Building Self-Reliance in an Era of Accessible Intelligent Tutors

Dinesh Ayyappan[*]
Georgia Institute of Technology
Atlanta, Georgia, USA
dayyappan3@gatech.edu

David A. Joyner
Georgia Institute of Technology
Atlanta, Georgia, USA
david.joyner@gatech.edu

Abstract

As Generative AI tools become increasingly capable of automating cognitively demanding tasks, it is critical to design tools that augment—rather than replace—human thinking. To this end, we propose that tools for thought should make explicit the roles and levels of intervention in human-AI partnerships, especially in educational contexts. Through an exploratory study of 71 high school students using an AI writing tutor with explicitly-labeled modes of assistance, we investigate how structured interfaces influence learners' interactions with intelligent tutors. Our findings suggest that interface design may affect different aspects of self-regulation and student perceptions of AI assistance. We discuss implications for designing educational AI systems that support metacognitive development while preserving student agency. This work contributes to understanding how to balance AI support with the development of independent thinking skills in educational contexts.

CCS Concepts

• **Computing methodologies → Intelligent agents**.

Keywords

Tools for thought, educational technology, human-AI collaboration, metacognition, generative AI

ACM Reference Format:
Dinesh Ayyappan and David A. Joyner. 2025. Building Self-Reliance in an Era of Accessible Intelligent Tutors. In *Proceedings of Workshop on Tools for Thought at CHI 2025 (Tools for Thought at CHI 2025)*. ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Educational technology has often presented both challenges and opportunities for teachers and students. From calculators that transformed mathematics instruction to machine translation tools that revolutionized language learning, each technological advance has required thoughtful integration into educational practice [6]. Large language models (LLMs) have become a new chapter in this arc that, unlike previous technologies, can generate entire essays, revise text, and support writers of different skill levels [16]. These new capabilities challenge teachers of writing, because they muddle notions of authorship, risk diminishing student agency and self-reliance [5], and complicate detection of ineffective collaborations or plagiarism [6, 8, 10, 20].

As Danielle Allen reminds us, "life moves faster than science" [1]. Teachers have had to make decisions about integrating LLMs before science could provide guidance, so it is both urgent and important for the research community to develop appropriate design patterns and effective practices for using LLMs and other Artificial Intelligence (AI) tools in educational contexts.

Meanwhile, students must also learn to leverage AI assistance while developing their own capabilities [12, 14]. However, current AI interfaces often blur the boundaries between human and AI roles, making it difficult for students to understand when and how to appropriately rely on AI support [2]. Recent work in human-AI collaboration suggests that role clarity and explicit intervention levels are crucial for productive partnerships [24].

This paper explores an approach where a novel AI tool makes explicit its role and mode of intervention, which allows students to make informed choices about AI assistance. This tool, which was co-designed with writing teachers for its specific learning context, bridges recent work in human-AI collaboration [9] and educational scaffolding [22] to ask this research question: *How can structured interfaces influence learners' interactions with intelligent tutors in an open-ended, analytical writing task?*

2 Related Work

Recent work in human-AI collaboration has highlighted the importance of clear role definitions in the design space[11]. In an experiment allowing users to co-write argumentative essays with LLMs, Padmakumar et al. [19] found that writing with language models can reduce content diversity between different authors. More concerning, in an experiment where participants wrote blog posts with AI assistants pre-configured to two different opinions, Jakesch et al. [13] found that users' views and writing patterns were significantly affected without their awareness. These experiments surface some dangers of allowing LLMs to directly modify or influence users' writing.

The challenge of maintaining appropriate boundaries around authorship becomes particularly acute in educational settings where developing students' voice and thinking skills are critical. In a survey at an American university, Barrett et al. [2] found significant disagreement between students and teachers about appropriate use of AI in writing tasks. This reflects an urgent need for clearer frameworks around AI assistance. Structured writing support is an active research area which includes tools like VISAR [25], a prototyping system for the iterative development of argumentative

[*]Corresponding author

Tools for Thought at CHI 2025, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

writing. However, this is a tool focused on developing and revising a text rather than teaching the underlying cognitive skills of argumentation.

In teaching and learning contexts, there is also a tension between AI assistance and student autonomy. In an AI tutoring experiment with nearly 1000 high school math students, Bastani et al. [4] found that students with the AI tutor performed better during practice than students without the AI tutor. However, when the AI tutor was removed for an exam, those students performed no better than the students who had no AI tutor at all. Students who had access to a base (non-tutor) LLM actually performed worse on the exam than the students with no AI tutor. Since the wrong kind of AI assistance can have negative effects on learning, we must carefully ensure that AI tutoring and assistance in learning contexts preserve student agency [7] and strengthen, rather than diminish, critical thinking skills [16].

In computing education, Kazemitabaar et al. [15] demonstrated one approach that used carefully designed guardrails to prevent over-reliance while maintaining helpful support. Reviewing AI writing assistance from the perspective of second-language acquisition, Ingley et al. [12] suggested that writers using AI tools will learn only when there is a small enough gap between their language skills and AI-generated feedback and not when AI tools are simply 'fixing' their flawed writing. This suggests that writing assistants should provide personalized feedback that supports the development of individual student capabilities. Personalization can have other benefits as well. Recent work by Yeh et al. [24] on Ghostwriter, an AI-powered writing assistant that allows users to specify style and context, showed that agency-supporting features can improve engagement with AI writing tools.

While there has been a lot of research on AI writing assistants, there is a natural divide between productivity-focused applications where friction in the user experience is generally negative and educational ones where friction is sometimes the heart of the matter and where the deepest learning happens. The design of human-AI interactions for writing is a uniquely difficult and complex space that requires holistic consideration of dimensions like task, user, and ecosystem [17, 18, 23].

Tools that focus on fostering agency and perceptions of personalization would be most successful with users who have developed their own style and preferences, but students are often still developing their sense of self as writers when they are thrown into this AI writing ecosystem. So, we ought to have different expectations and supports for student users than for professionals.

These dimensions intersect richly in the secondary school writing classroom where students have already reached language proficiency but are still developing their voice and the analytical skills required for university-level and professional writing. In this research, we work with teachers to design a task-specific, AI-powered writing assistant that helps us investigate how we can support the development of students' self-regulation and metacognitive skills when working with intelligent tutors.

3 Exploratory Study

To investigate how role clarity in a structured interface could affect learning behaviors and and self-reliance, we conducted an IRB-approved study with two teachers in an advanced English course at an international school in Singapore. This course was chosen due to its focus on rhetorical analysis skills, which are widely taught and have a stable, well-established curriculum in this context. The specific unit of teaching included the rhetorical analysis of speeches, so students were already acquainted with the frameworks and concepts for their analysis. With the goal of making this research more likely to be transferable and impactful, we sought an additional skill that would be more agnostic to content and may apply across a wider range of ages and disciplines. Keeping the scope to collaborative human-AI relationships, the educators identified that students sometimes rely on LLMs to "do the thinking work" for them, which can be detrimental to their own learning. We brainstormed ways to nudge students to be more self-reliant when performing cognitively demanding tasks, and so we designed the following experience.

3.1 Interaction Design

After working with a teacher to iterate on the tutor, the final flow of the interaction was as follows:

- **Student experience (Figure 1):** Students selected from several texts, chosen by their teachers. They started with identifying the Speaker, but they could move freely between that and the other elements of the mnemonic for rhetorical situation: SOAPSTONE (Occasion, Audience, Purpose, Subject, and Tone). At any time, students could interact with the AI Tutor, and at each interaction, the LLM gave rubric-aligned feedback on the student's current identification and illuminated the letter with a coded circle (red, yellow, or green). After 5 non-reds, the interfaces changed to allow longer-form analysis instead of just identification, but the conversational interaction with the AI tutor remained the same.
- **Backend:** Student interactions were handled through an Express router that classified inputs and generated appropriate responses. The system used Anthropic's Haiku for fast classification and Sonnet for deeper feedback. When students submitted identifications of SOAPSTONE elements, the backend evaluated their answers with a three-step process: 1) classified the input into on- or off-topic; 2) determined correctness (Yes/Almost/Not yet); and 3) provided tailored feedback. The system included rich few-shot examples for each element to guide responses. In full analysis mode, it evaluated students' rhetorical analysis essays against a rubric framework. The backend maintained conversation history to provide context-aware responses and included redirect strategies for off-topic inquiries. All of this was passed to a React frontend which updated visuals and communicated the AI tutor's feedback.

To use this framework for an experiment, we compared two interfaces: A role-explicit interface (Treatment) and a traditional LLM chat interface (Control). The Treatment condition provided explicit indicators of the level of thinking support being provided by the language model, with options *Tell me the answer* (for direct assistance)

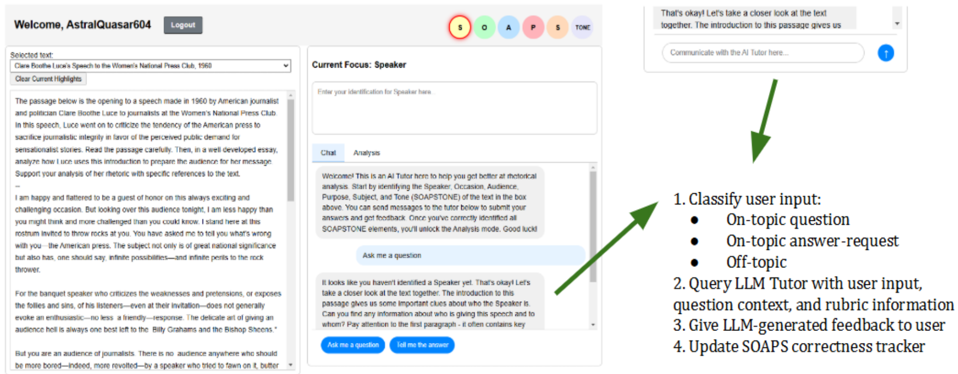

Figure 1: The task-specific AI tutor interface and experiment. On the left is a full screenshot from the Treatment interface; in the top-right is the Control interface.

and *Ask me a question* (for Socratic-style questioning to elicit student thinking). The less structured (Control) LLM interface allowed students to interact with the LLM with a traditional, open-ended chat box without explicit guidance on the level of support.

Both interfaces had the same underlying prompts and software infrastructure, and the interface allowed the selection and highlighting of text elements to support rhetorical analysis. One class section from each teacher was randomly assigned to the treatment condition, while the other section was assigned to the control condition. This balanced design helped to mitigate potential teacher effects, and using the same underlying prompts minimized differences in educational assistance between conditions.

3.2 Study Design

Out of 71 high school students ages 15-18 enrolled in four classes taught by two teachers, we obtained full consent (including parents) from 33. Of those students, 18 used he/him pronouns, and 14 used she/her pronouns; 31 self-identified as at or above grade-level expectations in reading and writing in English. When asked about their use of AI chatbots, 28 used them at least once per week, with 4 of those reporting daily use. Students completed pre- and post-surveys measuring self-regulation [21] and perceptions of AI assistance [3]. Students' interactions with the LLM, including the frequency and nature of their queries were logged, and teachers were interviewed for their reflections.

4 Preliminary Findings

Our analysis revealed several interesting, though limited, patterns that suggest directions for future research.

4.1 Self-Regulation Indicators

Analysis of pre- and post-survey responses (Table S1, Appendix) revealed no convincing difference between conditions on 10 of the 11 questions given to students.For the self-monitoring statement "While doing a task, I ask myself how well I am doing" (Q7), the control group showed significant improvement ($p < 0.05$), while the treatment group was unchanged.

4.2 Perceptions of the AI Tutor

Examining students' perceptions of the AI tutor revealed minor differences between conditions and some variation with the amount of student use, as not all students used the tutor equally. (Figure 2).

In the control group, there were significant negative correlations between the number of LLM interactions and ratings of the AI tutor's competence ($r = -0.31$, $p < 0.05$) and intelligence ($r = -0.28$, $p < 0.05$). This pattern did not appear among users of the structured, role-explicit (treatment) interface, where perceptions remained stable with regards to number of LLM interactions.

However, students did perceive the control interface tutor as more kind than the tutor in the treatment interface (Table 1), even though the underlying tutor had the same functionality, model, and prompts.

4.3 Teacher Perceptions

In structured interviews, the participating teachers (T1, T2) shared their observations and reflections on how students used the AI tutor. Both teachers noticed that students in the control group (traditional, open-ended interface) were quicker to get started. Students using the treatment interface asked questions like "How do I interact with these buttons?" (T2), suggesting a better tutorial process could have reduced starting friction.

Once students got going though, T1 observed that many students were "in the zone" and found it "really helpful being able to get such instant feedback". T2 emphasized the value of feedback as well, noting that their students "love feedback", wanting "constant feedback" and "more feedback". T1 highlighted that the quick, specific, and personal feedback made it "more obvious that [writing] was an iterative process of fixing, tweaking, trying again." This experience is different from the traditional interaction with teacher feedback, which usually comes at a time when students are not actively rewriting their work.

5 Discussion

Our exploratory study provides initial insights into how interface design might influence students' interactions with AI tutors, though our findings are limited in scope and statistical significance.

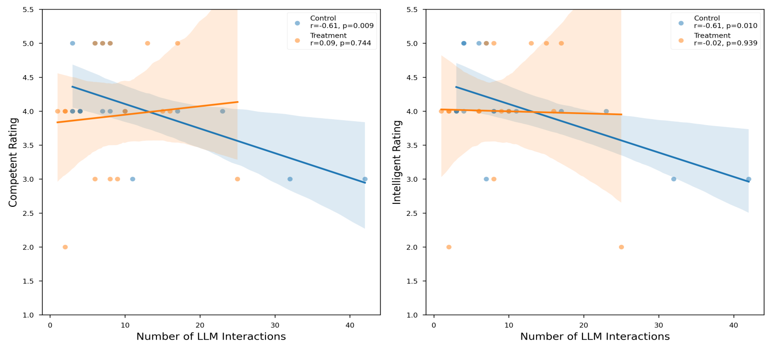

Figure 2: In the control group, increased LLM interactions correlated with lower ratings of the AI tutor's competence

Metric	Control Mean (SD)	Treatment Mean (SD)	Difference	p-value
easy to use	4.06 (0.97)	4.00 (0.89)	-0.06	0.857
like	4.06 (0.66)	3.81 (0.75)	-0.25	0.323
competent	4.06 (0.66)	3.94 (0.93)	-0.12	0.667
friendly	4.59 (0.51)	4.06 (1.00)	-0.53	0.063
knowledgeable	4.00 (0.71)	4.25 (0.77)	+0.25	0.340
intelligent	4.06 (0.66)	4.00 (0.97)	-0.06	0.839
kind	4.65 (0.49)	3.88 (1.02)	-0.77	**0.009**
sensible	3.94 (0.75)	4.06 (0.85)	+0.12	0.667
nice	4.41 (0.62)	4.00 (0.89)	-0.41	0.132

Table 1: Comparison of AI Perception Ratings Between Groups

The different patterns observed between groups— the treatment group's slight positive trend in persistence and the control group's improvement in self-monitoring—suggest that interface design may influence different aspects of learners' interactions with AI tutors. While these findings align with previous work on writer autonomy [5], the limited statistical significance ($p < 0.05$ only for self-monitoring in the control group) indicates a need for further investigation rather than definitive conclusions.

The significant difference in perceived "kindness" between interfaces, despite identical underlying LLM prompts regarding tone and response format, reveals an important trade-off that mirrors classroom experience: helping students stay 'on task' may conflict with helping them feel comfortable.

Towards shaping student self-regulation, our findings suggest that while interfaces can make AI capabilities transparent, supporting students in making informed choices about AI use remains challenging. This includes balancing opportunities for productive struggle with appropriate AI assistance.

Interface design for educational AI tools should consider how to clearly signal intervention levels without being rigid, support metacognitive decision-making without forcing it, and preserve student voice throughout the process. These goals provide some focus towards our research question about how structured interfaces can influence learners' interactions with intelligent tutors.

When asked about the future of AI-assisted tools in the classroom, teachers brought attention to the relational side of teaching, noting that totally outsourcing the feedback process could make it harder for them to "know [students] as people and learners" (T2) and that "Sometimes kids don't get things unless you're sitting next to them pointing to something" (T1). T2 noted that the use of automated feedback also raises equity concerns among a department of teachers, requiring norm-setting around the use of such tools. Looking forward, T2 hoped that teachers could help shape and prioritize what feedback to give students who might have many areas for improvement to avoid overwhelming them.

5.1 Limitations

Our exploratory work has several important limitations. With a small sample size and limited statistical significance in most measures, our findings should be viewed as preliminary rather than conclusive. We lacked first-hand interviews with students that could lend credibility to our interpretations of their behavior. Additionally, the study's short duration makes it difficult to assess meaningful changes in metacognitive development or agency, which typically emerge over longer periods.

The control group's improved self-monitoring could be particular to this population or task, suggesting that the effects of interface design may vary across different contexts. These limitations point to the need for more robust research designs that can better capture the complex relationship between interface design and student agency.

6 Future Directions

Future research should develop more sophisticated frameworks for understanding role clarity in educational AI, including better mapping of assistance levels to learning objectives. Longer-term studies with larger, more diverse student populations could help validate and refine our preliminary findings across different subject domains and types of learning tasks. Most importantly, future work should develop more robust measures of actual student behavior, learning outcomes, and metacognitive development to better answer our central question: how structured interfaces influence learners' interactions with intelligent tutors in ways that support rather than diminish human agency and capability.

Acknowledgments

Special thanks to Nick Lytle for feedback and testing, teachers Leaf and Alice for their collaboration and participation, and the host school for the opportunity.

References

[1] Danielle Allen. 2018. Foreword. In *The Human Condition* (2nd ed.). University of Chicago Press, xvi.

[2] Alex Barrett and Austin Pack. 2023. Not quite eye to A.I.: student and teacher perspectives on the use of generative artificial intelligence in the writing process. *null* (2023). doi:10.1186/s41239-023-00427-0

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (Jan. 2009), 71–81. doi:10.1007/s12369-008-0001-3

[4] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2024. Generative AI Can Harm Learning. (2024). doi:10.2139/ssrn.4895486

[5] Oloff C. Biermann, Oloff C. Biermann, F. Ning, Ning F. Ma, Dongwook Yoon, and Dongwook Yoon. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. *Conference on Designing Interactive Systems* (2022). doi:10.1145/3532106.3533506

[6] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. (2022). arXiv:2108.07258 [cs.LG] https://arxiv.org/abs/2108.07258

[7] Beatriz Borges, Niket Tandon, Tanja Käser, and Antoine Bosselut. 2023. Let Me Teach You: Pedagogical Foundations of Feedback for Language Models. *arXiv.org* (2023). doi:10.48550/arxiv.2307.00279

[8] Binglin Chen, Colleen M Lewis, Matthew West, and Craig Zilles. 2024. Plagiarism in the age of generative ai: cheating method change and learning loss in an intro to CS course. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*. 75–85.

[9] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. 2023. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. *International Conference on Human Factors in Computing Systems* (2023). doi:10.1145/3544548.3580969

[10] Alina Deriyeva, Jesper Dannath, and Benjamin Paaßen. 2024. Case study: Using LLMs to assist with solving programming homework assignments. In *Proceedings of DELFI Workshops 2024*. Gesellschaft für Informatik eV, 10–18420.

[11] Steffen Holter and Mennatallah El-Assady. 2024. Deconstructing Human-AI Collaboration: Agency, Interaction, and Adaptation. *Computer Graphics Forum* 43, 3 (2024). doi:10.1111/cgf.15107

[12] Spencer J. Ingley and Austin Pack. 2023. Leveraging AI tools to develop the writer rather than the writing. *null* (2023). doi:10.1016/j.tree.2023.05.007

[13] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and . 2023. Co-Writing with Opinionated Language Models Affects Users' Views. *International Conference on Human Factors in Computing Systems* (2023). doi:10.1145/3544548.3581196

[14] David Joyner, Zoey Anne Beda, Michael Cohen, Melanie Duffin, Amy Garcia Fernandez, Liz Hayes-Golding, Jonathan Hildreth, Alex Houk, Rebecca Johnson, Kayla Matchek, et al. 2024. When Chatting Isn't Cheating: Mining and Evaluating Student Use of Chatbots and Other Resources During Open-Internet Exams. In *Proceedings of the 17th International Conference on Educational Data Mining*. 143–156.

[15] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Z. Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs. *arXiv.org* (2024). doi:10.1145/3613904.3642773

[16] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*. ACM. https://www.microsoft.com/en-us/research/publication/the-impact-of-generative-ai-on-critical-thinking-self-reported-reductions-in-cognitive-effort-and-confidence-effects-from-a-survey-of-knowledge-workers/

[17] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, S. B. Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md. Naimul Hoque, Yewon Kim, Seyed Parsa Neshaei, Agnia Sergeyuk, A. Shibani, Disha Shrivastava, Lila Shroff, Jessi Stark, S. Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia H. Rho, Shannon Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. *arXiv.org* (2024). doi:10.1145/3613904.3642697

[18] Mina Lee, Mina Lee, Percy Liang, Percy Liang, Qian Yang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset forExploring Language Model Capabilities. *null* (2022). doi:10.1145/3491102.3502030

[19] Vishakh Padmakumar and He He. 2023. Does Writing with Language Models Reduce Content Diversity? *arXiv.org* (2023). doi:10.48550/arxiv.2309.05196

[20] Diana M Popescu and David A Joyner. 2024. Novelty, Rigidity, and Complexity: Toward Developing AI-Resistant Assessments in an Introductory Computer Science Class. In *2024 IEEE Digital Education and MOOCS Conference (DEMOcon)*. IEEE, 1–6.

[21] Laura Jonker Marieke J.G. van Heuvelen Tynke Toering, Marije T. Elferink-Gemser and Chris Visscher. 2012. Measuring self-regulation in a learning context: Reliability and validity of the Self-Regulation of Learning Self-Report Scale (SRL-SRS). *International Journal of Sport and Exercise Psychology* 10, 1 (2012), 24–38. doi:10.1080/1612197X.2012.645132 arXiv:https://doi.org/10.1080/1612197X.2012.645132

[22] Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes. *null* (2023). doi:null

[23] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376301

[24] Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. GhostWriter: Augmenting Collaborative Human-AI Writing Experiences Through Personalization and Agency. *arXiv.org* (2024). doi:10.48550/arxiv.2402.08855

[25] Qing Zhang and Anjali Ragupathi. 2023. Argue Better: Using Large Language Models to Generate Better Examples for Ineffective Persuasive Essay Arguments. *null* (2023). doi:null

A Supplementary Material

Question	Group	Pre (SD)	Post (SD)	Change	p-value
Q1: I can solve most problems if I invest the necessary effort.	Control	3.47 (0.51)	3.18 (0.53)	-0.29	0.096
Q1: I can solve most problems if I invest the necessary effort.	Treatment	3.62 (0.50)	3.44 (0.51)	-0.19	0.270
Q2: I check how well I am doing when I solve a task	Control	3.24 (0.66)	3.35 (0.70)	0.12	0.496
Q2: I check how well I am doing when I solve a task	Treatment	3.25 (0.68)	3.12 (0.50)	-0.12	0.432
Q3: I work as hard as possible on all tasks.	Control	3.00 (0.71)	3.06 (0.24)	0.06	0.718
Q3: I work as hard as possible on all tasks.	Treatment	3.25 (0.77)	3.06 (0.68)	-0.19	0.270
Q4: I put forth my best effort when performing tasks.	Control	3.18 (0.64)	3.00 (0.35)	-0.18	0.269
Q4: I put forth my best effort when performing tasks.	Treatment	3.38 (0.62)	3.25 (0.58)	-0.12	0.164
Q5: I always manage to solve difficult problems if I try hard enough.	Control	3.06 (0.66)	2.94 (0.75)	-0.12	0.431
Q5: I always manage to solve difficult problems if I try hard enough.	Treatment	3.31 (0.70)	3.38 (0.72)	0.06	0.751
Q6: I don't give up even if the task is hard.	Control	3.00 (0.61)	2.71 (0.69)	-0.29	0.096
Q6: I don't give up even if the task is hard.	Treatment	3.25 (0.58)	3.50 (0.52)	0.25	0.164
Q7: While doing a task, I ask myself, how well I am doing.	Control	2.47 (1.01)	3.18 (0.81)	0.71	**0.009**
Q7: While doing a task, I ask myself, how well I am doing.	Treatment	3.00 (0.73)	3.12 (0.62)	0.12	0.580
Q8: I work hard to do well even if I don't like a task.	Control	2.94 (0.83)	2.94 (0.66)	0.00	1.000
Q8: I work hard to do well even if I don't like a task.	Treatment	2.88 (0.72)	2.88 (0.81)	0.00	1.000
Q9: I work hard on a task even if it is not important.	Control	2.18 (0.81)	2.47 (0.87)	0.29	0.172
Q9: I work hard on a task even if it is not important.	Treatment	2.50 (0.82)	2.50 (0.73)	0.00	1.000
Q10: It is easy for me to concentrate on my goals and to accomplish them.	Control	2.82 (0.53)	3.12 (0.33)	0.29	0.096
Q10: It is easy for me to concentrate on my goals and to accomplish them.	Treatment	3.31 (0.70)	3.38 (0.62)	0.06	0.806
Q11: If I persist on a task, I'll eventually succeed.	Control	3.41 (0.62)	3.41 (0.62)	0.00	1.000
Q11: If I persist on a task, I'll eventually succeed.	Treatment	3.62 (0.50)	3.50 (0.52)	-0.12	0.333

Table S1: Pre/Post Survey Results for All Questions