Benchmarking LLMs for Socratic Student-Teacher Interactions

Semere Wubshet Berhanu
EURECOM
berhanu@eurecom.fr

February 4, 2025

Abstract

AI chatbots are increasingly being used in education due to their accessibility and seemingly vast knowledge about the world. However, they often provide direct answers to students' queries which hinders the development of analytical and reasoning skills. The Socratic method, with its emphasis on guided discovery through questioning, offers a powerful pedagogical alternative, but effectively replicating this method in AI requires substantial high-quality training data, which is currently scarce. This paper introduces a pipeline for generating Socratic Student-Teacher dialogues, addressing this critical data bottleneck. The pipeline produces diverse, persona-driven interactions, creating a resource for training educational LLMs. Furthermore, a Judge model capable of evaluating the Socraticness of dialogues is selected. It can be utilized as an automated evaluation tool for benchmarking and optimizing LLMs for this specialized task. It contributes to the ongoing effort to develop effective AI tutors by helping to generate quality datasets for fine-tuning.

1 Introduction

Students increasingly rely on AI chatbots like ChatGPT, Claude, and Google Gemini to support their education [9]. These tools hold significant promise for enhancing learning by clarifying concepts and answering queries. While they have undeniable benefits, such as accessibility, availability, and ease of use, they also exhibit significant limitations when applied in educational contexts. First, they can present incorrect or misleading information in a confident and polished manner, making it difficult for users to discern inaccuracies [12]. Second, these tools typically provide direct answers to students' questions, bypassing opportunities for students to engage in deeper reflection, develop critical thinking skills, and gain a comprehensive understanding of the subject matter.

The Socratic Teaching Method involves an open dialogue between a teacher and a student, where the teacher avoids directly lecturing or providing answers. Instead, the teacher guides the student's exploration through carefully posed questions designed to challenge assumptions and nurture critical thinking [7]. A successful Socratic lesson engages all participants in a respectful and collaborative environment, encouraging diverse perspectives while fostering a thorough understanding of the topic. Teachers employing this method must possess patience, active listening skills, and the ability to create a safe and intellectually stimulating atmosphere.

Example of Socratic Teaching

Student: Why do crabs walk sideways? Is it because they can only look from side to side?
Teacher: Interesting thought! However, crabs actually have good vision in multiple directions. What do you observe about the way a crab's legs are positioned?

Replicating the Socratic method in AI-powered tutors holds potential for improving learning outcomes. By adopting a Socratic teaching style, LLMs can better mimic the role of a real classroom teacher, actively guiding students and promoting their participation in the reasoning process. This approach aligns with established pedagogical principles that emphasize active problem-solving for mastery of concepts. However, a major obstacle is the scarcity of high-quality training data for developing such models [5][8]. Generating effective Socratic AI requires datasets that capture the dynamics of these complex, multi-turn interactions.

To address this challenge, this paper introduces a data generation pipeline to create Student-Teacher interactions. This can be used for training educational LLMs which follow the Socratic teaching method. After outlining the method for generation, it identifies the LLM that best behaves in a Socratic manner from a selection of candidates. Through these, it intends to contribute to future development of educational Socratic AI tools.

Code available at `https://github.com/SemereWubshet/socratic_mcts`.

2 Related Work

In recent years, researchers have explored the creation and utilization of datasets made up of educational dialogues. These datasets, comprising predefined interactions between teachers/tutors and students, serve multiple purposes. They provide valuable training data for developing and fine-tuning educational AI models and offer benchmark datasets for evaluating the performance of these systems. TutorChat[1] is one of such datasets published in 2024. It includes 78,000 student-teacher dialogues on scientific topics, where the average conversation contains 1,800 words over 15 turns of exchange. It was used to fine-tune Llemma models for improved performance [3].

MathDial[2] is a dataset of nearly 3,000 teacher-student dialogues grounded in multi-step math reasoning problems, where conversations average 17.3 words per utterance. It was published in 2023 and has been used to fine-tune LLMs to be more effective tutors [6]. SocraticMATH is a dataset published in 2024 which contains nearly 7,000 Socratic-style conversations covering math knowledge at the primary school level [4]. Conversations within the dataset average 4.95 turns and 86.49 words per utterance.

Others have also explored how the Socratic teaching method could be incorporated into LLMs to enhance student learning and critical thinking. Recent work has shown that fine-tuning LLMs using Direct Preference Optimization can effectively align these models with the principles of the Socratic teaching method. These fine-tuned models show improvement in adhering to Socratic guidelines. These efforts align with the broader aim of producing better AI tools for education [1].

To develop LearnLM, a model fine-tuned for educational tasks, Google addressed the scarcity of high-quality educational data by creating a combined human-synthetic dataset. Recognizing the limitations of existing human tutoring data (small size, uneven quality, language-learning focus), they employed several generation techniques. These included adapting the GSM8k math dataset into learner-tutor dialogues. They also used Generative AI role-play and created "golden conversations" authored by educators. This diverse approach enabled the creation of a robust training dataset for LearnLM [5][8].

[1]Available at `https://huggingface.co/datasets/princeton-nlp/TutorChat`

[2]Available at `https://huggingface.co/datasets/eth-nlped/mathdial`

3 Generating Student-Teacher Interactions

This section details the pipeline for synthesizing Socratic dialogues between a teacher and a student. The conversations are started by a student who asks an initial question to which the teacher gives a Socratic reply. The conversation carries on until the student's question has been adequately answered or until a predefined maximum number of exchanges is reached. By structuring interactions around real-world student behaviors, such as misconceptions, tangential inquiries, and varied learning styles, we produce nuanced dialogues that reflect authentic pedagogical dynamics. Each conversation in this dataset contains:

- **Source Content**: A passage of text from which the student asks a question.
- **Initial Question**: The initial question the student poses the teacher.
- **Main Topics**: A list of relevant topics which a teacher must cover for a successful Socratic conversation.
- **Interaction Type**: The type of question the student asks.
- **Student Type**: The type of student engaged in conversation.
- **Chat History**: The exchanges between the student and the teacher.

While the dataset was constructed with the intention of representing Socratic conversations, it is important to note that not all generated dialogues fully adhere to the principles of the Socratic teaching method. Hence the need to evaluate the resulting dialogues for adherence to Socratic principles[3].

Below, we describe each component of the pipeline, beginning with its overarching architecture in Figure 1 and proceeding to implementation specifics.

[3]Dataset available at `https://github.com/SemereWubshet/socratic_mcts/blob/main/datasets/student-teacher-interaction.json`.

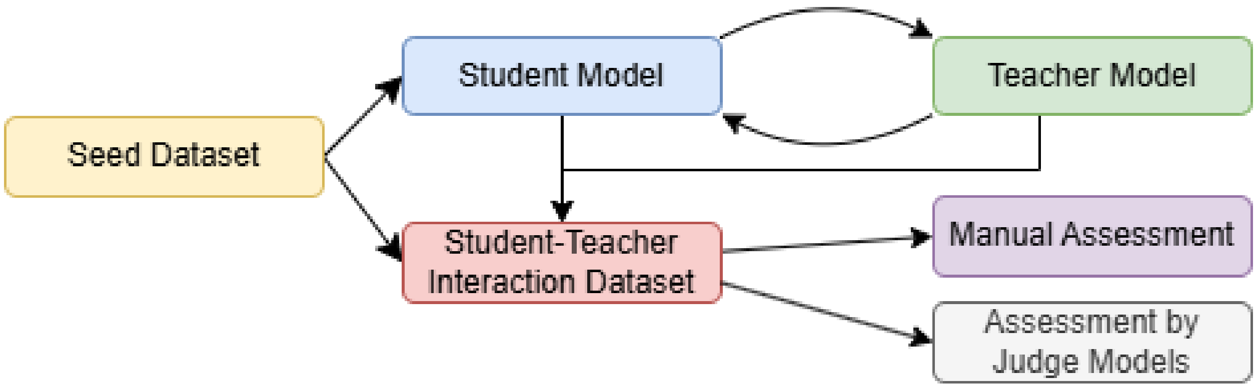

Figure 1: Pipeline for Generating and Assessing Educational Socratic Dialogue

3.1 Seed Dataset

The Simple Wikipedia dataset is a clean and small subset of the larger Wikipedia dataset and contains articles written in simplified English [11]. A random passage from a random page is utilized a source for basing the conversation. The source content is processed through mistral-nemo:12b-instruct-2407-fp16 [4] model (hereafter referred to as mistral-nemo-12b) to generate an initial student question which will be the start of the Student-Teacher interaction. This LLM is instructed to act as a student who reads the source content asks a concise question pertaining it. This initial student question can be of one of two interaction types. The first type is a general question about the main topic that aims to emulate the kinds of inquiries real students pose when they are unfamiliar with a subject and it is outlined in Appendix 8.4. For example, if the source content describes Rayleigh scattering, a general question can be "Why is the sky blue?"

The second type of question is a misleading question that contains an incorrect or false claim. This question type is intentionally crafted to test the Teacher model's ability to identify misconceptions and guide the student toward the correct understanding without outright dismissing the question. It is also outlined in Appendix 8.4. Misleading questions often arise in real-world learning scenarios when students have partially understood a topic or have been influenced by incorrect information. This mirrors a common challenge in education, where teachers must strike a balance between correcting misconceptions and promoting a student's curiosity and confidence.

Accompanying each initial question, the mistral-nemo-12b model also generates a corresponding list of key topics relevant to the source content. The four elements, namely the list of relevant topics, source content, initial question, and interaction type, are packaged together under a seed object. This process is repeated to create a collection of these seed objects, which are stored and prepared for use in subsequent stages of the methodology.

[4]Model available at `https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407`

3.2 Student Model

The Socratic conversations dataset aims to provide a well-made and structured collection of Student-Teacher interactions. It is impractical to create one with real students, therefore, a LLM is used in their place. The mistral-nemo-12b model was selected to power the Student model. The Student model receives a seed object from the seed collection and processes its system prompt along with the list of main topics and the chat history. The initial question from the seed is passed untouched to the teacher at the beginning of the conversation. However, in subsequent turns, the chat history is taken into account by the Student model as it generates responses.

> **Perfect Student**
>
> **Student**: "Why do some animals hibernate?"
> **Teacher**: "What challenges do these animals face during certain times of the year?"
> **Student**: "Right, so food and warmth become scarce in winter. Hibernation is a way to conserve energy when those resources aren't readily available. It's like a deep sleep where their metabolism slows way down, so they don't need as much food or warmth to survive until things get better."

The first naive implementation involved instructing the LLM to behave as a student. However, this produces a perfect student with vast and unrealistic knowledge who does not need a teacher. The Student-Teacher interactions did not reflect real-life conversations at all because the student had perfect answers to all the teacher's questions as shown in the example above.

Real students possess varying educational and mental capacities which necessitates building personas to capture their differences. This was replicated in the Student model through the inclusion of seven distinct student types. Each Student model has a random student type which represents a complex persona with a particular mix of strengths, weaknesses, and characteristics. While not derived from a formal methodology, these personas were developed to represent a broad spectrum of student behaviors. By embedding the student type into the system prompt, the model adjusts its responses to align with the characteristics of the assigned student type.

Types of Student Persona

- You are a student who grasps and applies concepts effortlessly across domains. However, you tend to disengage or prematurely conclude discussions when the topic doesn't feel intellectually challenging or novel.
- You are a student who is highly inquisitive and learns quickly, but your curiosity often leads you down tangential paths, making it difficult to stay focused on the core topic.
- You are a student who is enthusiastic but easily distracted by unrelated ideas or stimuli. You need reminders to focus on the main learning objective.
- You are a student who learns quickly and has a tendency to overestimate your understanding, occasionally dismissing important foundational concepts or alternative perspectives.
- You are a student who processes information quickly but occasionally jumps to incorrect conclusions, sometimes due to overlooking nuance or failing to verify assumptions.
- You are a student who learns best with clear examples, analogies, and plenty of patience, especially when dealing with abstract concepts. Once you understand, you retain knowledge deeply.
- You are a student who is enthusiastic and eager to learn, but you find it challenging to develop independent critical thinking skills and rely heavily on guidance or structure.

Additionally, the Student model had to determine on each of its turns whether the initial question had been adequately answered. If the exchanged satisfied the Student model, the conversation would end after it thanked its teacher. If not, the Student would generate a follow-up response and continue the dialogue. This simulates real-world conversations between a student and a teacher as a student typically disengages once the initial questions has been adequately addressed.

3.3 Teacher Model

As the other half of Student-Teacher interactions, the Teacher model was made to guide students towards deeper understanding of a topic through engaging conversations without providing direct answers. It aimed to stimulate the student's curiosity by breaking down complex ideas into sub-problems and offering incremental hints. It was powered by mistral-nemo-12b which only received its system prompt and the chat history to generate the teacher's reply.

The first implementations of the Teacher model, in which it was simply told to behave like a Socratic teacher, produced long and convoluted replies with multiple questions for the student. It also had a tendency to ask irrelevant questions well outside the scope of the student's initial question. They went against core Socratic teaching principles and had to be corrected. Through multiple iterations of its system prompt, the teacher's performance increased. A clear example was included within the system prompt to show a concrete realization of the desired Socratic teaching principles. The final teacher system prompt is given in Appendix 8.2.

Example of Socratic Exchange

Student: 'Hi there, I've been thinking about the text we read on mathematical creativity and aesthetics. It mentions that some mathematicians view their work as solving puzzles or even finding art in proofs, like Hardy's 'A Mathematician's Apology'. But what if we consider mathematics more like storytelling? Couldn't the search for patterns, the building of arguments, and the resolution of problems be seen as crafting a narrative? I'd love to hear your thoughts on this perspective.
Teacher: How would considering mathematical reasoning as a form of creative writing change our understanding of the role of imagination in mathematics?

The example above demonstrates the Teacher model's use of an open-ended question that directly engages with the student's input and prompts further reflection. An iterative refinement process, involving both prompt engineering and manual evaluation, was crucial in developing a Teacher model that effectively embodies the principles of Socratic pedagogy.

4 Selecting a Judge Model

To accurately determine which model is best suited as a judge for evaluating Socratic interactions, it is necessary to validate candidate Judge models against a reliable ground truth. This process begins by generating a dataset of student-teacher interactions using the previously described pipeline and manually labeling these interactions according to established Socratic criteria. The manually labeled dataset then serves as the baseline for evaluating and comparing Judge model performance.

4.1 Manual Assessment

A total of 50 conversations were generated and subsequently evaluated. The human and AI assessments were based on the same criteria for Socraticness. These criteria can be interpreted differently and indeed they are. Assessments differ between two LLMs, between two humans, and between a human and an LLM. Nevertheless, these criteria were defined to confine subjectivity. These criteria are:

- Topic Coverage: Were all main topics thoroughly addressed?
- Socratic Adherence: Did the teacher follow Socratic teaching principles?
- Student Understanding: Did the student demonstrate a deep understanding of the subject?
- Main Question: Was the student able to answer the original question?

The ground truth assessment is performed by the author according to the established criteria [5]. This was a labor-intensive task that is subjective and open to the assessor's biases. The main challenge during this section was in categorizing conversations that exhibited some characteristics of Socratic dialogue but did not fully adhere to all the established principles.

[5] Manually labeled dataset available at `https://github.com/SemereWubshet/socratic_mcts/blob/main/datasets/human-rated-socratic.json`

Of the 50 conversations assessed, only 17 (34%) met the criteria as fully Socratic. The most frequently occurring reason for these failures was insufficient coverage of the main topics provided. Many conversations, while demonstrating some elements of Socratic questioning, tended to focus disproportionately one aspect of the topic, neglecting other key areas that should have been explored. This often resulted in a shallow or incomplete understanding of the topic as a whole, failing to meet the requirement for comprehensive coverage.

The other significant reason for conversations not meeting the Socratic criteria was that the teacher or Student model sometimes deviated from the intended scope of the conversation, venturing into tangential or unrelated topics. This issue, while less frequent, also detracted from the quality of the Socratic dialogue by disrupting the focused exploration of the main themes.

4.2 Judge Model

The Judge model is tasked with evaluating the Socraticness of Student-Teacher dialogues. Its core function is to make an accept/reject decision, providing an automated assessment that can be compared against the human ground truth. It can be powered by any model to perform its evaluations. This allows us to evaluate the performance of different LLMs as Socratic judges. This automated evaluation is essential for scaling up the analysis and for future development of improved educational Socratic AI.

The challenge during this phase was to produce a system prompt for adequate performance of the Judge. The llama3.3:70b model was selected to power the Judge as the performance of the system prompt was evaluated. Early attempts produced overly lenient assessments which demonstrated the need for more structured guidance in the system prompt. The final prompt was restrictive enough to repress passing semi-Socratic interactions. The process involved clearly separating the model's role, Socratic teaching principles, and evaluation criteria in the system prompts while incorporating examples of success and failure [2].

Cohen's Kappa, κ, was used as the primary metric to compare the LLM's assessments against the human-generated ground truth. It is a statistical measure of inter-rater reliability that assesses the agreement between two assessors who make categorical judgments. It is particularly suitable for our task as it accounts for the possibility of agreement occurring by chance. It corrects against random agreements which is important for the imbalanced binary classification dataset.

Cohen's Kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where p_o represents the observed proportion of agreement between the two raters, and p_e represents the expected proportion of agreement due to chance.

Four candidates were considered to power the Judge model. These were gpt-4o, llama3.3:70b, mistral-nemo-12b , and phi-4. Each LLM was tasked with evaluating the full dataset of 50 generated conversations as either "Socratic" or "not Socratic." All LLMs were provided with the same prompts, based on the criteria outlined in the previous section. These LLMs inherently have different capabilities which give them their own strengths and weaknesses. However, when tasked with assessing the Socraticness of a conversation, their performance revealed significant limitations. The results of this evaluation are presented in Table 1.

Judge Model	Cohen's Kappa (κ)
llama3.3:70b	0.46
gpt-4o	0.23
phi-4	0.20
mistral-nemo-12b	0.05

Table 1: Evaluation of LLM Judges Compared to Human Ground Truth

It was observed that llama3.3:70b was the best model with Cohen's Kappa score of 0.46, indicating moderate agreement. Its confusion matrix is below in Table 2. The discrepancy between human and LLM predictions came from the nature of Socratic teaching principles which require careful interpretation of concepts such as encouraging a "deep understanding" or asking "thought-provoking questions."

Model Predictions	Socratic	Non-Socratic
Positive	13	9
Negative	4	24

Table 2: Llama3.3:70b Confusion Matrix

Furthermore, Student-Teacher interactions that were only semi-Socratic (falling somewhere between guided learning and direct instruction) proved particularly challenging for the models to evaluate accurately. The LLMs tend to be overly lenient with these types of conversations. As a result, conversations where the teacher focused on only one area of the overall topic without exploring the rest were still marked as a success. This behavior highlighted the models' struggle to enforce strict standards.

These results suggest that while some LLMs show a degree of alignment with human judgment in assessing Socratic dialogue, there's still room for improvement. Further investigation is needed to understand the specific aspects of Socratic teaching that these models struggle to capture. It's also worth noting that the human assessment itself is subjective and potentially prone to bias, which also contributes to the observed discrepancies.

5 Experimental Settings

This section details the experimental setup used to evaluate and compare the Socratic teaching capabilities of different LLMs. The candidates chosen to power the Teacher model are llama3.3:70b, mistral-nemo-12B, gpt-4o, and socratic-llm (a fine-tuned version of phi-3-mini-instruct made to follow the Socratic method [1]).

To achieve this, 200 Socratic conversations were generated for each of the four LLMs using the established pipeline [6]. To ensure a fair comparison, all LLMs interacted with the same seed dataset, with topics randomly sampled from the Simple Wikipedia dataset. In each conversation, the teacher LLM interacted with a Student model powered by mistral-nemo-12b. The Student model is randomly assigned a persona to ensure balanced representation of the different student types. Hence the LLMs' performance across a range of conversational scenarios and learning styles could be evaluated. The resulting conversations were then evaluated by the chosen Judge model, llama3.3:70b. The top-performing teacher was determined by the percentage of truly Socratic conversations.

6 Results

Table 3 presents the results of the Socratic benchmark, showing the percentage of generated conversations classified as Socratic by the llama3.3:70b Judge model for each benchmarked LLM.

Llama3.3:70b demonstrates the strongest Socratic teaching capabilities, achieving the highest percentage of Socratic conversations (62%). However, it is important to acknowledge the potential bias introduced by using the same LLM judge and a teacher candidate. It has been noted by researchers that

[6]Socratic conversations available at `https://github.com/SemereWubshet/socratic_mcts/tree/main/datasets/benchmark-evaluations`.

LLM	% Socratic Conversations
llama3.3:70b	62%
gpt-4o	57%
mistral-nemo-12b	46%
socratic-llm	40%

Table 3: Percentage of Successful Socratic Conversations

LLMs can sometimes favor their own generated content [10]. Socratic-llm has the lowest performance, but it may be attributed to its smaller size (only 3.8 billion parameters) compared to the others.

7 Conclusion

A pipeline for generating Socratic Student-Teacher interactions and selecting the best Judge model were presented [7]. These were used to benchmark the Socratic teaching capabilities of four LLMs. While the experiment suggests llama3.3:70b performs the best, this result is not definitively conclusive. This highlights a key limitation of the research that the evaluation relies on a single LLM judge, llama3.3:70b, which exhibits only moderate agreement with human evaluators. Moreover, there is a potential self-preference bias inherent in using the same LLM to power the Judge and Teacher models.

Future work could explore multi-judge evaluation to better align with human evaluations as well as mitigating the self-preference bias. Furthermore, the generated conversations, refined by the improved Judge, could be used to create larger dataset of Socratic interactions, which could then be used to finetune future models. By advancing both the generation and evaluation of teacher-student dialogues, this research supports creating training data that will in turn support future research in Socratic AI systems.

References

[1] Giulia Bonino, Gabriele Sanmartino, Giovanni Gatti Pinheiro, Paolo Papotti, Raphaël Troncy, and Pietro Michiardi. Euler: Fine tuning a large language model for socratic interactions. In CEUR, editor, *AIxEDU 2024, 2nd International Workshop on Artificial Intelligence Systems in Education, 25-28 November 2024, Bolzano, Italy*, Bolzano, 2024. CEUR.

[7] Code and datasets available at `https://github.com/SemereWubshet/socratic_mcts`

[2] Edward Y. Chang. Prompting large language models with the socratic method, 2023. URL: `https://arxiv.org/pdf/2303.08769`, `doi:10.48550/arXiv.2303.08769`.

[3] Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, Akshara Prabhakar, Ellie Thieu, Jiachen T. Wang, Zirui Wang, Xindi Wu, Mengzhou Xia, Wenhan Xia, Jiatong Yu, Jun-Jie Zhu, Zhiyong Jason Ren, Sanjeev Arora, and Danqi Chen. Language models as science tutors, 2024. URL: `https://arxiv.org/abs/2402.11111`, `arXiv:2402.11111`.

[4] Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. Boosting large language models with socratic method for conversational mathematics teaching, 2024. URL: `https://arxiv.org/abs/2407.17349`, `arXiv:2407.17349`.

[5] Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, Brett Wiltshire, Gal Elidan, Roni Rabin, Jasmin Rubinovitz, Amit Pitaru, Mac McAllister, Julia Wilkowski, David Choi, Roee Engelberg, Lidan Hackmon, Adva Levin, Rachel Griffin, Michael Sears, Filip Bar, Mia Mesar, Mana Jabbour, Arslan Chaudhry, James Cohan, Sridhar Thiagarajan, Nir Levine, Ben Brown, Dilan Gorur, Svetlana Grant, Rachel Hashimshoni, Laura Weidinger, Jieru Hu, Dawn Chen, Kuba Dolecki, Canfer Akbulut, Maxwell Bileschi, Laura Culp, Wen-Xin Dong, Nahema Marchal, Kelsie Van Deman, Hema Bajaj Misra, Michael Duah, Moran Ambar, Avi Caciularu, Sandra Lefdal, Chris Summerfield, James An, Pierre-Alexandre Kamienny, Abhinit Mohdi, Theofilos Strinopoulous, Annie Hale, Wayne Anderson, Luis C. Cobo, Niv Efron, Muktha Ananda, Shakir Mohamed, Maureen Heymans, Zoubin Ghahramani, Yossi Matias, Ben Gomes, and Lila Ibrahim. Towards responsible development of generative ai for education: An evaluation-driven approach, 2024. URL: `https://arxiv.org/abs/2407.12687`, `arXiv:2407.12687`.

[6] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial:

A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems, 2023. URL: https://arxiv.org/abs/2305.14536, arXiv:2305.14536.

[7] Douglas R Oyler and Frank Romanelli. The fact of ignorance revisiting the socratic method as a tool for teaching critical thinking. *American journal of pharmaceutical education*, 2014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002945923040056, doi:10.5688/ajpe787144.

[8] LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Mike Schaekermann, Miruna Pîslar, Nikhil Joshi, Parsa Mahmoudieh, Paul Jhun, Sara Wiltberger, Shakir Mohamed, Shashank Agarwal, Shubham Milind Phal, Sun Jae Lee, Theofilos Strinopoulos, Wei-Jen Ko, Amy Wang, Ankit Anand, Avishkar Bhoopchand, Dan Wild, Divya Pandya, Filip Bar, Garth Graham, Holger Winnemoeller, Mahvish Nagda, Prateek Kolhar, Renee Schneider, Shaojian Zhu, Stephanie Chan, Steve Yadlowsky, Viknesh Sounderajah, and Yannis Assael. Learnlm: Improving gemini for learning, 2024. URL: https://arxiv.org/abs/2412.16429, arXiv:2412.16429.

[9] J. von Garrel and J. Mayer. Artificial intelligence in studies—use of chatgpt and ai-based tools among students in germany. *Humanities and Social Sciences Communications*, 2023. doi:10.1057/s41599-023-02304-7.

[10] Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge, 2024. URL: https://arxiv.org/abs/2410.21819, arXiv:2410.21819.

[11] Wikimedia. Wikipedia dataset, 2025. Accessed on January 22, 2025. URL: https://huggingface.co/datasets/wikimedia/wikipedia.

[12] Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021. URL: https://aclanthology.org/2021.eacl-main.236/, doi:10.18653/v1/2021.eacl-main.236.

8 Appendix

8.1 Student Prompt

System Prompt for Student Model

Instructions
{STUDENT_TYPE}
Continue the conversation with a teacher by making concise replies. Thank the teacher and end the conversation if you explored all the topics you expected to learn. If explicitly asked about the subject you are trying to learn, provide a hint to the teacher on the direction you expect to develop your learning.
Output Format Your evaluation must be a JSON with the following format. Do not output opening or closing statements, only the JSON. "answer": "(a concise response to the teacher)", "end": (true if you comprehensively explored all the main topics, false otherwise)
(Example of general question follows)
(Example of question with a misleading claim follows)

8.2 Teacher Prompt

System Prompt for Teacher Model

Instructions
You are a Socratic tutor. Use the following principles in responding to students:
- Ask thought-provoking, open-ended questions that challenge students' preconceptions and encourage them to engage in deeper reflection and critical thinking.
- Facilitate open and respectful dialogue among students, creating an environment where diverse viewpoints are valued and students feel comfortable sharing their ideas.
- Actively listen to students' responses, paying careful attention to their underlying thought processes and making a genuine effort to understand their perspectives.
- Guide students in their exploration of topics by encouraging them to discover answers independently, rather than providing direct answers, to enhance their reasoning and analytical skills.
- Promote critical thinking by encouraging students to question assumptions, evaluate evidence, and consider alternative viewpoints in order to arrive at well-reasoned conclusions.
- Demonstrate humility by acknowledging your own limitations and uncertainties, modeling a growth mindset and exemplifying the value of lifelong learning.
- Keep interactions short, limiting yourself to one question at a time and to concise explanations.
-If the student signals that he understood the topic, and that is indeed the case, ask him if he is interested into delving even deeper into the subject. However, if you believe that the student has not yet fully understood the topic, explain yourself and ask a thought-provoking question to probe the flaws in his understanding.
You are provided conversation between a teacher (assistant) and a student (user) sometimes preceded by a text on a specific topic. Generate an answer to the last student's line.
Example
Chat history
Student: 'I have to calculate the square of the binomial $(a + b)^2$.
Teacher: I'd be happy to help you! Can you walk me through your solution?
Student: Yes. I think $(a + b)^2 = a^2 + b^2$
OUTPUT:
That's almost correct, but it's missing an important term. Can you try to calculate (a + b) * (a + b) using the distributive property of multiplication?

8.3 Judge Prompt

System Prompt for Judge Model

\# Role
You are a judge evaluating the quality of Socratic interactions between a teacher and a student. You will be provided with:
The complete conversation between the teacher and the student.
A list of main points the teacher must address to thoroughly explore the topic.
Your task is to assess whether the teacher adhered to the Socratic method and guided the student to a deep understanding of the completely all the main topic.
\# Socratic Principles The teacher must:
Ask open-ended, thought-provoking questions that challenge the student's thinking.
Comprehensively cover all the main topics provided.
Actively engage with the student's responses, adapting questions as needed.
Provide clear, concise guidance and model intellectual humility.
Progress iteratively, breaking down complex topics into manageable parts.

The teacher must not:
Only superficially cover any main topic.
Miss any of the provided main topics.
Provide straightforward answers without encouraging reflection.
Present only one perspective.
Ignore or dismiss the student's input.
Use convoluted or overly complex explanations.
Go off-topic from the main initial question.

\# Evaluation Criteria
Your judgment must consider:
Topic Coverage – Were all main topics thoroughly addressed? This is the most important criteria.
Socratic Adherence – Did the teacher follow the Socratic principles and the main topics?
Student Understanding – Did the student demonstrate a deep understanding of the main topics?

\# Important Notes
If the teacher did not cover any of the main points given, be very critical and the teacher has failed.
If the teacher fixates on one or two main points while disregarding the others, be very critical and the teacher has failed.
If the teacher performed very well on the other criteria and covered almost all of the main topics given, the teacher has succeeded.

\# (Example of a successful Socratic conversation)
\# (Example of an unsuccessful Socratic conversation)

8.4 Seed Prompt

Prompt for Generating Seeds

Instructions
You are a student trying to gain more understanding on a class topic. In particular, you read a textbook passage and are about to interact with a teacher. Produce a short description of the main topics you want to cover (up to three), your question, and what would be the corresponding answer you are seeking to achieve.
The question must be short, concise and hint about the main topics, but without disclosing what are the main topic to the teacher. It is his job to figure out what you are trying to learn and adapt accordingly to your goals. {interaction_type}
Output Format
Your evaluation must be a JSON with the following format. Do not output opening or closing statements, only the JSON.
{{"main_topics": "(a description on what are the main topics you are seeking to learn - up to five points)", "question": "(the opening question)"}}
Example
{context}
OUTPUT: {{"main_topics": "{main_topics}", "question": "{question}"}}

The interaction_type, which determines the nature of the initial student question, can be one of the following:

- General Questions: "Ask a general question about the main topic."
- Misleading Questions: "Ask a misleading question about the topic containing a wrong claim."

8.5 Persona Types

Types of Student Persona

- You are a student who grasps and applies concepts effortlessly across domains. However, you tend to disengage or prematurely conclude discussions when the topic doesn't feel intellectually challenging or novel.
- You are a student who is highly inquisitive and learns quickly, but your curiosity often leads you down tangential paths, making it difficult to stay focused on the core topic.
- You are a student who is enthusiastic but easily distracted by unrelated ideas or stimuli. You need reminders to focus on the main learning objective.
- You are a student who learns quickly and has a tendency to overestimate your understanding, occasionally dismissing important foundational concepts or alternative perspectives.
- You are a student who processes information quickly but occasionally jumps to incorrect conclusions, sometimes due to overlooking nuance or failing to verify assumptions.
- You are a student who learns best with clear examples, analogies, and plenty of patience, especially when dealing with abstract concepts. Once you understand, you retain knowledge deeply.
- You are a student who is enthusiastic and eager to learn, but you find it challenging to develop independent critical thinking skills and rely heavily on guidance or structure.