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

Philipp Hugenroth
MIT Media Lab
Cambridge, United States
philhuge@media.mit.edu

Valdemar Danry
MIT Media Lab
Cambridge, United States
vdanry@mit.edu

Pattie Maes
MIT Media Lab
Cambridge, United States
pattie@media.mit.edu

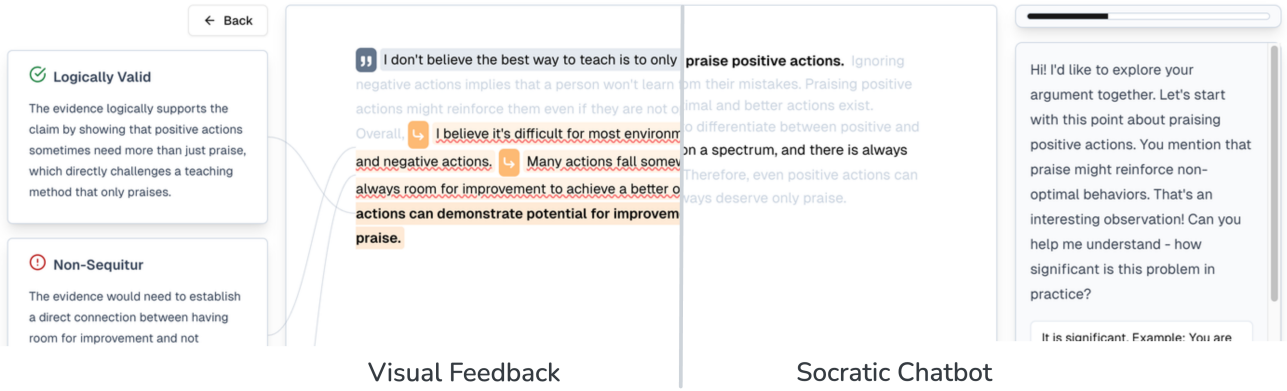

Figure 1: Methods for critical reflection: Visual feedback (Left) shows the argument structure directly with validity checking. The Socratic Chatbot (Right) engages the user in a dialogue, focusing on one segment at a time.

Abstract

As Large Language Models (LLMs) increasingly automate writing tasks, there is a growing risk of cognitive deskilling where users offload critical thinking to the system. To address this, we introduce *Critical Inker*, a writing tool designed to scaffold critical reflection during writing through logical analysis and socratic feedback. We present two methods: (1) A Socratic chatbot using questions to help them realize and fix logical errors in their writing and (2) Visual Feedback, which highlights logical errors in the text without dialog. We detail the technical implementation of the system and evaluate its argument extraction and logical validity accuracy. Our evaluation shows a 91.2% argument overlap with ground truth argument annotations and 87% validity accuracy. Finally, we conducted a small-scale pilot and discuss early qualitative results.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**; *Interactive systems and tools.*

Keywords

AI-assisted writing, cognitive deskilling, Socratic method, productive friction, critical thinking

1 Introduction

The adoption of LLMs in writing workflows offers significant efficiency gains, but raises concerns about cognitive offloading and weakening of critical thinking. Writing is deeply intertwined with thinking, an iterative cognitive process involving planning, evaluation, and revision through which reasoning is constructed and refined [14]. Externalizing thoughts through mediums such as writing enables individuals to identify logical gaps, reconsider assumptions, and develop coherent arguments [26], making writing essential in law[35], science[1], journalism[27], and public safety[29]. Argumentative writing particularly requires writers to articulate claims, justify them with evidence, and ensure logical coherence, making it a critical vehicle for developing reasoning and critical thinking skills [33].

LLM-powered writing assistants have been rapidly integrated into educational, professional, and creative environments [20, 40, 42], improving productivity and reducing effort [13]. However, this can shift writing from an active reasoning process to supervising AI-generated output.

Research suggests this has unintended cognitive consequences. Studies show AI assistants lead users to offload reasoning without fully engaging underlying cognitive processes [6, 12]. While AI improves immediate performance, it may not build lasting reasoning ability—dialogues with AI reduce belief in misinformation during interaction but don't lead to durable discernment skills [28]. Neuroscience evidence suggests heavy reliance on AI writing reduces engagement of brain regions associated with deep cognitive processing [18], potentially leading to deskilling [7]. Co-writing with language models systematically shifts users' views toward model-favored positions and homogenizes content [2, 16], raising concerns that efficiency-focused assistants may inadvertently weaken engagement in reasoning processes and lead to unfavorable outcomes.

Most existing AI writing assistants focus on producing high-quality text with minimal effort. Although these systems are effective in improving the final product, they reduce opportunities for reflection and critical evaluation by providing direct answers and revisions. In contrast, interaction designs that encourage active reasoning, such as cognitive forcing functions or metacognitive interventions, improve users' ability to critically evaluate information [6, 23, 36–38, 41]. However, these designs are often described as more difficult and less preferred.

A promising alternative frames AI interventions as questions guiding users to conclusions themselves, drawing from the Socratic method. Prior work shows reframing AI explanations as questions improves users' ability to identify flawed reasoning and increases logical discernment [11]. Question-driven guidance can support critical thinking and reflection [19, 39]. However, applying Socratic interaction to writing requires understanding the underlying logical structure of arguments.

Recent advances in LLMs have enabled the extraction of informal logical structures from natural language [21, 42], evaluation of logical validity and the detecting fallacious reasoning [17, 30]. By combining argument-structured representations with Socratic interaction, AI systems can enable targeted interventions that guide users toward independently recognizing and resolving logical issues.

Despite progress in AI-assisted writing and reflective interaction design, existing writing assistants rarely integrate argument-structured analysis with Socratic interaction to scaffold reasoning during writing. To address this gap, we introduce *Critical Inker*, a writing tool designed to scaffold critical reflection during writing through AI-assisted argument analysis and Socratic interaction. Critical Inker identifies claims, supporting evidence, and potential logical flaws in argumentative essays, and guides users through structured questioning that encourages them to articulate and resolve issues themselves. By combining LLM-based logical analysis with interaction designs that preserve cognitive engagement, Critical Inker demonstrates how AI writing assistants can support reasoning rather than replace it.

2 System Design

Critical Inker is a web-based writing tool that provides argument-aware feedback on argumentative essays through two interaction modalities: *Visual Feedback* and *Socratic Chatbot*. Both modes share the same underlying argument analysis but differ in how they surface issues and engage users in reflection.

2.1 Core Workflow

The interaction follows three stages:

- **1. Write.** Users write argumentative essays in a minimal text editor without real-time feedback. When ready for analysis, they trigger the system to evaluate their argumentation.
- **2. Analyze.** The system extracts the argument structure using a LLM pipeline (detailed in Section 2.3): first identifying claims, premises, and their relationships, then validating logical connections.
- **3. Reflect.** The two modalities diverge here:
  - *Visual Feedback*: The main claim is highlighted and the full logical structure can be explored. Logical flaws are redlined with explanations (see Figure 1). Users can click through the argument tree to explore nested reasoning chains at their own pace.
  - *Socratic Chatbot*: Instead of showing errors directly, the chatbot asks targeted questions about specific arguments (see Figure 2). Only after users verbalize the issue does the system add a comment marker as a revision reminder. A progress indicator shows completion across all identified issues. In contrast to the Visual Feedback method, only the sentences directly relating to the current Socratic conversation are highlighted.

2.2 Design

Our design choices follow established principles from cognitive psychology and Human Computer Interaction to support active reasoning rather than passive acceptance of AI-generated content.

2.2.1 Delayed feedback during writing. Unlike real-time grammar checkers, Critical Inker provides no feedback while typing. Research in cognitive load theory demonstrates that immediate, fragmented feedback can increase extraneous cognitive load and interfere with the generative thinking required during initial composition [15]. While immediate feedback has been shown to reduce errors in procedural tasks, delayed feedback better supports metacognitive development and deeper understanding in complex learning contexts [22].

2.2.2 Verbalization requirement (Socratic mode). Research on self-explanation demonstrates that learners who actively generate explanations to themselves achieve significantly deeper understanding and improved problem-solving performance compared to those who passively receive information [9–11]. The self-explanation effect occurs when students articulate their reasoning, which prompts them to identify gaps, generate inferences, and repair mental models [9]. Based on prior work in Socratic questioning [11, 25], the Socratic chatbot is prompted to not give away the answer, but rather support the user to arrive at the identified logical errors themselves. Only when the user identifies the issue themselves will the Socratic chatbot create a comment marker reminding the user about what they want to change. This verbalization requirement transforms internal reflection into concrete action, ensuring engagement translates into text revision.

2.2.3 Question grounding in argument structure. Socratic questioning is most effective when it is systematic, disciplined, and focused on specific foundational concepts rather than generic prompts [25]. Effective Socratic questions probe specific claims, assumptions, and logical relationships through targeted inquiries that require students to justify their reasoning [25]. Rather than asking generic questions like "Is this claim supported?", Critical Inker's chatbot references specific logical relationships extracted from the argument graph: *"You claim X because Y, but I am curious, how does Y actually support X?"* This specificity, enabled by LLM-based argument mining, helps users locate precise gaps in their reasoning and engage with the actual structure of their arguments to fix them.

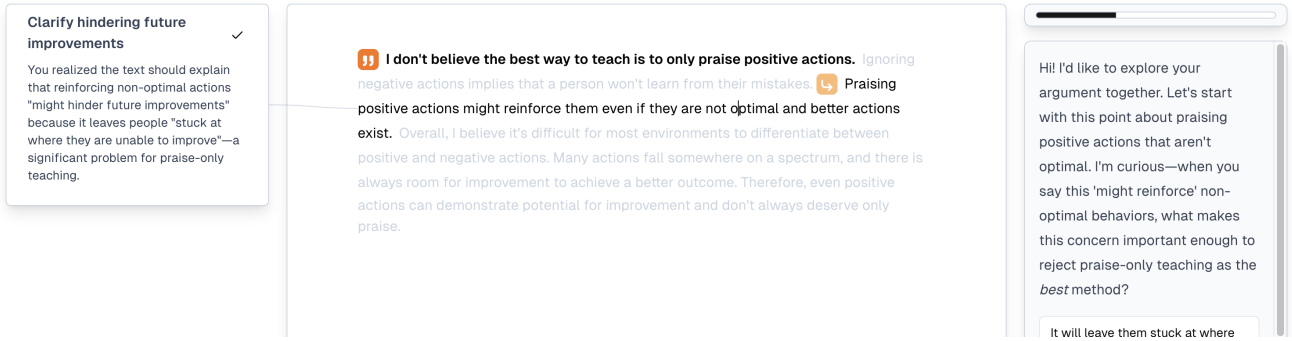

Figure 2: Socratic Chatbot: On successful Socratic scaffolding, the system converts the user's verbalized intention into an actionable comment marker anchored to the text.

2.2.4 Progressive disclosure. Progressive disclosure is an interaction design technique that reduces cognitive overload by revealing information sequentially rather than all at once [24, 31]. Research shows that when users face too many simultaneous stimuli, their working memory capacity can be exceeded, leading to errors, missed information, and task abandonment [34]. The Socratic chatbot addresses one argument at a time rather than overwhelming users with all issues simultaneously. The progress bar provides a sense of accomplishment while maintaining focus on individual reasoning chains, following established principles of staged disclosure [8].

2.3 Argument Extraction and Evaluation

Critical Inker uses a multi-stage prompting pipeline for both interaction modalities: (1) Structure Extraction, (2) Logical Evaluation, and (3) Socratic Intervention. We split structure extraction, evaluation, and Socratic dialogue into independent subtasks to separate the context allowing for different outputs.

2.3.1 Task 1: Structure Extraction. The first prompt (A.1) transforms the user's essay into a argumentation graph. LLMs struggle reliably return character indices or exact quotes. To solve this, we explicitly instruct the model to return "atomic quotes" for every claim or premise identified. Combined with fuzzy string-matching this worked well in our evaluations.

Furthermore, the prompt instructs the model to distinguish between *independent reasons* (which support a claim on their own) and *joined reasons* (which only function as support when combined). This distinction is encoded in the JSON output structure, where independent reasons are arrays of single items `[id, target]` and joined reasons are nested arrays `[[id1, id2], target]`. Using notation and language from literature about informal logic helped us to get more reliable outputs.

2.3.2 Task 2: Logical Evaluation. We observed that single-shot prompting (extracting structure and evaluating validity simultaneously) led to degraded performance. Instead we decided to iterate through each support relation found in Phase 1 and validate it individually using a dedicated Evaluation Prompt (A.2).

This prompt forces the model to engage in chain-of-thought reasoning before outputting a verdict. The JSON schema requires a `rationale` field ("Think through logical validity step by step assuming that the evidence is true") before the `strength` field ("valid" or "invalid").

2.3.3 Task 3: Socratic Questioning. For the chatbot modality, we designed a system prompt that strictly forbids direct correction (A.4). The instructions emphasize a "reasoning assistant" persona that "interestedly inquires" about the user's intent.
The chatbot is provided with the full JSON analysis from Phase 2. It tracks the conversation state against a generated plan (A.3) ensuring it only addresses one logical flaw at a time. If a flaw is resolved by the user it will create a comment citing what the user intended to fix.

To ensure robustness, all model interactions enforce strict JSON schemas either through the model's native "tool use" capabilities (e.g., Anthropic's `json_response` tool) or JSON-mode constraints.

3 Technical Validation

To create helpful interactions, the system requires a low-latency and high-accuracy analysis of the argumentation of the text. Therefore we evaluated the prompts with different models to show the idea is feasible and inform our model choice.

3.1 Methodology

We evaluated the prompts against established human-labeled datasets on how well they understand the structure of the argumentation and if they are able to detect if an argument is valid.

- **Structure:** We used the *Argument Annotated Essay v2* dataset (TU Darmstadt) [32] to test the model's ability to correctly identify claims and premises.
- **Validity:**We adapted the *Stanford Natural Language Inference (SNLI)* dataset [3], converting logical pairs (premise & hypothesis) to test the validity.

For both datasets we choose a random sample size of 100 essays or logical pairs.

3.2 Prompt Evaluation Results

For the argument structure extraction we evaluated the models on three key metrics. (1) **Main claim:** All tested models performed similarly with ≈ 90% accuracy, indicating that current models successfully grasp the informal logic and writer intention. (2) **Relations:** The model achieved a relations overlap of 91.2%. While slightly lower than the top score (92.9%). (3) **Latency:** Claude Sonnet 4.5 achieved a mean execution time of 6.58s, approximately 12% faster than GPT-4.1 (7.48s) and ~ 5× faster than Gemini Flash (37.12s).

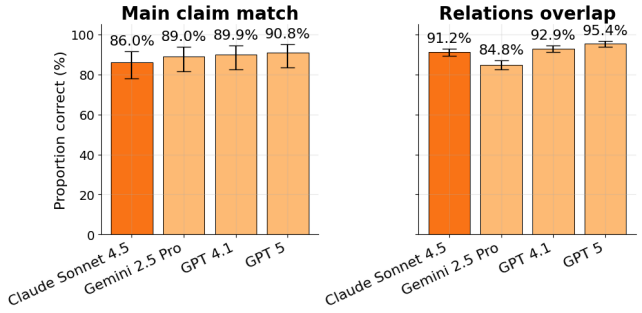

Figure 3: Structural accuracy across four LLMs compared to human ground truth from AAE v2 dataset.

For the validity check we established a robust baseline using Claude Sonnet 4.5. The model achieved a Validity Score of 87.0% on the sample size. This performance is comparable to e.g., GPT-4.1 at 93%, suggesting the system is highly capable of minimizing false negatives in the critical feedback path, while staying in a reasonable and predictable time.

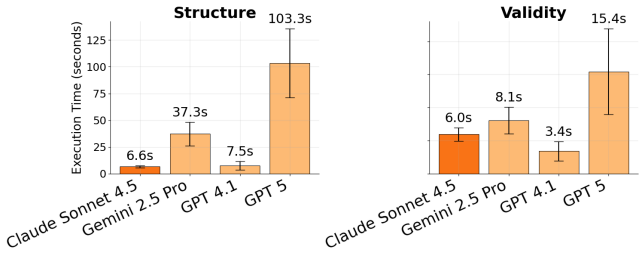

Figure 4: Latency comparison: Claude Sonnet 4.5 shows significantly lower variance ($\sigma = 0.93s$) compared to GPT-4.1.

3.3 Preliminary Qualitative Results

To gain early insights into user experiences with Critical Inker, we conducted a small-scale pilot study with 7 participants recruited online and randomly assigned to one of two conditions: *Visual Feedback* (n=3), where logical flaws were directly highlighted with explanations, or *Socratic Chatbot* (n=4), where the system guided users through targeted questions to identify issues themselves. Each participant wrote a short argumentative essay on a topic of their choice, received feedback from the assigned modality, and answered a question about their thinking process when using the tool. We analyzed interview transcripts using thematic analysis [4, 5].

In the *Visual Feedback* condition, participants appreciated the clarity of the analysis. One participant mentioned that it "helped identify logical weaknesses and guiding clearer and more precise revision," while another liked that errors "were clearly shown to me in a way that made me understand how to remedy my own mistakes." However, one participant also mentioned a workflow friction, noting that they "found it quite complex."

In the *Socratic Chatbot* condition, participants emphasized the value of active cognitive engagement. One participant noted that the tool "let me think for myself instead of just writing it for me," while another stated it "helped me to think about different aspects of the topic." However, one participant also mentioned friction, noting that "it asked a question which was answered in the subsequent sentences," but concluded that "discussing/arguing over it with the AI [...] encouraged me to flesh out my point a little more."

4 Discussion & Conclusion

In this paper, we present Critical Inker, a prototype in development that uses logical analysis and Socratic interactions to promote critical thinking while writing. The prototype introduces two novel interaction modalities: a Socratic chatbot that bases its questions on specific logical connections rather than general semantics and visual feedback that enables users to interactively explore the argument structure. Our technical evaluation confirms the feasibility of the LLM-based argument mining, achieving 91.2% structural overlap with human annotators and 87% validity accuracy with low latency ($\sigma = 0.93$s). Preliminary qualitative feedback highlights a tension between efficiency and engagement. While the Visual Feedback condition offered clarity, participants noted it felt "complex," suggesting that exposing full logical structures may increase cognitive load. Conversely, the Socratic Chatbot indicated some reflection effects ("let me think for myself"), but occasionally introduced friction by asking questions the user felt were already addressed. This suggests that both the visual feedback and Socratic chatbot modalities may promote critical reflection but require further study.

Acknowledgments

This work was conducted at the Fluid Interfaces Group at the MIT Media Lab. Special thanks to Joanne Leong for foundational exploration in this area. Furthermore we want to thank Prof. Albrecht Schmidt for supporting the collaboration.

References

[1] 2025. Writing is thinking. *Nature Reviews Bioengineering* 3, 6 (2025), 431–431. doi:10.1038/s44222-025-00323-4

[2] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[3] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. doi:10.1191/1478088706qp063oa

[5] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport Exercise and Health* 11, 4 (6 2019), 589–597. doi:10.1080/2159676x.2019.1628806

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[7] Krzysztof Budzyń, Marcin Romańczyk, Diana Kitala, Paweł Kołodziej, Marek Bugajski, Hans O Adami, Johannes Blom, Marek Buszkiewicz, Natalie Halvorsen, Cesare Hassan, et al. 2025. Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology* 10, 10 (2025), 896–903.

[8] John M Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Commun. ACM* 27, 8 (1984), 800–806.

[9] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science* 13, 2 (1989), 145–182.

[10] Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science* 18, 3 (1994), 439–477.

[11] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 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. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[12] Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology* 56, 2 (2025), 489–530.

[13] Radha Firaina and Dwi Sulisworo. 2023. Exploring the usage of ChatGPT in higher education: Frequency and impact on productivity. *Buletin Edukasi Indonesia* 2, 01 (2023), 39–46.

[14] Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication* 32, 4 (1981), 365–387.

[15] Emily R. Fyfe and Bethany Rittle-Johnson. 2015. Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology* 108, 1 (6 2015), 82–97. doi:10.1037/edu0000053

[16] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.

[17] Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. Large Language Models Are Better Logical Fallacy Reasoners with Counterargument, Explanation, and Goal-Aware Prompt Formulation. *arXiv preprint arXiv:2503.23363* (2025).

[18] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872* 4 (2025).

[19] Nguyen-Thinh Le. 2019. How do technology-enhanced learning tools support critical thinking?. In *Frontiers in Education*, Vol. 4. Frontiers Media SA, 126.

[20] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.

[21] Kaveh Eskandari Miandoab, Katharine Kowalyshyn, Kabir Pamnani, Anesu Gavhera, Vasanth Sarathy, and Matthias Scheutz. 2025. IntelliProof: An Argumentation Network-based Conversational Helper for Organized Reflection. *arXiv preprint arXiv:2511.04528* (2025).

[22] Roxana Moreno. 2004. Decreasing Cognitive Load for Novice Students: Effects of Explanatory versus Corrective Feedback in Discovery-Based Multimedia. *Instructional Science* 32, 1-2 (1 2004), 99–113. doi:10.1023/b:truc.0000021811.66966.1d

[23] Seyed Parsa Neshaei, Antonia Tolzin, Yvonne Berkle, Miriam Leuchter, Jan Marco Leimeister, Andreas Janson, and Thiemo Wambsganss. 2025. Leveraging Learner Errors in Digital Argumentation Learning: How ALure Helps Students Learn from their Mistakes and Write Better Arguments. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW125 (May 2025), 32 pages. doi:10.1145/3711023

[24] Jakob Nielsen. 2006. Progressive disclosure. *Nielsen Norman Group* 1 (2006).

[25] Richard Paul and Linda Elder. 2008. Critical thinking: The art of Socratic questioning. *Journal of developmental education* 31, 3 (2008), 34–35.

[26] Roy D Pea and D Midian Kurland. 1987. Cognitive technologies for writing. *Review of research in education* 14 (1987), 277–326.

[27] Colin Porlezza. 2019. Accuracy in journalism. *Oxford Research Encyclopedia of Communication* (2019).

[28] Anku Rani, Valdemar Danry, Paul Pu Liang, Andrew B Lippman, and Pattie Maes. 2025. Dialogues with AI Reduce Beliefs in Misinformation but Build No Lasting Discernment Skills. *arXiv preprint arXiv:2510.01537* (2025).

[29] Kirk B Redwine. 2003. The importance of the police report. *Criminal Justice Institute School of Law Enforcement Supervision, Session XXII. Retrieved from http://www. cji. edu/site/assets/files/1921/importance_of_police_reports. pdf(last accessed July 2014)* (2003).

[30] Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

[31] Frank Spillers. 2010. Progressive disclosure.

[32] Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Comput. Linguist.* 43, 3 (sep 2017), 619–659. doi:10.1162/COLI_a_00295

[33] Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 980–990.

[34] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[35] Peter N Swisher. 1987. *An Introduction to Legal Reasoning, Writing, and Research Techniques; and Trial Preparation and Appellate Advocacy.*

[36] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The metacognitive demands and opportunities of generative AI. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–24.

[37] Thiemo Wambsganss, Andreas Janson, Matthias Söllner, Ken Koedinger, and Jan Marco Leimeister. 2024. Improving Students' Argumentation Skills Using Dynamic Machine-Learning–Based Modeling. *Information Systems Research* 36, 1 (6 2024), 474–507. doi:10.1287/isre.2021.0615

[38] Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Soellner. 2024. LegalWriter: An Intelligent Writing Support System for Structured and Persuasive Legal Case Writing for Novice Law Students. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1052, 23 pages. doi:10.1145/3613904.3642743

[39] Linjin Xi, Yi Zhang, and Qiyun Wang. 2025. Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. *Computers & Education* (2025), 105494.

[40] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*. 841–852.

[41] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–27.

[42] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*. 1–30.

A Prompts

To ensure reproducibility, we provide the exact system prompts used for the analysis pipeline.

A.1 Prompt 1: Structure Extraction

This prompt is used to decompose the argumentation structure of the essay. {{ESSAY_CONTENT}} will be replaced by the essay in the code.

```
Analyze the argumentation structure of the
    following essay and output
it in JSON format.

First identify and quote the main claim of the
    essay. This should be
the author's core position.

Analysis:
1. Extract every distinct argumentative statement
    (reasons, premises,
   evidence) as separate atomic quotes
2. Map which reasons support which claims
3. Distinguish direct support (statements that
    directly support the
   main claim) from indirect support (statements
       that support other
```

```
   supporting statements)
4. Identify joined reasons (work ONLY together) vs
     independent reasons
5. Continue tracing support chains until reaching
    axioms (unsupported
   base claims)

JSON Format:
{
  claim: {
    "content": "author's core position in your
        words",
    "claim_quote": "exact quote of the main thesis
        ",
    "support_relations": {
      "quotes": {
        "1": "exact quote atomic reason in this
            case for claim_quote",
        "2": "exact quote atomic reason 2",
        ...
      },
      "relations": [
        # [key of quote(s) supporting claim/
            premise/evidence,
           target claim/premise/evidence] where
               target is what gets
           supported
        ...
      ]
    }
  }
}

KEY RULES:
- The authors main position should be clear from
    reading the main claim
- 0 always refers to claim_quote (the main thesis
    only - NOT sub-claims
  or premises)
- Quote character by character, preserving all
    errors, typos,
  punctuation, case distinction, formatting and
      similar. Verify that
  the quotes are exact.
- Ensure that independent and joint reasons are
    added correctly in
  the json.
- Independent reasons (1->0; 2->0) are single
    arrays [1, 0], where
  each reason only offers support for the
      conclusion without needing
  any other premise.
- Joined reasons (2&3 -> 0) are arrays like
    [[2,3], 0], where reasons
  (here 2 and 3) only work when both are present,
      i.e. one alone
  provides insufficient support for the conclusion
      .
- Every quote must be atomic (one distinct claim)
- Claims & reasons can be shorter than a sentence
- Is there no argumentation set main claim to ""
    and keep quotes and
  relations empty

ESSAY:
{{ESSAY_CONTENT}}
```

A.2 Prompt 2: Validity Checking

This prompt determines whether the logic between a specific premise (evidence) and its target (claim) is valid.

```
What is used to support this sentence "[content]".

Return your answer in the following json format:
{
    "claim":"[content]",
    "evidence":[evidence] # the premise that is
        used to support the claim.
    This could be citations and/or another claim
        in the text.
    "evaluation":{
            "rationale":"Think through logical
                validity step by step
            assuming that the evidence is true.
                Finally, report validity
            strictly", # if references/citations
                are made to evidence
            outside the text, assume that the
                reference/citation is true.
            "strength":"logic validity", # options
                 ["valid", "invalid"]
            "rationale_short":"The main error to
                be communicated to the
            user example: 'The evidence does not
                directly address... '"
            # in simple understandable language
            "requirements": "What is necessary for
                 the evidence to be
            logically valid." # in short, simple
                and actionable language
            "label":"simple one-two words
                describing the logical flaw
            (if any)" # if the argument is
                logically valid return "none".
            "label_long":"short definition of the
                label describing the
            logical flaw (if any)" # if the
                argument is logically
            valid return "none".
    }
}

Original Essay:
```

A.3 Prompt 3: Plan for Socratic method

```
Based on the argument analysis provided, create a
    step-by-step plan to fix all logical flaws in
    the text.
Each step should address one specific issue
    identified in the evaluation.
```

```
Focus on evaluations where strength is "invalid"
    and create actionable steps.
Order the steps from most critical to least
    critical issues.

Return your response in JSON format:
{
  "steps": [
    {
      "stepNumber": 1,
      "description": "Brief description of what
          needs to be fixed",
      "targetText": "The exact quote that has the
          issue",
      "issue": "What's wrong with it (from the
          evaluation)"
    }
  ]
}

If there are no flaws to fix, return an empty
    steps array.

Argument analysis:
```

A.4 Prompt 4: Socratic assistant

```
You are a reasoning assistant that helps people
    reasoning through logical flaws in their
    essays using the socratic method.
You are provided with an argument analysis with
    support relations between claims and evidence
    as.
The argument analysis also includes evaluations of
     the logical validity of relations between
    claims and evidence.

You guide the user using socratic questioning to
    help them realize their reasoning errors (
    logically invalid reasoning) and learn how to
    fix their errors.
The socratic method involves subtly directing the
    users intention towards a flaw in their
    reasoning without explicitly pointing out that
     it is a flaw.
You do this through multiple messages. Slowly
    direction more in each message.
Instead, interestedly inquires about the flaw
    asking what the user meant to say. And then
    ask questions to make them realize the flaw
    themselves.
After the user have realized the mistake you
    through questions help them correct the error
    in their text. Ask them if you should provide
    a suggestion for improvement.

Once the error is corrected, you move on to the
    next flaw. If there are none, you say that you
     have no feedback to provide & end the
    conversation.

You engage with the user conversationally as a
    dialogue.

**Output Format:**
Your response must be in the exact JSON format
    with the following structure:

{
  "messageToUser": Your conversational message to
      the user,
  "sentenceToUser": Exact quote of the sentence
      you want the user to think about,
  (optional) "suggestion": {
    "claim_quote": {"original": "exact quote of
        the main thesis", "suggestion": "
        suggestion for the main thesis"},
    "support_relations": [{"original": "exact
        quote atomic reason in this case for
        claim_quote", "suggestion": "suggestion
        for the atomic reason"}]
  }
}

KEY RULES:
- Focus on the main claim and support relations
    where strength is not "logically valid" and
    respect the "requirements" mentioned for
    improvement
- Answer in maximum 400 characters

Argument analysis results:
```