From Consumers to Critical Users: Prompty, an AI Literacy Tool for High School Students

Deepak Varuvel Dennison*[1], Raycelle C. C. Garcia*[1], Parth Sarin[1], Jacob Wolf[2], Christine Bywater[1], Benjamin Xie[1], Victor R. Lee[1]

[1]Stanford University
[2]Harvard University
{deepakvd, raycellegarcia, psarin, cbywater, benjixie, vrlee}@stanford.edu, jwolf@g.harvard.edu

Abstract

In an age where Large Language Models (LLMs) expedite the generation of text, the skills for critically evaluating and creating meaningful text using these models are often lacking. To help classroom teachers address this, we introduce *Prompty*, a specialized teaching tool co-designed to facilitate both critical and effective use of LLMs. *Prompty* serves multiple learning goals: it allows students to critically evaluate text generated by LLMs, aids in their writing practice, and provides a deeper understanding of how LLMs function—all within a student-friendly environment secured by essential guardrails. *Prompty* was co-designed in collaboration with high school teachers as part of CRAFT, an initiative by Stanford University to promote AI literacy. It was pilot-tested in a high school English class to serve as an AI writing assistant, focusing on the critical evaluation of machine-generated text. This trial yielded preliminary evidence that attests to the tool's effectiveness in fulfilling its educational goals. The findings from the pilot study indicate that easy-to-use tools like *Prompty* have great potential. These tools can be adapted to fit the goals of individual teachers. They can help in achieving subject-specific learning goals while serving as an effective way to teach AI concepts in high school.

Prompty at a Glance

Prompty (Figure 1) is a web-based learning tool that equips students to have an informed engagement with Large Language Models (LLMs) through a specially designed interface that supports the exploration and critical evaluation of machine-generated text.

Target age group: High school students (ages 14-18)

Setup and resources needed: For classroom implementation, teachers and students will need access to a web browser on a laptop or Chromebook. A one-to-one student-to-laptop ratio was implemented in this study and is recommended.

AI concepts addressed: The definition of LLMs, prompt engineering, and biases in LLMs.

Expected learning outcomes: Students will be able to cultivate informed engagement with LLMs, enabling them to critically evaluate machine-generated text effectively. It also provides a platform for students to harness LLMs as digital writing assistants.

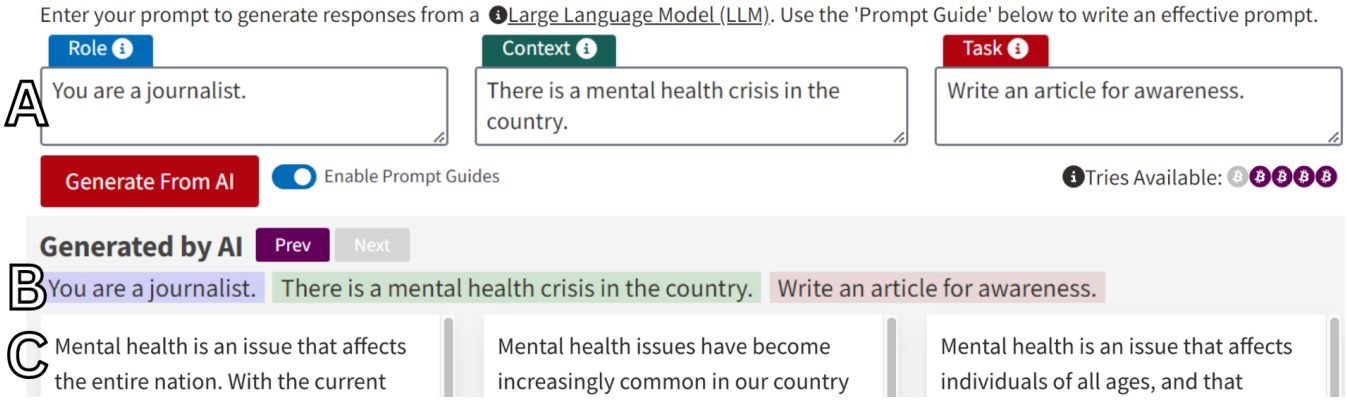

Figure 1: The *Prompty* Application User Interface. Users interact with an LLM in a learning environment designed for students. (A) Students use prompt guide to scaffold their development of a prompt. (B) Prompt is shown with different components highlighted (e.g., "Role" in blue). (C) LLM generates three outputs from the given prompt for students to compare.

* These authors contributed equally

Introduction

Since late 2022, applications built upon LLMs, such as ChatGPT, have swiftly ascended to cultural prominence, captivating the public with their powerful natural language processing abilities. Their low threshold interfaces, which enable the generation of voluminous text through a simple text box, have further fueled their allure. This heightened interest has not only triggered a stark increase in LLM-related research, as seen by a spike in arXiv publications (Zhao et al. 2023), but has also been hailed as a broader paradigm shift in artificial intelligence technologies by multiple scholars (Henriksen, Woo, and Mishra 2023; Kissinger et al. 2023; Pullar-Strecker 2023). ChatGPT's extraordinary user adoption trajectory—reaching one million users in just five days and expanding to 100 million users within two months (Chartr 2022; Paris 2023)—attests to its impact.

However, such rapid adoption has illuminated certain complexities and challenges. OpenAI, the organization behind ChatGPT, has publicly acknowledged limitations, such as the model's tendency to produce "plausible-sounding but incorrect or nonsensical answers" (OpenAI 2022). Moreover, numerous studies have flagged additional ethical and practical concerns, such as the models' inherent biases and potential for spreading misinformation (Doshi, Bajaj, and Krumholz 2023). In this context, there exists a critical need for educational paradigms that empower students to use these AI technologies both effectively and critically. Within this challenging landscape, global efforts to enhance AI literacy are gaining momentum. A systematic literature review by Casal-Otero et al. (2023) delves into these efforts, revealing that despite progress, effectively incorporating AI literacy education into K-12 settings remains a daunting task. To contribute to overcoming these obstacles, our work is focused on co-designing an extensive array of educational resources for AI literacy with high school teachers. These resources aim to provide a holistic understanding of various AI technologies, including LLMs.

We introduce *Prompty*—a versatile learning tool designed to cultivate an informed engagement with LLMs among students to support an array of related learning goals. This includes enabling students to critically evaluate the machine-generated text and exploring the scope of LLMs as digital assistants to hone their writing skills. It also serves to deepen understanding of LLM's behaviors—all within a secure and student-centric digital ecosystem. This paper will detail *Prompty's* genesis, elaborate on its design elements, share preliminary findings from its initial pilot implementation, and discuss the implications of these findings in the context of the larger repository of AI literacy tools under development. We will also outline future avenues, emphasizing the tool's potential for wider applications and its role in advancing AI literacy across various educational settings.

Background

Large Language Models

Language is often considered a capability that is intrinsic to humans, developing in early childhood and evolving throughout a lifetime (Hauser, Chomsky, and Fitch 2002). The quest to create machines that can read, write, and communicate like humans dates back to Alan Turing's seminal question in 1950, "Can machines think?" Since then, researchers in the field of artificial intelligence (AI) have tirelessly strived to imbue machines with natural language abilities (Goldstein and Papert 1977; Mehta and Devarakonda 2018).

One of the most notable advancements in AI in recent years has been the development of LLMs (Liang et al. 2022). The advent of LLMs has brought about a revolution in the AI community, with applications like ChatGPT and Bard having a profound impact on various fields. Some speculate that GPT-4 may represent the basis for an early version of an Artificial General Intelligence system, showcasing a wide range of intelligent abilities, including reasoning, planning, and the capacity to learn from experience, potentially at or above human-level (Bubeck et al. 2023). However, despite the rapid progress and uptake of LLMs, the fundamental principles underlying their functioning remain relatively unexplored. More research is essential to comprehend their emergent abilities fully, train highly capable LLMs, and align them with human values and preferences (Ferrara 2023; Harrer 2023; Wei et al. 2022). Additionally, there have been growing concerns among scholars regarding the ethical and social risks associated with LLMs (Weidinger et al., 2021).

Learning Tools for AI Literacy

As AI technologies continue to rise in prominence, there is a growing global interest in educating students about AI. In 2018, AAAI and CSTA jointly developed national guidelines for teaching AI to K-12 students in the form of five big ideas, which have been influential in shaping the AI literacy tools (Touretzky et al. 2019). Long and Magerko (2020) have identified a set of core competencies and proposed various design considerations to aid AI developers and educators in developing learner-centered AI literacy resources.

These considerations are demonstrated in the diverse array of learning tools currently being created by numerous initiatives to promote AI literacy among students. Some efforts include AI FOR K-12, aiEdu, CRAFT, MIT RAISE, and TeachAI among others. Within the domain of AI ethics, Williams et al. (2022) have introduced a curriculum designed to equip students with a critical perspective for comprehending AI systems and their societal impacts.

Various efforts craft unique learning experiences that intersect AI with other domains. These include endeavors

such as AI and cybersecurity (Broll and Grover 2023), AI in the context of sports (Kumar and Worsley 2023), AI in STEM (Lee and Perret 2022), and AI in English Language Arts (Chao et al. 2023).

Certain initiatives aim to familiarize students with specific AI technologies. Touretzky and Gardner-McCune (2023) have developed an approach to educate students about speech recognition and language intricacies within the AI context. Similarly, DiPaola et al. (2023) have introduced an interactive tool designed to facilitate comprehension of rule-based image recognition, a critical concept in computer vision. In addition, "Teachable Machine," as presented by Carney et al. (2020), serves as a resource to demystify the inner workings of machine learning. Research has shown that the utilization of digital tools can significantly improve learning outcomes in technology-related subjects (Hillmayr et al. 2020). As the demand for technology literacy continues to rise, these digital resources play an essential role in assisting students in achieving their crucial learning objectives.

Methods

Design Question

Aligning with the goals of the overarching curriculum project CRAFT, we sought solutions that are accessible, relevant, and adoptable by teachers from different disciplines. Addressing this uncharted and timely need to learn about LLMs in AI K-12 education, our team delved into the Design Question: How might we support teachers across disciplines in facilitating learning experiences that enable students to be informed and critical users of LLMs?

Participants and Procedures

While literature surrounding LLMs is still sparse in this emerging area of AI literacy to design an effective learning tool, our approach centered around pedagogical practices validated through learning sciences research and iterated through co-design with a high school teacher, Ms. L. Through co-design we aim to design features that meet the values of users (Van Mechelen et al. 2017), as well as give agency to K-12 teachers while expounding upon their wealth of knowledge and experiences (Lin and Van Brummelen 2021).

Ms. L, an English Language Arts (ELA) teacher at a charter school in California, was selected as a participant in the larger set of curricular co-design sessions for the overarching project. Over the course of 3 months, we conducted three online design and preparation sessions that lasted between 20 to 60 minutes, which led to in-person class implementations of *Prompty* by Ms. L. This implementation, conducted in 2 ELA high school classes with students of age 14-16, integrated *Prompty* as a writing assistant tool for a multi-part writing project that spanned 3 weeks. The key textual reference for this project was the book "The How of Happiness" by Sonja Lyubomirsky, which the class had already been assigned. Students had one-to-one access to laptop devices in the lessons and accessed *Prompty* through a web browser. Ms. L presented a mix of physical paper worksheets and digital worksheets as part of the lesson tasks alongside interacting with *Prompty*. As Ms. L taught in a school located far from the research team, we were unable to observe the class in person. Therefore, Ms. L shared her perspective as a teacher and her perceived views of the students' experiences in a final debrief session over Zoom.

Learning Objectives

Based on our informed understanding of LLMs, our team developed initial learning objectives towards AI literacy that were emulated in the design of *Prompty*. These were structured in the form of 3 questions:

1. How can I write utilizing generative AI as a resource?
2. How do I create effective prompts for generative AI?
3. What should I consider to be an expressive, responsible writer with AI?

We refined our initial objectives through discussions and insights from co-design sessions, continuously updating core learning objectives and lesson plans, which in turn shaped the user experience of *Prompty*.

Iterating from these objectives, Ms. L highly valued the goal of teaching students how to use generative AI effectively in their lives, an important skill as she feels that "writing as a practice will change dramatically in the next five years." Furthermore, she emphasized key ELA skills covering California state standards (California Department of Education 2013) to integrate into the lessons, including: comparing and evaluating different texts, writing and revising for different audiences and purposes, and using textual evidence to justify claims.

Based on these goals and the context of her classroom, Ms. L selected the following overarching learning goal and corresponding learning objectives in her implementation:

- Compose a creative writing piece using AI-generated responses to a prompt.
- Build and iterate on prompts to reach the desired output
- Compare and evaluate the results of generative AI

In the lessons, students generated pieces for different audiences of their choice using *Prompty*. Students first practiced with initial ungraded preparatory work, followed by a final piece that graded both the process of iterating generated text and the final written product using *Prompty*.

Tool Development

Technical Features

Technology

Prompty is a React-based web application powered by OpenAI's GPT-3 LLM model, known as *text-davinci-003*. We opted for text-davinci-003 due to its longer context window, enabling it to process up to 4097 tokens (Raf 2023). In AI models, a "token" is a basic unit of text, often representing a word or character. *text-davinci-003's* capacity to handle 4097 tokens in its context window can accommodate more extensive and intricate inputs, resulting in more detailed responses.

Before sending user prompts to the LLM, *Prompty* takes precautionary steps. It employs OpenAI's Moderation endpoint to ensure that the prompts adhere to content guidelines and prevent the generation of inappropriate responses. This Moderation endpoint helps developers to monitor and filter content, classifying it as hate, harassment, self-harm, sexual content, and violence, helping ensure compliance with OpenAI's usage policies (OpenAI 2023).

To further ensure the generated responses are student-friendly and devoid of explicit content, *Prompty* augments prompts with additional information before generating responses from the LLM. The moderation process includes guardrails, which are explained in the following section.

Additionally, *Prompty* logs successful text-davinci-003 generations in a Firestore database, associating them with the corresponding user IDs. This logging mechanism serves the dual purpose of retaining data across different sessions and monitoring the number of generations to encourage mindful usage.

Implementation of the Guardrails

Prompty places a significant emphasis on prioritizing the safety of its student users. Recognizing the potential curiosity of our target audience, particularly teenagers, our team deemed it crucial to implement guardrails to prevent the generation of inappropriate responses and ensure a secure user experience. This is achieved through the implementation of guardrails at two levels: the prompt level and the response level. At the prompt level, a critical safeguard is employed when a user submits a request. The submitted prompt undergoes an evaluation by OpenAI's Moderation endpoint, which employs advanced GPT-based classifiers (Markov et al. 2023). This system is designed to detect content falling into various undesired categories, including hate speech, harassment, self-harm, sexual content, and violence. When a user's text is evaluated using the Moderation endpoint, it returns scores for different content categories. These scores represent the model's confidence in associating specific content with a particular category. For instance, a score of 0.7 for violence indicates that the model is 70% confident that the content contains violent elements. To maintain a secure environment, prompts receiving a score greater than 0.5 in any of these categories are flagged. Users are then notified and asked to align their prompt with the platform's content policy, which prohibits content related to hate, harassment, self-harm, sexual content, or violence.

While the Moderation endpoint effectively guards against various inappropriate content, there may be cases where it does not identify potential risks adequately. In such situations, the LLM could generate content that includes sexually explicit material or profane language not suitable for a high-school audience. To address this concern, before the LLM is prompted for text completion, the user's input is modified on our server. Specifically, the prompt is expanded by adding a message: “Response should be suitable for young students. Strongly refrain from using any profanity, hate, or sexual references. Following is the prompt:" This modification ensures that the LLM operates within strict boundaries, emphasizing the importance of producing safe and appropriate content for classroom use.

Design Features

Design for Contrasting Cases

A learning mechanic that is both widely studied and considered to be effective by several learning sciences researchers is Contrasting Cases (Lin-Siegler, Shaenfield, and Elder 2015; Rittle-Johnson and Star 2007; Sidney, Hattikudur, and Alibali 2015). Contrasting cases is recognized as a valuable cognitive tool, as it enables a more profound comprehension of an object's attributes when juxtaposed with subtly different alternatives (Gentner and Namy 1999; Schwartz, Tsang, and Blair 2016). Contrasting Cases allows a learner to generate more ideas and make insights that they might not have made otherwise. We have utilized this learning mechanic as one of the main design features in *Prompty*.

In *Prompty*, when a user submits a prompt, the platform requests the LLM to generate three distinct responses, which are then presented side by side in the user interface as shown in Figure 2.

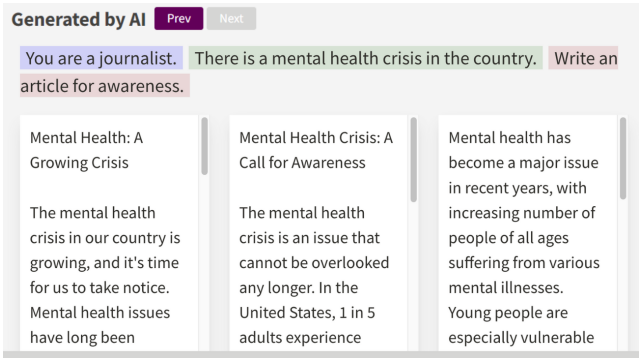

Figure 2: The *Prompty* Contrasting Cases Responses. For each prompt submission, three responses are presented to the user.

This design decision was implemented after a co-design session where Ms. L highlighted the importance of producing multiple results simultaneously:

> The thing that got me really excited was that when you gave it the prompts it responded back to you with like 4 or 5 options...[Otherwise] what my students are going to do is look at, try number one, and they're going to be like "great, I'm done. Why would I want to iterate it again?" … If they get 4 different ones, now, I'm like, which one is the best one? Why is that one the best one? Is there something about that one that you like better? What is it that is good or bad, about, or effective or not effective? That's the way that I'm going to frame it with them. What about this one versus that one? Are there pieces that you would want to combine? What do you want to emphasize?

Drawing from this insight, we see that Contrasting Cases offer several educational benefits. Firstly, it encourages active engagement with the learning content as students are nudged to compare and contrast the responses. This approach enhances comprehension by pushing the students to read between the lines and critically explore the accuracy and biases of the generated texts. Secondly, it encourages students to question the relevance of each response in relation to their specific objectives. Presenting three responses pushes students to identify relevant text segments that they can incorporate into their compositions more effectively. Furthermore, teachers are encouraged to play an active role in this process through lesson plans within *Prompty*, guiding students in critically comparing the responses. This collaborative approach fosters a rich learning environment that promotes deeper understanding through discussion and analysis. Finally, we note that *Prompty* has three responses due to UI screen space constraints and Ms. L's agreement in a subsequent co-design session that three responses were sufficient when combined with access to multiple prompt tries.

Prompt Guides for Distributed Cognition

In his seminal work, Pea (1993) has described the role of distributed intelligence in knowledge construction. Furthermore, as the concept of scaffolding became broadly applied in the field of education research, Pea (2004) posed an important distinction between scaffolding and distributed intelligence, namely by the characteristic of fading, the act of removing the scaffolds as the learner gains mastery. In exploring the definition of scaffolding and its underlying theories in relation to distributed intelligence and distributed cognition, Belland (2011) has provided an insightful alternative perspective. He has posited that distributed cognition can serve as a framework for understanding the impact of computer-based scaffolds, and thus how to optimize their use for enhanced learning. To achieve success in task completion, an individual must effectively employ a range of cognitive tools, skillfully leveraging them. According to the advocates of distributed cognition perspectives, there's no requirement for these cognitive tools to be confined solely within the individual's mind; rather, they can be distributed across multiple elements within distributed cognitive systems (Perkins 1995). At the core of distributed cognition is the notion that, over time, the distributed cognitive system can adapt, with one cognitive tool being replaced by others within the system (Hutchins 1995). This adaptation occurs as individuals develop schemas that incorporate processes supported by these cognitive tools. These schemas encompass mental representations of categories of concepts and general processes that individuals employ to comprehend the world and carry out actions. Consequently, the impact of computer-based scaffolds may manifest in students creating schemas that guide subsequent task completion approaches. As students engage in new tasks, their revised schemas offer guidance for their approaches.

In *Prompty*, we employ Prompt Guides as computer-based scaffolds to assist students in crafting well-structured prompts for the LLM. These Prompt Guides encourage students to approach prompt creation systematically, breaking it down into three key components: Role, Context, and Task. This is shown in Figure 3.

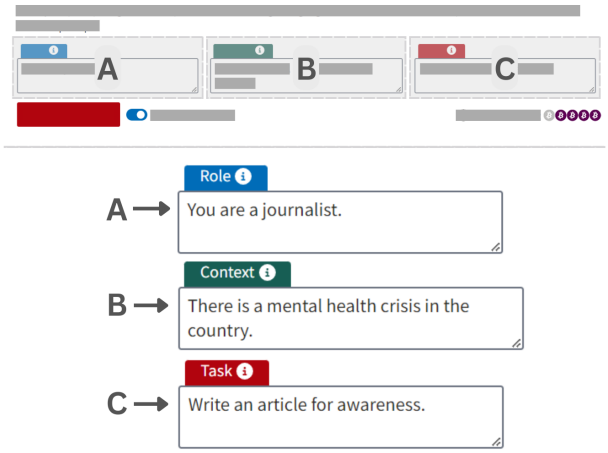

Figure 3: The *Prompty* Prompt Guides. Prompt writing is scaffolded into three key components: Role, Context, and Task.

In the Role component, students are prompted to specify the role with which the LLM's output should align. For example, they might be instructed to assume the role of a journalist. In the Context component, students are asked to provide additional context related to the content generation. For instance, they may be asked to describe a scenario such as "there is a growing mental health crisis in the country." In the Task component, students are guided to articulate the specific task they want the LLM to perform. For example, they might instruct the LLM to "write an article elaborating on the importance of mental health." By utilizing Prompt

Guides, students are enabled to develop the mental schema necessary for effective prompting of LLMs. Over time, this approach allows students to internalize the process, eventually enabling them to craft effective prompts without relying on the guides. Additionally, in *Prompty*, we offer an open-ended prompting interface that allows students to practice prompt generation without any scaffolding. This provides an opportunity for students to hone their skills and gain confidence in formulating prompts independently.

Limited Tries to Promote Conscious Usage

To encourage the responsible use of LLMs as a learning resource, we have implemented limitations on the number of times a student can use *Prompty* to generate responses during a session. This limitation is communicated to the user through the user interface (UI), making them aware of the number of attempts available, as shown in Figure 4.

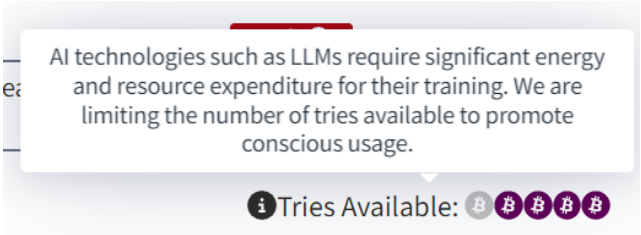

Figure 4: The *Prompty* Attempts Meter. Students have a limited number of tries to generate text.

The intention behind this limitation is to nudge the students to be more mindful of their usage and to encourage them to approach the prompting process with greater intentionality. The environmental impact is also an important factor that we wanted the students to be aware of. This is particularly pertinent since there is a growing concern about the environmental impact of artificial intelligence systems as LLM services become increasingly accessible (Dhar 2020; Luccioni, Viguier, and Ligozat 2022).

By setting a restriction on the number of tries, we aim to foster students in being directed, responsible, and reflective as they iterate different prompts while considering the environmental impact associated with extensive usage.

Preliminary Results

Since our team was unable to observe the implementation in person due to distance, our findings presented are from Ms. L's statements over a Zoom video debrief meeting after implementing *Prompty*. Her overall review was that teaching with *Prompty* was a positive learning experience that engaged students in both AI literacy as well as skills surrounding ELA standards. She perceived that her students were able to quickly understand how to use the *Prompty* interface after an initial exploration trying the tool, and adopt the process of prompting and revision to achieve a desired purpose. Ms. L highlighted how the simple interface was clean and effective for students to use. She also felt that this simple-to-use interface would allow more teachers to be able to adapt and implement *Prompty* in their classrooms.

Generating multiple responses at once was a key component of *Prompty* to drive the 'compare and evaluate' lesson objectives through Contrasting Cases. The scaffolding of Role, Context, and Task in the Prompt Guides worked well in helping students break down the components of a prompt. She noted that some of her students struggled to identify how to change the prompt based on weaknesses of their results, and suggested more guidance could be provided in supporting students on how to revise a prompt. She also highlighted a possible improvement would be to allow teachers to adjust the different scaffolding components for their students.

While Ms. L noted that several of her students requested more tries, she highlighted the value of limited tries for the learning experience, stating:

> I think that any time we put a limit on anything, that on one hand limits creativity a bit. But for somebody who's learning something new, it's a structure to play within…you only get 3 tries, and you really need to be intentional about what you're doing with each try. And you really need to pause and reflect between each try.

This reflects how the limit and structure were a helpful mix between open-ended exploration and intentional decision-making.

When incorporating the standard of citing textual evidence, while Ms. L initially intended for students to add citations manually to a produced generated text, she was interested to see students add quotes to the prompts themselves. She explained:

> The students ended up saying a lot to prompt the AI, so the kids that had the best responses ended up writing kind of a paragraph as their prompt. And so then, again, as an English teacher, cells are going off in my brain, [before] it's like actually, what's the product we're asking our students to produce? Now, the thinking is the prompting of the AI and the evaluation of the results, right?

While teaching with *Prompty*, Ms. L recognized disciplinary value in the process of prompt engineering in addition to the finished product itself.

Furthermore, Ms. L reflected on further skills students are developing while using *Prompty*, stating that the exercise is about "learn[ing] that the AI itself needs human attention and human evaluation and human judgment. Which, by the way, is like a DOK (Depth of Knowledge)…3, 4 [skill]. So to me, if they're doing that, I'm happy." Ms. L used the limitation of AI as a means to teach the skills students need to effectively use generative AI and encourage higher levels of

thinking according to the 4 levels of Webb's (2002) Depth of Knowledge framework.

Discussion

These preliminary findings from Ms. L's implementation perspectives validate the design features of *Prompty* as a learning tool. Namely, the use of Contrasting Cases nudged students towards comparison and evaluation, the limited number of tries created a structure that prompted intentionality and reflection, and the Prompt Guides scaffolded students to consider what goes into a prompt and how that might be reflected in the results. Together this supported the learning outcome to use AI in crafting a creative writing piece. Further guidance may be needed in supporting meaningful iterations, such as connecting the prompt guide components and how that is reflected in the results of the AI to further pinpoint how to iterate on the prompt.

An interesting shift in perspective observed was in the emphasis on product and process to attain disciplinary learning goals. Initially, during the design of the *Prompty* learning experience, Ms. L's disciplinary learning goals strongly focused on the written product of the AI, while the process focused on AI literacy goals on prompting. This cohesively joined to attain her learning objective of preparing students to use generative AI effectively to write for a specific purpose. While implementing *Prompty*, her experience with her students incorporating textual evidence expanded her views on how the prompting process itself facilitated opportunities to meet disciplinary objectives and standards. This integration is both a reflection on the discipline-specific skills that are prevalent throughout the process of interfacing with AI, as well as the wider perspective on how generative AI is evolving the writing process and skills associated with it.

Overall, the co-design process was mutually beneficial in designing a learning experience that facilitated learning objectives towards both AI literacy and within the context of the ELA discipline. Through *Prompty*, we exemplified how a learning resource can be developed that is readily usable by teachers from a background outside computer science and can meet their disciplinary objectives. In both the design process and implementation of *Prompty*, Ms. L could highlight, develop, and observe ways that integrated the teaching of AI literacy skills with her disciplinary learning objectives. This supports the integration of AI literacy by augmenting existing lesson structures, in keeping with our goal to incorporate AI literacy into classrooms of different disciplines without overwhelming teachers or cutting into limited lesson time.

Implications and Future Work

The positive learning outcomes of this initial co-design and implementation of *Prompty* is a promising step towards AI literacy tools that are classroom-ready and can be adapted by teachers of different disciplines to meet their own goals. Our experience with Ms. L reflects how the simple-to-use yet effective interface can lend itself to different discussion topics that interweave AI and ELA learning goals.

As we look towards wider implementation of *Prompty* by other teachers, we do bear in mind that Ms. L, as part of the co-design efforts of the wider curriculum project, may have more experience with AI literacy compared to other ELA teachers. As such, her ready integration of disciplinary objectives may be more difficult for other teachers to implement. This may affect the level of support and resources needed for teachers to successfully use *Prompty* in their classrooms. Additionally, the simple interface of *Prompty* allows us flexibility to pursue further topics and implementations beyond ELA to support AI literacy within other contexts. We are considering other use cases to explore these possibilities:

- To understand the workings of LLM-based Chatbots: While *Prompty*'s interface is not ideal for chat, the current UI could be exploited to teach how LLM-based chat applications work. By nudging students to sequentially stack the responses generated by the LLM, they can gain a clear understanding of how subsequent responses build upon their predecessors. This hands-on exercise provides insights into the fundamental mechanics of LLM chatbots, shedding light on their core functionality behind the scenes.
- To explain the parameter settings in LLMs: We can add an advanced prompting mode by incorporating additional configurable options like temperature, context window, and penalty. This allows students to experiment and see how different parameters affect LLM behavior, enhancing their understanding of LLMs and their practical application in various subjects.
- To experiment with different LLMs: While *Prompty* currently allows the students to engage only with the *text-davinci-003* model, it could be extended as a sandbox to allow the students to generate responses from different LLM models. This expansion would provide students with a safe platform for diverse experimentation, enabling them to explore the capabilities and differences among various LLMs.

In continuing work to enhance *Prompty's* adaptability over various topics, we strive to develop an effective learning tool that builds AI literacy within contexts that are relevant to teachers and support students in readily adapting to AI in their high school studies.

Acknowledgements

This material is based upon work supported by Stanford Digital Education, Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford Accelerator for Learning, and the McCoy Family Center for Ethics in Society as part of the Stanford Classroom-Ready Resources About AI For Teaching (CRAFT) initiative. We thank the teachers who co-designed curriculum with us. Authors declare no conflicts of interest.

References

Belland, B. R. 2011. Distributed Cognition as a Lens to Understand the Effects of Scaffolds: The Role of Transfer of Responsibility. *Educational Psychology Review* 23(4): 577-600. doi.org/10.1007 /s10648-011-9176-5

Broll, B.; and Grover, S. 2023. Beyond Black-Boxes: Teaching Complex Machine Learning Ideas through Scaffolded Interactive Activities. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 37(13): 15990-15998. doi.org/10.1609/aaai.v37i13. 26898

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; and Nori, H. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712 [cs.CL]. Ithaca, NY: Cornell University Library.

California Department of Education. 2013. California common core state standards: English language arts & literacy in history/social studies, science, and technical subjects. California: Authors.

Carney, M.; Webster, B.; Alvarado, I.; Phillips, K.; Howell, N.; Griffith, J.; Jongejan, J.; Pitaru, A.; and Chen, A. 2020. Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing system*, 1-8. doi.org/10.1145/3334480.3382839

Casal-Otero, L.; Catala, A.; Fernández-Morante, C.; Taboada, M.; Cebreiro, B.; and Barro, S. 2023. AI literacy in K-12: a systematic literature review. *International Journal of STEM Education*, 10(1). doi.org/10.1186/s40594-023-00418-7

Chao, J.; Ellis, R.; Jiang, S.; Rosé, C.; Finzer, W.; Tatar, C.; Fiacco, J.; and Wiedemann, K. 2023. Exploring Artificial Intelligence in English Language Arts with StoryQ. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15999-16003. doi.org/10.1609/aaai.v37i13.26899

Chartr. 2022. ChatGPT: the AI bot taking the tech world by storm. https://www.chartr.co/stories/2022-12-09-1-chatgpt-taking-thetech-world-by-storm. Accessed: 2022-09-05.

Dhar, P. 2020. The carbon impact of artificial intelligence. *Natural Machine Intelligence*, 2(8): 423-425. doi.org/10.1038/s42256-020-0219-9

DiPaola, D.; Malachowsky, P.; Blair Black, N.; Alghowinem, S., Du, X.; and Breazeal, C. 2023. An Introduction to Rule-Based Feature and Object Perception for Middle School Students. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 16004-16010. doi.org/10.1609/aaai.v37i13.26900

Doshi, R. H.; Bajaj, S. S.; and Krumholz, H. M. 2023. ChatGPT: temptations of progress. *The American Journal of Bioethics*, 23(4): 6-8. doi.org/10.1080/15265161.2023.2180110

Ferrara, E. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738 [cs.CL]. Ithaca, NY: Cornell University Library.

Gentner, D.; and Namy, L. L. 1999. Comparison in the development of categories. *Cognitive development*, 14(4): 487-513. doi.org/10.1016/S0885-2014(99)00016-7

Goldstein, I.; and Papert, S. 1977. Artificial intelligence, language, and the study of knowledge. *Cognitive science*, 1(1): 84-123. doi.org/10.1016/S0364-0213(77)80006-2

Harrer, S. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90. doi.org/10.1016/j.ebiom.2023.104512

Hauser, M. D.; Chomsky, N.; and Fitch, W. T. 2002. The faculty of language: what is it, who has it, and how did it evolve?. *Science*, 298(5598): 1569-1579. doi.org/10.1126/science.298.5598.1569

Henriksen, D.; Woo, L. J.; and Mishra, P. 2023. Creative Uses of ChatGPT for Education: a Conversation with Ethan Mollick. *TechTrends*, 1-6. doi.org/10.1007/s11528-023-00862-w

Hillmayr, D.; Ziernwald, L.; Reinhold, F.; Hofer, S. I.; and Reiss, K. M. 2020. The potential of digital tools to enhance mathematics and science learning in secondary schools: A context-specific meta-analysis. *Computers & Education*, 153, 103897. doi.org/10.1016/j.compedu.2020.103897

Hutchins, E. 1995. *Cognition in the Wild*. MIT Press. doi.org/10.7551/mitpress/1881.001.0001

Kissinger, H.; Schmidt, E.; and Huttenlocher, D. 2023. ChatGPT heralds an intellectual revolution. *Wall Street J*.

Kumar, V.; and Worsley, M. 2023. Scratch for Sports: Athletic Drills as a Platform for Experiencing, Understanding, and Developing AI-Driven Apps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 16011-16016. doi.org/10.1609/aaai. v37i13.26901

Lee, I.; and Perret, B. 2022. Preparing High School Teachers to Integrate AI Methods into STEM Classrooms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12783-12791. doi.org/10.1609/aaai.v36i11.21557

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; and Newman, B. 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 [cs.CL]. Ithaca, NY: Cornell University Library.

Lin, P.; and Van Brummelen, J. 2021. Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1-12. doi.org/10.1145/3411764.3445377

Lin-Siegler, X.; Shaenfield, D.; and Elder, A. D. 2015. Contrasting case instruction can improve self-assessment of writing. *Educational Technology Research and Development*, 63: 517-537. doi.org/10.1007/s11423-015-9390-9

Long, D., and Magerko, B. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1-16. doi.org/10.1145/3313831.3376727

Luccioni, A. S.; Viguier, S.; and Ligozat, A. L. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. arXiv preprint arXiv:2211.02001 [cs.CL]. Ithaca, NY: Cornell University Library.

Markov, T.; Zhang, C.; Agarwal, S.; Nekoul, F.E.; Lee, T.; Adler, S.; Jiang, A.; and Weng, L. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 15009-15018. doi.org/10.1609/aaai.v37i12.26752

Mehta, N.; and Devarakonda, M. V. 2018. Machine learning, natural language programming, and electronic health records: The next step in the artificial intelligence journey?. *Journal of Allergy and Clinical Immunology*, 141(6): 2019-2021. doi.org/10. 1016/j.jaci.2018.02.025

OpenAI, 2022. Chatgpt: Optimizing language models for dialogue. OpenAI. https://openai.com/blog/chatgpt. Accessed: 2023-

09-05.

OpenAI. 2023. "Moderation - OpenAI API". https://platform.openai.com/docs/guides/moderation. Accessed: 2023-09-05.

Paris, M. 2023. ChatGPT hits 100 million users, google backs claude bot maker and CatGPT goes viral. https://www.forbes.com/sites/martineparis/2023/02/03/chatgpt-hits-100-million-microsoftunleashes-ai-bots-and-catgpt-goes-viral/. Accessed: 2023-09-05.

Pea, R. D. 1993. Practices of distributed intelligence and designs for education. *Distributed cognitions: Psychological and educational considerations*, 11: 47-87.

Pea, R. D. 2004. The Social and Technological Dimensions of Scaffolding and Related Theoretical Concepts for Learning, Education, and Human Activity. *The Journal of the Learning Sciences*, 13(3): 423-451. doi.org/10.1207/s15327809jls1303_6

Perkins, D. 1995. *Outsmarting IQ: The emerging science of learnable intelligence*.

Pullar-Strecker, T. 2023. Artificial intelligence: World at 'tipping point', says Sir Peter Gluckman. https://www.stuff.co.nz/business/131393822/artificial-intelligence-world-at-tipping-point-says-sir-peter-gluckman. Accessed 2023-09-05.

Raf, 2023. How do text-davinci-002 and text-davinci-003 differ?. OpenAI. https://help.openai.com/en/articles/6779149-how-do-text-davinci-002-and-text-davinci-003-differ. Accessed: 2023-09-05.

Rittle-Johnson, B.; and Star, J. R. 2007. Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, 99(3): 561. doi.org/10.1037/0022-0663. 99.3.561

Schwartz, D. L.; Tsang, J. M.; and Blair, K. P. 2016. *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*. WW Norton & Company.

Sidney, P. G.; Hattikudur, S.; and Alibali, M. W. 2015. How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learning and Instruction*, 40: 29-38. doi.org/10.1016/j.learninstruc.2015.07.006

Touretzky, D.; Gardner-McCune, C.; Martin, F.; and Seehorn, D. 2019. Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI conference on artificial intelligence*, 33(01): 9795-9799.

Touretzky, D. S.; and Gardner-McCune, C. 2023. Guiding Students to Investigate What Google Speech Recognition Knows about Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13):16040-16047. doi.org/10.1609/aaai. v37i13.26905

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind*, 59(236): 433–460. doi.org/10.1093/mind/LIX.236.433

Van Mechelen, M.; Derboven, J.; Laenen, A.; Willems, B.; Geerts, D.; and Abeele, V. V. 2017. The GLID method: Moving from design features to underlying values in co-design. *International Journal of Human-Computer Studies*, 97:116-128. doi.org/10.1016 /j.ijhcs.2016.09.005

Webb, N. L. 2002. Depth-of-knowledge levels for four content areas. *Language Arts*, 28: 1-9.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; and Chi, E.H. 2022. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 [cs.CL]. Ithaca, NY: Cornell University Library.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; and Kenton, Z. 2021. Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359 [cs.CL]. Ithaca, NY: Cornell University Library.

Williams, R.; Ali, S.; Devasia, N.; DiPaola, D.; Hong, J.; Kaputsos, S.P.; Jordan, B.; and Breazeal, C. 2022. AI+ ethics curricula for middle school youth: Lessons learned from three project-based curricula. *International Journal of Artificial Intelligence in Education*, 1-59. doi.org/10.1007/s40593-022-00298-y

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; and Du, Y. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223 [cs.CL]. Ithaca, NY: Cornell University Library.