# RetLLM-E: Retrieval-Prompt Strategy for Question-Answering on Student Discussion Forums

**Chancharik Mitra*, Mihran Miroyan*, Rishi Jain*, Vedant Kumud, Gireeja Ranade, Narges Norouzi**

University of California, Berkeley
Berkeley, CA, USA
cmitra@berkeley.edu, miroyan.mihran@berkeley.edu, rishiraij@berkeley.edu, vkumud@berkeley.edu, ranade@eecs.berkeley.edu, norouzi@berkeley.edu

## Abstract

This paper focuses on using Large Language Models to support teaching assistants in answering questions on large student forums such as Piazza and EdSTEM. Since student questions on these forums are often closely tied to specific aspects of the institution, instructor, and course delivery, general-purpose LLMs do not directly do well on this task.

We introduce RetLLM-E, a method that combines text-retrieval and prompting approaches to enable LLMs to provide precise and high-quality answers to student questions. When presented with a student question, our system initiates a two-step process. First, it retrieves relevant context from (i) a dataset of student questions addressed by course instructors (Q&A Retrieval) and (ii) relevant segments of course materials (Document Retrieval). RetLLM-E then prompts LLM using the retrieved text and an engineered prompt structure to yield an answer optimized for the student question.

We present a set of quantitative and human evaluation experiments, comparing our method to ground truth answers to questions in a test set of actual student questions. Our results demonstrate that our approach provides higher-quality responses to course-related questions than an LLM operating without context or relying solely on retrieval-based context. RetLLM-E can easily be adopted in different courses, providing instructors and students with context-aware automatic responses.

## Introduction

In the past several years, Large Language Models (LLMs) have progressed rapidly in their capabilities to answer questions. Two main challenges exist in developing LLM-based educational Q&A systems. The first is that the content for every class can be specific, varying between institutions, instructors, and even semesters. LLMs are trained on trillions of tokens of data from unstructured internet sources: this pre-training knowledge, while useful for general conceptual Q&A tasks, is not directly transferrable to answering specific questions about a course's content or assignments.

The second challenge comes from the fact that the relevance and factuality of responses are of utmost importance in educational settings. Since the primary objective of most LLMs is to produce "human-like" text via next-token prediction, *likely* responses will be prioritized over *correct ones*. Perhaps even more concerning is that LLMs often present answers that are syntactically cogent when the actual content may be semantically incorrect, a phenomenon called hallucination (McKenna et al. 2023).

Instruction- and prompt-tuning methods (Gupta et al. 2022; Liu et al. 2022) combined with other methods have shown promise in conditioning LLMs like ChatGPT (OpenAI 2023) and LLaMA-2-chat (Touvron et al. 2023) to reason and be more conversational. In this work, we explore **Ret**rieval-based prompting methods that yield higher quality **LLM** answers to **E**ducational questions.

In an educational context, we would like LLM responses to be thorough and explain their reasoning. Such answers have been elicited from models using in-context (Zhou et al. 2023) and Chain-of-Thought (Wei et al. 2022b) prompting methods. At the same time, in the context of answering homework questions, it is essential that a response not entirely give away the answer to the problem but instead guide a student towards finding the correct answer themselves. Our system prompts the LLM by using (1) document retrieval from course material and (2) past answers from TAs to similar questions. Importantly, our system allows us to control the sources that the LLM draws from. Thus, we can ensure the use of only high-quality text and prevent it from using sources such as homework solutions (which may lead to inadvertently disclosing more information than we intend).

When tasked with producing high-quality answers to student questions, our prompt-engineered query with the context of past Q&A information and relevant course material affords our method an advantage over the simple querying of an LLM with no context. Our key contributions include the following:

- RetLLM-E An end-to-end LLM prompting methodology that leverages **Q&A retrieval** and **document retrieval** to provide high-quality answers to student questions.
- A demonstration of the pedagogical benefit of our method by evaluating our method (following the metrics from (Jia et al. 2022a)) and comparing it to ground-truth teaching staff answers.

*These authors contributed equally.

# Related Work

## Prompting Methods

In response to the growing computational demands of fine-tuning LLMs for specialized tasks (Wei et al. 2022a), prompting has emerged as a compelling alternative. One work finds that prompting can be worth hundreds of fine-tuning data points when adapting a pre-trained LLM for a specific new task (Le Scao and Rush 2021). Consequently, we delve into various state-of-the-art prompting techniques and their benefits.

Among purely prompt-based adaptation methods, "in context learning" is a straightforward and versatile approach, using complete examples of expected behavior to prime the LLM to answer similarly (Brown et al. 2020; Dong et al. 2023). Chain-of-thought prompting is a prompting method that engages an LLM to reason and expound upon its answers, both in few-shot (Wei et al. 2022c) and zero-shot scenarios (Kojima et al. 2023). Moreover, additional complexity has been introduced through tree-like (Yao et al. 2023) and graph-like (Yao, Li, and Zhao 2023; Besta et al. 2023) prompting methods to enhance response quality and reasoning abilities.

Our work looks to improve responses to student questions by using Q&A retrieval as part of the prompt, similar to in-context learning and Chain-of-Thought prompting. With this, we hope to motivate LLMs to give answers identical in quality and reasoning to those of teaching assistants and other course staff.

## Retrieval Methods

Some of the earliest examples of Information Retrieval (IR) leveraged Boolean methods that attempted to encapsulate the "closeness" of a document to a query through matching keyword combinations and modeling word dependencies (Salton, Fox, and Wu 1983). Since then, research has shown that probabilistic vector space models outperform this traditional approach (Salton and Buckley 1988). In recent years, much progress has been made in neural methods for document retrieval (Reimers and Gurevych 2019; Su et al. 2023; Wu et al. 2023). Work in this area focuses on several fronts: (i) pre-training methods on unannotated data (Fan et al. 2022; Gao, Yao, and Chen 2021; Norouzi and Mazaheri 2023); (ii) zero-shot methods (Liang et al. 2020; Ma et al. 2021; Sachan et al. 2023); (iii) generalization through transfer learning (Mokrii, Boytsov, and Braslavski 2021) and knowledge distillation methods (Lin, Yang, and Lin 2021; Hinton, Vinyals, and Dean 2015); (iv) multi-lingual (Sheridan and Ballerini 1996; Sun and Duh 2020; Lawrie et al. 2023) and multi-modal (Srinivasan et al. 2021; Zhang 2021) retrieval methods. We leverage and build on this work to use document and Q&A retrieval to generate appropriate prompts for LLMs to generate better answers.

## Educational Support through Language Modeling

One of the main domains LLMs are used is in question-answering (Lu et al. 2022; Hudson and Manning 2019), in which LLMs have been specifically tuned for using the prompting and fine-tuning methods. The capabilities of LLMs are leveraged in the educational domain; specifically, LLMs were successfully tested on responding to student help requests (Hellas et al. 2023), and identifying problems and making suggestions for improvement related to student submissions (Jia et al. 2022b).

Furthermore, models like GPT-4 have demonstrated passing and near-perfect scores on exams such as the SAT and AP tests (OpenAI 2023) and even highly-advanced assessments such as the United States Medical Licensing Exam (Kung et al. 2023).

Another pedagogically practical application of language models is semantic text similarity identification (Chandrasekaran and Mago 2021, 2022; Gao, Yao, and Chen 2021). This can be used for specific educational tasks such as duplicate question identification (Mass et al. 2020; Rücklé, Moosavi, and Gurevych 2019), saving valuable teaching-staff time.

While there are clear opportunities for automated question-answering with LLMs, risks, and challenges such as computational constraints, privacy, and educational domain expertise are pressing concerns (et. al. 2022). We seek to build a system that addresses these concerns while improving the student forum experience for both TAs and students.

# Data

We evaluate our system in the context of the EdSTEM discussion forum for DATA 100 (Principles and Techniques of Data Science), a large upper-division data science course at the University of California, Berkeley. The enrollment in DATA 100 is around 1,200 students each semester.

Data is central to our research as we rely on using Q&A student forum *and* retrieved course document data in generating higher-quality responses. (Jain et al. 2024) demonstrates the importance of both Q&A and document data in a strict retrieval task without using an LLM. Our work also leverages task-specific educational data retrieval to engineer an LLM prompt that yields better answers to student questions. For large-scale classes such as DATA 100, the ample history of instructor answers to student questions and documents related to the course content (notes, textbook, worksheets, etc.) can be used as data sources for retrieval. We describe our data sources in more detail below.

## Historical Q&A Pair Data

DATA 100 covers the general upper-division Data Science curriculum. The students taking the class come from a variety of majors and backgrounds. The dataset for Q&A retrieval uses historical data from the Fall 2022 and Spring 2023 class offerings, which involved approximately 1,200 students each; the cleaned and processed dataset contains 6016 QA pairs[1]. The question set includes both content- and logistics-related questions. We use questions asked in the Spring 2023 offering of the same class for testing. The course's anonymized EdSTEM student forum dataset offers valuable meta-data for categorization and filtering.

[1]Data collection, storage, and processing protocols were approved under OPHS Protocol ID: 2023-07-16571.

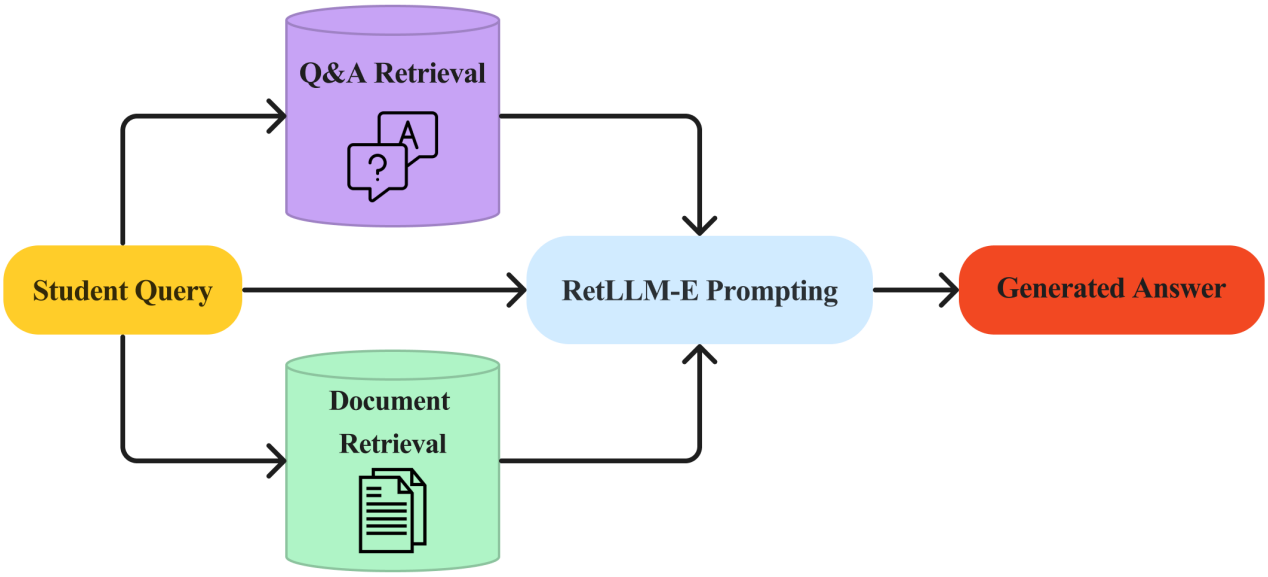


Figure 1: Full Pipeline. This figure shows the 3 main parts RetLLM-E (blue). The (i) question (yellow) (ii) the document retrieval module (green), and (iii) the Q&A retrieval module (purple) are incorporated into a prompt-engineered format for the LLM to generate a response (red).

The meta-data includes user meta-data, private tags, thread types (post, announcement, and question), and category tags (homework, lab, discussion, logistics, etc.). To maintain the context of threaded/chained conversations, we store questions and answers in a nested, dictionary-like data structure.

We formatted the raw anonymized content to optimize the data for retrieval tasks. Each structured data point, or Q&A pair, comprises the student question, the corresponding staff response, and key post details from EdSTEM (the post title and the associated context from prior conversations under the same thread). To safeguard student privacy, we excluded sensitive posts, especially those containing private questions or the following keywords: "extenuating," "DSP," "extension," "personal," and "health."

### Course Materials for Document Retrieval

Our document dataset includes instructor-curated course materials: course notes, textbook, homework questions, course syllabus, and course policy documents. We intentionally excluded homework solutions from this dataset to ensure our system does not inadvertently provide direct solutions to students.

To tailor the course materials for document retrieval, we classified documents in a manner congruent with the EdStem forum categories. The categories we adopted are:

1. General (2 documents): course notes, course textbook,
2. Logistics (3 documents): course syllabus, course policies,
3. Homeworks (8 documents): all homework assignments,
4. Discussions (13 documents): all discussion assignments,
5. Labs (14 documents): all lab assignments,
6. Projects (4 documents): all projects, and
7. Exam-preps (12 documents): all exam preparation worksheets.

### Test Set Description

To structure the conversation on the EdSTEM forum, questions in DATA 100 were organized in a specific way. In particular, specific 'mega-threads' are created for each homework problem to allow students to discuss questions about a specific homework problem. Additionally, students may ask questions about logistics or other conceptual questions as 'stand-alone' questions. We evaluate our system performance on 'stand-alone' questions.

We conducted human evaluations on a random sample of 100 questions from question threads. There are 46 logistics- and 54 content-related questions in the test set. It should be noted that the Q&A retrieval search is still performed across all the questions in the data, including mega-thread comments, to find the most relevant context.

## Methods

The high-level overview of our methodology for generating an LLM response is illustrated in Figure 1.

- Each student question is passed to both Q&A retrieval and document retrieval modules. The data sources described in Section  are the information sources in the document retrieval module.
- The Q&A retrieval module uses **multi-qa-mpnet-base-dot-v1** (Reimers and Gurevych 2019) as a backbone, as described in (Jain et al. 2024). It maps Q&A pairs to 768-dimensional embeddings and was trained on over 215M

pairs across multiple datasets. We use the nearest neighbor approach to identify the top three Q&A pairs closest to the original student question from the bank of processed Q&A pairs.

- We use a document retrieval pipeline with an **Instructor–XL** (Su et al. 2023) backbone, as described in (Jain et al. 2024). It has 335 million parameters and was trained across 15 datasets for information retrieval. We use this because it's state-of-the-art in over 70 benchmarks. It yields the top three document slices closest to the original student question from the processed course-related documents.
- We leverage prompt engineering to incorporate both the Q&A and document context into a prompt template designed to enhance the pedagogical value of the generated response. This is described in detail in Section .
- We prompt the LLaMA-2-13B-chat model with weighted sampling from the top five tokens (Touvron et al. 2023). As we are working with sensitive, real student data, we needed the LLM to be of a size feasible to be hosted locally on a server protected by institutional authentication. LLaMA-2-13B-chat is a smaller-size model that has demonstrated impressive performance on various LLM evaluations (Touvron et al. 2023).

### Prompt Engineering Details

Here, we describe our prompt engineering methodology. We also lay out the structure of our prompting method in Figure 2. Recent work has shown that LLMs can generate better responses and simulate an expert's behavior when preconditioned on a specific task or subject (Xu et al. 2023; Park et al. 2023). Thus, the first part of our prompting strategy is to have the LLM adopt the identity of an appropriate expert, which in this case is a teaching staff in DATA 100. We also precondition the LLM to give answers relevant to the general topics of the class. Finally, we precondition the LLM to respond in a "clear, helpful, and positive tone."

In addition to pre-conditing the LLM to take on the identity and style of experienced teaching staff, we tell the LLM not to respond if it is unsure of the answer. We hypothesize that this pre-conditioning will reduce any incorrect responses that may be generated.

We prompt the model with retrieved context from the document retrieval (backbone: Instructor–XL (Su et al. 2023)) and Q&A retrieval (backbone multi-qa-mpnet-base-dot-v1 (Reimers and Gurevych 2019)) pipelines. Finally, we include the student question in the prompt with a tag of "Response: ", following which the LLM will generate its response given the prompt.

## Evaluation

We perform a quantitative and qualitative evaluation of our system. We evaluate RetLLM-E and the following three baselines on a test set of 100 question-answer pairs from different question categories (logistics, assignments, etc.).

1. **No-LLM**: Our first baseline evaluates the quality of the Q&A and document retrieval. We do not evaluate no-LLM on quantitative metrics due to the large difference in linguistic structure from staff answers.
2. **Zero-shot**: Our second baseline considers the zero-shot LLM response to only the student question.
3. **No-retrieval**: Finally, to evaluate the added benefit of the *retrieval element* of RetLLM-E, we also evaluate the responses from prompting the LLM with the student question plus prompt engineering as in RetLLM-E (i.e., Expert Prompting, Task Pre-conditioning, and Tone Preconditioning, i.e., the gray box in Figure 2), but *without any of the retrieval context that is unique to our pipeline* (i.e., the purple and green box in Figure 2).

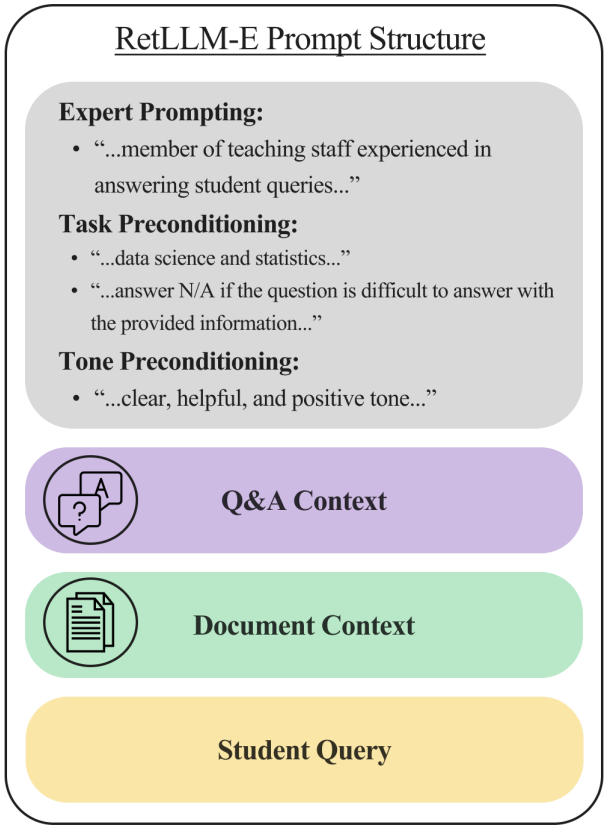


Figure 2: RetLLM-E Prompt. The prompt structure of RetLLM-E is outlined. In addition to the key elements of Q&A context, document context, and the student question, we include prompt-engineering methods to provide tailored responses to students' questions and not reinforce misconceptions if the LLM is unsure of the answer.

Thus, through this careful ablation, we will be able to delineate the relative value of (i) the LLM, (ii) RetLLM-E as a whole, and (iii) just the novel context of QA and document retrieval by itself.

### ROGUE & BERTScore

We compare RetLLM-E, zero-shot, and no-retrieval using the ROUGE metric (Lin 2004) and BERTScore (Zhang et al.

| Score | Relevance | Readability | Positive Tone | Factuality |
|---|---|---|---|---|
| 1 | The response is not relevant to and does not address the student's question. | The response is incomprehensible. | The response has a negative tone. | None of the statements in the response are correct. |
| 2 | The response is somewhat relevant to and partially answers the student's question. | The response is somewhat coherent and fluent. | The response has a neutral tone. | Some of the statements in the response are correct. |
| 3 | The response is relevant to and completely answers the student's question. | The response is coherent and fluent. | The response has a positive tone. | All the statements in the response are correct. |

Table 1: Criteria for conducting human evaluations on generated responses and retrieved documents. All the metrics - relevance, readability, tone, and factuality - are graded on a scale of 1 to 3.

2020), where staff responses serve as the ground truth. The ROGUE-N is a metric for comparing N-gram matches between generated text and a ground truth text. We evaluate uni-gram and bi-gram ROGUE F1-scores for RetLLM-E and all baselines. BERTScore is an embedding similarity metric. We compare the cosine similarity between BERT embeddings of our baselines and RetLLM-E to the ground truth staff responses.

### Human Evaluation

Our quantitative rubric for human evaluation is adapted from (Jia et al. 2022a), and is described in detail in Table 1. Our quantitative human evaluation examines *Relevance*, *Readability*, *Tone*, and *Factuality* of the responses. Since (Jia et al. 2022a) focuses on providing feedback to student submissions instead of answering student questions, we remove the *Suggestions* and *Problems* criteria from their evaluation criteria and replace them with *Relevance*.

To limit subjectivity in human evaluation, we consider all factors on a three-point scale: a score of one indicates a low evaluation of the metric, and a score of three indicates a high evaluation. The evaluation was performed by four co-authors of this paper. Two of these have experience serving on the teaching staff of a large-scale computer science or data science course. Each evaluator was simultaneously presented with responses from RetLLM-E and the three baselines (no-LLM, zero-shot, no-retrieval) in a blinded, unlabeled manner. We note that we do not evaluate the no-LLM case for *Readability*, *Tone*, and *Facutuality* since it is just returning a set of documents and Q&A pairs and is not crafting a response to the question. A summary of results is in Table 2.

Finally, in the cases where the baseline or RetLLM-E responses scored high on *Factuality*, i.e., scored a 2 or a 3, we compared their responses with the staff answers from our test dataset to identify which was preferred. We say the generated response is *better* than the staff answer if it is either (1) more comprehensive, or (2) presents more/better examples from course notes and previous assignments to solidify student understanding. Likewise, the generated answer is *worse* than the staff answer if the staff answer is (a) more factual, (b) more concise, or (c) the generated response gave the solution away. If we cannot decide, we declare a *tie*. We treat all the responses with a score of 1 in the *Factuality* metric to be worse than the staff response.

## Results

### RetLLM-E Outperforms Baselines on ROUGE and BERTScore

We use ROUGE and BERTScore metrics to compare the LLM-generated responses of RetLLM-E and baselines to staff answers; the corresponding results are presented in the first three columns of Table 2. Although modest, the results demonstrate that RetLLM-E can use retrieved context from course documents and student-forum Q&A data to generate answers more linguistically similar to staff answers than all other baselines.

### Human Evaluation: RetLLM-E Responses Are More Relevant and Factual

The results of human evaluations based on the metrics from Table 1 are presented in Table 2. From our assessment of no-LLM, we find that the average *Relevance* score of retrieved context (both Q&A and a document) is 2.02. This suggests that, in general, the retrieved context is fairly relevant to the question, with only some questions or documents being irrelevant. Both no-retrieval and RetLLM-E score fairly well on *Relevance* compared to the other two cases.

We find that RetLLM-E outperforms the zero-shot case on every aggregated metric. Prior work (Le Scao and Rush 2021; Xu et al. 2023; Park et al. 2023) has demonstrated that expert prompting, as well as cues like details about the topic and tone of questioning, can improve the quality of LLM responses. Our results show that this also holds for the task of educational Q&A. Additionally, significant gains in all four evaluation metrics demonstrate that providing context to LLM is pedagogically beneficial.

Finally, we come to the no-retrieval baseline. We see that RetLLM-E outperforms no-retrieval modestly in *Relevance* while significantly improving in answer *Factuality*. We note that we see minor decreases in *Readability* and *Tone*. Our results specifically indicate that the novel combination of retrieved course documents and student Q&A is responsible for increasing the *Factuality* and *Relevance* of LLM answers for course questions with little to no degradation in

| | AUTOMATED | | | HUMAN EVALUATION | | | |
|---|---|---|---|---|---|---|---|
| **Model** | **ROUGE_1** | **ROUGE_2** | **BERTScore** | **Relevance** | **Readability** | **Tone** | **Factuality** |
| no-LLM | - | - | - | 2.020±0.69 | - | - | - |
| zero-shot | 0.136 | 0.030 | 0.806 | 2.330±0.78 | 2.500±0.66 | 2.390±0.62 | 1.920±0.84 |
| no-retrieval | 0.165 | 0.038 | 0.830 | 2.680±0.60 | **2.800**±0.43 | **2.820**±0.41 | 1.860±0.80 |
| RetLLM-E | **0.193** | **0.045** | **0.838** | **2.690**±0.66 | 2.770±0.49 | 2.780±0.48 | **2.040**±0.80 |

Table 2: With RetLLM-E, we observe a significant improvement in average *Factuality* while having little to no degradation in *Relevance*, *Readability*, and *Tone* when compared to the baseline responses. We also present the standard deviation for each measurement over the 100 question-answer pairs.

*Readability* and *Tone*.

It is worth noting that the significant variance in the results from Table 2 comes more from how response quality differs on the wide variety of question types (e.g., content, logistical, and assignment-specific). Thus, the average human evaluation metric is the most meaningful. We qualitatively disentangle some of the response variability to different types of questions in Section , and we leave it to future work to find methods for limiting this type of variation when answering different question types.

## Human Evaluation: Qualitative Observations

To fully understand the types of responses RetLLM-E and the other baselines generated, we first aggregate some rough qualitative observations.

**Cases Where RetLLM-E Response Is Equal or Better Than staff Answers:** As shown in Figure 3, there were several questions where the RetLLM-E response to a question was deemed of overall higher quality than the staff response. Qualitatively, we found that RetLLM-E outperforms staff answers on clearly written conceptual and logistical questions. For straightforward conceptual questions, this is understandable, as these questions are the least specific to the course (i.e., data science and statistics concepts not particular to the class). For logistical questions, the clarity in handling the retrieved documents from the syllabus was well demonstrated when the response seamlessly incorporated information like lecture times or attendance policies.

**Traits of High-Quality RetLLM-E Responses:** Overall, the equal or higher quality answers that were generated by RetLLM-E had several traits. Their *Tone* was positive, and their writing was readable, always encouraging students to ask any further questions they had at the end of the response, for example. This was not common in the zero-shot responses. The content of RetLLM-E responses also went beyond simply answering the question and would give *examples relevant to specific course problems*. For example, when asked about bar plots and histograms, RetLLM-E's answer would not only clearly explain the use cases of the two categories of plots but would also use the question setup from a past assignment to provide a supporting example. Of course, while the no-retrieval and (to a lesser degree) the zero-shot responses occasionally included examples, the specificity to the course was not present, except in rare cases when the question contained the necessary context. Qualitatively, these results help explain the importance of the retrieved context for improving *Relevance* and *Factuality* as well as of the prompt engineering methods for improving response *Readability* and *Tone*.

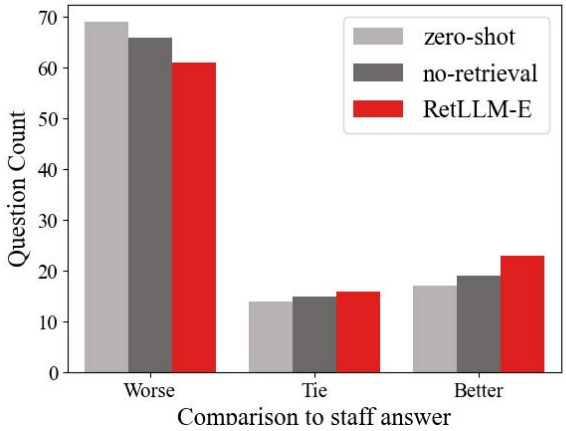


Figure 3: We compare the LLM responses to the staff answers. We say the generated response is *better* than the staff answer if it is either (1) more comprehensive or (2) presents more/better examples from course notes and previous assignments. Likewise, the generated answer is *worse* than the staff answer if the staff answer is (a) more factual, (b) more concise, or (c) the generated response gave the solution away. If we cannot decide, we declare a *tie*.

**Content Hallucination:** We noticed some responses demonstrating high *Relevance* but lacking *Factuality* and vice versa, which indicates that the LLM is hallucinating an answer. However, both *Relevance* and *Factuality* are required for an answer to be *correct*. Using this paradigm, we can see from our results in Table 2 that zero-shot had significantly lower *Relevance* scores but higher *Factuality* scores than no-retrieval. During evaluation, we noticed that in these cases the zero-shot baseline was often hallucinating. Without any prompt guidance, zero-shot responses were more free to answer questions without being relevant, however,

they were correct information. These hallucinated examples include ones where zero-shot (i) fabricated questions as part of the response, (ii) responded to questions as if they were emails or messages, or (iii) appeared to respond to questions with answers on topics completely unrelated to the task. We note that when evaluating, we took *Factuality* as just evaluating the contents of the answer to reduce the correlation between *Factuality* and *Relevance*.

The prompt engineering used in both no-retrieval and RetLLM-E gave the LLM valuable context about its role and the subject matter of questions it would encounter. This ensured that all responses from these two models were related to the student question. Compared to no-retrieval, RetLLM-E's responses contained did better on both *Relevance* and *Factuality*. This intuitively is the benefit of including the retrieved context as part of prompting. In fact, no-retrieval would at times hallucinate course policies (e.g., saying that specific submissions would earn credit when they would not) and course resources (e.g., referring to a learning management system when one was not utilized for the course). In cases where the retrieval failed to identify relevant documents and Q&A pairs, RetLLM-E responded using the incorrect documents or Q&A pairs. This is a critical issue to address in future work.

## Limitations

A limitation of our current research is the evaluation of RetLLM-E exclusively within the confines of a single class, and thus more work is needed to ensure that our results generalize. Furthermore, we evaluate the performance of RetLLM-E on stand-alone questions. We do not evaluate performance on mega-threads, which are specialized threads dedicated to answering questions about a specific homework problem. These mega-threads often have nuanced conversations happening in them that are difficult for our model to pick up on, and further work is required to achieve high-quality responses in the case of mega-threads.

Having automated responses to student questions in discussion forums can certainly make the job of a teacher easier. However, this needs to be done with care. In particular, one needs to be thoughtful about the privacy of student data. In our case, this means we hosted the LLM on a local server, which is certainly expensive and not necessarily scalable.

## Future Work

In the future we aim to explore the application of RetLLM-E across various academic disciplines and courses. In addition, future research could focus on enhancing the interaction between the LLM and students. Currently, RetLLM-E generates a single response to a student's question, which may inadvertently provide direct answers rather than guide the student toward the solution. Future iterations could implement more nuanced interactions, such as a back-and-forth discussion style, offering hints and leading the student toward the correct answer. This could be achieved through advanced fine-tuning and prompt engineering techniques.

Finally, we must ensure that human biases that might be present in historical data are not replicated by our system. Doing this is necessarily tricky and something we are actively considering.

## Conclusion

In AI-enabled educational applications, efficient question-answering mechanisms are pivotal. Our research demonstrates that RetLLM-E, a novel LLM prompting method, exhibits enhanced performance in educational question-answering compared to existing methodologies. A unique feature of RetLLM-E is the integration of document and Q&A retrieval modules within the prompting pipeline. The combined retrieval approach not only harnesses the content-rich nature of educational materials but also leverages insights from prior student Q&A. Subsequent evaluations, based on automated metrics such as ROUGE and BERTScore, coupled with human assessments, affirmed the efficacy of RetLLM-E in generating more relevant and factual answers than zero-shot querying and prompt engineering.

From a pedagogical standpoint, RetLLM-E's advancements have significant implications. An optimized question-answering tool can provide valuable real-time feedback to students, facilitating their learning process. Furthermore, it can serve as a supplementary resource for educators.

## References

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Podstawski, M.; Niewiadomski, H.; Nyczyk, P.; and Hoefler, T. 2023. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv:2308.09687.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chandrasekaran, D.; and Mago, V. 2021. Evolution of Semantic Similarity—A Survey. *ACM Computing Surveys*, 54(2): 1–37.

Chandrasekaran, D.; and Mago, V. 2022. Evolution of Semantic Similarity – A Survey. *ACM Computing Surveys*, 54(2): 1–37.

Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Wu, Z.; Chang, B.; Sun, X.; Xu, J.; Li, L.; and Sui, Z. 2023. A Survey on In-context Learning. arXiv:2301.00234.

et. al., B. 2022. On the Opportunities and Risks of Foundation Models. *arXiv*.

Fan, Y.; Xie, X.; Cai, Y.; Chen, J.; Ma, X.; Li, X.; Zhang, R.; and Guo, J. 2022. Pre-training Methods in Information Retrieval. arXiv:2111.13853.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Gupta, P.; Jiao, C.; Yeh, Y.-T.; Mehri, S.; Eskenazi, M.; and Bigham, J. P. 2022. InstructDial: Improving Zero and Few-shot Generalization in Dialogue through Instruction Tuning. arXiv:2205.12673.

Hellas, A.; Leinonen, J.; Sarsa, S.; Koutcheme, C.; Kujanpää, L.; and Sorva, J. 2023. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. *ArXiv*, abs/2306.05715.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. arXiv:1902.09506.

Jain, R.; Miroyan, M.; Mitra, C.; Kumud, V.; Ranade, G.; and Norouzi, N. 2024. Elevating Learning Experiences: Leveraging Large Language Models as Student-Facing Assistants in Discussion Forums. Poster presented at ACM Technical Symposium on Computer Science Education (SIGCSE) V. 2.

Jia, Q.; Young, M.; Xiao, Y.; Cui, J.; Liu, C.; Rashid, P.; and Gehringer, E. 2022a. Insta-Reviewer: A Data-Driven Approach for Generating Instant Feedback on Students' Project Reports. In Mitrovic, A.; and Bosch, N., eds., *Proceedings of the 15th International Conference on Educational Data Mining*, 5–16. Durham, United Kingdom: International Educational Data Mining Society. ISBN 978-1-7336736-3-1.

Jia, Q.; Young, M.; Xiao, Y.; Cui, J.; Liu, C.; Rashid, P.; and Gehringer, E. 2022b. Insta-Reviewer: A Data-Driven Approach for Generating Instant Feedback on Students' Project Reports. In Mitrovic, A.; and Bosch, N., eds., *Proceedings of the 15th International Conference on Educational Data Mining*, 5–16. Durham, United Kingdom: International Educational Data Mining Society. ISBN 978-1-7336736-3-1.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2023. Large Language Models are Zero-Shot Reasoners. arXiv:2205.11916.

Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; and Tseng, V. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2): 1–12.

Lawrie, D.; Yang, E.; Oard, D. W.; and Mayfield, J. 2023. Neural Approaches to Multilingual Information Retrieval. In *European Conference on Information Retrieval*, 521–536. Springer.

Le Scao, T.; and Rush, A. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2627–2636. Association for Computational Linguistics.

Liang, D.; Xu, P.; Shakeri, S.; dos Santos, C. N.; Nallapati, R.; Huang, Z.; and Xiang, B. 2020. Embedding-based Zero-shot Retrieval through Query Generation. arXiv:2009.10270.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, S.-C.; Yang, J.-H.; and Lin, J. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 163–173. Online: Association for Computational Linguistics.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68. Dublin, Ireland: Association for Computational Linguistics.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Ma, J.; Korotkov, I.; Yang, Y.; Hall, K.; and McDonald, R. 2021. Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation. arXiv:2004.14503.

Mass, Y.; Carmeli, B.; Roitman, H.; and Konopnicki, D. 2020. Unsupervised FAQ Retrieval with Question Generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 807–812. Online: Association for Computational Linguistics.

McKenna, N.; Li, T.; Cheng, L.; Hosseini, M. J.; Johnson, M.; and Steedman, M. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. arXiv:2305.14552.

Mokrii, I.; Boytsov, L.; and Braslavski, P. 2021. A Systematic Evaluation of Transfer Learning and Pseudo-labeling with BERT-based Ranking Models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Norouzi, N.; and Mazaheri, A. 2023. Context-aware analysis of group submissions for group anomaly detection and performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15938–15946.

OpenAI. 2023. GPT-4 Technical Report. *arXiv*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. arXiv:2304.03442.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rücklé, A.; Moosavi, N. S.; and Gurevych, I. 2019. Neural Duplicate Question Detection without Labeled Training Data. 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)*, 1607–1617. Association for Computational Linguistics.

Sachan, D. S.; Lewis, M.; Joshi, M.; Aghajanyan, A.; tau Yih, W.; Pineau, J.; and Zettlemoyer, L. 2023. Improving Passage Retrieval with Zero-Shot Question Generation. arXiv:2204.07496.

Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5): 513–523.

Salton, G.; Fox, E. A.; and Wu, H. 1983. Extended Boolean Information Retrieval. *Commun. ACM*, 26(11): 1022–1036.

Sheridan, P.; and Ballerini, J. P. 1996. Experiments in multilingual information retrieval using the SPIDER system. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 58–65.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2443–2449.

Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.-t.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1102–1121. Toronto, Canada: Association for Computational Linguistics.

Sun, S.; and Duh, K. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4160–4170.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022c. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Wu, X.; Ma, G.; Lin, M.; Lin, Z.; Wang, Z.; and Hu, S. 2023. ConTextual Masked Auto-Encoder for Dense Passage Retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4): 4738–4746.

Xu, B.; Yang, A.; Lin, J.; Wang, Q.; Zhou, C.; Zhang, Y.; and Mao, Z. 2023. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. arXiv:2305.14688.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv:2305.10601.

Yao, Y.; Li, Z.; and Zhao, H. 2023. Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Large Language Models. arXiv:2305.16582.

Zhang, H. 2021. Voice keyword retrieval method using attention mechanism and multimodal information fusion. *Scientific Programming*, 2021: 1–11.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zhou, W.; Jiang, Y. E.; Cotterell, R.; and Sachan, M. 2023. Efficient Prompting via Dynamic In-Context Learning. arXiv:2305.11170.