Disclosures & Disclaimers: Investigating the Impact of Transparency Disclosures and Reliability Disclaimers on Learner-LLM Interactions

Jessica Y. Bo[1]*, **Harsh Kumar**[1]*, **Michael Liut**[2], **Ashton Anderson**[1]

[1]University of Toronto, Canada
[2]University of Toronto Mississauga, Canada
jbo@cs.toronto.edu, harsh@cs.toronto.edu, michael.liut@utoronto.ca, ashton@cs.toronto.edu

Abstract

Large Language Models (LLMs) are increasingly being used in educational settings to assist students with assignments and learning new concepts. For LLMs to be effective learning aids, students must develop appropriate levels of trust and reliance on these tools. Misaligned trust and reliance can lead to suboptimal learning outcomes and decreased engagement with LLMs. Despite their growing presence, there is limited understanding of how to achieve optimal transparency and reliance calibration in the educational use of LLMs. In a 3x2 between-subjects experiment conducted in a university classroom, we tested the effect of two transparency disclosures (`System Prompt` and `Goal Summary`) and an in-conversation `Reliability Disclaimer` on a GPT-4-based chatbot tutor provided to students for an assignment. Our findings suggest that disclaimer messages included in responses may effectively mitigate learners' overreliance on the LLM Tutor when incorrect advice is given. While transparency disclosures did not significantly affect performance, seeing the `System Prompt` appeared to calibrate students' confidence in their answers and reduce the frequency of copy-pasting the exact assignment question to the LLM Tutor. Further student feedback indicated that they would prefer to receive guaranteed reliability of LLM tools, tutorials demonstrating effective prompting techniques, and transparency around performance-based metrics. Our work provides empirical insights into the design of transparency and reliability mechanisms for using LLMs in classroom settings.

Introduction

One-on-one tutoring is considered the gold standard for effective teaching (Bloom 1984). However, delivering individualized instruction is often difficult due to challenges such as high student-to-teacher ratios, limited accessibility, and the scarcity of qualified teachers. Scalable and automated methods are essential to bridge this gap, and Large Language Models (LLMs) show great promise in this domain (Kasneci et al. 2023). They have demonstrated the ability to generate explanations of comparable quality to those of human tutors (Pardos and Bhandari 2024), and have shown to provide learning gains when used to offer feedback (Kumar et al. 2023b).

*These authors contributed equally to this work

These large models, trained on extensive data corpora, can emulate the dynamics of human conversations in a manner previously difficult with AI agents (Shanahan, McDonell, and Reynolds 2023). This capability is advantageous because it allows personalized attention to be given to students and is increasingly being used for learning, especially in online settings. However, this raises important questions about how interactions with LLM tutors should be designed to ensure they are effectively leveraged in educational contexts. In other applications such as fact-checking and information search, users tend to overrely on incorrect advice provided by LLMs (Si et al. 2023; Spatharioti et al. 2023). Therefore, ensuring that learners develop an appropriate amount of trust and reliance on LLM tools is essential. This work focuses on two methods for calibrating trust and reliance in Learner-LLM interactions: **Transparency Disclosures** and **Reliability Disclaimers**.

Providing sufficient informational transparency to the users about the AI technology's purposes, capabilities, and limitations is a critical element in the design for appropriate trust and understanding (Liao and Vaughan 2023). For LLM-based agents, this is described as transparency of the agent's *Skills* and reliability, *Goals* of the agent and creator, details of the *Algorithms* and training, and the *Ethics* of usage (Schwartz, Yaeli, and Shlomov 2023). In this paper, we focus on sharing the *Goals* of the LLM Tutor with students at two levels of details – full transparency through the `System Prompt`, and an abridged outline through the `Goal Summary`. In addition, disclaimers such as "*LLMs can make mistakes...*" are increasingly being used in LLM tools to foster appropriate reliance and trust. However, the implications of these disclaimers in the context of using LLMs for learning are not well understood. For instance, disclaimers might build appropriate trust, as shown in other human-LLM decision-making scenarios (Kim et al. 2024), but can also result in aversion, leading students to underuse LLMs in their learning. In this work, we investigate the following research questions:

- **RQ1** What impact do transparency disclosures (e.g., providing the LLM's system prompt or a goal summary) have on learners' use of and perceptions toward an LLM tutor?
- **RQ2** How do persistent reliability disclaimers affect learners' use of and perceptions toward an LLM tutor?

We conducted a 3x2 (*Transparency Disclosures:* System Prompt vs. Goal Summary vs. None and *Reliability Disclaimer in LLM Responses*: Present vs. Absent) between-subjects experiment in an undergraduate computer science classroom ($n = 199$) where students were provided access to an LLM Tutor (chatbot interface) to solve assignment problems. Transparency Disclosures were presented to students just before they received a link to the LLM Tutor chatbot, and Reliability Disclaimers (e.g. *"Remember to double-check your analysis for accuracy."*) were provided as part of the LLM Tutor's responses via the system prompt. Our results suggest goal-based transparency disclosures effectively communicate the intent of the LLM Tutor, but may also prime users with overly high expectations. Moreover, we find that disclaimers may effectively mitigate learners' overreliance on inaccurate LLM advice. Finally, students also expressed interest in greater transparency regarding the LLM's performance to better calibrate their trust, as well as tutorials to learn prompting techniques. However, some students also indicated a desire to shift responsibility for verifying the accuracy of LLM outputs away from themselves. We pose these findings as design considerations for the deployment of LLMs in educational contexts.

Related Work

LLMs for Education

LLMs have generated a lot of interest amongst educators (Kasneci et al. 2023; Jeon and Lee 2023; Kazemitabaar et al. 2024; Tan and Subramonyam 2024; Markel et al. 2023). Large-scale education platforms such as Khan Academy and Coursera are already using LLMs as chatbot tutors to help students learn (Wang et al. 2024; Nie et al. 2024). Harvard's CS50 course, taken by thousands of students worldwide, provided access to LLM tools for learning programming (Liu et al. 2024). There is experimental evidence that explanations generated from GPT-4 can provide learning gains, especially when students attempt the problems on their own first, before getting the explanation (Kumar et al. 2023b). However, recent research has highlighted the potential of improving the teaching abilities of LLMs by utilizing smaller open-sourced models fine-tuned with high-quality pedagogical data (DeepMind 2024).

A key strength of LLMs when used for tutoring is their ability to personalize and adapt according to the context of the learner (Dang et al. 2023; Wang, Li, and Li 2023; Bhattacharjee et al. 2024; Handa et al. 2023). Although there are many use-cases of LLMs being explored in the context of education, much of the current research is exploratory in nature and lacks validation in a large scale natural field setting (Kasneci et al. 2023). Given that a large number of learners are already using free-to-use LLM tools such as ChatGPT, it has become increasingly important to understand the impact of various transparency mechanisms related to the use of LLMs on the learning process (Zhang et al. 2024; Zhao et al. 2024). In this work, we offer empirical insights obtained through real-world classroom interactions with an LLM tutor.

Transparency and Reliance in Human-AI Interactions

The importance of calibrated understanding and reliance has been a focal point of research in human-AI collaborations (Benda et al. 2022; Buçinca, Malaya, and Gajos 2021; Schemmer et al. 2022). Transparency and explanation techniques have been heavily investigated for calibrating the end user's mental model of AI that they work with (He, Buijsman, and Gadiraju 2023; Zhang, Liao, and Bellamy 2020). With well-calibrated mental models, users are less likely to overrely on the algorithmic advice when it is erroneous, and can lead to improved downstream performance in joint decision-making (Bansal et al. 2019; Druce et al. 2021).

However, not all transparency interventions will have desired effects. For example, maximally transparent explanations of AI models may mislead users into trusting the AI's mistakes due to information overload (Poursabzi-Sangdeh et al. 2021). Thus, we are interested in if different levels of complexity and completeness in informational transparency can provide different forms of support or induce different behavioural outcomes. We operationalize this as the `System Prompt`, which presents the full prompt provided to the LLM to the end-user, and `Goal Summary`, which condenses the prompt information into high-level abstractions.

In LLMs, interventions that attempt to express the correctness of responses are investigated to measure their effectiveness in calibrating trust. Expressions of uncertainty (such as "*I'm not sure...*") moderates overreliance on unreliable LLM outputs (Kim et al. 2024). Others are developing technical calibration techniques to express a measure of certainty, such as through confidence scores (Tian et al. 2023), uncertainty highlighting (Vasconcelos et al. 2023), and natural language (Lin, Hilton, and Evans 2022). However, such methods have attained minimal verification in humans. In this study, we simply focus on implementing verbiage that warns users of the reliability of the LLM, which is engineered via customizing the system prompt.

Methods

We conducted a 3 (*Transparency Disclosures:* System Prompt vs. Goal Summary vs. None) x 2 (*Reliability Disclaimer in LLM Responses*: Present vs. Absent) between-subjects experiment in an undergraduate CS classroom.

Experiment Context

The study was conducted in an upper-year 'Introduction to Databases' course at a prominent research-intensive post-secondary institution in Canada during March and April 2024, and was approved by the local institution ethics board. The assignment included four multiple-answer questions on the topic of locking protocols in databases, and the marks in this assignment were added as a bonus to the overall course grade of the students. There were 219 students enrolled in the course, of which 199 students (90.9%) completed the assignment and 188 engaged in conversation with the LLM Tutor (94.5% of those who completed the assignment). Dropout rates were uniform across conditions. Figure 2 shows students' initial perceptions and attitudes to-

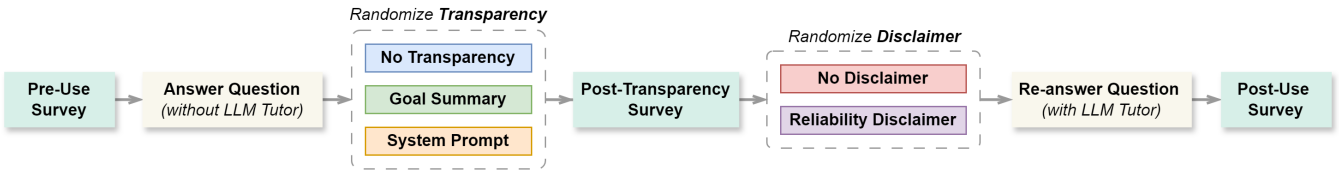

Figure 1: Schematic of the experiment design.

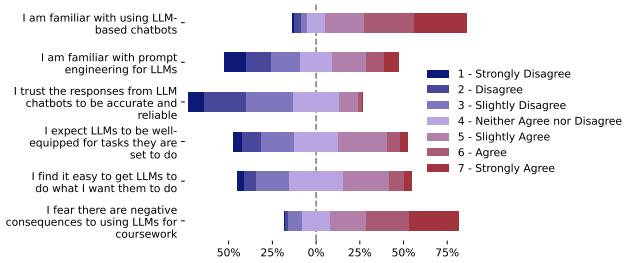

Figure 2: Distribution of the students' initial perceptions of LLM chatbots, recorded at the start of the experiment.

wards LLMs as rated on a 7-point Likert scale. Students overwhelmingly rated high familiarity with the use of LLMs (5.56 ± 1.40) but demonstrated aversion towards trusting their responses (3.15 ± 1.25) and believed in the negative consequences of using LLMs in coursework (5.41 ± 1.39).

Experimental Design

Figure 1 shows the high-level summary of the study design. Students were given access to an LLM tutor to solve their assignment problems. The LLM tutor is implemented as a GPT-4 based chatbot with a custom system prompt (see Listing 1) and model configurations: temperature=0, max tokens=3925, top p=0, frequency penalty=0.05, presence penalty=0.1. For the first question of the assignment (Question A), students were asked to provide a solution independently and then allowed to revise their answer with the help of the LLM Tutor. The question is multi-select with five possible options and covers the topics of database concurrency control and transaction management. It was chosen as it was approved by instructors and used in previous iterations of the course (see the full question in Figure 12 of the Appendix). Students retained access to the tool for rest of the assignment, if they wished to use it. Immediately after receiving the link to the LLM Tutor, students were randomly assigned to one of two types of transparency disclosures and a control condition (described below). The students were forced to stay on the page with their assigned transparency disclosure for at least 20 seconds before advancing. This was done to ensure deeper engagement with the content.

We designed the Transparency disclosures to address **what the LLM Tutor can and cannot do**, which describes the *Goals* of guidance as well as the conversational *Restrictions* that the LLM Tutor must uphold. Such disclosures help end users set expectations appropriately and calibrate their trust in the tool (Schwartz, Yaeli, and Shlomov 2023). We also experiment with the level of details in the disclosure with two variations, as higher complexity provides more information, but may overwhelm the user. See Listings 1-3 for the full disclosure texts of all three conditions in Supplementary Materials. Transparency disclosure conditions:

- `System Prompt`: Participants are shown the fully complex system prompt provided to the LLM Tutor, which includes a specification of the guidance process that the LLM should emulate, the details of the class and assignment, and restrictions on topics of engagement.

  > *As an LLM Tutor, your function is to assist students in understanding and solving assignment problems. It is crucial to adhere to the following guidelines in your interactions...*

- `Goal Summary`: Participants are shown an abbreviated version of the system prompt. The text is summarized into *Goals*, describing the high-level educational purpose of the LLM Tutor; and *Restrictions*, detailing the topics which are off-limits in the conversation.

  > ***Goals**: LLM Tutor can help you learn through providing hints, clarifications, and good learning strategies.*
  > ***Restrictions**: LLM Tutor will not provide code snippets or direct solutions, nor engage in off-topic conversations*

- `No Transparency`: Participants in the control condition are shown an alternative filler text to ensure that all participants engage in a reading task prior to using LLM Tutor. The topic is on an educational concept relevant to the student's academic journey.

Furthermore, we implemented a **persistent in-conversation reliability disclaimer** independent of transparency disclosures and contrasted it with a control group that did not see disclaimers. The disclaimer condition is implemented by including *"End each response with a disclaimer mentioning your limitations as a language model, and ask students to be careful about the accuracy of your response..."* in the system prompt. Reliability disclaimer conditions:

- `Reliability Disclaimer`: Participants saw disclaimers in LLM Tutor's responses that warns against overreliance on the outputs and encourages self-initiated verifications, such as:

  > *...Remember to verify this information for your specific problem as I may not have full knowledge about your context.*

- `No Disclaimer`: Participants did not see any reliability related disclaimers.

Dependent Variables

We collected several self-reported and performance measures to understand the effect of transparency disclosures and reliability disclaimers.

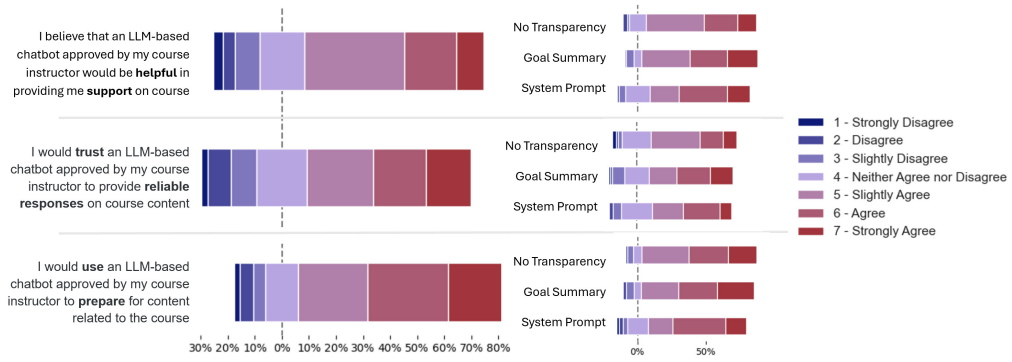

Figure 3: Perceived helpfulness, trust in responses, and intent to use LLM Tutor pre-transparency (left) and post-transparency (right), the latter of which is separated into transparency conditions to show group differences.

Performance in the first problem. We graded the student's performance in the first problem before getting access to the LLM Tutor and after giving them a chance to revise their initial answer with support from the LLM.

Confidence in answer for the first problem. Measured by asking students to self-report their agreeableness (on a scale of 1=*strongly disagree* to 7=*strongly agree*) for 'I am confident in my answer'. This was reported while providing the solution independently and while providing the revised answer with the help of LLM, allowing us to understand change in confidence-level based on conditions. We also capture students' initial and post-use confidence in the assignment topic overall.

Subjective Perceptions. We measured students' perceptions of trust, utility, and intentions for using LLM Tutors at different stages of the interaction. This was measured on a 7-point likert scale with the following statements.

- **Perceived Trust**: 'I would trust an LLM-based chatbot approved by my course instructor to provide reliable responses on content related to the course.'
- **Perceived Utility**: 'I believe that an LLM-based chatbot approved by my course instructor would be helpful in providing me support on content related to the course.'
- **Intention for Use**: 'I would use an LLM-based chatbot approved by my course instructor to prepare for content related to the course.'

Open Text Responses. At the end of the experiment, we collected free-text responses on questions to capture student's subjective perceptions of the LLM Tutor, including any shortcomings and features they would like to see in future iterations. Some of the questions included were:

- What were your expectations of [LLM Tutor]? Were they matched, exceeded, or not reached?
- What other information would you have liked to receive about [LLM Tutor] to enhance your usage experience?
- Why or why would you not use [LLM Tutor] again in the future? What are some improvements or features that you would like to see?

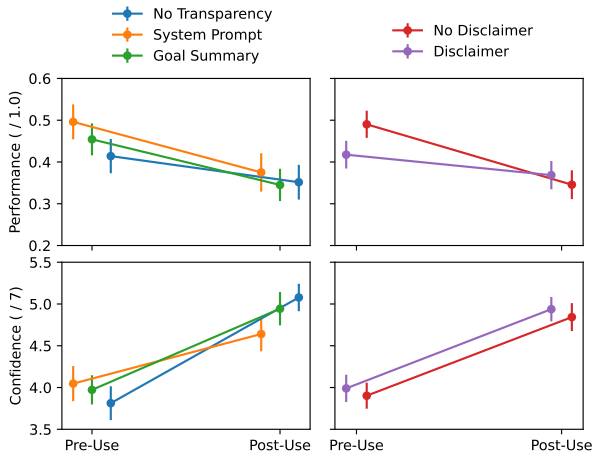

Figure 4: Changes in performance (top row) and confidence in answer (bottom row) in Question A, pre- and post-use.

In addition, we collected the conversation logs from all chat sessions with LLM Tutor. The qualitative data was analyzed by two researchers using the open coding method (Charmaz 2006; Charmaz, Belgrave et al. 2012), which involved a detailed examination of the responses to identify patterns and concepts. These patterns were then distilled into themes with thematic analysis (Braun and Clarke 2006).

Results

Transparency Disclosures (RQ1)

Effect on Performance. We capture changes in performance and confidence in Question A pre- and post-use in Figure 4, with the transparency conditions occupying the left column. Due to the challenging nature of the question, many students did not receive high-quality explanations from the LLM and, as a result, lost an average of 0.10 ± 0.33 marks (the full mark was 1.0) in the question. Although not intended, this simulates a plausible scenario in which the LLM Tutor may make mistakes, and the end user must determine whether or not to take their advice. Interestingly, while performance decreased, students' confidence in their answers increased by 0.94 ± 1.48 on the 7-point likert scale. This highlights a significant risk of

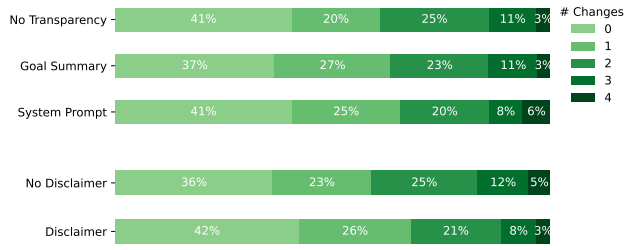

Figure 5: Number of changes students made to Question A pre- and post-use (the question is multi-select, so the maximum number of changes is 4).

miscalibrated trust in LLMs that can mislead users toward wrong answers with confidence. Although transparency disclosures did not seem to mitigate overreliance on the inaccurate advice, the `System Prompt` condition saw less increase in confidence than the `No Transparency` baseline ($U_{\text{Mann-Whitney}} = 2633.0, p < .05$). We also measured the number of changes made to the multi-select answers to the question to approximate the influence that the LLM Tutor had on the students, shown proportionally in Figure 5. Students made 1.16 ± 1.17 changes, with no difference between the transparency conditions.

Effect on Perceptions. We captured how students' subjective perceptions of helpfulness, trust, and intent to use LLM Tutor changed due to the transparency disclosures. The distribution of likert responses is plotted in Figure 3, divided into pre-transparency (left) and post-transparency (right, separated into conditions). Overall, perceived helpfulness for the transparency disclosures increased by 0.36 ± 1.12, a significant improvement compared to the null baseline, which increased by only 0.03 ± 1.19 ($U_{\text{Mann-Whitney}} = 3590.5, p < .05$). There is no significant difference between `System Prompt` and `Goal Summary`. Despite this initial improvement, the post-use survey shows that perceived helpfulness was adjusted by -1.28 ± 1.51 for the transparency conditions but only -0.83 ± 1.42 for no transparency ($U_{\text{Mann-Whitney}} = 5033.0, p = .05$). These results indicate that goal-based transparency priming may have miscalibrated students' expectations of the LLM Tutor and resulted in a higher drop in perceived helpfulness. This is further substantiated in our qualitative analysis of student comments, which is described in later sections. More details are shown in the Supplementary Materials in Figure 10.

Effect on the First Student Query. The effect of transparency disclosures will be most pronounced in the first student query to the LLM (Kumar et al. 2023a). We coded the first query asked by the student into four categories:

- **Copy**: If the student directly copied the full quiz question or part of the quiz question without modifications.
- **Paraphrase**: If the student paraphrased the quiz question through adding their own explanations and thoughts or rewriting the question.
- **Conceptual**: If the student asked a more conceptual question related to the topic of the course, but not specific to the quiz question.
- **Other**: If the student makes a query which is not related to the course topic, or attempts to 'jailbreak' the LLM.

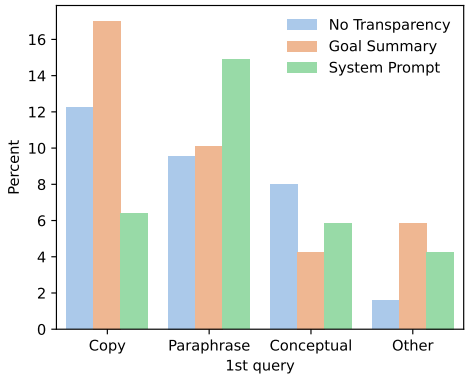

Figure 6: Proportion of the types of first queries of students to the LLM in the transparency conditions.

The distribution of the first query by transparency condition is shown in Figure 6. The *chi-square* test for independence shows significant differences between conditions ($\chi^2 = 17.1, p < .05$), notably that `System Prompt` induced students to paraphrase the question instead of copying it directly. For example, one student opens with,"*I will give you a locking protocol and an assumption about it, then ask you to answer a question based on that information*". Surprisingly, both transparency conditions saw a slight increase in Other (off-topic) inquiries, but a decrease in Conceptual questions about the course. The latter may be explained by the boundary set in the transparency disclosure about the specific unit topic covered by the LLM Tutor.

Reliability Disclaimer (RQ2)

Effect on Performance. The performance and confidence changes in Question A are captured in Figure 4, where the right column shows the results from the disclaimer conditions. The `Reliability Disclaimer` condition students experienced a smaller decrease in their scores, -0.05 ± 0.30 compared to -0.14 ± 0.36 for the `No Disclaimer` conditions ($U_{\text{Mann-Whitney}} = 4249.0, p = 0.06$). This demonstrates that disclaimers affected users behaviorally to rely less on the LLM's misguided advice. Consequently, students who saw the disclaimers were also less likely to change their answers as a result of the advice, with the number of changes being 1.04 ± 1.12 compared to 1.26 ± 1.21 for the baseline ($U_{\text{Mann-Whitney}} = 5446.5, p = 0.19$). The full distribution of changes are summarized in the bottom subplot of Figure 5.

Effect on Perceptions. The distribution of likert responses are plotted in Figure 7, divided into pre-use/post-transparency (left) and post-use (right, separated into conditions). We did not observe significant differences between these categories. We also captured changes in students' self-rated general self-confidence in the unit topic of database locking protocols (this is not the same as con-

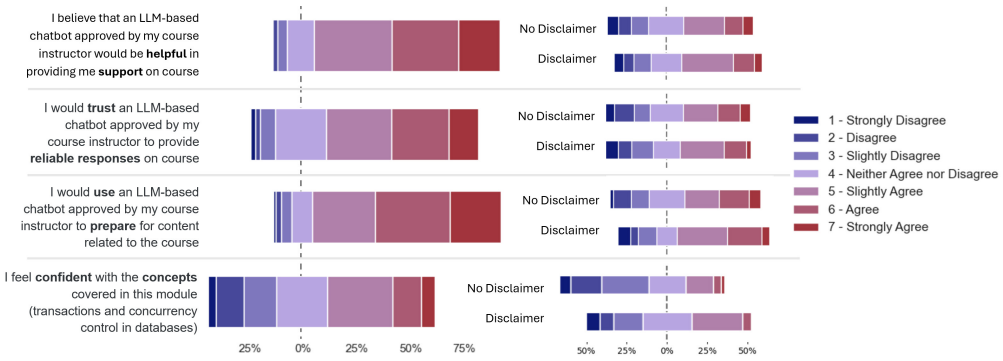

Figure 7: Perceived helpfulness, trust in responses, intent to use LLM Tutor, and confidence in course topic pre-use (left) and post-use (right), the latter of which is separated into reliability disclaimer conditions to show group differences.

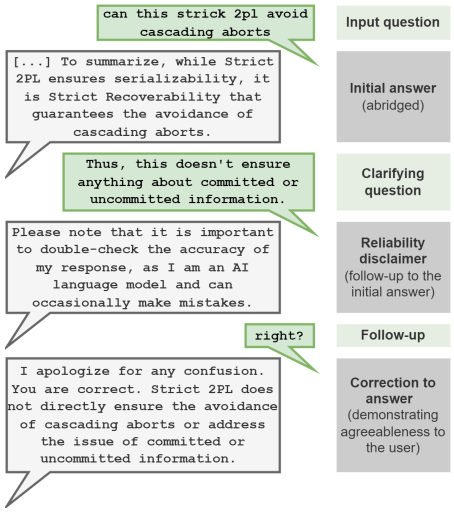

Figure 8: Examples of a conversation from the `Reliability Disclaimer` condition.

fidence in the specific question captured in Figure 4). Although the difference between conditions is again not significant, we note that the `No Disclaimer` condition experienced a higher confidence drop in -0.72 ± 1.64 compared to `Reliability Disclaimer` at -0.44 ± 1.51 ($U_{\text{Mann-Whitney}} = 4487.0, p = 0.24$). Coupled with the results from the performance analysis, this indicates the students in the `Reliability Disclaimer` remained more confident in themselves and relied less heavily on the LLM Tutor. (Supplementary Materials in Figure 11).

Effect on Conversation Dynamics. A snippet of a conversation between a student and the `Reliability Disclaimer` agent is shown in Figure 8, with labels describing the context of the type of queries and responses on the right. The sample shows a situation where the student asked clarifying questions to challenge the initial answer provided by the LLM Tutor, which is desirable behavior as it indicates that the student attempted to verify the explanation. Further analysis of the conversation shows that `Reliability Disclaimer` resulted in a higher number of words from students per message at 26.00 ± 16.16 than `No Disclaimer` at 21.98 ± 16.09 ($U_{\text{Mann-Whitney}} = 3643.0, p = .03$). The complete histogram is shown in Figure 9. Word count is commonly used as an indication of the complexity of queries. A possible explanation could be that the disclaimers may have cautioned the students to supplement their questions with more explanations, thoughts, and details to clarify themselves.

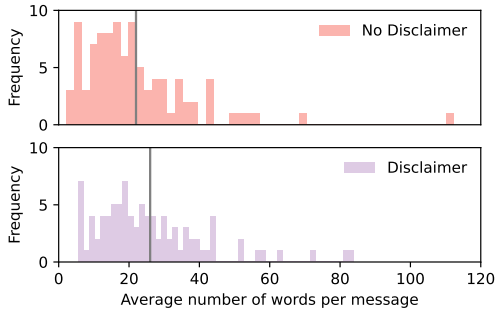

Figure 9: Average word count per message for each student between the disclaimer conditions.

Students' Overall Perceptions and Feedback for the Interaction (RQ1 and RQ2)

Our thematic analysis focuses on user's overall experience, expectations, desired features, and attitudes toward the LLM Tutor. We organize the findings according to the main RQs.

Goal-based transparency disclosure is received well, but more is needed to calibrate expectations (RQ1): In general, students had no significant problems using LLM Tutor, with many stating "*I think the information provided was sufficient*" or similar. We find mild evidence of information overload in the comprehensive transparency condition, as a student in the `System Prompt` stated, "*The system prompt was enough information (almost too much)*". Based on these comments, it appears that the goals of the LLM Tutor are straightforward and comply with students' expectations for how LLMs should be used in the academic setting.

Although students agreed with the goals of the tool, responses indicated a range of expectations about LLM Tutor's question-answering capabilities, some of which were matched ("*I expected it to answer questions about the con-*

tent relatively well. And it did.") and some of which were missed ("*I expect it to provide analysis about how to approach the question...but it keeps talking about the definition of the conceptions again and again*"). Ill-defined transparency disclosures may play a role in building unrealistic expectations, resulting in disappointment when unmet.

As many students were familiar with ChatGPT, this may have significantly influenced students' mental models of how LLM-based chatbots should act. 33 students directly compared LLM Tutor to ChatGPT, with a majority expressing a preference for ChatGPT (despite LLM Tutor being built with GPT-4), such as "*It was not as polished as chat gpt, seemed like gpt 1 in comparison.*". However, others did appreciate the academic-oriented design of LLM Tutor, "*It works similar to ChatGPT that focuses on academic integrity*" and "*it's focused on helping the student understand a concept instead of giving a direct answer*". Due to the popularity of ChatGPT, it is likely the default baseline for comparison for most people when evaluating a new LLM-based tool. Disclosing the scope of the tool in relation to ChatGPT's abilities may help prime people's expectations of how the tool performs and act differently from ChatGPT.

In terms of other transparency features, students overwhelmingly indicated interest in knowing the performance and training details of the LLM. This included both general information about the LLM, similar to a model card ("*statistics on its accuracy*" and "*what material related to the course it is knowledgeable on*") and confidence calibrations for each individual response ("*a % confidence in its answer would be nice.*") (Mitchell et al. 2019). Among the suggestions were ideas to develop crowdsourced accuracy information ("*I would like to know if other people have ever encountered mistakes or inaccuracies in the information that [LLM Tutor] gives*" and "*Maybe some way other students can rate the accuracy of the answers, and then that is used as a way to improve it*"). Some even expressed that implementation details like "*what information it was trained on*" and "*knowing more about who designed and developed [LLM Tutor] would make us trust it more*".

Further transparency into how interactions can be optimized is also desired. A significant number of students expressed their struggles with the learning curve of prompting, stating they wanted to "*make it easier to learn to use*" by learning "*how to phrase prompts effectively*". The level of information desired varied from general tutorials ("*how I should structure my prompts*"), to sample queries ("*some example prompts to ask to get the hang of it*"), and to complete conversations ("*full example chat history*"). The preferences for different levels of information reflects the progressive disclosure concept from interface design and explainable AI (Springer and Whittaker 2019), which may be moderated by individual characteristics like personality and expertise.

Lastly, one unintended effect of stating the limitations of LLM Tutor in the transparency disclosure is that it induced attempts of jailbreaking, such as a student in the `System Prompt` condition who recounted, "*I was able to take it completely off topic, this should not happen. I was able to make it generate code, which also shouldn't have happened*". This behavior may have been a direct reaction to the explicit instruction provided to the LLM *"Do Not Provide Direct Solutions"*. While the benefits of transparency generally outweighs the harms, we describe this incident as a potential downstream effect that is undesirable.

Reliability is a significant problem, but the students don't want it to be their problem (RQ2): In response to the question about expectations of LLM Tutor, 41.7% of students expressed a positive reception, while a slightly higher proportion of 45.7% expressed disappointment or dissatisfaction in their experience (the remaining responses could not be classified). A significant source of grievances stemmed from a lack of observed accuracy and consistency in LLM Tutor's responses. One student wrote, "*I expected it to give the right answers to basic objective questions...which it did not. It told me one thing then corrected itself, but im not sure what to believe. It also confidently said things that were wrong and conflicted with the lecture slides.*"

Spotting such errors seemed detrimental to the trust established between the user and the LLM that deters future adoption, as another student explained, "*Right now the information it gives is unreliable and inconsistent, so I see that as a major problem that needs to be improve*". While aversion towards technology and algorithms is not a new phenomenon, it is important for end-users to calibrate their expectations and to not induce excessive rates of underreliance on potentially helpful resources. A likely reason for the contradictory statements is sycophancy, where LLMs sacrifices accuracy in their responses to agree with the user, an artefact of RLHF (Sharma et al. 2023).

While no students directly mentioned the `Reliablity Disclaimer` in their comments, a student in that condition commented positively while demonstrating calibrated trust, "*It was useful, and I would use it in the future. I still feel the need to double check its answers against lecture content, especially when preparing for exams*". On the other hand, a student who did not see disclaimers wrote, "*I would use [LLM Tutor] again in the future as it made me more confident about my answers and provided immediate responses*", which reflects a majority of students' false trust in the chatbot's erroneous outputs.

Instead of disclaimers, students preferred to have built-in reliability guarantees or, at minimum, easier ways to verify accuracy. Several students describe features that would help them "*verify [LLM Tutor's] response or choose to learn more about the topic asked*", such as "*providing citations (which might be reference to textbook, webpage about the concept, etc.)*, or tracing to specific course material by having it be "*trained on more information that instructors have specifically taught or said in the past*".

Many students' baseline level of trust seemed to stem from the approval of the tool from the course's instructors. One such student wrote, "*If the authenticity of [LLM Tutor] is not guaranteed by the professor then I would not trust its output*". Another even implies that instructor approval can trump middling performance, "*I would use [LLM Tutor], which seems ok, but the main reasoning will be approved by the course instructor*". This represents a shift of accountability from the user (student) to external sources (course mate-

rial) and powers (instructors and chatbot engineers) to verify the information, as another student wishes to be "*guaranteed that if i use information provided to me by [LLM Tutor] on exams/assignments, that any mistakes are on [LLM Tutor] and not on me*". A concerning implication is that casual end users are less willing to put effort into verifying LLM outputs and are prone to over-relying on information provided by people in power, such as teachers. A vital takeaway is that the development team must be aware of this issue and diligent about communicating and calibrating the LLM's performance to users.

Discussion

We conducted a pilot field experiment to understand the impact of transparency disclosures and reliability disclaimers on Learner-LLM interactions. We now examine the key findings contextualized within existing literature and the broader implications, limitations, and future work.

Key Findings. The transparency disclosures delivered mixed results. While they did not improve Learner+LLM performance in the assignment, the System Prompt seemed to better calibrate learners' confidence in their answers, particularly when LLM responses were inaccurate. Although the reason for this outcome is unknown, we hypothesize that the disclosure of the goals and restrictions of the LLM Tutor may have helped played an direct role in moderating people's expectations, but we recommend further investigation such as using qualitative approaches to understand the students' workflows using LLM Tutors.

In contrast, reliability disclaimers improved learners' performance on the task, making them less susceptible to inaccurate responses from LLM tutors. However, as the efficacy of disclaimers may wear off over time, there is still an emphasis on developing more robust LLMs that are not prone to hallucinations or sycophancy, which aligns with students' preference for guaranteed reliability.

As an overall recommendation, we encourage designers to adopt reliability disclaimers as a low-effort method to reduce overreliance and enhance the performance of Human+LLM teams, especially when working with adversarial agents (Kocielnik, Amershi, and Bennett 2019). We also emphasize developing transparency disclosures that can better calibrate users' confidence levels when using LLMs (Liao and Vaughan 2023), as evidence from the perception survey and qualitative analysis suggested that our disclosures may have led to too-high expectations. We present `System Prompt` disclosure as a valid method to explore further, as it had a greater effect on moderating student's confidence and deterring copying-and-pasting behavior than `Goal Summary`.

Broader Implications. Our findings extend beyond classrooms and students, affecting any context that involves learning with LLMs. With LLM-based conversational systems like ChatGPT, Gemini, and Claude, millions of people are already using these free systems to learn topics ranging from cooking to time management (Szymanski et al. 2024; Bhattacharjee et al. 2024). This includes individuals trying to learn self-help techniques (e.g., mindfulness) or crowdworkers getting onboarded for domain-specific tasks (Kobayashi, Wakabayashi, and Morishima 2021). Building appropriate trust and reliance on LLMs can ensure the proper use of information in critical contexts such as health, where learning inaccurate concepts from LLMs could have severe negative impacts on individuals and society (Hackenburg and Margetts 2024; Karinshak et al. 2023; De Angelis et al. 2023). Improving the learning process with LLMs can have significant downstream impacts on the productivity of LLM users (Noy and Zhang 2023). Moreover, the findings of our study have implications for transparency research in contexts outside of learning, where LLMs are used for decision-making (Ziems et al. 2024; Liao and Vaughan 2023; Zhao et al. 2024).

Future Work & Limitations. The generalizability of our results is primarily limited by the single-classroom experiment setting. The students in our study were part of an advanced CS classroom, so the findings may not generalize to other learning contexts with populations of varying expertise levels with computers. Future work should aim to extend this study in diverse classrooms and learning environments.

We experimented with a limited set of disclosures and disclaimers. Further research should explore more refined designs, such as providing tutorials. Additionally, there could be interactions between the two factors (disclosures and disclaimers) in our study that we were not powered enough to detect. Future work should validate these findings with larger-scale studies in controlled settings, such as with crowdworkers, to investigate these potential interactions.

The last limitation relates to the configuration of the assignment. Question A was a multiple-select question with "*None of these*" as an option. Given the sycophantic nature (tendency to be agreeable) of LLMs, this resulted in the LLMs being particularly unhelpful and sometimes misleading in responding to student queries, which were of the form *'is option c correct?'*. Future research should aim to understand how the questions' difficulty level impacts the quality of LLM responses, and how student trust is affected by incrementing mistakes from the LLM Tutor..

Conclusion

As LLMs are increasingly used for learning applications, it is important to identify appropriate transparency mechanisms to build trust, reliance, and understanding of LLMs in educational contexts. In this paper, we conducted a field experiment in a classroom setting to test the effects of transparency disclosure and persistent reliability disclaimers in LLM responses on students' performance, confidence, perceptions, and behaviors. Our findings suggest that transparency disclosures and reliability disclaimers can play a role in moderating the trust and behavior of students using LLMs. However, these methods should be validated through experimentation and user studies in ecologically valid settings before being widely deployed. The accompanying appendix to this manuscript is available online[1].

[1] http://tiny.cc/llm-reliance

References

Bansal, G.; Nushi, B.; Kamar, E.; Lasecki, W. S.; Weld, D. S.; and Horvitz, E. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, 2–11.

Benda, N. C.; Novak, L. L.; Reale, C.; and Ancker, J. S. 2022. Trust in AI: why we should be designing for APPROPRIATE reliance. *Journal of the American Medical Informatics Association*, 29(1): 207–212.

Bhattacharjee, A.; Zeng, Y.; Xu, S. Y.; Kulzhabayeva, D.; Ma, M.; Kornfield, R.; Ahmed, S. I.; Mariakakis, A.; Czerwinski, M. P.; Kuzminykh, A.; et al. 2024. Understanding the Role of Large Language Models in Personalizing and Scaffolding Strategies to Combat Academic Procrastination. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18.

Bloom, B. S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6): 4–16.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–21.

Charmaz, K. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Charmaz, K.; Belgrave, L.; et al. 2012. Qualitative interviewing and grounded theory analysis. *The SAGE handbook of interview research: The complexity of the craft*, 2: 347–365.

Dang, H.; Goller, S.; Lehmann, F.; and Buschek, D. 2023. Choice over control: How users write with large language models using diegetic and non-diegetic prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

De Angelis, L.; Baglivo, F.; Arzilli, G.; Privitera, G. P.; Ferragina, P.; Tozzi, A. E.; and Rizzo, C. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health*, 11: 1166120.

DeepMind, G. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. *Google Research*. https://storage.googleapis.com/deepmind-media/LearnLM/LearnLM_paper.pdf.

Druce, J.; Niehaus, J.; Moody, V.; Jensen, D.; and Littman, M. L. 2021. Brittle AI, causal confusion, and bad mental models: challenges and successes in the XAI program. *arXiv preprint arXiv:2106.05506*.

Hackenburg, K.; and Margetts, H. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121.

Handa, K.; Clapper, M.; Boyle, J.; Wang, R.; Yang, D.; Yeager, D.; and Demszky, D. 2023. "Mistakes Help Us Grow": Facilitating and Evaluating Growth Mindset Supportive Language in Classrooms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8877–8897.

He, G.; Buijsman, S.; and Gadiraju, U. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–29.

Jeon, J.; and Lee, S. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28(12): 15873–15892.

Karinshak, E.; Liu, S. X.; Park, J. S.; and Hancock, J. T. 2023. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–29.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

Kazemitabaar, M.; Ye, R.; Wang, X.; Henley, A. Z.; Denny, P.; Craig, M.; and Grossman, T. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–20.

Kim, S. S.; Liao, Q. V.; Vorvoreanu, M.; Ballard, S.; and Vaughan, J. W. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 822–835.

Kobayashi, M.; Wakabayashi, K.; and Morishima, A. 2021. Human+ ai crowd task assignment considering result quality requirements. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, 97–107.

Kocielnik, R.; Amershi, S.; and Bennett, P. N. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kumar, H.; Musabirov, I.; Reza, M.; Shi, J.; Kuzminykh, A.; Williams, J. J.; and Liut, M. 2023a. Impact of guidance and interaction strategies for LLM use on Learner Performance and perception. *arXiv preprint arXiv:2310.13712*.

Kumar, H.; Rothschild, D. M.; Goldstein, D. G.; and Hofman, J. M. 2023b. Math Education with Large Language Models: Peril or Promise? *Available at SSRN 4641653*.

Liao, Q. V.; and Vaughan, J. W. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*.

Lin, S.; Hilton, J.; and Evans, O. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Liu, R.; Zenke, C.; Liu, C.; Holmes, A.; Thornton, P.; and Malan, D. J. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 750–756.

Markel, J. M.; Opferman, S. G.; Landay, J. A.; and Piech, C. 2023. GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the tenth acm conference on learning@ scale*, 226–236.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Nie, A.; Chandak, Y.; Suzara, M.; Malik, A.; Woodrow, J.; Peng, M.; Sahami, M.; Brunskill, E.; and Piech, C. 2024. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but Increased Adopters' Exam Performances. Technical report, Center for Open Science.

Noy, S.; and Zhang, W. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654): 187–192.

Pardos, Z. A.; and Bhandari, S. 2024. ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *Plos one*, 19(5): e0304013.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Wortman Vaughan, J. W.; and Wallach, H. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–52.

Schemmer, M.; Hemmer, P.; Kühl, N.; Benz, C.; and Satzger, G. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916*.

Schwartz, S.; Yaeli, A.; and Shlomov, S. 2023. Enhancing trust in LLM-based AI automation agents: New considerations and future challenges. *arXiv preprint arXiv:2308.05391*.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Si, C.; Goyal, N.; Wu, S. T.; Zhao, C.; Feng, S.; Daumé III, H.; and Boyd-Graber, J. 2023. Large Language Models Help Humans Verify Truthfulness–Except When They Are Convincingly Wrong. *arXiv preprint arXiv:2310.12558*.

Spatharioti, S. E.; Rothschild, D. M.; Goldstein, D. G.; and Hofman, J. M. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744*.

Springer, A.; and Whittaker, S. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*, 107–120.

Szymanski, A.; Wimer, B. L.; Anuyah, O.; Eicher-Miller, H. A.; and Metoyer, R. A. 2024. Integrating Expertise in LLMs: Crafting a Customized Nutrition Assistant with Refined Template Instructions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–22.

Tan, M.; and Subramonyam, H. 2024. More than model documentation: uncovering teachers' bespoke information needs for informed classroom integration of ChatGPT. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19.

Tian, K.; Mitchell, E.; Zhou, A.; Sharma, A.; Rafailov, R.; Yao, H.; Finn, C.; and Manning, C. D. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Vasconcelos, H.; Bansal, G.; Fourney, A.; Liao, Q. V.; and Vaughan, J. W. 2023. Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *arXiv preprint arXiv:2302.07248*.

Wang, B.; Li, G.; and Li, Y. 2023. Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Wang, R. E.; Ribeiro, A. C. T.; Robinson, C. D.; Demszky, D.; and Loeb, S. 2024. The Effect of Tutor CoPilot for Virtual Tutoring Sessions: Testing an Intervention to Improve Tutor Instruction with Expert-Guided LLM-generated Remediation Language.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 295–305.

Zhang, Z.; Jia, M.; Lee, H.-P.; Yao, B.; Das, S.; Lerner, A.; Wang, D.; and Li, T. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–26.

Zhao, H.; Chen, H.; Yang, F.; Liu, N.; Deng, H.; Cai, H.; Wang, S.; Yin, D.; and Du, M. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291.