questions without any prior knowledge of the concept.
3. **Dynamic Student with Lesson:** The student generates questions after initially receiving the static lesson.

These interaction types allow us to examine different facets of conversational learning and address four key research questions:

- **RQ1:** Can students effectively learn concepts in a non-interactive, static setting?
- **RQ2:** Can the student model, through questioning, elicit enough information to match quiz performance from a static lesson?
- **RQ3:** Do the questions posed by the student effectively seek new information, leading to a deeper understanding of the concept?
- **RQ4:** What patterns or features emerge in the questions generated by the student model?

3.4 Models

We use the `gpt-3.5-turbo` models for our teacher and student LLMs, with the exception of image-based tasks where we use `gpt-4o` as the student and teacher LLMs. These models are chosen for their strong language understanding and generation capabilities, which are vital for conversational learning. For the static lesson, we prompt the teacher model to generate a comprehensive lesson, distilling the main content into a concise summary. In the static setting, the student model is provided with this lesson, if available, to answer the quiz questions. During dynamic interactions, after every dialogue turn, the student model is prompted to integrate the newly acquired information from the ongoing conversation and, if available, the prior static lesson. The student then uses this consolidated knowledge to answer quiz questions. The temperature is set to 1.0 when generating dynamic dialogues and 0 when generating quiz answers for fair comparison. All experiments are repeated across three seeds.

3.5 Evaluation Metrics

To evaluate the effectiveness of LLMs in each interaction setting, we employ the following metrics to assess concept learning progression.

Quiz Performance. Our primary metric is the accuracy of the student model's responses in concept quizzes, measured as the fraction of quiz questions answered correctly. This metric quantifies how well the student has internalized the concept discussed during the interactions.

Cumulated Information Coverage. While quiz performance provides a broad measure of learning, it doesn't capture the interaction dynamics or conversation quality. To address this, we introduce the *Cumulative Information Coverage* (CIC) metric, which evaluates how well the student's questions cover relevant information from the context. For instance, in the case of movie plots, CIC measures how effectively the student's questions encompass details from the Wikipedia page.

CIC in built on the idea of *concept elicitability*, which assesses how well a question draws out relevant context or how comprehensively an answer reflects it. Using a natural language inference (NLI) model, with the question as the premise and the context as the hypothesis, we calculate *concept elicitability* using the entailment score to indicate how well the context answers the question. If q_j is a question from the j-th turn in a conversation and s_k is the k-th sentence of the context, then CIC for the conversation c until turn i is defined as:

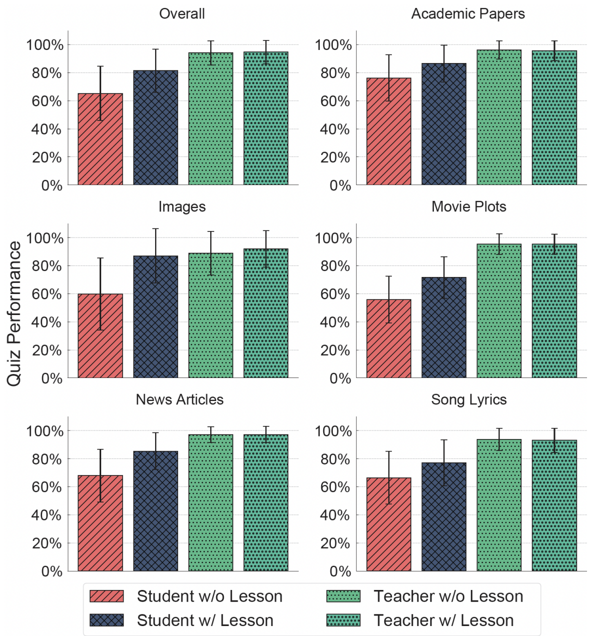

Figure 2: Average Quiz Performance of student and teacher LLMs across different domains.

$$\mathrm{CIC}_i(c) = \frac{1}{K}\sum_{k=1}^{K}\max_{j=1,\ldots,i}\sigma(q_j, s_k) \qquad (1)$$

In other words, this metric measures the maximum information from the context covered by all questions or responses up to that point. We use a $\max$ operation because an earlier question might cover more information than a later one.

4 Experiment Results

4.1 RQ1: Can students learn concepts effectively in a non-interactive setting?

We investigate the limitations of LLMs in non-interactive settings, where students lack the ability to proactively question a teacher for comprehensive concept learning. Specifically, we measure the quiz performance of the student LLM in two scenarios: (1) without any prior knowledge of the concept or its content, and (2) after receiving a *static lesson* from a teacher. To highlight the disparity in learning outcomes with limited concept knowledge, we further compare the student's quiz performance to that of a teacher with access to (a) the relevant concept context, and (b) both the context and the *static lesson* generated by the teacher.

Results. Figure 2 shows the quiz performance of student and teacher LLMs across various domains from our dataset, based on their access to

Domain	Student Questions
Academic Papers	• Could you elaborate on the methodology employed in the academic paper to achieve its objectives? • What was the approach used in the paper to investigate and analyze v-palindromes in different number bases?
Images	• What is the dominant color scheme used in the image? • Can you describe the main action or event taking place in the image?
Movie Plots	• What is the central conflict that drives the plot of the movie? • What specific events lead to Margot's initial attraction to Robert at the beginning of the movie?
News Articles	• Who are the key figures mentioned in the news article regarding the California budget deficit? • What are the main events discussed in the news article?
Song Lyrics	• How does the use of metaphor enhance the exploration of authenticity and vulnerability in the song "Actress" by Maya Delilah? • How does the artist convey the theme of loyalty in "Back From That"?

Table 1: Examples of student questions generated in the Student w/o Lesson setup for each domain

information. Interestingly, students with no specific knowledge of new concepts perform above chance, likely relying on pre-existing knowledge. This effect is particularly pronounced in the Academic Papers and News Articles domains. When provided with a *static lesson*, student performance improves significantly ($p < 0.01$) across all domains, though it remains notably lower ($p < 0.01$) than that of teachers with full concept knowledge in all domains except Images. As expected, the teacher's performance remains consistent, regardless of incorporating the static lesson. These findings highlight the substantial learning gap when LLMs rely solely on static information, underscoring the potential of our dataset as a benchmark for studying the benefits of dynamic, conversational learning approaches to enhance LLM capabilities.

4.2 RQ2: Can dynamic student models match the performance from static lessons?

In our second research question, we analyze the accuracy of concept learning when a student model engages in dialogue with a teacher to elicit information. We compare this to the performance of a student model that receives a static lesson from the teacher without any interaction.

Study Design. We measure concept learning accuracy through quiz performance across different methods. Additionally, we compare the student's performance when learning via dialogue with that of the teacher, who has complete knowledge of the concept, establishing an upper bound for quiz performance. We also track the student model's quiz performance at the end of each conversational turn to gain insights into the progression of learning in interactive settings.

Main Results. Figure 3 shows the quiz performance of various approaches across five domains: Academic Papers, Movie Plots, News Articles, Song Lyrics, and Images.[2] The student model without a lesson shows noticeable improvements in all domains, outperforming the static student with a lesson in four out of five domains. This suggests that the student model asks sufficiently comprehensive questions during the conversation. Table 1 shows some examples of questions generated by the student model during conversations in the student models without a lesson setup.

To address whether a dynamic student starting from a knowledge of the static lesson is more effective at eliciting additional knowledge that translates into improved quiz performance (compared to the dynamic student starting from *tabula rasa* in the scenario just described above), we evaluate the student's ability to gather further information from the teacher after being conditioned on the initial lesson. Generally, adding conversational capabilities to the student with a static lesson leads to slight, statistically significant improvements ($p < 0.01$) over the student without a lesson, except in the Song Lyrics and Images domains. However, the student's performance in this scenario remains significantly lower ($p < 0.01$) than that of the teacher in all domains except Images, indicating that LLMs may require additional guidance to effectively learn concepts through interaction.

Overall, our findings demonstrate that *while student models are capable of learning through interaction, the extent of knowledge acquired via this method significantly lags behind a teacher that acts with complete knowledge of the concept*.

4.3 RQ3: Can questions posed by the student models effectively seek new information?

In the previous sub-section, we observed that student models learning through interaction with a teacher still lag behind the teacher's performance in quiz accuracy. However, it remains unclear whether this gap is due to the student asking uninformative

[2]Detailed variance across domains is provided in Table 2.

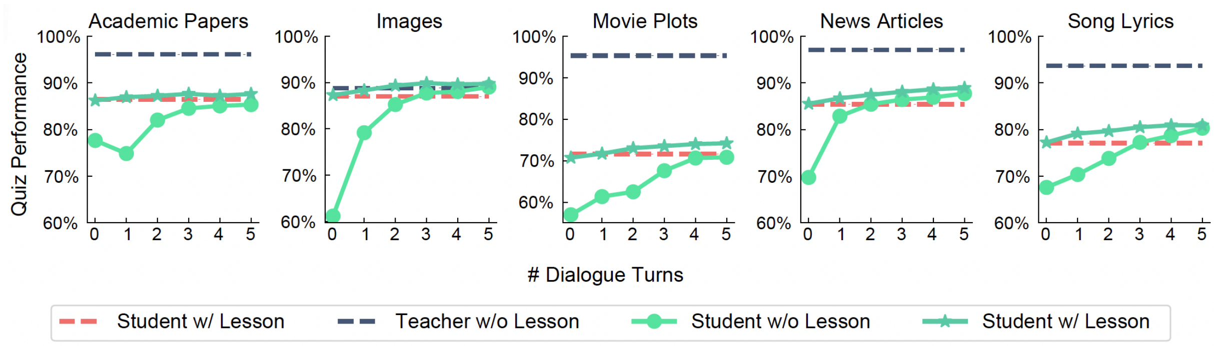

Figure 3: Performance of student LLMs across various dynamic evaluation settings along with static baselines.

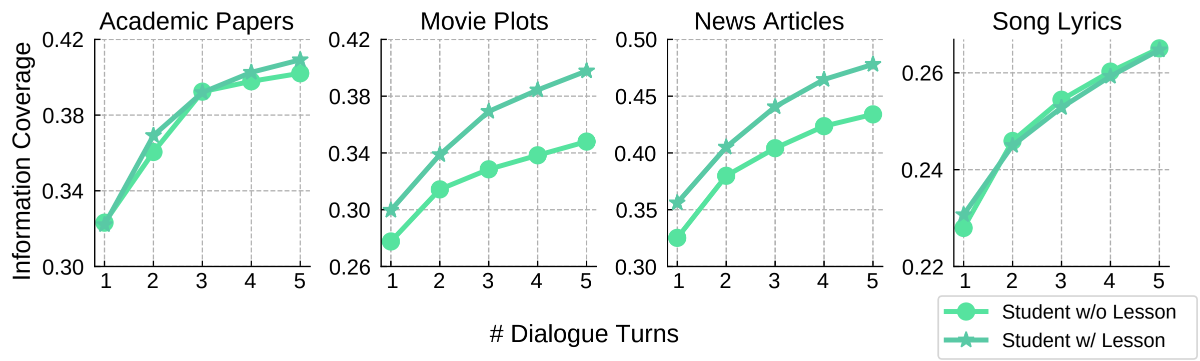

Figure 4: Average CIC of questions asked by student LLMs across various dynamic evaluations.

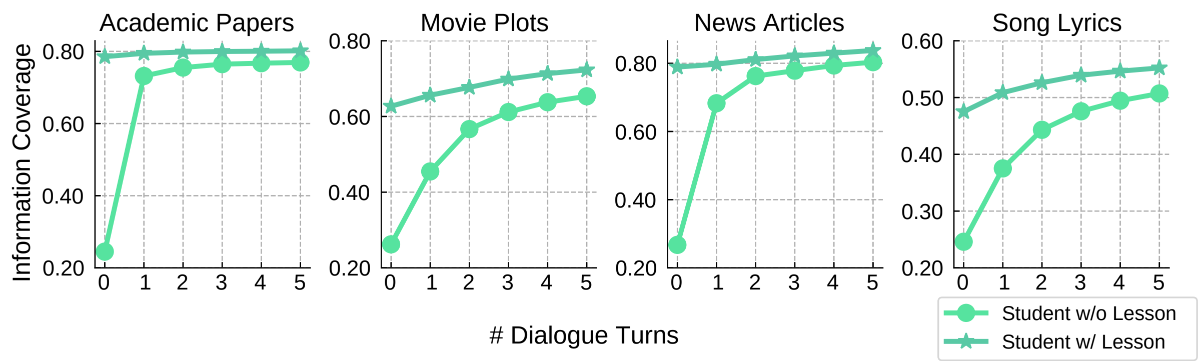

Figure 5: Average CIC of answers generated by teacher LLMs across various dynamic evaluations.

questions. This research question aims to measure the quality of questions posed by the student model.

Study Design To evaluate question quality, we utilize our Cumulative Information Coverage (CIC) metric (§3.5). CIC measures the extent of relevant information elicited by each student question relative to the teacher's complete concept knowledge. Conversely, for teacher responses, CIC quantifies the amount of information covered by each response. We compute CIC after each question in the dialogue, allowing us to assess how much of the context is addressed by the questions or responses at the conclusion of each dialogue turn. For this, we exclusively focus on text-based domains (i.e., Movie Plots, News Articles, Academic Papers, Song Lyrics) since the NLI model operates only on text.

Main Results Figure 4 shows the average CIC scores for the questions posed by student models throughout the dialogue. Across most domains, the information coverage of questions asked by the Student with Lesson forms an upper bound. However the gap between Student w/ Lesson and Student w/o Lesson is not significant for the Academic Papers and Song Lyrics domains ($p > 0.01$). Questions posed for news articles are the most comprehensive, while those for song lyrics are the least.

Figure 5 displays the average CIC scores for answers generated by teacher models. In all domains, information coverage of answers in the Student with Lesson setup forms an upper bound. A significant gap ($p < 0.01$) between Student w/o Lesson and Student w/ Lesson setup is observed, indicating that presenting a lesson before questioning leads to more informative questions and responses. While student questions show a trend of increasing information coverage, the coverage of teacher responses tends to saturate after a few questions, similar to the quiz performance results.

Our findings demonstrate that (1) questions posed by student LLMs and teacher responses cover more information with each turn, and (2) presenting an initial static lesson produces more comprehensive questions and responses.

4.4 RQ4: What patterns emerge in questions asked by the student models?

While previous research questions focus on learning effectiveness, it remains unclear which factors contribute to the learning gains of the student model. In this fourth research question, we investigate whether specific features of the questions posed by the student model are associated with better learning outcomes.

Study Design We explore the relationship between learning gains and predefined features of the student's questions. Learning gain is measured as the increase in quiz performance compared to the previous turn. We examine four key features that may correlate with learning gains in both the Student w/ Lesson and Student w/o Lesson setups: (1) question length, (2) maximum depth of the question's syntax tree, (3) total count of named entities in the question, and (4) position of the question within the dialogue (represented as a binary feature). We calculate the Pearson Correlation Coefficient between these features and learning gain, using Stanza (Qi et al., 2020) for syntax parsing and entity extraction.

Main Results Overall, we do not observe strong correlations between any of the predefined features and learning gains. Generally, the influence of these features diminishes when the student model is provided with a lesson. Named entity count shows a relatively stronger correlation in the news articles domain, where entities like people, locations, and organizations are central. Question length is the most correlated feature in the absence of a lesson across most domains. Interestingly, being the first question in a dialogue has a negative correlation with learning gain, unlike other positions. Appendix A (Figure 6 in) includes detailed analysis of feature correlations for each domain.

Qualitative Analysis Although none of the features strongly correlate with learning gains, distinct patterns emerge across dialogues. For example, in movie plots, the student model typically begins by asking about the central conflict, then progresses to questions about character development, key themes, and setting. A similar pattern is observed in academic papers, where questions generally follow themes such as objectives, methodology, key findings, limitations, and motivation. Despite the temperature being set to 1.0 during question generation, these patterns might suggest a lack of diversity in the questions, which may contribute to the observed performance saturation after a few questions.

5 A Future of Conversational Learning

While our research demonstrates the promise of interactive learning for LLMs, it also highlights the challenges and opportunities that lie ahead. Our findings show that dynamic, conversational interactions with a teacher enable LLMs to gain more comprehensive understanding across domains than static lessons. However, despite the benefits of interactive learning, student models still lag behind teachers with full concept knowledge, indicating the need for further advancements in this area. Addressing these limitation through more sophisticated question generation techniques could improve the models' ability to explore concepts from multiple perspectives.

Future work can explore extensions of existing machine learning theories, such as active learning, to better analyze and optimize interactive learning methods. By treating active learning as a special case, these extensions could lead to new theoretical frameworks that capture the complexities of real-time, adaptive learning. Additionally, expanding interactive learning to include multimodal scenarios, such as audio and video, could provide richer educational experiences and better simulate real-world learning environments. Investigating long-term retention of knowledge acquired through interaction, as well as the ethical implications of deploying AI in educational settings, will also be critical. Such conversational learning systems would not only learn but also teach effectively, potentially transforming both AI and education technologies.

Limitations

Our study on interactive learning with Large Language Models (LLMs) has several key limitations that warrant consideration.

Firstly, our exclusive use of closed-source GPT models limits the generalizability of our findings. Different architectures or training paradigms might yield varying results in interactive learning scenarios. This limitation extends to the persistent performance gap between student models and teachers with full concept knowledge, suggesting our current approach, while promising, falls short of enabling LLMs to fully grasp complex concepts through dialogue alone.

Another significant limitation lies in our evaluation methodology. Our primary metrics of quiz performance and Cumulative Information Coverage, while informative, may not capture all aspects of concept understanding. These metrics might overlook nuanced comprehension or the ability to apply learned concepts in novel contexts. Moreover, our focus on immediate concept acquisition leaves open questions about long-term retention and integration of knowledge gained through interactive learning. More comprehensive evaluation methods could offer a more holistic picture of LLM learning, including assessments of reasoning ability, knowledge transfer, and conceptual integration over time.

Lastly, the scalability of our approach to larger datasets, longer conversations, or more complex concepts remains untested. As the complexity of tasks increases, the computational resources required for extended dialogues could become prohibitive, potentially limiting practical applicability in real-world settings. This scalability challenge is closely tied to ethical considerations, particularly regarding the deployment of AI in educational contexts. Important issues such as AI transparency, potential biases in learning outcomes, and the impact on human learning processes when interacting with AI teachers remain unaddressed.

Addressing these limitations will be crucial for realizing the full potential of conversational AI in educational and knowledge acquisition contexts. Future work should focus on diversifying model selection, developing more comprehensive evaluation metrics that include long-term retention, addressing scalability challenges, and thoroughly examining the ethical implications of AI in education.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Forough Arabshahi, Kathryn Mazaitis, Toby Jia-Jun Li, Brad A Myers, and Tom Mitchell. 2020. Conversational learning. *Preprint on webpage at https://forougha. github. io/paperPDF/Conversational Learning. pdf*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations*.

David A. Cohn, Les E. Atlas, and Richard E. Ladner. 2004. Improving generalization with active learning. *Machine Learning*, 15:201–221.

Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. 2008. Towards scalable dataset construction: An active learning approach. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 86–98. Springer.

Mihail Eric and Christopher Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 468–473, Valencia, Spain. Association for Computational Linguistics.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers.

Belinda Z Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *arXiv preprint arXiv:2310.11589*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach? teacher explanations improve student performance via personalization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1527–1536, Copenhagen, Denmark. Association for Computational Linguistics.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2019. Learning to ask for conversational machine learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4164–4174, Hong Kong, China. Association for Computational Linguistics.

Alex Tamkin, Dat Pham Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. 2022. Active learning helps pretrained models learn the intended task. In *Advances in Neural Information Processing Systems*.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Appendix

A Additional Results

A.1 Static Baseline Results

Baseline student LLMs, relying solely on pretraining knowledge, achieve relatively low scores across all domains, with performance ranging from 24.53% in the images domain to 74.18% in the science sub-domain of news articles. However, when provided with structured lessons from teacher LLMs, student performance improves significantly. For example, in the academic papers domain, student performance increases from an average of 75.70% to 85.57% with lessons, demonstrating a 9.87 percentage point increase. Teacher LLMs consistently outperform student LLMs, with their direct access to original material providing them with comprehensive contextual knowledge. Their near-perfect scores, ranging from 91.60% to 98.55% across domains, set a high bar for student LLMs. When teacher LLMs receive additional lessons, their performance improves only marginally, with an average increase of 0.46 percentage points. This slight improvement suggests that, while summaries are beneficial, the original material already covers the essential information comprehensively, and

Domain	Sub Domain	Student w/o Lesson	Student w/ Lesson	Teacher w/o Lesson	Teacher w/ Lesson
Academic Papers	**q-bio**	$81.48_{(17.03)}$	$92.22_{(7.57)}$	$\mathbf{98.33}_{(5.14)}$	$98.33_{(5.14)}$
	q-fin	$77.92_{(18.24)}$	$86.25_{(13.02)}$	$96.67_{(4.99)}$	$\mathbf{97.50}_{(4.63)}$
	econ	$70.77_{(19.37)}$	$84.62_{(11.27)}$	$\mathbf{96.92}_{(4.80)}$	$96.15_{(5.97)}$
	cs	$77.20_{(16.19)}$	$86.40_{(13.19)}$	$\mathbf{96.00}_{(5.77)}$	$95.60_{(7.68)}$
	math	$72.13_{(19.08)}$	$82.13_{(15.39)}$	$92.67_{(9.33)}$	$\mathbf{92.93}_{(10.64)}$
	eess	$78.67_{(15.83)}$	$86.13_{(16.22)}$	$\mathbf{96.00}_{(6.45)}$	$95.47_{(6.06)}$
	stat	$76.27_{(12.98)}$	$88.80_{(13.21)}$	$\mathbf{96.53}_{(6.11)}$	$94.80_{(7.33)}$
	physics	$75.47_{(15.53)}$	$85.73_{(11.44)}$	$\mathbf{97.07}_{(5.47)}$	$96.53_{(4.85)}$
Images	**-**	$59.82_{(25.69)}$	$86.98_{(19.31)}$	$88.80_{(15.50)}$	$\mathbf{91.96}_{(12.90)}$
Movie Plots	**-**	$55.96_{(16.62)}$	$71.62_{(14.78)}$	$95.31_{(7.34)}$	$\mathbf{95.38}_{(7.20)}$
News Articles	**entertainment**	$69.38_{(15.05)}$	$87.92_{(10.23)}$	$97.43_{(4.41)}$	$\mathbf{98.06}_{(4.00)}$
	business	$67.07_{(19.12)}$	$85.45_{(11.27)}$	$\mathbf{96.59}_{(6.56)}$	$96.50_{(6.19)}$
	science	$72.97_{(16.44)}$	$87.88_{(12.65)}$	$98.36_{(4.32)}$	$\mathbf{98.85}_{(4.63)}$
	us-news	$74.31_{(16.66)}$	$90.65_{(10.11)}$	$\mathbf{97.84}_{(4.61)}$	$97.52_{(5.57)}$
	sports	$50.45_{(19.94)}$	$77.76_{(15.26)}$	$\mathbf{95.42}_{(7.43)}$	$95.12_{(7.86)}$
	politics	$71.30_{(15.62)}$	$85.12_{(12.70)}$	$\mathbf{96.36}_{(5.58)}$	$96.36_{(5.91)}$
	world-news	$74.95_{(14.88)}$	$84.81_{(13.60)}$	$97.50_{(5.05)}$	$\mathbf{98.15}_{(4.26)}$
	oddities	$61.53_{(18.23)}$	$86.81_{(12.02)}$	$\mathbf{97.64}_{(6.70)}$	$97.08_{(5.49)}$
Song Lyrics	**-**	$66.47_{(18.87)}$	$77.07_{(16.47)}$	$\mathbf{93.65}_{(7.95)}$	$93.10_{(8.63)}$
Overall		$70.22_{(17.44)}$	$84.97_{(13.14)}$	$96.06_{(6.50)}$	$\mathbf{96.07}_{(6.58)}$

Table 2: Concept Quiz Performance of student and teacher LLMs that are privy to varying levels of information (i.e., with or without lesson) across the different sub-domains in our proposed datasets. Numbers reported are the mean and standard deviation of performance across available documents for each domain/sub-domain. Bold numbers are the best method for a domain/sub-domain.

teacher LLMs' access to detailed source material is crucial to their high performance. Overall, the substantial underperformance of the student LLM compared to the teachers highlights the challenge posed by our datasets to LLMs, leaving room for effective guidance by teachers.

A.2 Features of Questions Posed by Student Model

Figure 6 shows average correlations between learning gains from questions posed by student LLMs and the predefined set of features: (1) length of the question, (2) the maximum depth of the syntax tree of the question, (3) total count of named entities in the question, and (4) position of the question in the dialogue. Results are presented for both Student w/ Lesson and Student w/o Lesson setups.

A.3 Example Questions Posed by Student LLM

B Dataset Creation

B.1 Data Collection

Datasets were compiled by scraping a wide array of sources to obtain song lyrics, movie plots, news articles, and academic papers, all published after July 2023.[3] In addition to textual data, the Visual Question Answering (VQA) dataset from the COCO image collection was utilized to add a multimodal dimension to the context preparation, further challenging the instructional capabilities of the models under study.

Automated Scraping Python scripts using libraries such as `requests` and `BeautifulSoup` were employed to streamline the data collection process. These scripts fetched the necessary data by navigating to the relevant URLs, parsing the HTML content, and extracting the required information. The following general approach was adopted:

- **Concurrent Processing:** The use of `ThreadPoolExecutor` from the `concurrent.futures` module enabled concurrent downloading and processing of data, significantly speeding up the data collection process.
- **Error Handling and Retries:** Robust error handling mechanisms were implemented to manage network issues and server errors, including retries for failed requests.

[3]This temporal criterion was strategically chosen to ensure that the data used was not previously encountered by the GPT-3.5 model, thus eliminating potential biases or prior knowledge that could influence the model's performance in teaching and learning scenarios.

B.2 Context Preparation

Textual Contexts Textual data including movie plots, song lyrics, and news articles were retained in their plain text format to facilitate easy processing. Academic papers, typically presented in PDF format and characterized by their extensive length, were converted to text. However, given their voluminous nature, only the first 1500 words of each document were used. This limitation ensured that there was ample but manageable content for generating accurate instructional materials and assessments, while avoiding reaching context limits for the used models.

Image Processing For the processing of images, the GPT-4-Vision model was employed, as GPT-3.5 does not support image inputs. Each image was converted into a base64 encoded string, a format suitable for model processing. The GPT-4-Vision model was then used to generate a set of five multiple-choice questions per image. This decision was based on the consideration that asking the model to generate a larger number of questions, such as ten, could lead to redundancy and a decline in question quality. The limited content inherent to single images typically does not support the generation of a large number of high-quality, diverse questions without compromising the depth or relevance of the content being tested.

Justification for Using Older Image Data While the textual contexts were specifically required to be post-July 2023 to avoid GPT-3.5's pre-existing knowledge influencing the study, the use of older images from the COCO dataset does not present the same risk. Images, unlike text, do not carry direct semantic content that could be memorized or specifically encoded in a language model's training data. Therefore, the age of the images is inconsequential to the model's ability to analyze and interpret visual information. This distinction allows for the inclusion of a broad range of visual contexts, enhancing the multimodal aspect of the study without compromising the integrity of the experimental results.

Question Generation Each textual context was processed using custom Python functions that lever-

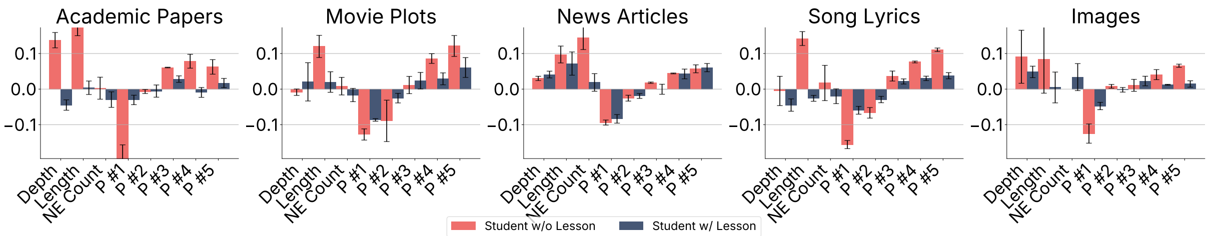

Figure 6: Average correlations between learning gains from questions posed by student LLMs and the predefined set of features

Domain	Example Quiz Questions
Academic Papers	Question 6: During what periods was the coefficient of performance found to be 41% higher? A) When there was a low cooling demand B) During high cooling demand periods C) When the filter was not used D) During off-peak hours
Images	Question 2: What color are the flowers in the right garden bed? A) Blue and yellow B) Red and yellow C) Green and white D) Pink and orange
Movie Plots	Question 3: Who criticizes Barbie for encouraging unrealistic beauty standards in the real world? A) Ken B) Gloria C) Weird Barbie D) Sasha
News Articles	Question 3: Which group of migrants does the new rule primarily target? A) Families with children B) Individuals seeking better job opportunities C) Migrants with criminal records or terrorist links D) Refugees fleeing war
Song Lyrics	Question 5: What does the singer refer to as their addiction in the pre-chorus? A) Coffee B) Reading C) Exercise D) Someone

Table 3: Example quiz questions for each domain

aged `gpt-3.5-turbo` on the OpenAI API to generate a set of questions and their respective answers. This method ensured that each question was directly relevant to the context it was derived from, simulating realistic scenarios where a teacher generates quiz material based on specific content. For the image-based contexts, similar functions were employed to generate descriptive and inferential questions, thus testing the model's ability to integrate visual information with textual instruction. Table 3 shows example quiz questions for each domain.

C Prompt Templates

Table 5 provides a legend for prompts used in the static and dynamic settings of our study.

Listing 1: Lesson Generation Prompt given Concept. We list the different prompts used for different domains in the same listing for brevity.

```
System:
    Movie Plots: "Prepare a student for any quiz on this movie plot, by explaining its storyline, character
        ↪ arcs, themes, and significant scenes. Your explanation should cover all essential aspects,
        ↪ enabling the student to confidently answer questions on any part of the movie."

    Images: "Prepare a student for any quiz on this image by providing a detailed analysis of its elements,
        ↪ composition, and context. Highlight the key features and underlying messages, ensuring the
        ↪ student can address questions related to any aspect of the image."

    Academic Papers: "Prepare a student for any quiz on this academic paper by summarizing its objectives,
        ↪ methodology, findings, and significance. Your summary should comprehensively cover the paper's
        ↪ content, preparing the student to tackle questions on any part of the study."

    News Articles: "Prepare a student for any quiz on this news article by outlining the main events, key
        ↪ figures, and the article's context. Ensure your summary is thorough, allowing the student to
        ↪ respond to questions on any detail of the article."

    Song Lyrics: "Prepare a student for any quiz on these song lyrics by dissecting the narrative, themes,
        ↪ and expressive techniques used. Provide a complete understanding, enabling the student to engage
        ↪  with questions on any aspect of the lyrics."

User:
    {concept}
```

Listing 2: Quiz Generation Prompt for Concept

```
System:
    "Generate {number_of_questions} multiple-choice questions based on the provided {domain}, each with 4
        ↪ options (A, B, C, D). After each question, immediately provide the correct answer, preceded by '
        ↪ Correct Answer: '. The format should be strictly followed for each question and answer pair.
        ↪ Here is an example of how each question and answer should be formatted:

    Question 1: [Question text]
    A) Option A
    B) Option B
    C) Option C
    D) Option D
    Correct Answer: A

    Please adhere to this format for all {number_of_questions} questions and their corresponding answers."

User:
    {concept}
```

Listing 3: Prompt for Student w/o Lesson

```
System:
    "You will be given a set of {number_of_questions} multiple-choice questions regarding a {domain}. Please
        ↪  provide your answers in the following format:

    1. A single string of {number_of_questions} capital letters (A, B, C, or D) representing your choices
        ↪ for each question.
    For example: ABCDABCDAB

    OR

    2. A numbered list with the question number followed by a closing parenthesis or a dot, a space, and
        ↪ then the capital letter (A, B, C, or D) representing your choice.
    For example:
    1) A
    2) B
    3) C
    ...

    Even if you feel you lack context, make an educated guess for each answer. You must provide exactly {
        ↪ number_of_questions} answers, one for each question, and use only the specified formats."

User:
    "Questions: {questions}"
```

Listing 4: Prompt for Student w/ Lesson

```
System:
    "You will be given a set of {number_of_questions} multiple-choice questions regarding a {domain}. Please
        ↪  provide your answers in the following format:

    1. A single string of {number_of_questions} capital letters (A, B, C, or D) representing your choices
        ↪ for each question.
    For example: ABCDABCDAB

    OR

    2. A numbered list with the question number followed by a closing parenthesis or a dot, a space, and
        ↪ then the capital letter (A, B, C, or D) representing your choice.
    For example:
    1) A
    2) B
    3) C
    ...

    Even if you feel you lack context, make an educated guess for each answer. You must provide exactly {
        ↪ number_of_questions} answers, one for each question, and use only the specified formats."

User:
    "Lesson: {lesson}
    Questions: {questions}"
```

Listing 5: Prompt for Teacher w/o Lesson

```
System:
    "You will be given the original information of a {domain} and a set of {number_of_questions} multiple-
        ↪ choice questions based on it. Please provide your answers in the following format:

    1. A single string of {number_of_questions} capital letters (A, B, C, or D) representing your choices
        ↪ for each question.
    For example:
    ABCDABCDAB

    OR

    2. A numbered list with the question number followed by a closing parenthesis or a dot, a space, and
        ↪ then the capital letter (A, B, C, or D) representing your choice. For example:
    1) A
    2) B
    3) C
    ...

    You must provide exactly {number_of_questions} answers, one for each question, and use only the
        ↪ specified formats."

User:
    "Original Information: {concept}
    Questions: {questions}"
```

Listing 6: Prompt for Teacher w/ Lesson

```
System:
    "You will be given the original information of a {domain} and a set of {number_of_questions} multiple-
        ↪ choice questions based on it. Please provide your answers in the following format:

    1. A single string of {number_of_questions} capital letters (A, B, C, or D) representing your choices
        ↪ for each question.
    For example:
    ABCDABCDAB

    OR

    2. A numbered list with the question number followed by a closing parenthesis or a dot, a space, and
        ↪ then the capital letter (A, B, C, or D) representing your choice. For example:
    1) A
    2) B
    3) C
    ...

    You must provide exactly {number_of_questions} answers, one for each question, and use only the
        ↪ specified formats."

User:
    "Original Information: {concept}
    Lesson: {lesson}
    Questions: {questions}"
```

Listing 7: Prompt for Student in Dynamic Conversation

```
System:
    Movie Plots: "To learn more about the movie plot known only to the teacher and get prepared for any quiz
        ↪  on that, ask questions on its storyline, character arcs, themes, and significant scenes. Ask
        ↪ diverse questions encompassing plot progression, character actions, involvement, thematic
        ↪ exploration, and character motivations. Include questions seeking specific details such as
        ↪ character names, objects, settings, and dates. Include questions that prompt thorough analysis
        ↪ of the plot and a deeper comprehension of its unfolding events. Ensure questions are diverse and
        ↪  comprehensive, covering all facets of the movie. Also, feel free to ask detailed questions
        ↪ whenever necessary. If you run out of questions, always think of and come up with more creative
        ↪ and detailed questions! Ask one question at a time! NEVER PROMPT TEACHER TO ASK ANY QUESTION!"

    Academic Papers: "To learn more about the academic paper known only to the teacher and get prepared for
        ↪ any quiz on that, ask questions on its objectives, methodology, findings, and significance. Ask
        ↪ diverse questions encompassing experiments, its relation to prior studies, limitations,
        ↪ motivation and key takeaways. Include questions seeking specific details such as experimental
        ↪ setup. Include questions that prompt thorough analysis of the paper and a deeper understanding
        ↪ of its broader contributions. Ensure questions are diverse and comprehensive, covering all
        ↪ aspects of the paper. Also, feel free to ask detailed questions whenever necessary. If you run
        ↪ out of questions, always think of and come up with more creative and detailed questions! Ask one
        ↪  question at a time! NEVER PROMPT TEACHER TO ASK ANY QUESTION!"

    News Articles: "To learn more about the news article known only to the teacher and get prepared for any
        ↪ quiz on that, ask questions on the main events, key figures, and the article's context. Ask
        ↪ diverse questions encompassing background stories and broader implications. Include questions
        ↪ seeking specific details such as names of individuals, events, actions, and dates. Include
        ↪ questions that prompt thorough analysis of the article and a deeper comprehension of unfolding
        ↪ events. Ensure questions are diverse and comprehensive, covering all aspects of the article.
        ↪ Also, feel free to ask detailed questions whenever necessary. If you run out of questions,
        ↪ always think of and come up with more creative and detailed questions! Ask one question at a
        ↪ time! NEVER PROMPT TEACHER TO ASK ANY QUESTION!"

    Song Lyrics: "To learn more about song lyrics known only to the teacher knows about and get prepared for
        ↪  any quiz on that, ask questions on its narrative, themes, and expressive techniques used. Ask
        ↪ diverse questions encompassing emotions, individuals, events, involvement, themes and references
        ↪  to other content. Include questions that prompt thorough analysis of the lyrics and a deeper
        ↪ comprehension of its meaning. Ensure questions are diverse and comprehensive, covering all
        ↪ facets of the lyrics. Also, feel free to ask detailed questions whenever necessary. If you run
        ↪ out of questions, always think of and come up with more creative and detailed questions! Ask one
        ↪  question at a time! NEVER PROMPT TEACHER TO ASK ANY QUESTION!"

    {chat_history}

Chat History:
    Teacher:
        {lesson} You can ask me any question about the {context}.
    Student: ...
    Teacher: ... Do you have any other questions?
    ...
```

Listing 8: Prompt for Teacher in Dynamic Conversation

```
System:
    Movie Plots: "Prepare the student comprehensively for any quiz on this movie plot, by answering
         ↪ questions on its storyline, character arcs, themes, and significant scenes. Content: {content}\n
         ↪  Do not ask any questions to the student, only answer the questions! Generate long and detailed
         ↪ answers! Include specific event and character-related details in your answers like what happened
         ↪ , who performed specific actions, and who was involved."

    Academic Papers: "Prepare the student comprehensively for any quiz on this academic paper, by answering
         ↪ questions on its objectives, methodology, findings, and significance. Content: {content}\n Do
         ↪ not ask any questions to the student, only answer the questions! Generate long and detailed
         ↪ answers!"

    News Articles: "Prepare the student comprehensively for any quiz on this news article, by answering
         ↪ questions on the main events, key figures, and the article's context. Content: {content}\n Do
         ↪ not ask any questions to the student, only answer the questions! Generate long and detailed
         ↪ answers! Include specific event-related details in your answers like what happened, who
         ↪ performed specific actions, and who was involved."

    Song Lyrics: "Prepare the student comprehensively for any quiz on this movie plot, by answering
         ↪ questions on the narrative, themes, and expressive techniques used. Content: {content}\n Do not
         ↪ ask any questions to the student, only answer the questions! Generate long and detailed answers
         ↪ !"

    {chat_history}

Chat History:
    Teacher:
        {lesson} You can ask me any question about the {context}.
    Student: ...
    Teacher: ... Do you have any other questions?
    ...
```

Listing 9: Prompt for Student Evaluation in Dynamic Conversation

```
System:
    "You will be given a lesson on a specific topic. Please review the lesson carefully.\nLesson:{lesson}

    {chat_history_lesson_removed}

    You will be given a set of {number_of_questions} multiple-choice questions regarding a {context}. Please
         ↪  provide your answers in the following format:


    1. A single string of {number_oof_questions} capital letters (A, B, C, or D) representing your choices
         ↪ for each question. For example: ABCDABCDAB


    OR


    2. A numbered list with the question number followed by a closing parenthesis or a dot, a space, and
         ↪ then the capital letter (A, B, C, or D) representing your choice. For example:

    1) A
    2) B
    3) C
    ...


    Even if you feel you lack context, make an educated guess for each answer. You must provide exactly {
         ↪ number_of_questions} answers, one for each question, and use only the specified formats. Based
         ↪ on the discussion, please answer the following questions to evaluate your understanding.

User:
    Questions: {questions}
```

Domain	Count
Images	150
Movie Plots	179
Song Lyrics	417
Academic Papers	164
Computer Science	25
Economics	13
Electrical Eng. & Systems Sci.	25
Mathematics	25
Physics	25
Quantitative Biology	18
Quantitative Finance	8
Statistics	25
News Articles	412
Business	41
Entertainment	48
Oddities	24
Politics	54
Science	55
Sports	67
US News	51
World News	72
Total	1,322

Table 4: Composition of the Dataset

Objective	Reference	Setting
Lesson Generation	Listing 1	Static
Quiz Generation	Listing 2	Static
Student w/o Lesson	Listing 3	Static
Student w/ Lesson	Listing 4	Static
Teacher w/o Lesson	Listing 5	Static
Teacher w/ Lesson	Listing 6	Static
Student	Listing 7	Dynamic
Teacher	Listing 8	Dynamic
Student Evaluation	Listing 9	Dynamic

Table 5: Legend for prompts used in the various stages of our study, including both static and dynamic experiments.