INTERACT: Enabling Interactive, Question-Driven Learning in Large Language Models

Aum Kendapadi* **Kerem Zaman*** **Rakesh R. Menon*** **Shashank Srivastava**
UNC Chapel Hill
aumken@alumni.unc.edu, {kzaman, rrmenon, ssrivastava}@cs.unc.edu

Abstract

Large language models (LLMs) excel at answering questions but remain passive learners—absorbing static data without the ability to question and refine knowledge. This paper explores how LLMs can transition to interactive, question-driven learning through student-teacher dialogues. We introduce INTERACT (INTERactive learning for Adaptive Concept Transfer), a framework in which a "student" LLM engages a "teacher" LLM through iterative inquiries to acquire knowledge across 1,347 contexts, including song lyrics, news articles, movie plots, academic papers, and images. Our experiments show that across a wide range of scenarios and LLM architectures, interactive learning consistently enhances performance, achieving up to a 25% improvement, with 'cold-start' student models matching static learning baselines in as few as five dialogue turns. Interactive setups can also mitigate the disadvantages of weaker teachers, showcasing the robustness of question-driven learning. [1]

1 Introduction

Large language models (LLMs) are impressive creatures. They have become fluent at summarizing texts, assisting users, and tackling complex reasoning problems. Yet, their training remains largely static[2], relying on fixed datasets rather than interactive processes. In contrast, humans naturally refine their understanding by asking questions, prodding teachers, and poking holes in their explanations until the world makes sense — strategies that help them learn new concepts effectively (see Figure 1) (Vygotsky and Cole, 1978; Ram, 1991).

Infusing LLMs with this kind of interactive, question-driven inquiry can be valuable for knowledge-intensive domains. Instead of passively absorbing data, an LLM could engage in a dialogue: requesting clarifications, seeking missing details, and testing its evolving comprehension. In education, for example, an AI "student" could interact with a "teacher" model, focusing on a learner's trouble spots rather than delivering the same summary to everyone. In professional contexts such as medicine or scientific research, iterative questioning would let AI systems refine diagnoses, refine hypotheses, and illuminate overlooked in collaboration with human experts.

To advance this vision, we introduce INTERACT (INTERactive learning for Adaptive Concept Transfer), a framework that simulates teacher-student dialogues for LLM-based interactive learning. We ask the question: *How effectively can LLMs learn new concepts through conversational interactions?* Instead of simply consuming summarized information, the "student" model actively questions the "teacher", iteratively building knowledge through inquiry. Much as a human learner hones understanding through persistent, well-placed questions, the student LLM can surface ambiguities, verify assumptions, and guide the conversation toward deeper conceptual clarity.

We evaluate INTERACT on 1,347 unseen contexts spanning movie plots, academic papers, news articles, song lyrics, and visual descriptions—carefully curated to exclude pretraining overlap. This ensures that the student LLM must genuinely acquire new information rather than rely on memorized patterns. We compare two teaching modes: *static lessons*, where a teacher provides a condensed summary of key content, and *dynamic interactions*, where a student LLM engages by asking questions. We evaluate the effectiveness of these approaches by simulating student-teacher interactions. To measure student learning, we test the LLM's understanding with a quiz either after the static lesson or after each dialogue turn in the dynamic inter-

* Equal contribution

[1] Our code and dataset are available at https://github.com/aumken/interact.

[2] Despite alignment methods such as preference tuning (Ouyang et al., 2022; Rafailov et al., 2023).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8992–9024

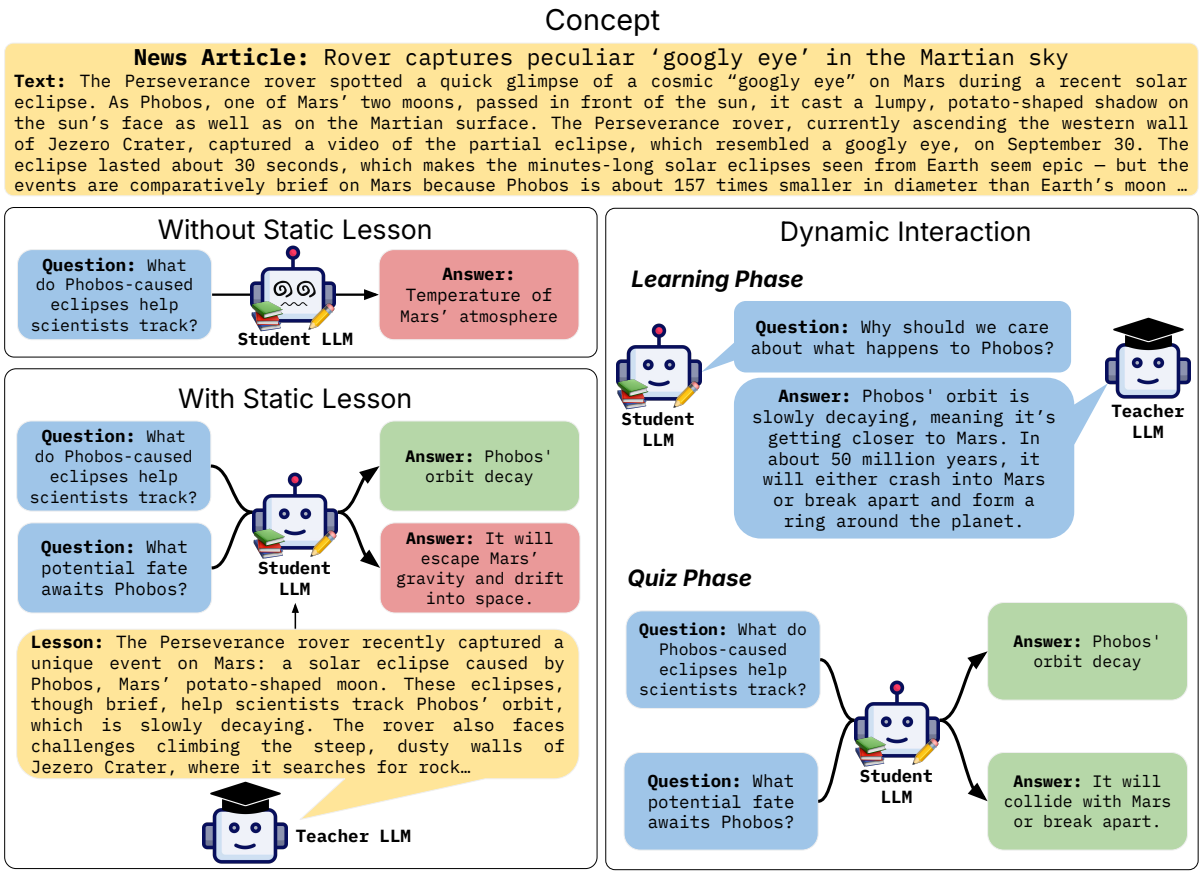


Figure 1: Overview of the INTERACT framework for concept learning in LLMs. Given a new concept, here, a news article from a time period outside of the LLM pretaining data, non-interactive approaches such as zero-shot prompting (*left-top*) and static lessons (*left-bottom*) fail due to lack of information or intricacies in the concept. Through dynamic interaction with a teacher (*right*), a student can learn about a concept more comprehensively.

action setting. The latter enables a study of LLM questions that most help understanding new concepts. We find that conversational interactions consistently enhance learning achieving **up to a 25% improvement in quiz performance**, sometimes matching learning from static data in as few as five dialogue turns. Dynamic interactions also mitigate the advantages of stronger teachers. However, despite the benefits of interaction, student models still underperform teacher LLMs, highlighting the need for improved dialogue strategies.

INTERACT's implications are wide-ranging. In education, interactive AI tutors could probe students' knowledge, helping instructors pinpoint misconceptions and personalize their teaching. The same approach can guide trainees in developing their own questioning strategies or assist domain experts in fields like medicine or research by unearthing critical missing links. Rather than functioning as static encyclopedias, LLMs can become proactive collaborators. Our contributions are:

- A framework for enabling LLMs to learn concepts via teacher-student interactions.
- A benchmark of 1,347 unseen contexts, spanning movie plots, academic papers, news articles, song lyrics, and images.
- Empirical evidence showing that interactive learning improves concept acquisition.
- Analyses showing the importance of adaptive questioning and uncovering insights to improve interactive learning strategies.

2 Related Work

Conversational Machine Learning. Early work on language-guided machine learning often used single-turn instructions and limited examples (Srivastava et al., 2017; Hancock et al., 2018; Arabshahi et al., 2020; R. Menon et al., 2022). To improve comprehension, researchers have explored active learning (Collins et al., 2008; Tamkin et al., 2022) and language-based clarifications (Rao and Daumé III, 2018; Srivastava et al., 2019). Our ap-

proach follows the latter, but unlike methods that rely on templated question generation, we let LLMs produce contextually relevant queries. Our teacher-student paradigm is consonant with knowledge distillation (Hinton, 2015), though here the "student" acquires knowledge by active questioning rather than passively receiving information. Our approach is also related to work on document-grounded dialogue (Feng et al., 2020, 2021), which model multi-turn interactions grounded in text.

Interactive Learning with LLMs. Recent work shows that LLMs benefit from explanations (Wei et al., 2022; Lampinen et al., 2022), self-generated feedback (Madaan et al., 2023; Chen et al., 2024), and tuning on larger-model outputs (Ho et al., 2023). While these focus on teacher-provided information, we highlight the student's role in shaping the discourse. Other studies examine learning human preferences via dialogue (Li et al., 2023; Handa et al., 2024) or conversational QA (Abbasiantaeb et al., 2024), but our focus is on student-driven inquiry and eliciting deeper explanations.

Adaptive Learning. Adaptive teaching often involves detecting misconceptions and customizing examples (Ross and Andreas, 2024). In contrast, we emphasize a student-led approach, where queries guide the interaction to resolve uncertainties. Our work aligns with benchmarks like MediQ (Li et al., 2024), which use multi-turn questioning, but we extend our analyses beyond the medical domain. By allowing student-driven inquiry across varied contexts, we examine how proactive dialogue can enable effective learning.

Interactive Learning in Human Tutoring. Evidence from human learning shows that interactive settings consistently outperform passive instruction. Adaptive tutoring through feedback and dialogue yields substantial gains over classroom-based teaching (Bloom, 1984). These improvements stem not only from personalization, but from the structure of interaction itself, specifically, student-led questioning and responsive explanation (VanLehn, 2011). Our setup parallels this structure: the student model initiates queries to target its own uncertainties. Learning emerges through focused, multi-turn exchanges rather than static instruction.

3 Experimental Setup

In this section, we delineate our problem setup (§ 3.1), outline the creation of our datasets (§ 3.2), present the different interaction scenarios (§ 3.3), outline the different models evaluated and our evaluation metric (§ 3.4).

3.1 Problem Setup

In this work, a *concept* refers to a distinct unit of knowledge that captures ideas or information embedded in documents across various domains such as literature, sciences, and current world events. Practically, each concept is instantiated through a context document. For example, the concept of a given movie is represented by its Wikipedia plot, while a scientific concept is captured by a research paper excerpt. Our goal is to explore how a student LLM, ($\mathcal{S}$), can learn such concepts by interacting with a teacher LLM ($\mathcal{T}$).

The student $\mathcal{S}$ can ask any open-ended or information-seeking questions about a concept, while the teacher $\mathcal{T}$ has direct access to the ground-truth context for the concept which it can use to answer those questions. Although one might envision human experts as teachers, large-scale experimentation is more feasible with LLM teachers that can faithfully convey the necessary information. By equipping $\mathcal{T}$ with the source context and $\mathcal{S}$ with only the answers $\mathcal{T}$ provides, we isolate the effects of interactive, inquiry-driven learning.[3]

3.2 Datasets

Since LLMs gain extensive world knowledge from their pre-training on open web-text (Roberts et al., 2020), evaluating their learning abilities on concepts within their pre-training data can lead to ambiguous interpretations. To ensure a robust analysis of concept acquisition, we compiled datasets comprising a range of concepts that are previously unseen by the LLMs. For this, we both automatically scraped and manually compiled song lyrics, movie plots, news articles, academic papers and images, all from after December 2023 (since we tested LLMs pre-trained on data obtained before this period). These documents were collected from platforms such as Genius, Wikipedia, CNN, arXiv and COCO (Lin et al., 2014). This carefully curated dataset spans a range of complexity and information types, enabling a robust evaluation of LLMs' interactive learning performance across various scenarios.

[3]Listing 9 shows the instructions provided to the teacher while answering student questions.

Context	Source	# Contexts	Focus
Song Lyrics	Genius	467	Learning from figurative language
News Articles	CNN	346	Learning factual knowledge
Movie Plots	Wikipedia	214	Learning story elements: characters, events
Academic Papers	arXiv	170	Learning technical knowledge across disciplines
Images	COCO	150	Learning to analyze visual contexts

Table 1: Overview of the different content domains used for evaluation, including the number of contexts, sources, and primary focus areas. Each context type tests distinct capabilities of LLMs.

Concepts Dataset Composition Our evaluation dataset comprises of 1,347 contexts spanning multiple domains. This compilation includes song lyrics, news articles, movie plots, academic papers and images. Table 1 provides an overview of the domains and the number of concepts per domain in the dataset. Further details about the dataset composition can be seen in Appendix C.

Static Lesson Generation for Concepts. We create a 'static lesson' for each concept in our dataset by providing `gpt-4o-2024-08-06` with the context document for the concept, and prompting it to produce a lesson. The generated lesson serves as an initial information source the student might leverage during certain interaction scenarios described in § 3.3. Appendix D provides prompts for lesson generation and samples of lessons.

Quiz Generation for Concepts To measure learning performance, we generate a nine-question quiz per concept, with three levels of difficulty and three questions per level. Questions were crafted using `gpt-4o-2024-08-06`. We used an adversarial filtering strategy to exclude questions that could be answered by a `gpt-4o-mini` model without reference to the provided context. This ensured that each question required eliciting concept information from the provided context. A manual analysis of a random subset of questions by the authors showed that 97% of them satisfied three criteria: (a) good for testing student understanding, (b) answerable using the context, and (c) do not require deeper knowledge beyond the context. Appendix C includes details of the quiz generation process.

3.3 Student-Teacher Interaction Scenarios

In this work, we explore three scenarios to assess the conversational learning capabilities of LLMs:

1. **Static Student with Lesson:** The student only receives the static lesson (no dialogue) before answering quiz questions.
2. **Dynamic Student without Lesson:** The student begins with no prior knowledge and acquires information by asking questions.
3. **Dynamic Student with Lesson:** The student first receives the static lesson, then refines understanding through questions.

These scenarios let us investigate whether interactive questioning can complement or surpass static instruction, and how the quality of teacher and lesson information shapes learning outcomes. We explore five research questions:

- **RQ1:** How well can students learn concepts from static lessons?
- **RQ2:** How well can students learn concepts through interactions?
- **RQ3:** How does the quality of the teacher and lesson affect dynamic student performance?
- **RQ4:** Can borrowed interactions improve student performance?
- **RQ5:** What patterns or features emerge in the questions generated by the student model?

3.4 Models and Evaluation Metrics

Models. We evaluate a range of open and closed-source LLMs as both teachers and students. For text-based domains, we test `gpt-4o-mini` and instruction tuned versions of `LLaMA-3.1-8B/70B`, `Ministral-8B`, `Mistral-Nemo`, and `gemma-2-9B/27B`. For the image domain, we experiment with `gpt-4o-mini`, `Pixtral-12B-2409`, and `LLaMA-3.2-11B-Vision-Instruct` (multimodal LLMs). These models are chosen for their strong language understanding and generation capabilities, which are vital for conversational learning. Unless mentioned otherwise, we provide

the student model with the `gpt-4o` generated lesson in the static and dynamic settings.[4]

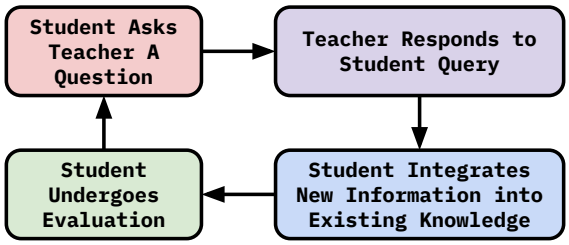


Figure 2: Workflow of the evaluation process in each dialogue turn of the *dynamic interactions* setting, starting with the student asking a concept-related question.

During dynamic interactions, after every dialogue turn, the student model is prompted to integrate the newly acquired information from the ongoing conversation and, when applicable, the prior static lesson. This integration is done through appending the conversation history to the student's context, allowing the model to consolidate knowledge incrementally. The student then uses this updated context to answer quiz questions, as illustrated in Figure 2. For fair comparison, dynamic dialogues are generated with a temperature value set to 1.0, while quiz answers are generated with a temperature of 0. All experiments are repeated across three seeds to ensure robustness.

Our metric for measuring concept learning performance 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.

4 Results

In this section, we present our findings. We begin by establishing a baseline for non-interactive scenarios (§4.1) before exploring how interactive questioning affects concept learning (§4.2). Subsequently, we analyze the role of teacher and lesson quality (§4.3), investigate whether borrowed interactions can substitute for active engagement (§4.4), and consider what factors correlate with successful conversational learning (§4.5).

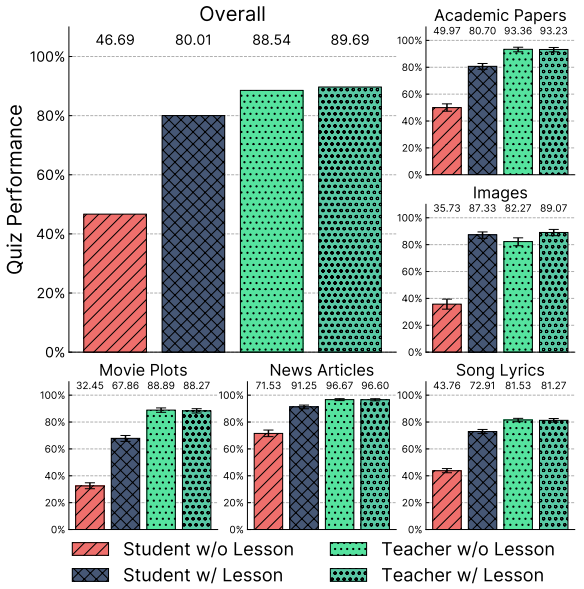


Figure 3: Average quiz performance of student and teacher `gpt-4o-mini` models across different domains. Errorbars indicate the 95% confidence interval calculated by bootstrap. Quiz performance with other LLMs can be found in Appendix Figure 5.

4.1 RQ1: How Well Can Students Learn Concepts from Static Lessons?

We first evaluate students' ability to learn concepts without interaction. Here, the student either receives (1) no knowledge about the concept, or (2) the *static lesson* for the concept. We then compare these students' quiz performance to the teacher's performance. The teacher is given direct access to the context for the concept (and optionally, also the static lesson) and serves as a conceptual upper bound for the student models' performance.

Results. Figure 3 shows the quiz performance of student and teacher `gpt-4o-mini` models across various domains from our dataset, based on their access to information (we see similar trends across other LLMs; see Appendix B). Students provided with no knowledge of new concepts perform above chance, likely relying on pre-training knowledge on the Academic Papers and News Articles domains. This is likely because new concepts build upon previously established theories and facts, respectively, for these domains. When provided with a *static lesson*, student performance improves sig-

[4]We omit evaluations with `gpt-4o` in the student-teacher setup primarily due to the high cost, which made extensive experimentation in dynamic settings infeasible.

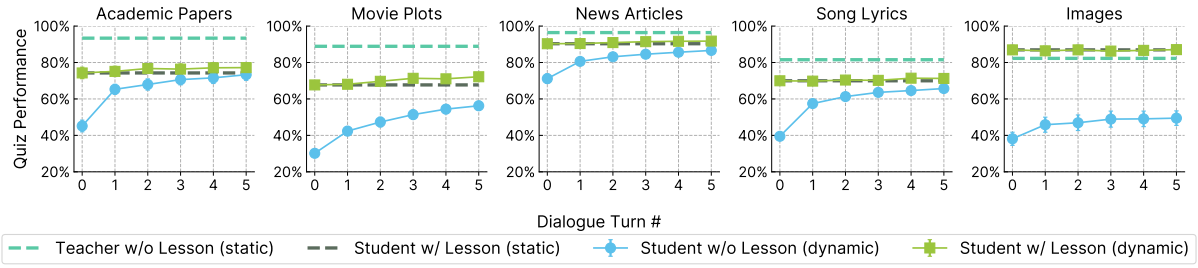


Figure 4: Performance of student `gpt-4o-mini` models across various static and dynamic evaluation settings over five interaction rounds. Errorbars indicate the 95% confidence interval calculated by bootstrap. Performance with other LLMs can be found in App. Fig. 6.

Model	Student w/o Lesson		Student w/ Lesson		Teacher	Quiz Recovery of Student w/o Lesson (%)	
	Start	End (Δ)	Start	End (Δ)	Performance	wrt S w/ L Start	wrt Teacher
gpt-4o-mini	47.91	73.68 (+25.77)	78.83	81.23 (+2.40)	90.05	91.23	81.47
LLaMA-8B	38.12	60.13 (+22.01)	70.88	72.43 (+1.55)	85.70	84.89	75.49
LLaMA-70B	60.34	76.81 (+16.47)	80.58	82.94 (+2.36)	91.24	93.81	85.13
Ministral-8B	33.82	59.66 (+25.84)	67.68	69.36 (+1.68)	82.38	85.14	72.33
Mistral-Nemo	46.81	70.61 (+23.80)	74.06	76.82 (+2.76)	82.05	91.81	81.90
Gemma-9B	47.19	63.74 (+16.55)	75.39	77.08 (+1.69)	89.15	83.21	71.75
Gemma-27B	52.83	65.17 (+12.34)	77.77	79.58 (+1.81)	89.83	86.72	73.47

Table 2: Aggregated performance for LLMs across text domains, showing average start/end performances and recovery percentages for students without/with lessons in the dynamic setting and the static teacher performance. Higher recovery percentages indicate stronger concept understanding as measured by quiz performance.

nificantly ($p < 0.01$[5]) across all domains, though it remains significantly lower ($p < 0.01$) than that of teachers with concept context access in all domains except Images. For this domain, the student with static lesson performs better than the teacher without the lesson, and we posit this is because of the stronger text-based training of `gpt-4o-mini` compared with its image-processing abilities. This trend does not hold with other vision language models like `Pixtral-12B-2409` (Figure 5h). As would be expected, for the text-based domains, teacher performance with or without the static lesson remains largely the same (<1% difference on average across text domains).

4.2 RQ2: How Well Can Students Learn Concepts through Interactions?

Next, we analyze the accuracy of concept learning when a student interactively engages in dialogue with a teacher. We track the learning progress over five rounds of interactions. We compare this to two approaches: (1) a student model that receives a static lesson from the teacher without interaction, and (2) the teacher model that has full knowledge of the concept, which serves as a conceptual upper bound. Notably, exceptions to this upper bound in the Images domain occur for `LLaMA-3.2-11B-Vision-Instruct` and `gpt-4o-mini`, where the student with a static lesson outperforms the teacher.

Main Results. Figure 4 shows that students without a lesson initially lag behind their static-lesson counterparts, but improve substantially after each question-answer exchange for `gpt-4o-mini`. These improvements hold across different language models (+22% absolute gain for LLaMA-8B, Table 2; Figure 6 in Appendix B). Despite steady improvement, most dynamic students do not surpass the static-lesson baseline within five rounds, particularly in text domains. Longer or more strategically guided interactions might close the gap further. In the Images domain, however, performance quickly saturates over the first couple of interaction rounds (see Figure 7 for `Pixtral-12B` and `LLaMA-3.2-11B` results).

[5]We use the student t-test to perform the hypothesis testing across all experiments.

Domain	Round	Question
Song Lyrics	round 1	What inspired Kylie Minogue to create "Hold On To Now," and what themes does the song explore?
Academic Papers	round 2	How do Convolutional Neural Networks (CNNs) process seismic data differently than traditional data analysis methods?
News Articles	round 3	What are the key legal arguments the district attorney has used in his previous clemency requests in other cases, if any?
Movie Plots	round 5	What are some specific scenes or moments from "Problemista" that exemplify Alejandro's unwavering determination in the face of adversity?

Table 3: Samples of Student Questions from Interactions with Teachers for `gpt-4o-mini`.

These results suggest that LLMs can generate comprehensive and domain-relevant questions during the conversation to learn about new concepts, with the potential for further refinement with more interaction rounds (see Figure 8). Table 3 shows examples of questions generated by the student `gpt-4o-mini` during conversations in the student without a static lesson setup. Additional examples for other student models are provided in Table 10.

When starting from a static lesson, adding interaction to the student leads to statistically significant improvements ($p < 0.01$) over the student without a lesson. However, student performance remains significantly lower ($p < 0.01$) than that of teachers in all domains except Images, indicating scope for better interaction strategies. In summary, *while students are capable of learning through interaction, the extent of knowledge acquired via this method significantly lags behind that of teachers*.

4.3 RQ3: Does Teacher and Lesson Quality Influence Dynamic Learning?

We now examine whether improving teacher quality or the initial lesson can enhance interactive learning. Are students paired with stronger teachers or given higher-quality initial lessons better poised to reach teacher-level understanding?

Study Design To evaluate the effects of both lesson quality and teacher strength, we design two complementary experiments:

Context	gpt-4o Lesson		LLaMA-8B Lesson	
	Static	Post-Int.	Static	Post-Int.
Academic Papers	73.02	71.80	70.18	72.59
Movie Plots	58.68	61.69	51.03	60.97
News Articles	87.18	88.20	81.29	88.43
Song Lyrics	64.65	64.63	59.20	64.65

Table 4: Quiz performance comparison for `LLaMA-3.1-8B-Instruct` in the static and dynamic (Post-Interaction) settings using gpt-4o and `LLaMA-3.1-8B-Instruct` static lessons.

1. **Effect of Static Lesson Quality:** We substitute the static `gpt-4o-2024-08-06`-generated lesson provided to a `LLaMA-8B` student with a lesson generated by a `LLaMA-8B` teacher. We then measure the student model's concept learning accuracy under two conditions: (a) static (no interaction) and (b) dynamic (after five interaction rounds with the `LLaMA-8B` teacher, building upon the initial static lesson).
2. **Effect of Teacher Strength:** We pair a `LLaMA-8B` student with a stronger `LLaMA-70B` teacher, which also provides the static lesson. To analyze learning behavior with a weaker teacher, we reverse the setup and have `LLaMA-70B` acts as the student and `LLaMA-8B` as the teacher.

Main Results. From Table 4, we find that a stronger teacher's static lesson (`gpt-4o-2024-08-06`) confers a 3-5% average improvement in static student performance compared to the `LLaMA-3.1-8B-Instruct`-generated lesson. However, after five interaction rounds, the difference narrows considerably, with final scores differing by about 1%. This suggests that while a high-quality initial lesson can boost static accuracy, the dynamic interaction process largely mitigates any initial lesson quality gap.

Student →	LLaMA-8B		LLaMA-70B	
Teacher →	L-8B	L-70B	L-70B	L-8B
Aca. Pap.	71.80	71.63 (-0.17)	83.97	84.31 (+0.34)
Mov. Plts	61.69	61.26 (-0.43)	76.31	76.99 (+0.68)
News Art.	88.20	87.71 (-0.49)	93.55	93.02 (-0.53)
Song Lyr.	64.63	66.16 (+1.53)	77.93	76.54 (-1.39)

Table 5: Post-interaction concept quiz performance comparison when using `LLaMA-3.1-8B-Instruct` and `LLaMA-3.1-70B-Instruct` as teacher models. Note, these are conducted in the student with lesson dynamic setting. (L=LLaMA)

Surprisingly, there is also a minimal difference

in performance from having a stronger teacher (Table 5). A weaker or stronger teacher does not consistently improve the student's final performance after dynamic interactions. This indicates that simply increasing teacher strength does not guarantee deeper student understanding or more effective question-asking behavior. Students fail to consistently capitalize on a teacher's superior knowledge through improved questioning strategies.

These suggest that *while stronger lessons and teachers provide a head start, the depth of student-driven inquiry during interactive learning is a key determinant of concept mastery*.

4.4 RQ4: Can Borrowed Interactions Substitute for Proactive Engagement?

Next we explore the following question: *Can weaker students benefit from previously generated, high-quality interaction transcripts from stronger students — effectively "borrowing" another student's dialogue—without actively participating?*

Study Design. We generate transcripts from an interaction between a strong teacher-student pair (both `LLaMA-70B` models) and provide them as context (similar to the teacher lesson) to a weaker student (`LLaMA-8B`) that never engaged in that particular dialog. We then measure if the weaker student's performance improves from this passive exposure alone.

Eval. LLM →	**LLaMA-8B**		**LLaMA-70B**	
Interactions →	**L-8B**	**L-70B**	**L-70B**	**L-8B**
Aca. Pap.	67.94	65.09 (-2.85)	80.43	81.55 (+1.12)
Mov. Plts	41.38	44.30 (+2.92)	65.98	61.83 (-4.15)
News Art.	76.29	77.04 (+0.75)	87.57	87.47 (-0.10)
Song Lyr.	56.91	59.75 (+2.84)	73.26	70.66 (-2.60)

Table 6: Concept quiz performance when using *student w/o lesson* interactions of `LLaMA-8B` and `LLaMA-70B` as context for `LLaMA-8B` and `LLaMA-70B` student models.

Eval. LLM →	**LLaMA-8B**		**LLaMA-70B**	
Interactions →	**L-8B**	**L-70B**	**L-70B**	**L-8B**
Aca. Pap.	71.80	72.42 (+0.62)	83.97	84.44 (+0.47)
Mov. Plts	61.69	61.67 (-0.02)	76.31	76.62 (+0.31)
News Art.	88.20	87.77 (-0.43)	93.55	93.38 (-0.17)
Song Lyr.	64.63	66.43 (+1.80)	77.93	77.15 (-0.78)

Table 7: Concept quiz performance when using *student w/ lesson* interactions of `LLaMA-8B` and `LLaMA-70B` as context for `LLaMA-8B` and `LLaMA-70B` student models.

Main Results. Tables 6 and 7 show that passive exposure to borrowed transcripts does not significantly improve performance. Conversely, we also observe that observing the interactions of a weaker student-teacher combination does not diminish performance considerably across most scenarios (with the exception of the Movie Plots and Song Lyrics domains in the student without lesson dynamic setting). This suggests that the benefits of interactive learning are not about conversation quality, but the student's capacity to appropriately ask questions. Thus, *passive exposure to high-quality content cannot substitute for pro-active engagement*.

4.5 RQ5: What Interaction Factors Predict Student Learning Gains?

Finally, we examine which aspects of teacher-student interactions drive better learning outcomes.

Study Design. We first collect a comprehensive set of 53 interaction-related features[6] that can be computed from the teacher-student transcripts. These factors include syntactic features of the student/teacher responses, statistics about tokens, metrics of linguistic complexity, semantic relatedness of questions and responses, among others. For each domain and configuration with `gpt-4o-mini` as student and teacher models, we extract these features from the recorded interactions and aggregate them into a feature matrix. We then train a random forest regressor, using best-found hyperparameters from cross-validation, to predict the learning gain of the student model after each round of interaction with a teacher. Performance is evaluated using the R^2 score on a held-out test set.

Main Results. In most domains, the predictive power of our feature set is low: R^2 scores on held-out data are often close to zero. However, for the song lyrics domain, R^2 scores reach up to 0.14, suggesting that meaningful signals can be extracted in specific conditions. The top contributing features are cumulative exposure (number of unique tokens), overlap between quiz questions and student questions, semantic alignment between student and quiz questions, response information density, and response correctness. This suggests that *our current feature sets and metrics are insufficient for robustly capturing the nuanced factors driving interactive learning success*. Identifying richer metrics remains a key challenge.

[6]We provide the full list of features in Appendix Table 12.

5 A Future of Conversational Learning

Our findings suggests promise for a new paradigm for learning in LLMs: by shifting from static data absorption to interactive, curiosity-driven dialogue. The INTERACT framework and curated dataset provide a fertile testing ground for refining this paradigm, where learning experiences are not delivered by a teacher but co-constructed by learners. Despite the benefits of interactive learning, current student models lag behind teachers, indicating the need for better interaction strategies.

Future work can explore extending existing machine learning theories, such as active learning, to 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. They also point to concrete applications: dynamic AI tutors humans with evolving lesson plans, accelerated scientific discovery, and enhanced reasoning with images, audio, or video data. Investigating methods for long-term retention of knowledge acquired through interaction will also be critical. At its heart, this line of research nudges us toward a future where machine learning systems not only learn from us, but learn with us.

Limitations

Our investigation here has some important limitations. A significant limitation lies in our evaluation method. While quiz performance is informative, it may not capture all aspects of concept understanding. This metric 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.

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.

6 Acknowledgements

This work was supported by the NSF grant DRL2112635. The authors would also like to thank anonymous reviewers for valuable feedback to improve the manuscript.

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.

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*.

Benjamin S. Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6):4–16.

Benjamin Samuel Bloom and David R. Krathwohl. 1966. Taxonomy of educational objectives. handbook i: Cognitive domain.

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*.

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.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. *CoRR*, abs/2109.12595.

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. *CoRR*, abs/2011.06623.

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.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z Li. 2024. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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.

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

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rakesh R. Menon, Sayan Ghosh, and Shashank Srivastava. 2022. CLUES: A benchmark for learning classifiers using natural language explanations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6523–6546, Dublin, Ireland. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Ashwin Ram. 1991. A theory of questions and question asking. *Journal of the Learning Sciences*, 1(3-4):273–318.

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.

Alexis Ross and Jacob Andreas. 2024. Toward in-context teaching: Adapting examples to students' misconceptions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13283–13310, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

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 Extended Related Work

Knowledge Distillation and Interactive Reasoning. Knowledge distillation methods traditionally train student models on outputs generated by teachers, often leading to discrepancies between training and inference data (Kim & Rush, 2016; Sanh et al., 2019). Generalized Knowledge Distillation (Agarwal et al., 2023) addresses this by incorporating self-generated sequences and teacher feedback. While these approaches align student outputs with teacher feedback, our work evaluates learning through dynamic interactions, focusing on how students refine their knowledge by engaging in conversation and inquiry. Interactive frameworks also highlight the importance of dialogue in improving LLM performance. Studies show that LLMs benefit from human explanations (Wei et al., 2022; Lampinen et al., 2022), self-generated feedback (Madaan et al., 2023; Chen et al., 2024), and fine-tuning on explanations from larger models (Ho et al., 2023). Our approach extends these findings by focusing on the student's ability to ask informative questions, enabling richer teacher explanations and deeper learning.

Dynamic Question Generation and Simulated Interaction. Dynamic question generation methods have demonstrated that pre-trained models can tailor questions based on a student's knowledge state (Srivastava & Goodman, 2021). Their LM-KT model personalizes questions to match student proficiency, enhancing learning outcomes compared to static question pools. While this approach centers on teacher-driven question generation, our work focuses on student-driven questioning, where students actively inquire to address their knowledge gaps. Simulated environments further highlight the role of interactive dialogue in assessing LLM capabilities. Frameworks like SOTOPIA (Zhou et al., 2024) and COBLOCK (Wu et al., 2024) evaluate social intelligence and collaboration through multi-turn interactions. These studies emphasize adaptability and communication in achieving shared goals. In contrast, our work explores how student-teacher dialogues facilitate concept learning, emphasizing the student's role in refining knowledge through inquiry across diverse domains.

B Additional Results

B.1 Interaction Utilization

As shown in Section 4.4, passive exposure to high-quality content is insufficient to replace proactive engagement. Specifically, providing transcripts of interactions between a stronger teacher-student pair to a weaker student does not lead to comparable performance. We also observe that interactions involving weaker students do not substantially degrade overall performance. To assess whether models genuinely leverage the observed interactions, we conduct an ablation study in which we replace meaningful interactions with irrelevant ones.

Student →	**w/o Lesson**		**w/ Lesson**	
Interactions →	**Orig.**	**Random**	**Orig.**	**Random**
Academic Papers	78.54	47.33 (-31.21)	82.56	47.17 (-35.39)
Movie Plots	57.85	34.11 (-23.74)	73.94	34.21 (-39.73)
News Articles	87.40	73.11 (-14.29)	92.52	72.95 (-19.57)
Song Lyrics	68.92	45.84 (-23.07)	74.94	45.53 (-29.41)
Images	49.47	23.73 (-25.73)	87.07	23.60 (-63.47)

Table 8: Quiz performance comparison for `gpt-4o-mini` with its original and random interactions for students without/with lessons. Random interaction results are averaged across 3 seeds.

Study Design. For each domain, we sample 100 concept interactions and replace them with interactions from a randomly selected, unrelated concept within the same domain. In this setup, the interactions are sourced from dialogues between `gpt-4o-mini` student and teacher models.

Main Results. Table 8 shows that exposure to irrelevant transcripts results in a significant performance decline. This indicates that students are indeed making effective use of the observed interactions.

B.2 Additional Static Results

Figure 5 shows the performance of various other LLMs on the static interactions. Teacher LLMs consistently outperform student LLMs, with their direct access to original material providing them with comprehensive contextual knowledge. Their near-perfect scores set a high bar for student LLMs. When teacher LLMs receive additional lessons, their performance improves only marginally. This suggests that, while summaries are beneficial, the

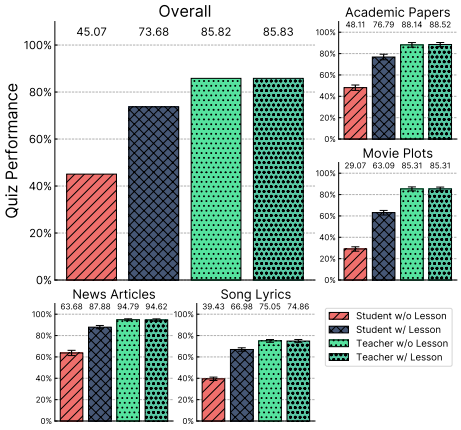
Table 9: Default hyperparameters for the dynamic setting experiments. Static configuration uses the same hyperparameters with the student quiz evaluation module alone.

Feature Name	Description	Value
Experiment Parameters		
num_interaction_rounds	Total number of interaction rounds.	5
seed	Random seed for reproducibility.	0
Student Question Generation		
max_tokens	Maximum tokens for question generation.	256
temperature	Sampling temperature for question generation.	1
Student Quiz Evaluation		
max_tokens	Maximum tokens for quiz evaluation.	10
temperature	Sampling temperature for quiz evaluation.	0
Student Summary Generation		
max_tokens	Maximum tokens for summary generation.	256
mode	Mode for summary aggregation.	concat
temperature	Sampling temperature for summary generation.	0.7
Teacher Answer Generation		
lesson_mode	Lesson provider.	`gpt-4o-2024-08-06`
max_tokens	Maximum tokens for answer generation.	512
temperature	Sampling temperature for answer generation.	0.7

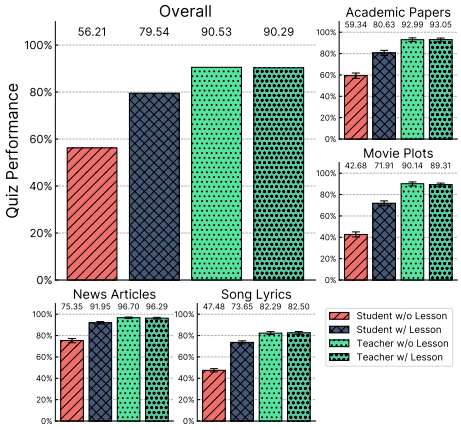
original material already covers the essential information comprehensively, and teacher LLMs' access to detailed source material is crucial to their high performance. Overall, the substantial underperformance of the student LLM compared to the teachers highlights the challenge posed by our datasets to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

B.3 Additional Dynamic Results

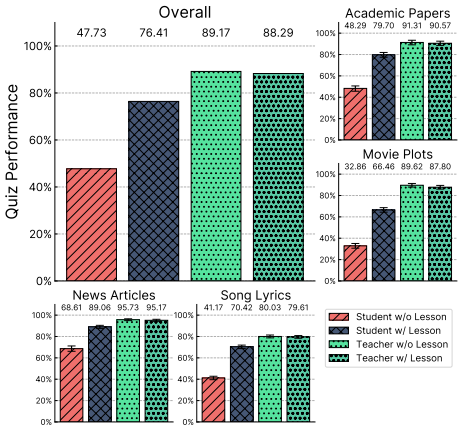
Figure 6 shows the performance of various other LLMs on the dynamics interactions. Table 10 provides some example questions the `gpt-4o` student LLM asked within dynamic interactions.



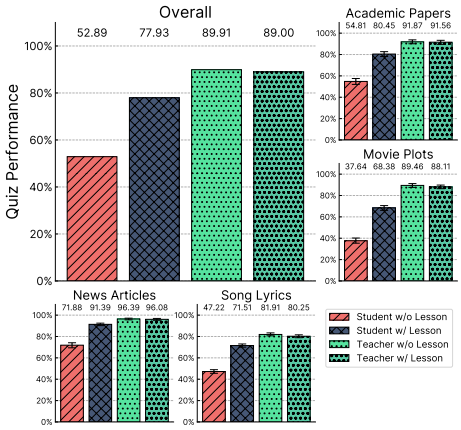
(a) **LLaMA-3.1-8B-Instruct**



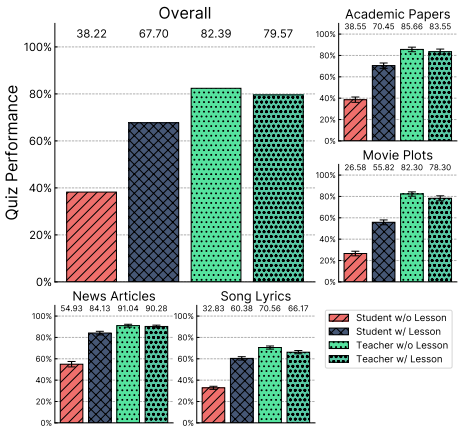
(b) **LLaMA-3.1-70B-Instruct**



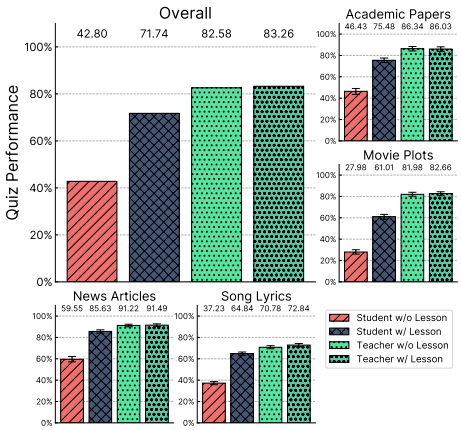
(c) **gemma-2-9B-IT**



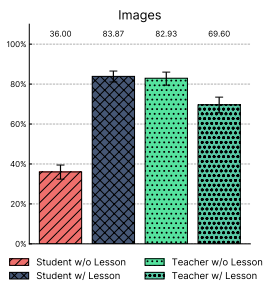
(d) **gemma-2-27B-IT**



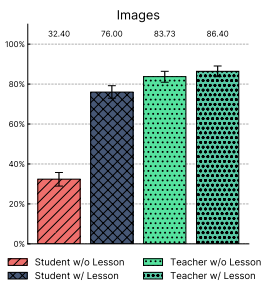
(e) **Ministral-8B-Instruct**



(f) **Mistral-Nemo-Instruct**

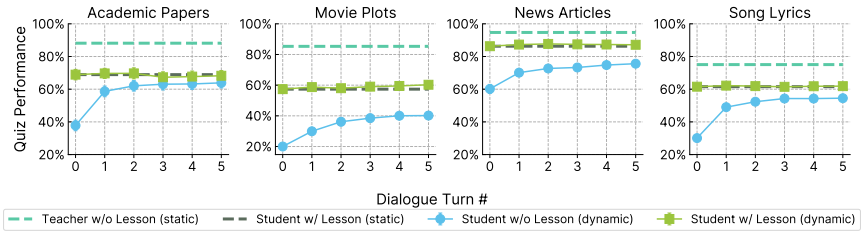


(g) **LLaMA-3.2-11B-Vision-Instruct**

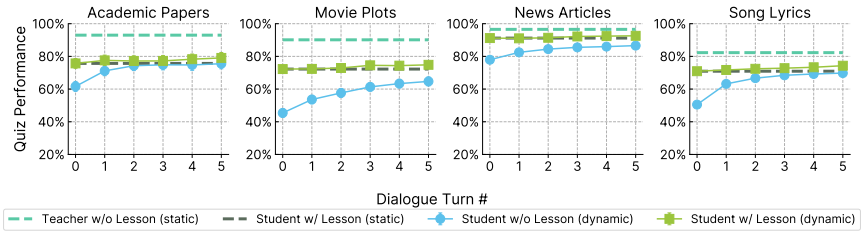


(h) **Pixtral-12B-2409**

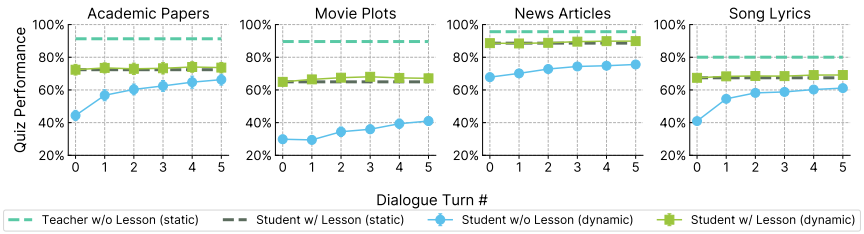
Figure 5: Average quiz performance of student and teacher LLMs across different domains. Errorbars indicate the 95% confidence interval calculated by bootstrap.



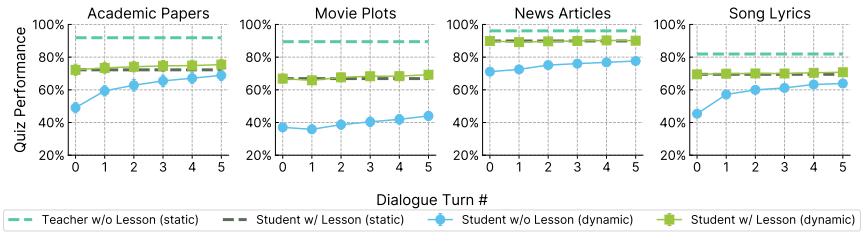
(a) **LLaMA-3.1-8B-Instruct**



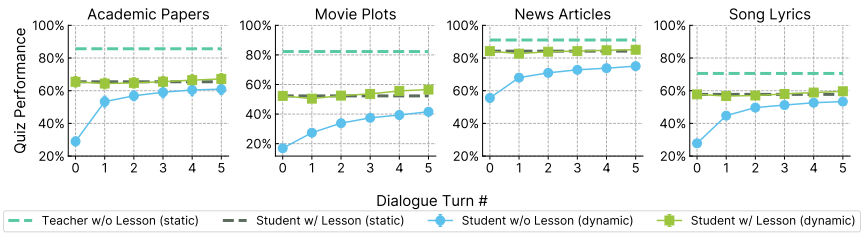
(b) **LLaMA-3.1-70B-Instruct**



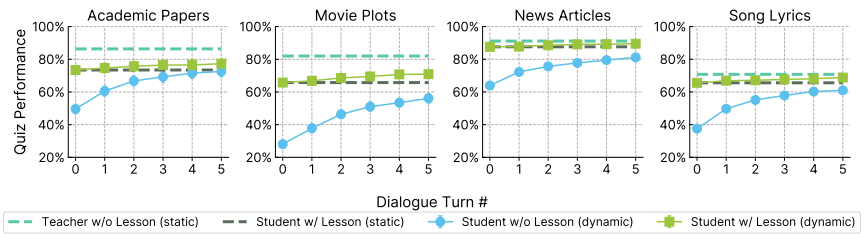
(c) **gemma-2-9B-IT**



(d) **gemma-2-27B-IT**

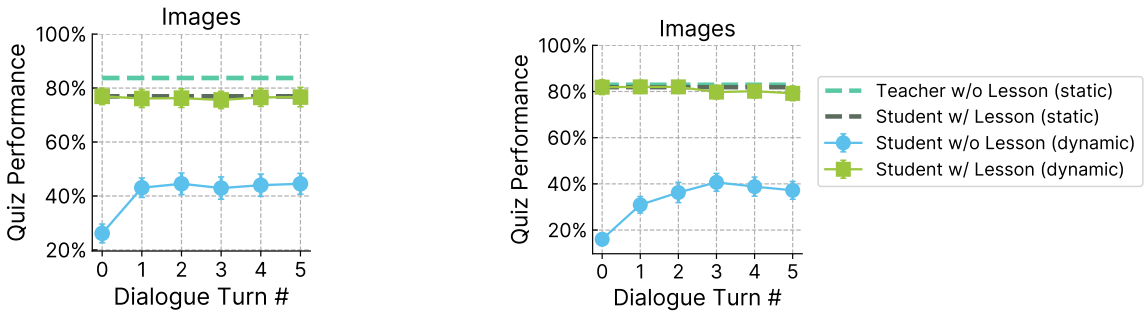


(e) **Ministral-8B-Instruct**



(f) **Mistral-Nemo-Instruct**

Figure 6: Performance of student LLMs across various static and dynamic evaluation settings. Errorbars indicate the 95% confidence interval calculated by bootstrap.



(a) **`Pixtral-12B-2409`** (b) **`LLaMA-3.2-11B-Vision-Instruct`**

Figure 7: Performance of student MLLMs across various static and dynamic evaluation settings. Errorbars indicate the 95% confidence interval calculated by bootstrap.

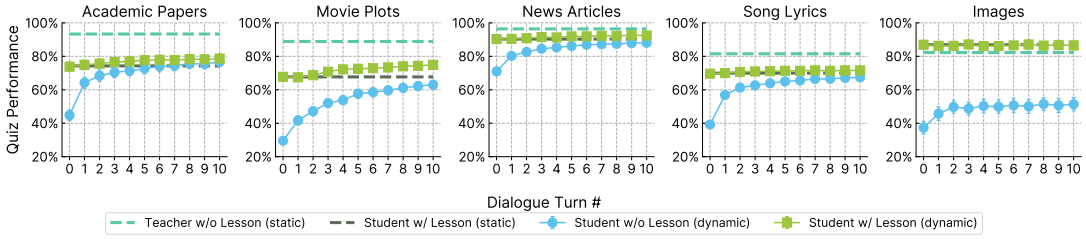


Figure 8: Performance of `gpt-4o-mini` across various static and dynamic evaluation settings when provided with more interaction rounds. Errorbars indicate the 95% confidence interval calculated by bootstrap.

Table 10: Sampled Student Questions from Interactions

Model	Domain	Round	Question
gpt-4o-mini	news articles	round 4	What health or safety considerations should individuals keep in mind when shaping their eyebrows, especially when adopting trends like thin brows?
	news articles	round 3	What specific actions or policies have been proposed or implemented by colleges in response to the House committee's findings on antisemitic harassment?
	movie plots	round 1	What is the central theme explored in "The Bikeriders," and how does it reflect the culture or lifestyle of biker communities?
	movie plots	round 5	What role does humor play in movies featuring "Space Cadet" characters, and how does it enhance the storytelling or character development?
	movie plots	round 3	What themes and motifs are commonly associated with hit man characters in films, and how do these elements enhance the storytelling?
	academic papers	round 2	How do Convolutional Neural Networks (CNNs) process seismic data differently than traditional data analysis methods?
	academic papers	round 3	What specific practical applications or investment strategies can be developed using "Dynamic Factor Allocation Leveraging Regime-Switching Signals"?
	academic papers	round 3	What were the key findings of the study regarding the relationship between the shape of brain connections and cognitive performance?
	song lyrics	round 1	What inspired Kylie Minogue to create "Hold On To Now," and what themes does the song explore?
	song lyrics	round 5	How does the title "Comatose" relate to the overall themes presented in the song, and what significance does it hold in the context of the lyrics?
LLaMA-3.1-8B-Instruct	news articles	round 5	How does the role of social media platforms, particularly in amplifying and spreading anti-vaccine misinformation, contribute to the decline in public confidence in vaccines, as seen during the Trump presidency?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	news articles	round 2	What motivated Nima Rinji Sherpa to pursue climbing the "eight-thousanders" despite being just 17 years old, and how has he maintained his interest and commitment to mountaineering throughout his journey?
	news articles	round 5	What specific actions or responses from others, if any, demonstrate the impact and effectiveness of D'Pharaoh Woon-A-Tai's bold Emmys statement in raising awareness about the Missing and Murdered Indigenous Women movement?
	movie plots	round 3	Considering the narrative's creative interpretation of Japanese folklore, how does the character of Hiiragi's transformation from a suppressed emotional state to becoming an Oni relate to the traditional Japanese concept of "Yūgen" (mysterious and profound sense of the beauty of the world), and what insights does it offer into the human experience?
	movie plots	round 2	What specific historical events or cultural movements might have influenced the meaning of "The Beautiful Game" to shift from a focus solely on the sport's elegance and skill to its positive social impact and role in promoting social change?
	movie plots	round 2	How does the concept of deception, such as Marisol's attempt to hide her husband's condition, relate to the broader theme of social dynamics during a zombie apocalypse, and what implications does it have for the characters and the story?
	academic papers	round 4	What specific contributions have Deborah Colleen and Steve Kol steril fl iterators Arthur Choi have made in the research of "Superpolynomials of Algebraic Links", and how have their contributions impacted the broader field of mathematics?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	academic papers	round 3	What are the typical CT scan characteristics or image features that facilitate the employment of AI algorithms, such as CNNs, in identifying and distinguishing between malignant and benign pulmonary nodules?
	song lyrics	round 3	What specific line or lyrics in the song "Find Myself Again" by Blanks is an example of expressing the tension between the desire for comfort and security, and the need for growth and change, as mentioned in the teacher's current understanding?
	news articles	round 5	What are some potential long-term consequences for high- income families making over $150,000 who are living paycheck to paycheck, and how might this impact their financial stability and overall well-being in the future?
LLaMA-3.1-70B-Instruct	news articles	round 3	How do the different portrayals of Joan of Arc in art, fashion, and film reflect and shape societal attitudes towards women, power, and identity, and what do these varied interpretations reveal about the cultural significance of her legend?
	movie plots	round 3	How does the historical context of the 1924 Paris Olympics and the societal attitudes towards women in sports during that era influence the narrative and themes presented in "Young Woman and the Sea"?
	movie plots	round 5	What is the significance of the Cheeks family circus in "Saving Bikini Bottom: The Sandy Cheeks Movie", and how does its final performance at the Krusty Krab relate to the overall themes of "Family and Support" and "The Power of Friendship"?
	movie plots	round 1	Can you explain the core theme or message that the concept of "Upgraded" typically conveys in movie plots, and how it relates to human emotions or societal commentary?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	academic papers	round 4	How do researchers using algebraic approaches to cosmological integrals ensure the accuracy and reliability of their computations and results, especially when dealing with complex and multivariable systems?
	academic papers	round 2	Can you explain how the integration of convolutional neural networks with Wiener filters addresses the limitations of traditional methods in characterizing seismic isolation systems, particularly in the context of gravitational wave detectors?
	academic papers	round 5	How do the results from coarse-grained modeling, specifically those related to gas adsorption at the nanoparticle level, contribute to the development of more efficient and selective metal-organic framework/polymer composite materials for gas separation applications?
	song lyrics	round 2	What role does the artists' use of language play in conveying the themes and emotions present in the song "Muppy", and how does their lyrical style contribute to the overall tone and message of the song?
	song lyrics	round 2	How does the artist, Quando Rondo, use the themes of loss and heartbreak to underscore the struggle with the consequences of one's actions in the song "Gotta Do Better & Pray"?
	news articles	round 1	1. What context or background information led to the discovery of a long-lost copy of the US Constitution in a filing cabinet?
	news articles	round 3	What emotional or musical characteristics were notable in the recently discovered Chopin waltz to distinguish it from other works of his?
`Ministral-8B-Instruct`	news articles	round 3	What are the key legal arguments the district attorney has used in his previous clemency requests in other cases, if any?
	movie plots	round 5	What role does comedy play in the story arc of "The People's Joker" in relation to the protagonist's personal journey and character development?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	movie plots	round 1	What is the primary theme of the movie "Road House"?
	movie plots	round 2	What are the key symbols and imagery used in "Demon Slayer: Kimetsu no Yaiba – To the Hashira Training"?
	academic papers	round 3	Could you explain the role of geography in this concept, and specifically how the Soviet Union's geographical interests and strategic positions influenced the propaganda's representation of cities? *Teacher, please respond before I ask the next question.*
	academic papers	round 2	What specific challenges does the survey identify as prominent in the application of large language models to algorithm design?
	academic papers	round 1	Which are the primary challenges addressed by the concept of "Privacy for Free in the Over-Parameterized Regime"?
	song lyrics	round 5	Why does the song "No Love" by Marjinal use Whatever attendance worksheet backwards by rachal owner lapie ?
	song lyrics	round 1	What are the primary themes and emotions conveyed in the song "Darling" by Ali Gatie?
	song lyrics	round 1	What are the primary themes present in the song "Or What" by NLE Choppa & 41?
	news articles	round 2	What was the reason given by Iranian authorities for executing Jamshid Sharmahd?
	news articles	round 2	What did the DNA analysis specifically reveal about the fate of the crew of Sir John Franklin's expedition?
`Mistral-Nemo-Instruct`	news articles	round 4	What are some unique interactive experiences or activities visitors can expect at this barbecue museum, besides learning about different styles and regions?
	movie plots	round 5	Can you analyze the role of trust in the relationship between the two "Wolfs" or fixers in the movie plots domain? How does their mutual trust affect their ability to work together and maintain their loyalty and honor?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	movie plots	round 5	Can you provide an example of a movie where "It's What's Inside" is explored in a non-fantastical or supernatural way, using a different plot device?
	movie plots	round 1	What is the origin story of Ricky Stanicky in the movie plot domain?
	academic papers	round 4	How does the presence of alternative folds affect AlphaFold's ability to predict the functional properties of proteins, such as their interactions with other biomolecules or their cellular localization?
	academic papers	round 5	What are the potential benefits of using a light-weight feature-attention module in the FACMIC model, aside from reducing communication costs?
	academic papers	round 2	How does the brain-aware readout layer improve interpretability compared to traditional approaches?
	song lyrics	round 4	What are the specific themes and messages conveyed in "Stuck Inside (CG5 Remix)" and how does the collaboration between Black Gryph0n, CG5, & The Living Tombstone enhance or alter these themes?
	song lyrics	round 0	N/A
	news articles	round 1	What is the geographic location of this shipyard and the potential impact of the fire on the surrounding environment?
	news articles	round 3	What are some historical examples of unexpected events that significantly impacted the jobs report and how did they influence the economic outlook at the time?
`gemma-2-9B-IT`	movie plots	round 5	What are some specific scenes or moments from "Problemista" that exemplify Alejandro's unwavering determination in the face of adversity?
	movie plots	round 3	Could you give some examples of movies that depict civil wars and how they exemplify those themes of loyalty and betrayal you mentioned?
	movie plots	round 5	What are some concrete examples of "societal control" and "individual autonomy" struggles depicted in movies that explore the concept of "Humane"?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	academic papers	round 3	What are some of the specific advantages of using "Fluorescence mass spectrometry" for detecting trapped O+2 ions compared to other detection methods?
	academic papers	round 3	Could you provide an example of a real-world application where the Improved Velocity-Verlet Algorithm proves particularly beneficial for the Discrete Element Method?
	academic papers	round 5	What specific performance metrics were used to evaluate the effectiveness of EViT-UNet for medical image segmentation, and how did it compare to existing methods in terms of accuracy, speed, and resource usage?
	song lyrics	round 5	How does the call-and-response element of the song contribute to the theme of unity and togetherness?
	song lyrics	round 2	How does DRAM use imagery and metaphors in "Marry Me" to convey the depth and intensity of his feelings?
	song lyrics	round 3	Can you tell me more about the musical structure of "Back To The Wall" and how it contributes to the overall message of the song?
	news articles	round 1	What specific historic milestone did Nikola Jokic achieve?
	news articles	round 5	How do the traditional Ayurvedic and homeopathic practices offered at Soukya International Holistic Health Centre align with King Charles' known interest in alternative medicine and sustainability?
`gemma-2-27B-IT`	movie plots	round 5	How does the film "Apartment 7A" use sound and music to create an atmosphere of suspense and unease?
	movie plots	round 1	What are some common themes or motifs explored in stories within the "My Oni Girl" genre?
	academic papers	round 3	What are some examples of specific applications where the Diffusion-PINN Sampler has been successfully used or is particularly well-suited?
	academic papers	round 5	What are some potential limitations or challenges associated with this method of modeling dynamic neural activity?

Continued on next page

Table 10 – continued from previous page

Model	Domain	Round	Question
	academic papers	round 3	What are the potential benefits and implications of using Large Language Models for algorithm design outlined in the survey?
	song lyrics	round 4	Can you provide any insights into the lyrical choices made by American Authors in their rendition of "Sweet Caroline" compared to Neil Diamond's original version?
	song lyrics	round 2	Can you provide examples of specific lyrics from the song that illustrate the themes of neighborhood violence and retaliation?

Domain Name	Context Count	Avg. # Tokens
Images	150	-
Movie Plots	214	671.3
Song Lyrics	467	296.9
Academic Papers	170	1560.8
Computer Science	23	1438.5
Economics	16	1424.2
Electrical Engineering	25	1443.3
Mathematics	23	1667.6
Physics	22	1781.7
Quantitative Biology	22	1752.1
Quantitative Finance	15	1462.4
Statistics	24	1507.1
News Articles	346	1163.7
Business	38	1068.5
Entertainment	30	627.7
Health	28	1372.8
Politics	43	1690.6
Science	27	1406.4
Sports	41	1232.1
Style	51	1023.0
Travel	10	1026.6
US News	28	1166.7
World News	50	981.7
Total	1,347	967.1

Table 11: Dataset composition including context counts across domains and average token counts.

C Dataset Creation

C.1 Data Collection

Datasets were compiled by using a mix of API, scraping, and manual collection of data starting from January 2024 to obtain song lyrics, movie plots, news articles, and academic papers.[7] 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. Table 11 shows a breakdown of the dataset distribution, showcasing the diversity and scope of the data collected.

[7]This temporal criterion was strategically chosen to ensure that the data used was not previously encountered by the GPT-4o model, thus eliminating potential biases or prior knowledge that could influence the model's performance in teaching and learning scenarios.

Movie Plots The dataset for movie plots was compiled by scraping Wikipedia pages under the Creative Commons Attribution-ShareAlike 3.0 License. This method complies with Wikipedia's robot policy, ensuring ethical scraping practices. Only movie plots released from January 2024 onwards were included to ensure data relevance and alignment with GPT-4o's latest knowledge cutoff date of October 2023. The scraping process adhered strictly to Wikipedia's terms, ensuring attribution is maintained, and derivatives follow the same licensing requirements. This inclusion ensures the experimental results remain unaffected by prior knowledge encoded in GPT-4o's training data.

Song Lyrics Song lyrics were collected using the Genius API via the LyricsGenius Python client. Previously, scraping was used, but this was transitioned to API usage to comply with Genius's Terms of Service, which explicitly prohibit scraping while permitting data retrieval via their API. Only lyrics from songs released after January 2024 were included, ensuring alignment with GPT-4o's knowledge cutoff. This transition to API usage guarantees the dataset's legality and ethical compliance, avoiding any terms-of-service violations while maintaining the integrity of the collected data.

News Articles News articles were collected from CNN during a one-day span in November 2024 by downloading raw HTML pages. Articles were categorized into topics such as politics, world, business, and entertainment. Only articles published from January 2024 onwards were included. CNN's Terms of Service permit automated content retrieval for academic purposes, provided it does not manipulate page views or server traffic. This ensured the legality of this data collection process. Furthermore, the inclusion of recent articles minimizes the risk of duplicating pre-existing knowledge in GPT-4o, ensuring up-to-date and unbiased context for research purposes.

Academic Papers Academic papers were sourced from arXiv using their API, in compliance with their Terms of Use, which allow retrieval and utilization of e-prints for research purposes. Only papers released in October 2024 were included, focusing on fields such as computer science, mathematics, and economics. The first 1,500 words of each paper were extracted using arXiv's

Table 12: Feature Names, Descriptions, and Details

Feature Name	Description	Details
Question-Level Features		
Question Length	Number of tokens or words in the student's question.	Token count via spaCy.
Question Complexity	Syntactic complexity via average parse tree depth.	Calculated using spaCy's dependency parsing.
Lexical Sophistication	Average word length as a proxy for rarity.	Based on average token length.
Named Entity Count	Number of named entities in the student's question.	Utilizes spaCy's Named Entity Recognition (NER).
Question Informativeness	Number of unique domain-relevant keywords.	Intersection with predefined domain keywords.
Question Directness	Presence of a question mark indicating clarity.	Binary: 1 if '?', else 0.
Politeness/Hedging	Count of politeness or hedging words like "maybe" or "could".	Uses a predefined set of hedging words.
Question Type	Categorizes by type (e.g., Who, What, Where).	Binary indicator for specific question starters.
Question Novelty	Semantic difference from previous questions.	Cosine similarity of embeddings (computed with a hash function).
Question Specificity	Focus based on named entities presence.	Binary: 1 if entities >0, else 0.
Bloom's Taxonomy	Corresponding Bloom's taxonomy level (Bloom and Krathwohl, 1966).	A binary feature for each level. Utilizes `gpt-4o` to decide the taxonomy.
Teacher-Response-Level Features		
Response Length	Number of tokens or words in the teacher's response.	Token count via spaCy.
Info Density	Ratio of informational tokens (nouns, proper nouns) to total tokens.	Based on POS tagging with spaCy.
Response Novelty	Amount of new content vs. previous responses.	Cosine similarity of embeddings (computed with a hash function).
Response Correctness	Factual correctness via QA model score.	Utilizes Hugging Face's QA pipeline (distilbert-base-cased-distilled-squad).
Response Completeness	Whether the response fully addresses the question.	Binary based on QA model score (>0.5 considered complete).
Response Complexity	Syntactic complexity of the teacher's response.	Calculated using spaCy's dependency parsing.
Response Sentiment	Sentiment score for the response.	Utilizes Hugging Face's text classification pipeline (clapAI/roberta-base-multilingual-sentiment)
Entity Diversity	Variety of named entities in the response.	Utilizes spaCy's NER to extract and count unique entities.
Temporal Positioning	Presence of chronological cues in the response.	Count of temporal keywords.
Use of Examples	Presence of illustrative phrases like "for example".	Binary based on phrases like "for example", "such as".
Interaction-Dynamics Features		
Turn Index	Current round number in the interaction.	Sequential indexing starting from 1.
Cumulative Exposure	Number of unique facts introduced so far.	Unique token count.
Student Adaptation	Change in question complexity from the previous round.	Difference in complexity scores.
Teacher Adaptation	Change in response complexity from the previous round.	Difference in complexity scores.
Information Gain	Semantic difference from the previous response.	Cosine similarity of embeddings (e.g., SentenceTransformers).
Topic Shifts	Shift to a new aspect of the concept.	Cosine similarity of embeddings.
Unanswered Queries	Count of unanswered questions.	Uses `gpt-4o` to determine whether a given response answers the corresponding question.
Progressive Elaboration	Degree of building upon earlier knowledge.	Based on response length trends.
Student Context Coverage	The cumulative overlap between question entities so far and context entities.	Utilizes spaCy's NER to extract and count unique entities.
Teacher Context Coverage	The cumulative overlap between response entities so far and context entities.	Utilizes spaCy's NER to extract and count unique entities.
Student Quiz Coverage	The cumulative overlap between question tokens so far and quiz question tokens.	The ratio of unique quiz tokens covered by the questions.
Teacher Quiz Coverage	The cumulative overlap between response tokens so far and quiz question tokens.	The ratio of unique quiz tokens covered by the responses.
Student Semantic Alignment	The average maximum similarity between each quiz question and the questions asked so far.	Cosine similarity of embeddings.
Teacher Semantic Alignment	The average maximum similarity between each quiz question and the responses so far.	Cosine similarity of embeddings.
Linguistic/Style Features		
Lexical Diversity (Student)	Type-token ratio in student questions.	Unique/total words ratio.
Lexical Diversity (Teacher)	Type-token ratio in teacher responses.	Unique/total words ratio.
Domain-Specific Terms	Frequency of domain keywords in text.	Intersection with predefined domain keywords.
Sentence Length Variability	Std. deviation of sentence lengths in responses.	Calculated using spaCy's sentence segmentation.
Readability Score	Readability via Flesch-Kincaid or similar.	Utilizes the `textstat` library.
Passive Voice Count	Number of passive constructions in the response.	spaCy dependency `auxpass` count.
Modal Language Count	Frequency of modal/uncertain words.	Based on a predefined set of modal words.
Semantic/NLP Features		
Similarity to Summary	Alignment with reference summary.	Cosine similarity of embeddings (computed with a hash function).
Coreference Complexity	Number of coreference chains.	Uses spaCy-coref pipeline.
Semantic Cohesion	Similarity with all previous responses.	Cosine similarity of embeddings (computed with a hash function).
Coverage of Key Plots	Fraction of key plot elements mentioned.	Presence of key plot terms.
Performance/Contextual Features		
Prior Knowledge Estimate	Initial knowledge based on pre-interaction quiz.	Initial quiz score.
Student Confidence	Quiz accuracy score (0-100).	Numeric value from `student_evaluation`.
Improvement in Questions	Trend in question clarity/specificity.	Difference from first round complexity.
Instruction-Following Score	Adherence to teacher instructions.	Uses reward model scores as proxy.
Redundancy in Answers	Fraction of repeated information.	Token overlap with previous responses.
Politeness/Social Cues	Presence of courteous language.	Predefined polite words like "please", "thank you".
Meta-Linguistic Feedback	References to previous turns.	Binary: Phrases like "as mentioned".

beta HTML renderer. In cases where the HTML renderer was unavailable, PDFs were processed instead. This method ensured textual consistency, avoided formatting issues, and complied with arXiv's API guidelines. By restricting papers to those published after January 2024, the dataset ensures it is free from pre-existing knowledge encoded in GPT-4o's training data.

Images The Visual Question Answering dataset was sourced from the COCO image collection, which is available under the Creative Commons Attribution 4.0 License. This dataset enhances the multimodal aspect of the study by integrating visual contexts alongside text. While the textual datasets were restricted to content published after January 2024 to avoid influencing GPT-4o with pre-existing knowledge, the inclusion 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 GPT-4o's ability to analyze and interpret visual information. This distinction justifies the inclusion of older images to expand the scope of visual contexts without compromising the experimental results.

C.2 Quiz Questions

Quiz Question Complexities. To ensure comprehensive evaluation of the LLM's learning capabilities, quiz questions were designed across three levels of complexity: Middle-School, College, and Graduate. Each level progressively increases in difficulty and depth, as described below.

- **Middle-School Level Questions:** These questions test foundational understanding by focusing on basic recall of facts, definitions, or direct observations. They are simple and factual, ensuring accessibility for beginners.
- **College Level Questions:** These questions assess intermediate conceptual understanding, requiring interpretation of logical relationships, main ideas, causes, effects, and motivations. They are more challenging and engage with the material at a deeper level.
- **Graduate Level Questions:** These questions evaluate in-depth analytical and critical thinking skills, focusing on symbolic or thematic interpretation, synthesis of ideas, and evaluation of broader themes or theories. They involve the highest complexity, demanding advanced problem-solving or theoretical applications.

Table 13 provides examples of the three types of questions generated for the evaluation process, as well as question generated across multiple domains.

Adversarial Quiz Question Generation. To ensure the robustness and relevance of quiz questions, an adversarial generation process was implemented. This aimed to eliminate questions that could be answered by models using pre-existing knowledge rather than learning exclusively from the provided context. Given that LLMs are trained on large-scale open web-text datasets (Roberts et al., 2020), they often possess world knowledge that can lead to misleading evaluations if the concepts are not sufficiently novel.

We observed that some generated questions were too easy for the models to answer due to several factors. Prior knowledge of the concepts being tested often allowed models to answer questions without relying on the provided context, as these concepts were part of the models' pre-training data. In some cases, the material in the context was sequential, serving as a follow-up to previously established information, making it easy for models to infer the answers. Additionally, some questions lacked sufficient cognitive challenge, failing to effectively test the model's concept acquisition abilities.

To address these issues, we implemented an iterative adversarial filtering strategy. For each context, a set of nine questions was initially generated using the `gpt-4o-2024-08-06` model. Each question was then tested with a smaller model, `gpt-4o-mini`, to identify questions that could be answered correctly without context. Those questions were filtered out and regenerated, repeating the process until the smaller model consistently missed the answer. To prevent infinite regeneration, we set a maximum of five attempts. If, after five iterations, the smaller model still answered the question correctly, we settled on the most recently generated version of the question.

This iterative adversarial approach ensured that each retained question effectively tested the model's ability to learn new concepts from the provided context, reducing reliance on prior knowledge.

Domain	Difficulty	Example Quiz Questions
Academic Papers	Middle-School	What does the acronym dMRI stand for in the context of the study?
	College	According to the document, what is the conjecture regarding every ribbon knot?
	Graduate	What overarching theme does the paper suggest through the study of DAHA and motivic superpolynomials?
Images	—	What color are the flowers in the right garden bed?
Movie Plots	Middle-School	Who is the demigoddess mentioned in "The Casagrandes Movie"?
	College	Why does Gary refuse money from Madison initially?
	Graduate	Analyze the symbolic significance of Mae's satellite decryption key within the narrative.
News Articles	Middle-School	Why was Ellie the Elephant created as the New York Liberty's mascot?
	College	How does the artwork "Comedian" draw parallels to Marcel Duchamp's urinal according to commentators?
	Graduate	Which of the following best analyzes the potential thematic motivations behind Bob Costas' decision to retire from baseball play-by-play commentary after 42 years?
Song Lyrics	Middle-School	What is happening when "the lights go off" according to the singer?
	College	What is the stylistic effect of repeating the chorus in the song lyrics?
	Graduate	From a psychological perspective, how can the metaphor "off switch" be interpreted in terms of defense mechanisms?

Table 13: Example quiz questions for each domain with different difficulty levels.

Quiz Question Validation. To ensure the quality and relevance of the generated quiz questions, a thorough manual evaluation process was conducted. This process aimed to validate the effectiveness of the questions in assessing the concept-learning abilities of student LLMs across various textual domains, including News Articles, Academic Papers, Song Lyrics, and Movie Plots.

A carefully selected sample of at least 50 questions was evaluated for each domain, ensuring both broad and balanced coverage. For domains with subdomains (e.g., News Articles and Academic Papers), this involved choosing sets of three questions—one at each difficulty level (Middle School, College, and Graduate)—from each subdomain and repeating this process until the target sample size was reached. For instance, the News Articles domain, composed of 10 categories, was sampled in two rounds of selection (3 questions × 10 categories × 2 rounds = 60 questions), and Academic Papers, with 8 subfields, was sampled in three rounds (3 questions × 8 subfields × 3 rounds = 72 questions). For domains without subdomains, like Song Lyrics and Movie Plots, 51 questions were chosen through a similar iterative approach, randomly selecting triplets that included all three difficulty levels each time. This method ensured that the evaluation set was both representative of content diversity and reflective of the full spectrum of cognitive challenges the quiz was designed to assess.

For domains without subdomains, such as Song Lyrics and Movie Plots, questions were selected randomly but still included a balanced mix of difficulty levels. This ensured diversity in the evaluation set while maintaining alignment with the domain's unique characteristics.

The evaluation focused on three key criteria to determine question quality: (1) whether the question was suitable for testing student understanding, (2) whether it was answerable using the provided

context alone, and (3) whether it avoided requiring knowledge beyond the provided context to grasp the concept. The results of this evaluation indicated that over 97% of the reviewed questions satisfied all three criteria, demonstrating their effectiveness and relevance in the quiz phase. See Table 14 for the breakdown of the quiz validation results across each domain and corresponding question criteria.

This manual verification process, combined with adversarial filtering during question generation, ensured that the quiz questions were of high quality and closely aligned with the intended learning objectives for each domain.

Category	# Evaluated	Yes (%)
Question 1: Suitable for testing student understanding?		
Academic Papers	72	70 (97.22%)
Movie Plots	51	51 (100.0%)
News Articles	60	58 (96.67%)
Song Lyrics	51	48 (94.12%)
Overall	234	227 (97.01%)
Question 2: Answerable using provided context?		
Academic Papers	72	70 (97.22%)
Movie Plots	51	51 (100.0%)
News Articles	60	59 (98.33%)
Song Lyrics	51	48 (94.12%)
Overall	234	228 (97.44%)
Question 3: Avoids requiring extra knowledge?		
Academic Papers	72	70 (97.22%)
Movie Plots	51	51 (100.0%)
News Articles	60	59 (98.33%)
Song Lyrics	51	48 (94.12%)
Overall	234	228 (97.44%)

Table 14: Breakdown of Quiz Validation Results Across Each Domain for Three Evaluation Questions.

D Prompt Templates and Generated Samples

Table 15 provides a legend for prompts used in the static and dynamic settings of our study.

Objective	Reference	Setting
Lesson Generation	Listing 1	Static
Quiz Generation	Listing 2	Static
Quiz Generation (Images)	Listing 3	Static
Student w/o Lesson	Listing 4	Static
Student w/ Lesson	Listing 5	Static
Teacher w/o Lesson	Listing 6	Static
Teacher w/ Lesson	Listing 7	Static
Student Question	Listing 8	Dynamic
Teacher Answer	Listing 9	Dynamic
Student Summarize Conversation	Listing 10	Dynamic

Table 15: Legend for prompts used in the various stages of our study, including both static and dynamic experiments.

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

```
User:
    Movie Plots: "Prepare the student comprehensively 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: "Equip the 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: "Enable the student to excel in 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 the 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: "Equip the 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."

    {concept}
```

Listing 2: Quiz generation prompt for text domains

```
You are an expert educational content creator specializing in crafting challenging multiple-choice questions
    ↪  that require specific contextual knowledge to answer correctly. Your task is to generate 3 multiple
    ↪ -choice questions each at middle-school, college, and graduate levels based on the provided context.
    ↪  Each question should have 4 options (A, B, C, D).

**Goals:**

- **Context Dependency:** Ensure that the correct answer can only be identified using specific information
    ↪ from the provided context.
- **Strong Distractors:** Create plausible distractors that are carefully crafted to appear correct to
    ↪ someone without the context, making the question unanswerable without it.
- **Difficulty Levels:** Questions should be designed to challenge the intended academic level,
    ↪ progressively increasing in complexity from middle-school to graduate level.

---

**Instructions for Each Academic Level:**

**1. Middle-School Level:**

- **Question Focus:**
  - Basic recall of key facts, definitions, or direct observations from the context.
  - Questions should be straightforward, encouraging foundational understanding.
- **Distractor Design:**
  - Distractors should reflect common misconceptions or mix-ups related to the context.
  - Use simple language appropriate for middle-school students.
  - Make distractors plausible by including slight alterations or errors in details.

**2. College Level:**

- **Question Focus:**
  - Emphasize conceptual understanding, interpretations, and logical relationships within the context.
  - Explore main ideas, causes, effects, motivations, and relationships.
- **Distractor Design:**
  - Distractors must be closely aligned with the correct answer but contain subtle inaccuracies or
      ↪ misinterpretations.
  - Use ambiguity in language or phrasing to make multiple options appear correct.
  - Include logical misinterpretations or overemphasis on secondary details.
  - Ensure the question is difficult to answer correctly without careful analysis of the context.

**3. Graduate Level:**

- **Question Focus:**
  - In-depth analysis, symbolic or thematic interpretation, and synthesis of ideas from the context.
  - Encourage analysis of underlying theories, abstract connections, or advanced problem-solving.
- **Distractor Design:**
  - Distractors must involve subtle conceptual misalignments or alternative interpretations.
  - Introduce complex, layered reasoning or speculation grounded in the context.
  - Avoid simple factual errors; frame distractors as plausible alternatives requiring expert-level
      ↪ understanding to evaluate.
  - Ensure all options are so plausible that the question is unanswerable without deep knowledge of the
      ↪ context.

---

**Formatting Guidelines:**

- **Number the questions from 1 to 9.**
- **For each question, strictly follow this format:**

Question [number]: [Question text]
A) Option A
B) Option B
C) Option C
D) Option D
Correct Answer: [A/B/C/D]
Difficulty Level: [Middle-School/College/Graduate]

- **Do not include any explanations or additional information beyond what is specified.**

---

**Here is the context:**

{plot}

Please generate the questions as per the instructions above.
```

Listing 3: Quiz generation prompt for Images domain

```
Generate 5 multiple-choice questions based on the provided image, 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 5 questions and their corresponding answers.
```

Listing 4: Prompt for student w/o lesson static evaluation

```
The following is a quiz about "{{ concept_name }}". Choose the best option that answers the question.


- Do not output anything other than the option as A/B/C/D
- Your answer has to be one of the options. If unsure, pick your best guess.

Question: {{ question }}
Options:
A. {{ option_A }}
B. {{ option_B }}
C. {{ option_C }}
D. {{ option_D }}
```

Listing 5: Prompt for student w/ lesson static evaluation

```
Given the following lesson from a teacher about "{{ concept_name }}", choose the best option that answers
    ↪ the question.

- Do not output anything other than the option as A/B/C/D
- Your answer has to be one of the options. If unsure, pick your best guess.

Lesson:
{{ lesson }}

Question: {{ question }}
Options:
A. {{ option_A }}
B. {{ option_B }}
C. {{ option_C }}
D. {{ option_D }}
```

Listing 6: Prompt for teacher w/o lesson static evaluation

```
Given the following context and lesson from a teacher about "{{ concept_name }}", choose the best option
    ↪ that answers the question.

- Do not output anything other than the option as A/B/C/D
- Your answer has to be one of the options. If unsure, pick your best guess.

Context:
{{ context }}

Question: {{ question }}
Options:
A. {{ option_A }}
B. {{ option_B }}
C. {{ option_C }}
D. {{ option_D }}
```

Listing 7: Prompt for teacher w/ lesson static evaluation

```
Given the following context about "{{ concept_name }}", choose the best option that answers the question.

- Do not output anything other than the option as A/B/C/D
- Your answer has to be one of the options. If unsure, pick your best guess.

Context:
{{ context }}

Lesson:
{{ lesson }}

Question: {{ question }}
Options:
A. {{ option_A }}
B. {{ option_B }}
C. {{ option_C }}
D. {{ option_D }}
```

Listing 8: Jinja-style prompt for student in dynamic conversation

```
To learn more about the concept "{{ concept }}" from the "{{ domain }}" domain, you can ask a teacher
    ↪ questions about its key aspects, characteristics, and underlying principles.

You are given the following information:
- **Concept Name**: "{{ concept }}"
{% if lesson -%}
- **Lesson (Current Understanding)**:
"{{ lesson }}"
{%- endif %}
{% if quiz_performance -%}
- **Quiz Performance**: {{ quiz_performance }}
{%- endif %}
{% if previous_questions -%}
- **Previous Questions Asked**: You have already asked the following questions:
{% for question in previous_questions -%}
  - {{ question }}
{%- endfor %}
{%- endif %}

Additional Instructions:
- **You have a total of {{ total_questions }} questions to ask about    {{ concept }}   ** of which you have
    ↪  asked {{ asked_questions }} questions so far.
- **Ask only one question at a time and wait for the teachers response before asking the next question.**
- Do not list all your questions at once.
- Ask diverse questions covering its origin, purpose, structure, function, context, and significance.
- Include specific questions about key details such as themes, motifs, emotions, processes, relationships,
    ↪ and examples relevant to the concept.
- Ensure questions are varied, thorough, and cover all facets of the concept.
- Ask **only** one question at a time to maintain focus and clarity.
- Think creatively if you run out of questions, aiming for in-depth, insightful inquiries!
- Your response should not contain any other information besides the **one** question.

Question:
```

Listing 9: Prompt for teacher answer in dynamic conversation

```
You are a teacher tasked with answering a student's question. You have access to a reliable source, `
    ↪ concept_knowledge`, which contains relevant information. Your goal is to provide a clear, concise
    ↪ answer to the students `question` using **only** the information from `concept_knowledge`.

### Process:

1. **Understand the Question**:
   - Identify what the student is asking.

2. **Find Relevant Information**:
   - Search `concept_knowledge` for the necessary details.
   - Address multiple parts of the question if applicable.
   - Do not copy verbatim; summarize the key points.

3. **Formulate the Response**:
   - Answer the question directly and clearly, based only on `concept_knowledge`.
   - Keep the response concise and focused on the student's needs.
   - Do not venture into information beyond the scope of `concept_knowledge` even if the student probes.

4. **Check the Response**:
   - Ensure the answer addresses the question and stays within the bounds of `concept_knowledge`.
   - If necessary, revise to make the response clearer and more concise.

### Final Response:
- Provide a direct answer to the question using information present in or closely relevant to `
    ↪ concept_knowledge` **only**.
- Do not reference `concept_knowledge` in your answer.
- Ensure the response is complete and accurate.
- Keep your responses concise and focused.

### Input:
- **concept_knowledge**: "{{ concept_knowledge }}"
- **question**: "{{ question }}"
```

Listing 10: Prompt for student conversation summarization in dynamic conversation

```
{%- if lesson is not none -%}
Lesson:
{{lesson}}
{%- endif %}

{% for content in conversation %}
{{ content['role'].title() }}: {{ content['content'] }}
{%- endfor -%}
```