Comparing Expert and ChatGPT-authored Guidance Prompts

Allison Bradford†
Berkeley School of Education
University of California, Berkeley
Berkeley, CA, USA
allison_bradford@berkeley.edu

Weiying Li
Berkeley School of Education
University of California, Berkeley
Berkeley, CA, USA
weiyingli@berkeley.edu

Libby Gerard
Berkeley School of Education
University of California, Berkeley
Berkeley, CA, USA
libbygerard@berkeley.edu

Marcia C. Linn
Berkeley School of Education
University of California, Berkeley
Berkeley, CA, USA
mclinn@berkeley.edu

ABSTRACT

Students bring a multitude of ideas and experiences to the classroom while they are reasoning about scientific phenomena. They often need timely guidance to refine build upon their initial ideas. In this study we explore the development of guidance prompts to provide students with personalized, real-time feedback in the context of a pedagogically grounded chatbot. In the current version of the tool, guidance prompts are authored by learning scientists who are experts in the content of the items and in Knowledge Integration pedagogy. When students engage with the chatbot, an idea detection model is used to determine the ideas that are present in a student explanation and then the expert-authored guidance prompts are assigned based on rules about which ideas are or are not present in the student explanation. While this approach allows for close attention to and control of the pedagogical intent of each prompt, it is time consuming and not easily generalizable. Further this rule-based approach limits the ways in which students can interact with the chatbot. The work in progress study presented in this paper explores the potential of using generative AI to create similarly pedagogically grounded guidance prompts as a first step towards increasing the generalizability and scalability of this approach. Specifically, we ask: using criteria from the Knowledge Integration Pedagogical Framework, how do ChatGPT 3.5-authored guidance prompts compare to human expert-authored guidance prompts? We find that while prompt engineering can enhance the alignment of ChatGPT-authored guidance prompts with pedagogical criteria, the human expert-authored guidance prompts more consistently meet the pedagogical criteria.

CCS CONCEPTS

• Applied computing – Education – Computer-assisted Instruction

L@S '24, July 18–20, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-0633-2/24/07.
https://doi.org/10.1145/3657604.3664669

KEYWORDS

Automated Guidance, Generative AI, Knowledge Integration Pedagogy

ACM REFERENCE FORMAT:

Allison Bradford, Weiying Li, Libby Gerard and Marcia C. Linn. 2024. Comparing Expert and ChatGPT-authored Guidance Prompts. In *Proceedings of Eleventh ACM Learning@Scale conference (L@S'24). July 18–20, 2024, Atlanta, GA, USA. ACM NY, NY, USA, 5 pages.* https://doi.org/10.1145/3657604.3664669

1 INTRODUCTION

Students bring a multitude of ideas and experiences to the classroom while they are reasoning about scientific phenomena [2, 10]. These ideas and experiences are powerful resources for learning, particularly when teachers leverage those ideas as starting points for constructing deeper understanding. As student begin learning about new science phenomena, their repertoire of ideas is a blend of vague, descriptive, and mechanistic ideas [2]. Research shows that when students are given guidance that supports them to build on their ideas and distinguish between them and new ideas raised during instruction, students can develop more coherent understanding [4, 8]. Conversely, when instruction fails to address the ideas students hold while introducing new concepts, students can develop a fragmented understanding rather than distinguishing which ideas hold explanatory power and constructing connections across their repertoire of ideas [2].

While the importance of timely guidance to support students to build upon and distinguish among their ideas is well documented, teachers are not always able to provide such feedback given the constraints of the classroom, including large class sizes and many standards to cover. Advanced natural language processing (NLP) techniques have the potential to alleviate the strain on teachers and increase the likelihood that all students receive personalized guidance to help them develop coherent understanding. For example, recent approaches enable the detection of individual ideas within each student's written science explanations [16] which has made it possible to provide students with adaptive guidance that

responds to the ideas in their explanations. Recent work has shown that idea-based guidance, when aligned to a pedagogical framework, can support students to raise additional ideas and write more coherent explanations when embedded in a curriculum-based chatbot [1, 3, 9]. Often, this pedagogically grounded approach to idea-based guidance relies heavily on expert design of guidance and a rule-based approach to assign prompts to student explanations which limits the generalizability and scalability of the approach. The work in progress study presented in this paper explores the potential of using generative AI to create similarly pedagogically grounded guidance prompts as a first step towards increasing the generalizability and scalability of this approach. Specifically, we ask: using criteria from the Knowledge Integration Pedagogical Framework, how do ChatGPT 3.5-authored guidance prompts compare to human expert-authored guidance prompts?

2 RELATED WORK

2.1 Automated Guidance

A common use of artificial intelligence (AI) in educational settings is to provide automated feedback and guidance to students on products of their learning like short essays [3, 4, 15] or scientific models [18] AI has also been used to individualize student learning experiences in intelligent tutoring systems [6, 7 ,13] and educational chatbots that provide direct guidance or hints to students as they engage in learning tasks in many domains [1, 14]. The impact of chatbots on learning outcomes is not yet clear [17]. Wollny et al. suggested that the lack of impact might reflect that "chatbot development in education is still driven by technology, rather than having a clear pedagogical focus of improving and supporting learning" [17, p.13]. A recent study comparing the impact on learning gains of hints generated by ChatGPT and hints generated by human tutors found that both sets of hints produced positive learning gains, though only the gains from human tutor-authored hints were statistically significant [12]. The findings illustrate the promise of leveraging ChatGPT for authoring guidance prompts. The present student seeks to incorporate a pedagogical lens both in an effort to enhance the guidance prompts authored by ChatGPT and to establish criteria by which to evaluate guidance prompts.

2.2 Knowledge Integration Guidance

This study draws on the Knowledge Integration (KI) perspective on learning and corresponding pedagogical framework [11]. KI recognizes that learners hold varied ideas that reflect their lived experiences. It advocates that supporting students to develop coherent understanding involves eliciting students' prior ideas, providing opportunities to discover new ideas, using evidence to distinguish among existing ideas and new ideas, and guiding students to make connections among their ideas to form an explanation or argument [11]. The KI perspective holds that students benefit from opportunities to build from their initial ideas. Students then can distinguish whether their ideas hold explanatory power. They can also connect their initial ideas to evidence. Leveraging the KI framework to design guidance prompts has resulted in students making learning gains they respond to that guidance, both in the context of guidance assigned based on a wholistic score of student explanations [4, 5] and of idea-based guidance [1, 3, 9].

3 METHODS

In this study we explore the development of guidance prompts to provide students with personalized, real-time feedback in the context of a pedagogically grounded chatbot. The aim of the guidance prompts is to support students to build on their initial ideas by eliciting more of their thinking before pushing them to connect their ideas in a revised explanation. In the current version of the tool, guidance prompts are authored by learning scientists who are experts in the content of the items and have been involved in the design of the curriculum in which the items are situated. We refer to these expert-authored guidance prompts at KI prompts. When students engage with the chatbot an idea detection model is used to determine the ideas that are present in a student explanation and then the KI prompts are assigned based on rules about which ideas are or are not present in the student explanation. While this approach allows for close attention to and control of the pedagogical intent of each prompt and leverages expert knowledge of how students progress in their ability to explain the phenomenon in the item, it is time consuming and not easily generalizable. Further this rule-based approach limits the ways in which students can interact with the chatbot. This study seeks to compare the nature of prompts generated by experts to prompts generated by ChatGPT 3.5 as an initial step towards improving the efficiency and flexibility of developing such chatbots.

3.1 Car Item and Idea Rubric

The *Car* item is an embedded assessment in a 6th grade science unit about global climate change. The car item asks students to consider the following scenario: "*On a cold winter day, Akbar is walking to his car that is parked in the Sun. It has not been driven for a week. How will the temperature inside the car feel? Colder than the outside air, warmer than the outside air, or exactly the same as the outside air. Explain.*" This question is an analogy for the greenhouse effect, and to answer it, students need to integrate ideas about how energy from the sun is transferred and transformed. To build an idea detection model and design guidance prompts, we first created an idea rubric that enumerates the ideas typically expressed in response to the item, including ideas rooted in personal experience, vague ideas, and non-normative ideas in addition to the ideas that comprise the mechanism targeted by the item [3]. For the purposes of this paper, we will only focus on designing guidance prompt that correspond to the ideas, not the idea detection model development. To develop the rubric, the authors reviewed approximately 1000 student responses to the item that were collected in prior research and identified an initial set of ideas used by students to explain the *Car* phenomenon. The set of ideas was reviewed and refined in partnership with other researchers and classroom teachers. Through this process, we generated a set of 24 ideas for the *Car* item

which included nine mechanistic ideas, seven vague or personal experience-based ideas, and eight nonnormative or off-track ideas. An example of a mechanistic idea is that "solar radiation transforms into heat energy when it is absorbed." An example of a vague idea is the idea that "the sun warms the car." Finally, an example of an inaccurate or off-track idea is the idea that "the car's heater makes the inside warm."

3.2 Procedure for Expert-authored Prompts

The first author of this paper authored KI prompts for each of the ideas in the idea rubric following the KI framework. Drawing on the KI framework, the KI prompts were designed to elicit more of the students' thinking about the idea they had expressed and support them to connect their thinking to accurate, mechanistic ideas. The aim of starting from the idea students initially expressed, rather than introducing a new idea they had not touched upon, was to support students to build a coherent explanation that connects their initial thinking to other relevant ideas. The KI prompts were also designed to be open-ended such that they did not provide the student with the answer to the *Car* item or include new science ideas or vocabulary. The KI prompts were reviewed and revised with the research team, who are all experts in the KI pedagogical framework and a variety of middle grade science topics.

3.3 Procedure for ChatGPT 3.5-authored Prompts

We used a prompt engineering approach to produce guidance prompts authored by ChatGPT. We refer to these guidance prompts as ChatGPT prompts. We used the freely available version, ChatGPT 3.5. Each time we prompted ChatGPT, we provided a set of instructions, the wording of the item, the set of 24 ideas, and criteria the guidance prompts should meet. In our first round of guidance prompt generation, referred to as ChatGPT v1, we provided the following criteria: "The guidance prompts should 1) respect and build off of the idea the student has, 2) elicit more details about the idea or elicit another idea the student has, 3) not include the answer to the question." We observed that the ChatGPT prompts were very long and included more sophisticated science vocabulary than students typically have. The prompts also introduced too many new ideas, some of which were repetitive, including an emphasis on thermal equilibrium and thermal energy transfer, despite thermal equilibrium representing a mechanism that is not central to explaining the temperature inside the car. The instructions to write guidance prompts that help the student build their explanation without giving away the answer may still encourage ChatGPT to include new, relevant ideas rather than building on student ideas.

In response to these observations, for the next round of guidance prompt generation, ChatGPT v2, we modified the criteria provided by saying ChatGPT should be concise (fewer than 30 words) and should not introduce new ideas. We observed that in this round, the ChatGPT prompts were much shorter but still used sophisticated language and rarely encouraged students to build from their own ideas towards accurate and mechanistic ideas.

For ChatGPT v3, we used the same criteria as in v2, but in the instructions told ChatGPT that the ideas came from 6th graders and that ChatGPT should behave like a 6th grade science teacher providing guidance to students. This did not improve the ChatGPT guidance prompts.

For ChatGPT v4, we added criteria that ChatGPT prompts should guide students to build off initial ideas towards accurate ideas. We also labeled each of the provided ideas as inaccurate/off track, vague, or accurate. We observed that these instructions produced very similar prompts to v2 and v3. We noticed that the ChatGPT prompts seemed constrained by the students' initial idea. ChatGPT did not generate prompts that acknowledged and built on the initial student idea while also encouraging students to consider related ideas and evidence.

3.4 Evaluation of the Guidance Prompts

We compiled a dataset consisting of the human expert-authored guidance prompts, the first set of ChatGPT 3.5- guidance (v1), and the final set of ChatGPT 3.5- guidance (v4) for each of the 24 ideas. Within each idea, we randomly ordered the guidance prompts and removed the label that indicates how the prompt was generated (expert, ChatGPT V1, and ChatGPT V4).

Then, another expert in KI who is familiar with the *Car* item but did not author the expert-authored prompts for it evaluated the prompts. The rater gave a binary rating to each guidance prompt as to whether or not it met the criteria. For each idea, the expert also ranked the guidance prompts in order of likelihood to be used by a human teacher when interacting with a student, with 1 being most likely to be used and 3 being least likely to be used. To do so, the evaluator read the student idea that was provided to both ChatGPT and the expert author and the three possible guidance prompts. The evaluator then assigned a rank to each of the possible guidance prompts. We compare the average ratings and ranking for the expert authored, ChatGPT V1 and ChatGPT V4 prompts. We also calculated the number of words and the number of ideas from the set of 24 detected in each guidance prompt. We use the number of ideas as a proxy to indicate how much information that would be found in an accurate explanation was included in the guidance prompt. Based on our criteria, there should only be one idea contained within a prompt. We compare the average number of words, average number of ideas, and average ratings for each set of guidance prompts.

4 RESULTS AND DISCUSSION

Examining the expert ratings of the guidance prompts, we noticed a few trends. In both v1 and v4, ChatGPT performs as well as the expert at respecting and building on the initial idea. ChatGPT v1 also performed as well as the expert at writing prompts that elicit student ideas. There was a slight decline in ability to elicit ideas as we refined instructions to ChatGPT (Table 1).

	Human Expert	*ChatGPT v1*	*ChatGPT v4*
Respect/Build on Student idea (1,0)	0.96	1.00	0.96
Elicit more details about this idea or another idea (1,0)	1.00	1.00	0.83
Support the student to connect the stated idea to accurate ideas (1,0)	1.00	0.88	0.75
Concise (<30 words) (1,0)	0.96	0.75	1.00
*Does not include new ideas or science terminology (1,0)	1.00	0.75	1.00
Sounds like teacher (1-3)	1.33	2.58	2.17
Average Word Count	18.21	24.29	9.21
Average Idea Count	0.63	0.83	0.46

Table 1. Average ratings for each category of guidance prompts (1.0 is best)

Revising the instructions to ChatGPT resulted in improved adherence to the criteria to be concise and to not include new ideas or science vocabulary. However, none of our revisions resulted in the generation of ChatGPT prompts that would support a student to connect their stated idea to accurate ideas. This often appeared to be the result of staying very close to the initial idea, even if it was vague or inaccurate/off-track. For example, in response to the idea that "the car's heater is making the car warm," the expert authored guidance prompt asks the student "Have you ever felt warm inside a car even when the heater wasn't on?" This prompt builds from the starting idea about the heater, but also nudges the student to consider other explanations without introducing new ideas for the student to use. ChatGPT v1 prompted, "You mentioned the car's heater. Can you explain how the car's heating system affects the temperature inside, especially considering that the car hasn't been driven for a week?" While this prompt begins to suggest that the student should attend to the fact that the car hasn't been driven recently, it still focuses the student on explaining how the heater affects temperature. ChatGPT v4 was more concise, but did not support the student to consider or build connections to any other ideas: "How might the car's heater impact temperature?"

Overall, the human expert-authored KI prompts adhered most closely to the pedagogical criteria and most frequently were ranked as being the prompt most likely to be used by a human teacher. The ChatGPT prompts were able to meet some criteria. Through prompt engineering, we were able to produce ChatGPT prompts that came closer to the pedagogical criteria.

5 LIMITATIONS AND NEXT STEPS

This work in progress has many limitations. One limitation is the nature of the ChatGPT prompt engineering. While we found we were able to use prompt engineering to see improvement along some criteria, we were not able to prompt ChatGPT to produce guidance prompts that consistently supported students to connect to accurate ideas. Further exploration of instructions for ChatGPT is needed. For example, we could examine the impact of provided ideal or example student responses to KI prompts. We could also provide examples of good KI prompts. We could also try providing more information about the KI framework to achieve better alignment to all criteria.

We could also expand the pool of raters for the KI and ChatGPT prompts to include both classroom teachers and additional experts. Additionally, we used ChatGPT, version 3.5. It is possible that later versions of ChatGPT or other generative AI platforms, perhaps using different LLMs would perform better.

Lastly, the best evaluation of prompts is the impact they have on students' ability to explain phenomena. Future work will compare the impact of ChatGPT-authored prompts to KI prompts.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2101669.

REFERENCES

[1] Allison Bradford, Weiying Li., Brian Riordan, Kenneth Steimel, and Marcia C. Linn. 2023. Adaptive Dialog to Support Student Understanding of Climate Change Mechanism and Who is Most Impacted. In *Proceedings of the 17th International Conference of the Learning Sciences-ICLS 2023, pp. 816-823.* International Society of the Learning Sciences.

[2] A.A. diSessa. 1988. *Knowledge in Pieces*

[3] Libby Gerard, Sarah Bichler, Allison Bradford, Marcia C. Linn, Kenneth Steimel, and Brian Riordan. 2022. Designing an Adaptive Dialogue to Promote Science Understanding. In C. Chinn, E. Tan, C. Chan, & Y. Kali (Eds.), Proceedings of the 16th International Conference of the Learning Sciences—ICLS 2022 (pp. 1653–1656). International Society of the Learning Sciences

[4] Libby Gerard, and Marcia C. Linn. 2022. Computer-based guidance to support students' revision of their science explanations. *Computers & Education* 176, 104351.

[5] Libby Gerard, Camillia Matuk, Kevin McElhaney, and Marcia C. Linn. 2015. Automated, adaptive guidance for K-12 education. *Educational Research Review* 15, 41-58.

[6] Neil Heffernan, and Cristina Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24, 470-497.

[7] Kenneth R. Koedinger, John R. Anderson, William H. Hadley, and Mary A. Mark. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8, 30-43.

[8] Leslie Rupert Herrenkohl, Tammy Tasker, and Barbara White. 2011. Pedagogical practices to support classroom cultures of scientific inquiry. *Cognition and instruction* 29, 1-44.

[9] Weiying Li, Libby Gerard, Jonathan Lim-Breitbart, Allison Bradford, Marcia C. Linn, Brian Riordan, and Kenneth Steimel. 2023. Explaining thermodynamics: Impact of an adaptive dialog based on a natural language processing idea detection model. In Blikstein, P., Van Aalst, J., Kizito, R., & Brennan, K. (Eds.), Proceedings of the 17th International Conference of the Learning Sciences - ICLS 2023 (pp. 1306-1309). International Society of the Learning Sciences.

[10] Marcia C. Linn. 2006 *The Knowledge Integration Perspective on Learning and Instruction.*

[11] Marcia C. Linn and Bat-Sheva Eylon. 2011. *Science Learning and Instruction.* Routledge.

[12] Zachary A. Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871*

[13] Zachary A. Pardos, Matthew Tang, Ioannis Anastasopoulos, Shreya K. Sheel, and Ethan Zhang. 2023. OATutor: An Open-source Adaptive Tutoring System and Curated Content Library for Learning Sciences Research. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581574

[14] Diana Pérez-Marín. 2021. A review of the practical applications of pedagogic conversational agents to be used in school and university classrooms. *Digital*, *1*(1), 18-33.

[15] Sadhana Puntambekar, Indrani Dey, Dana Gnesdilow, Rebecca J. Passonneau, and ChanMin Kim. 2023. Examining the effect of automated assessments and feedback on students' written science explanations. In Blikstein, P., Van Aalst, J., Kizito, R., & Brennan, K. (Eds.), Proceedings of the 17th International Conference of the Learning Sciences - ICLS 2023 (pp. 1865-1866). International Society of the Learning Sciences.

[16] Riordan, B., Bichler, S., Bradford, A., King Chen, J., Wiley, K., Gerard, L. and C. Linn, M. 2020. An empirical investigation of neural methods for content scoring of science explanations. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (Seattle, WA, USA → Online, Jul. 2020), 135–144.

[17] Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet? -a systematic literature review on chatbots in education. *Frontiers in artificial intelligence* 4, 654924