Critical Inker: Scaffolding Critical Thinking in AI-Assisted Writing Through Socratic Questioning

Philipp Hugenroth, LMU Munich
Valdemar Danry, MIT Media Lab
Pattie Maes, MIT Media Lab

Main theme(s): writing support / informal logic / feedback system / critical thinking

Target domain(s): education / writing / learning

Cognitive 'target(s)': critical thinking / reflection / learning

Type of contribution & main idea

With analyzing the argumentation of an essay using LLMs we enable new interaction modalities for writing support that help refining logical argumentation and scaffold critical thinking.

The AI tool's key characteristics

"Critical Inker" is a **goal-driven** tool providing feedback to a written argumentative essay. It is intended to enable the users to use the text understanding benefits of LLMs when refining a text, while preventing cognitive deskilling and potentially even improving critical thinking.

Different to common Chatbots the system analyzes the text, but **won't point out which arguments are wrong** and rewrite them. Instead the chatbot guides the user to **discover and verbalize logical fallacies** themselves using the **Socratic method**. No suggestions will be given and only if an issue is resolved a quote of what the user wants to fix is added as a comment.

Key characteristic of the tool is the analysis of the text using the LLM. First we will **analyze the structure and then the validity of each argument** in the structure. The Socratic chatbot will **receive the analysis and first creates a plan** on how to address the issues and then enters the conversation with the user. It will follow the plan iteratively until all issues are resolved by the user.

		Interaction design/usage strategy	Intermediary task outcomes	Final task outcomes
Interaction & outcomes	How people use and interact with the AI tool for a specific activity (first column) and what outcomes this provides, enables, or leads to (second and third column).	After writing the essay the user will click a button to analyze the text. The tool will break down the argumentation in the background. Visually the main claim and the current sentence the Socratic chatbot is talking about will be highlighted. The socratic chatbot shows up on the right with a text summary and a first question. When resolving an issue a comment will be added on the left side of the text editor. The user can jump out of the conversation edit the text and resolve comments	Intermediary outcome is realizing logical fallacies in your argumentation by yourself and verbalizing them in the Socratic chat. This can vary from small hints, where a person might have forgotten to mention something to structural issues in the argumentation.	Foremost the user will refine their essay to have a stronger underlying logical argumentation. The main outcome is a refined argumentative essay. At the same time we anticipate that using the tool might prevent deskilling or can even improve critical thinking.
Objective value	The 'objective' quality of the process and the outcomes – i.e. what makes them 'good' or desirable (e.g. critical thinking, understanding, new insights, learning).	A change to existing tools is that the chatbot knows of the logical structure and on purpose hides from the user what arguments are weak or flawed. We evaluated the LLM argument analysis and show in the workshop paper that it is comparable to human performance on the AAEv2 and SNIL dataset. To validate the design intervention we want to study it compared to regular chatbots as writing support [1][3].	Realizing mistakes by yourself and verbalizing them might feel helpful and could lead to avoiding them in the future [2]. This could be measured through engagement questionnaires, edits made to the text and measure the amount of logical issues before and after the dialectic conversation.	The Socratic method is considered as an approach to improve critical thinking [4]. Therefor the dialectic conversation in the tool might have a learning effect, which we want to study over time. For showing a learning effect we would need to run a longitudinal study where participants write using this method for several weeks. Critical thinking could be asses through the critical thinking rubric and letting people write without intervention after using the tool for a while.
Perceived value & UX	How the process of using the AI tool and its outcomes is experienced by users (e.g. how they experience the cognitive work they are doing).	User engage in the conversation and discover mistakes. They verbalize what they want to change and when they step out of the conversation change the text based on their intention. This could be assessed by engagement questionnaires, edits made to the text and measure the amount of logical issues before and after the dialectic conversation.	The interface creates friction delaying the user from refining an essay compared to a classical chatbot. If the user finds the interaction with the Socratic chatbot engaging and helpful after using it more than once it could indicate that they value the longterm benefits over the intermediary friction.	Desirable outcomes from the user perspective are essays with less logical issues and the feeling of improving their critical thinking over time. Undesirable outcomes include longer time needed to write and high affordance. If they can benefit from the potential learning effect in school or in their work setting it might be seen as valuable.

How do people's goals 'interact with' the AI tool's goals? Is there a tension?

The tool enables users to refine their writing and improve the logical argumentation. **If the users has the incentive of improving their text** (e.g. in education) the AI tool's goals align with the user goals.

At the same time the tool might create a tension as having to discover all issues through a dialogue feels **slower than directly prompting an LLM** to show the issues or even fix them.

For the AI tool and the user goals to align the user has to get a value out of improved critical thinking. On a broader level we could even argue that we **as a society want to prevent deskilling** the tool might be seen as societal valuable independent of the individual end user.

How do you expect people to continue using this AI tool?

I expect people to continue **using the tool when improving critical thinking** is in the interest of the user.

Adjusting the information communicated by the tool depending on the use case and only every other time intervening might improve the likelihood of continued usage. In our paper we discuss a feedback method that directly points out mistakes, which could be combined with the socratic method.

What you would like to discuss

We would like to discuss our learnings from building the project and the first feedback from user trials.

Additionally it would be interested in to start a conversation around **measuring learning effects in TFT projects.**

What would you like to take away from the workshop?

We would love to get **feedback on our prototype** and **find collaborators to explore different use cases** beyond argumentative essays (e.g. law, medicine, scientific writing).

Key references (e.g. of main theories, empirical evidence, measurement methods etc.)

[1]: Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287 **[2]:** Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023). Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3544548.3580672 **[3]:** Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544 **[4]:** Ho, Y., Chen, B., & Li, C. (2023). Thinking more wisely: using the Socratic method to develop critical thinking skills amongst healthcare students. *BMC Medical Education*, *23*(1), 173. https://doi.org/10.1186/s12909-023-04134-2

How to proceed with this work/idea?

For the workshop we want the **tool to be useable for participants** to try it out.

With finalizing the project work we will **open source all code and prompts** accommodating the final results. Input from the workshop can inspire collaboration with us on further development on "Tools for Thought".