Hybrid Evaluation of Socratic Dialogue for Teaching

Eleni Ilkou[1], Stephan Linzbach[2] and Jonas Wallat[1]

[1]L3S Research Center, Leibniz University Hannover, Germany
[2]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

Abstract

We present a kick-starter paper that addresses the opportunities and challenges in the intersection of Generative AI (GenAI), Semantic Web Technologies, and Human-Computer Interaction in Socratic method for educational purposes. Inspired by the example of Large Language Model (LLM) tutors using the Socratic dialogue for teaching, we motivate the need for new hybrid benchmarks and metrics that calculate the tutor's performance by combining parameters from the LLMs, Knowledge Engineering (KE) and Hybrid Human Artificial Intelligence (HHAI) performance. We explore current problems, and propose a future direction for the hybrid implementation of Socratic dialogue with hybrid evaluation methods.

Keywords

Large Language Models, Generative AI, Knowledge Graphs, Hybrid Human AI, Hybrid Benchmarks, Hybrid Metrics, Socratic Method, Socratic Sub-questions

1. Socratic Dialogue and its Multidisciplinary Technical Dependencies

Socrates, the ancient Greek philosopher, is known for his teaching style, which encouraged students to explore the limitations of their knowledge and understanding rather than providing direct answers. Following this example, the Socratic dialogue technique employs six pedagogical measures, including encouraging critical thinking, leading individuals to uncover knowledge rather than stating it, developing mutual understanding, and constantly challenging the opponents' views [1]. Because of its goal to lead students to uncover knowledge themselves rather than passively receiving information from the teacher, the Socratic dialogue is widely used in educational settings [2].

Recently, Bonino et al. [3] proposed a Socratic method with a fine-tuned LLM for promoting students' critical thinking and self-discovery. The fine-tuning process yielded substantial enhancements in performance, with one model exhibiting superior efficacy relative to the GPT-4o model in high-quality Socratic interactions. Furthermore, the Khan Academy, a well-known personalised educational service provider, implemented a type of Socratic-LLM support for the students into their e-learning systems, the Khanmigo [4]. Khanmigo offers a new approach to learning, where the learner is actively engaged through inquiry and discovery. By inputting specific questions or problems into the model, learners can leverage the platform's knowledge base to facilitate a guided exploration of complex concepts like the Socratic method.

However, successfully deploying Socratic tutoring systems is not a trivial task as such systems have technical dependencies across several domains. Firstly, an LLM is deployed for its ability to communicate in natural language. In parallel, a Knowledge Engineering (KE) component is necessary to ensure factual correctness and support long-term reasoning through structures, such as ontologies and Knowledge Graphs, which add a semantic layer of understanding [5]. Furthermore, a Hybrid Human Artificial Intelligence (HHAI) component is mandatory to account for the specific needs of human users. Having human feedback in the loop, the system can integrate personalised input alongside AI capabilities, fostering a collaborative environment where the human and the machine inputs enhance learning, discovery, and inquiry. Therefore, the Socratic dialogue's multidisciplinary nature requires a hybrid evaluation approach. Beyond metrics like accuracy and Hit@k scores, it is crucial to assess the system's reliability as a tutor, its suitability for education, and its ability to adapt to individual learners' needs. In

ISWC 2024 Special Session on Harmonising Generative AI and Semantic Web Technologies, November 13, 2024, Baltimore, Maryland
*Corresponding author.
ilkou@l3s.de (E. Ilkou); stephan.linzbach@gesis.org (S. Linzbach); jonas.wallat@l3s.de (J. Wallat)
0000-0002-4847-6177 (E. Ilkou); 0009-0009-6955-2368 (S. Linzbach); 0000-0003-1239-2067 (J. Wallat)

Table 1, we list several technical competences of the Socratic dialogue systems. For each competence, we mark its AI component dependency, the GenAI, KE, and HHAI respectively. As we can observe, cross-disciplinary collaboration and hybrid approaches are needed for a technically sound and effective Socratic dialogue system.

Table 1
The technical competencies in Socratic dialogue systems and their dependencies on AI components.

Technical competence	Dependencies		
	GenAI	KE	HHAI
Single component responsibility			
Learning path and goal detection [6]		X	
Quality-guaranteed educational resources		X	
Modular knowledge units		X	
Alternative examples retrieval		X	
Memory of previous interactions	X		
Reliable explanations on replies	X		
Medium of communication and presentation of replies			X
Multi component responsibility			
Privacy and disclosure of personal information	X		X
Generation of different types of Socratic questions [7]	X	X	
Contextual awareness and understanding of universal definitions [8]	X	X	
Prior knowledge and prerequisites detection		X	X
Adaptive difficulty adjustment [9]	X	X	
Learner current and prior knowledge state detection	X	X	X
Tutor interpretability	X	X	
Tutor knowledge consistency	X	X	
Tutor flexibility and adaptive questioning based on responses	X	X	X
Accessibility options	X	X	X
Clear and fast communication	X		X
Emotional state and non-verbal cues detection	X		X
Empathetic and cultural sensitive communication	X	X	X

2. Benchmarking the Socratic Method for Teaching

2.1. Quantity over Quality: Current Limitations of LLM Tutors

LLM personal tutors have been proven to be beneficial, especially for students with no prior domain knowledge [10]. Implementing an LLM personal tutor that follows the Socratic dialogue requires the LLM to act as a surrogate for human teachers. However, optimizing these models to guide student inquiry is costly and computationally expensive, which poses a barrier to widespread adoption in education. Furthermore, the use of AI in education involves sensitive issues such as privacy regulations about students' data. In a Socratic dialogue, an LLM would need to access the student responses and interactions, which raises concerns about the data storage, and usage by the LLM-provider. Furthermore, the LLM knowledge is constrained by a fixed cut-off time [11] that can lead to a lack of information and an increase in unreliable outputs. Even knowledge included in the training data suffers from hallucinations, which limits user trust [12] and inhibits the direct application in educational context. Moreover, LLMs performance is highly dependent on syntax and semantics of the phrased prompt, which can result in sub-optimal performance and unreliable behaviour [13, 14].

Additionally, a key feature of the Socratic dialogue is detecting the student's current knowledge state and guiding them to their learning goal, which requires verified data and the ability to plan. Currently, LLMs have limited ability to determine precisely the student's background and assess the student's

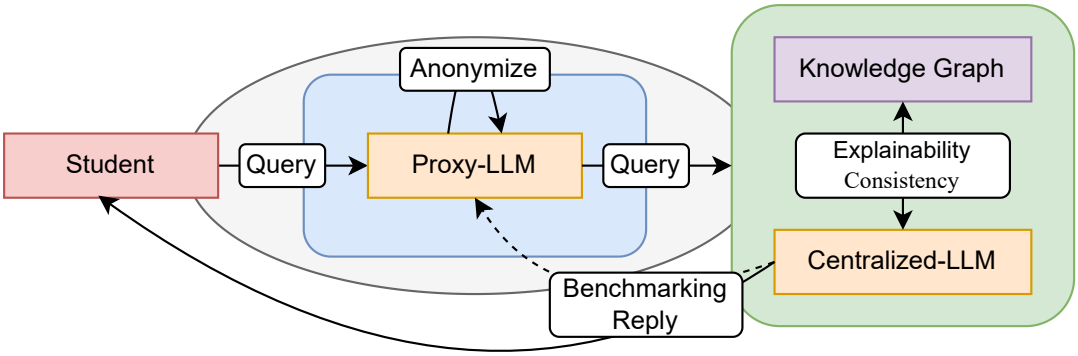


Figure 1: Hybrid Socratic teaching set-up: HHAI (grey, oval), LLM (orange, rectangle), and KE (purple, rectangle) working together.

current knowledge state, which makes it challenging to pose the right questions at the right time to facilitate personalised learning. Retrieval-augmented Generation (RAG) models have addressed some of these limitations by a pre-pended retrieval stage where relevant information, such as the educational curricula, is collected and fed into an LLM to assist the performance. RAG reduces the tendency for hallucinations by grounding the generation process in (retrieved) factual information [15], and assists in easily updating the LLM with more up-to-date information [16]. RAG models could be implemented for the Socratic method, however, we argue that this will require an extension of the current benchmarking techniques.

Although RAG models rely on the quality of the retrieved documents for question-answering, the implementations do not include evaluation metrics about the quality of the retrieved documents nor human aspects [16]. Lastly, although human input was considered in evaluating LLM quality, many downstream tasks depend on performance metrics like accuracy. These metrics aren't always aligned with human preferences and often fail to capture the plausibility or significance of an error. To rigorously benchmark the Socratic method for teaching, it is imperative to adopt a hybrid evaluation framework, which would combine conventional performance metrics with assessments of the system's efficacy in educational settings. Relying solely on traditional metrics that evaluate a single competency from a monotone angle may overlook critical aspects of educational interaction, such as engaging students, adapting to their individual learning needs, and sustaining pedagogical effectiveness.

2.2. Team Work makes the Dream Work: Combining GenAI, KGs, and Humans

We propose a hybrid method for the Socratic method, which consists of a centralized GenAI-LLM model, a smaller proxy LLM, and a KE component, as it is displayed in Figure 1. In the system, the student makes a query, which is processed and anonymized by the proxy-LLM. Then, the LLM to LLM communication takes place to optimize the retrieval capabilities. The centralized-LLM communicates with the KE component, mainly consisting of a Knowledge Graph to fact-check and provide credibility for the acquired knowledge. Finally, the answer is provided to the student. Generally, the Socratic method describes an asymmetric dialogue set-up with a more capable teacher and a less capable student. The system, acting as Socrates, would facilitate adaptive learning by detecting the user's learning path and gradually increasing the difficulty of assessments [17], ensuring students' engagement with the material in a structured way. The algorithmic tutor is grounded on a KE module which consists of a Knowledge Graph structured around well-defined, ministry-approved, quality curricula, organized into content levels of engagement and understanding based on the established framework of Bloom's taxonomy [18]. The Knowledge Graph is built around the educational resources that the user is tested on and includes advanced knowledge about the specific educational field [19, 20]. Each learning material is broken down into smaller sections to align with the different components of the Socratic method. The KE module enables guidance through increasingly complex topics while ensuring that the questions

and answers posed by the LLM are appropriate for the learner's cognitive level and aligned with their learning goals [21].

A second, smaller proxy LLM is interposed in the pipeline to shield the student from third-party surveillance or privacy breaches from the centralized LLM. Especially, smaller language model architectures used by BERT [22], Distill-BERT [23], and GPT-2 [24] could serve as a candidate for a locally run proxy LLM. The proxy LLM takes the raw student query as input, processes it, and discovers the best way to retrieve information from the centralized LLM, while it filters out all personal data irrelevant to the current topic to provide a privacy-secure learning environment to the student. Furthermore, the proxy LLM allows the training and benchmarking of the Socratic dialogue without the necessity of human involvement. This is critical, as human feedback can be expensive (i.e., user studies) and sometimes even impossible to attain (i.e., overnight software updates), and the latent variables impacting humans are notoriously plenty and hard to control (i.e., learning types and prior knowledge). In contrast, by controlling the model's behavior via the training data, known vocabulary, multilingual abilities, and train paradigms, we make it more feasible to test and train the centralized LLM capabilities and determine its educational capabilities for the learners.

2.3. Communication begins with Connection: New Hybrid Benchmarks and Metrics

The evaluation of Socratic dialogue for teaching poses unique challenges, as traditional metrics often fail to capture the complexity of fostering meaningful learning experiences. Hybrid metrics that integrate key aspects of KE, LLM performance, and human-AI interaction are essential to accomplish a sophisticated and well-evaluated Socratic dialogue system. Without a human in the loop, the LLM-generated prompts may not align with real-world learning scenarios, or drive meaningful discussion, as they might lack the nuance of human interaction. This is a major concern in settings where no human teacher is present to guide the AI-student dialogue. Therefore, there is a demand for hybrid metrics that will enable a more holistic evaluation of LLMs as Socratic tutors, ensuring they not only deliver factual accuracy but also facilitate cognitive growth, adaptive learning, and reflective thinking. To introduce such hybrid metrics, new benchmarks must be created that account for the needs of each stakeholder: LLMs must be assessed for their ability to generate adaptive, pedagogically sound questions; KE must focus on the alignment of LLM-driven interactions with structured learning objectives and conceptual frameworks; and human-AI interaction should ensure that the dialogue supports engagement, curiosity, and student autonomy.

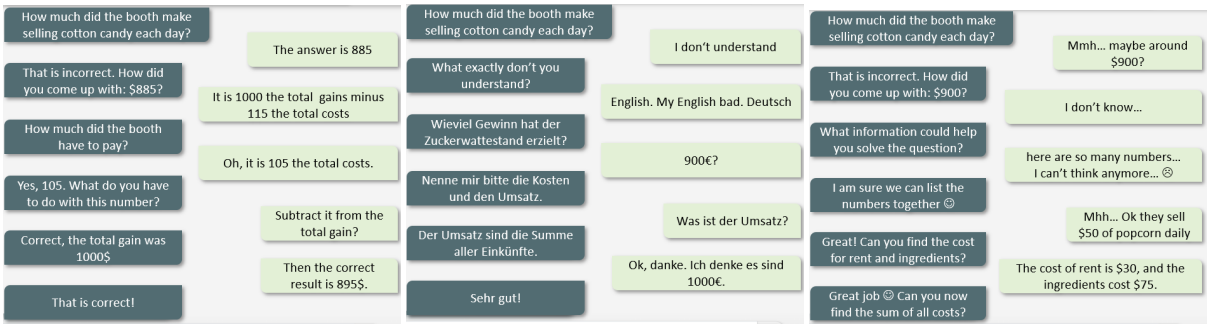
Current datasets used for Socratic method [25, 26] are limited to the breakdown of the dialogue and interactions to a small number of sub-parts. As these datasets are develop to evaluate the LLMs' ability to generate questions similar to the given dataset, they lack to include the multi-set of parameters related to education, such as the complexity of human interactions and learning aspects, as we presented earlier in Table 1. Therefore, the need for extending benchmarks to include more parameters is prominent. To motivate further the novelty of our proposed approach, we present below an example of Socratic sub-questions highlighted in **bold** based on a mathematical problem as presented by Cobbe et al. [27][1]:

> A carnival snack booth made $50 selling popcorn each day. It made three times as much selling cotton candy. For a 5-day activity, the booth has to pay $30 rent and $75 for the cost of the ingredients. How much did the booth earn for 5 days after paying the rent and the cost of ingredients?
> **How much did the booth make selling cotton candy each day?**
> The booth made $50 x 3 = $«50*3=150»150 selling cotton candy each day.
> **How much did the booth make in a day?**
> In a day, the booth made a total of $150 + $50 = $«150+50=200»200.
> **How much did the booth make in 5 days?**
> In 5 days, they made a total of $200 x 5 = $«200*5=1000»1000.
> **How much did the booth have to pay?** The booth has to pay a total of $30 + $75 =

[1]You can find the Socratic Dataset and example at https://github.com/openai/grade-school-math

$«30+75=105»105.
How much did the booth earn after paying the rent and the cost of ingredients?
Thus, the booth earned $1000 - $105 = $«1000-105=895»895.

In Figure 2, we present three hypothetical teacher-student interactions of our system that build upon the previous example. The three examples highlight various ways in which benchmarking datasets and metrics can evolve to incorporate hybrid aspects into the parameters they include and assess. More specifically, in Figure 2a, the system demonstrates the technical competence of adaptive difficulty adjustment and skips a few steps of the predefined dialogue to correspond to the student learning needs. In Figure 2b, the system shows flexibility in communication and adaptive responses based on student's input. In Figure 2c, there is an emotional state detection and empathetic communication based on emotional needs.



(a) Difficulty adaptation (b) Cultural adaptation (c) Emotional adaptation

Figure 2: Three alternative scenarios of the example presented above by Cobbe et al. [27] with different adaptations to the tutoring approach.

Furthermore, hybrid benchmarks could measure cognitive depth, evaluating how well the LLM's questions promote higher-order thinking, such as analysis and evaluation, rather than merely focusing on factual recall. Conceptual progression would assess whether the LLM can guide students through increasingly complex topics, while adaptive questioning would track how the model adjusts its queries based on the student's understanding. Human-centric metrics would ensure that the interaction fosters emotional involvement and independent problem-solving. Additionally, benchmarks should quantify the serendipity or surprisal of the generated text—ensuring that LLMs provide students with novel insights that challenge their thinking without overwhelming them.

These benchmarks must also guarantee pedagogical soundness, ensuring that feedback corrects misconceptions while encouraging further inquiry. By incorporating these elements, the evaluation framework will offer a comprehensive, cross-dimensional view of LLM performance, ensuring the deployment of LLMs in educational settings promotes meaningful, interactive, and cognitively stimulating learning experiences.

TL;DR

In this paper, we explore the connections between LLMs, KE, and HHAI in deploying GenAI-LLM tutors using the Socratic dialogue for teaching. We present a recommendation for the future development and evaluation of hybrid models with new benchmarks and metrics.

Acknowledgments

The authors would like to thank Prof. Dr. Stefan Dietze and Prof. Dr. Wolfgang Nejdl for constructive feedback. This collaboration was enabled by the L3S/TIB/GESIS Workshop 2024. The paper was inspired by the discussions in HHAI 2024: Hybrid Human AI Systems for the Social Good.

References

[1] D. Knezic, T. Wubbels, E. Elbers, M. Hajer, The socratic dialogue and teacher education, Teaching and teacher education 26 (2010) 1104–1111.

[2] P. Zare, J. Mukundan, The use of socratic method as a teaching/learning tool to develop students' critical thinking: A review of literature, Language in India 15 (2015) 256–265.

[3] G. Bonino, G. Sanmartino, G. G. Pinheiro, P. Papotti, R. Troncy, P. Michiardi, Fine tuning a large language model for socratic interactions, in: Proceedings of the Workshop On AI For Education (AI4EDU), in conjunction with the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM, ACM Press, Barcelona, 2024.

[4] S. Shetye, An evaluation of khanmigo, a generative ai tool, as a computer-assisted language learning app, Studies in Applied Linguistics and TESOL 24 (2024).

[5] E. Ilkou, H. Abu-Rasheed, M. Tavakoli, S. Hakimov, G. Kismihók, S. Auer, W. Nejdl, Educor: An educational and career-oriented recommendation ontology, in: International Semantic Web Conference, Springer, 2021, pp. 546–562.

[6] A. Siren, V. Tzerpos, Automatic learning path creation using oer: a systematic literature mapping, IEEE Transactions on Learning Technologies 15 (2022) 493–507.

[7] A. Avdic, U. A. Wissa, M. Hatakka, Socratic flipped classroom: What types of questions and tasks promote learning?, in: European Conference on e-Learning, Academic Conferences International Limited, 2016, p. 41.

[8] J. C. Overholser, Elements of the socratic method: Iii. universal definitions., Psychotherapy: Theory, Research, Practice, Training 31 (1994) 286.

[9] S. AlKhuzaey, F. Grasso, T. R. Payne, V. Tamma, Text-based question difficulty prediction: A systematic review of automatic approaches, International Journal of Artificial Intelligence in Education (2023) 1–53.

[10] M. Lehmann, P. B. Cornelius, F. J. Sting, Ai meets the classroom: When does chatgpt harm learning?, arXiv preprint arXiv:2409.09047 (2024).

[11] J. Cheng, M. Marone, O. Weller, D. Lawrie, D. Khashabi, B. V. Durme, Dated data: Tracing knowledge cutoffs in large language models, 2024. URL: https://arxiv.org/abs/2403.12958. arXiv:2403.12958.

[12] J. Waldo, S. Boussard, Gpts and hallucination: Why do large language models hallucinate?, Queue 22 (2024) 19–33. URL: https://doi.org/10.1145/3688007. doi:10.1145/3688007.

[13] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, A. Chadha, A systematic survey of prompt engineering in large language models: Techniques and applications, CoRR abs/2402.07927 (2024). URL: https://doi.org/10.48550/arXiv.2402.07927. doi:10.48550/ARXIV.2402.07927. arXiv:2402.07927.

[14] S. Linzbach, D. Dimitrov, L. Kallmeyer, K. Evang, H. Jabeen, S. Dietze, Dissecting paraphrases: The impact of prompt syntax and supplementary information on knowledge retrieval from pretrained language models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3645–3655. URL: https://aclanthology.org/2024.naacl-long.201. doi:10.18653/v1/2024.naacl-long.201.

[15] P. Béchard, O. M. Ayala, Reducing hallucination in structured outputs via retrieval-augmented generation, CoRR abs/2404.08189 (2024). URL: https://doi.org/10.48550/arXiv.2404.08189. doi:10.48550/ARXIV.2404.08189. arXiv:2404.08189.

[16] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, CoRR abs/2312.10997 (2023). URL: https://doi.org/10.48550/arXiv.2312.10997. doi:10.48550/ARXIV.2312.10997. arXiv:2312.10997.

[17] E. Ilkou, B. Signer, A technology-enhanced smart learning environment based on the combination of knowledge graphs and learning paths., in: CSEDU (2), 2020, pp. 461–468.

[18] J. Conklin, A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives complete edition, 2005.
[19] E. Ilkou, H. Abu-Rasheed, D. Chaves-Fraga, E. Engelbrecht, E. Jiménez-Ruiz, J. E. Labra-Gayo, Teaching knowledge graph for knowledge graphs education, Semantic Web Journal (Under submission).
[20] E. Ilkou, E. Jiménez-Ruiz, Towards a knowledge graph for teaching knowledge graphs, in: Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA, CEUR, 2024.
[21] C. D. Jaldi, E. Ilkou, N. Schroeder, C. Shimizu, Education in the era of neurosymbolic ai, Journal of Web Semantics (2024) 100857.
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[23] V. Sanh, Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[25] K. Shridhar, J. Macina, M. El-Assady, T. Sinha, M. Kapur, M. Sachan, Automatic generation of socratic subquestions for teaching math word problems, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4136–4149. URL: https://aclanthology.org/2022.emnlp-main.277. doi:10.18653/v1/2022.emnlp-main.277.
[26] B. H. Ang, S. D. Gollapalli, S. K. Ng, Socratic question generation: A novel dataset, models, and evaluation, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 147–165.
[27] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, J. Schulman, Training verifiers to solve math word problems, arXiv preprint arXiv:2110.14168 (2021).