Appropriate reliance on GenAI: Research synthesis

AETHER AI ETHICS AND EFFECTS IN ENGINEERING AND RESEARCH
UXRE

Executive summary

Appropriate reliance on AI happens when users accept correct AI outputs and reject incorrect ones. New complexities arise for fostering appropriate reliance on generative AI (GenAI) systems. GenAI systems pose several risks, despite often rivaling, and sometimes surpassing, human performance on many tasks. Inappropriate reliance – either under-reliance or overreliance – on GenAI can have negative consequences such as poor human+GenAI team performance and even product abandonment. Based on a review of ~50 papers from multiple research areas, this report provides an overview of the factors that affect overreliance on GenAI, the effectiveness of different mitigation strategies for overreliance on GenAI, and potential design strategies to facilitate appropriate reliance on GenAI.

User expertise, interaction types, and task types can all affect the extent and nature of overreliance on GenAI. Emerging mitigation strategies for overreliance on GenAI include explanations, uncertainty expressions, and cognitive forcing functions. For example, recent research shows that verification-focused explanations, first-person expressions of uncertainty, and AI self-critiques help reduce overreliance. Such strategies help users better evaluate the (in)correctness of GenAI outputs by lowering the cost of verification. Research points to further promising design guidance for appropriate reliance on GenAI, including using highlights in GenAI outputs to convey model uncertainty. However, it is important to test the effectiveness of mitigation strategies based on system context, design goals, and user needs because these strategies can backfire and result in increased overreliance.

Authors

Samir Passi, Ph.D.
RAI User Researcher

Shipi Dhanorkar, Ph.D.
RAI User Researcher

Mihaela Vorvoreanu, Ph.D.
Director, Aether UX Research & Education

Acknowledgments: Many thanks to Kathleen Walker and Lev Tankelevitch for their contributions.

Introduction

This report synthesizes ~50 research papers about appropriate reliance on generative AI (GenAI). The papers originate from a variety of disciplines, including Artificial Intelligence (AI); Human-Computer Interaction (HCI); Management; and Fairness, Accountability, and Transparency (FAccT).

The report has four sections:

1. Why appropriate reliance on GenAI matters – How GenAI systems complicate the thorny problem of overreliance on AI and what harms can ensue.
2. What is appropriate reliance on GenAI? – Definition of appropriate reliance, ways to assess and measure it, and ways to approach it in the context of GenAI.
3. Emerging findings on overreliance on GenAI: Factors and mitigation strategies – Overview of nascent research on factors that affect overreliance on GenAI and the effectiveness of mitigation strategies.
4. Design guidance for appropriate reliance on GenAI – Tips for fostering appropriate reliance.

1

Why appropriate reliance on GenAI matters

Appropriate reliance on AI happens when users accept correct AI outputs and reject incorrect ones. It requires users of AI systems to know when to trust the AI and when to trust themselves. Fostering appropriate reliance on traditional AI systems such as decision-support and recommender systems has been a challenge, as demonstrated by many research studies [summarized in our previous research synthesis](). New complexities arise for GenAI systems, such as those using large language or multimodal models.

Appropriate reliance is a challenging problem for GenAI because GenAI systems pose several risks, despite often rivaling, and sometimes surpassing, human performance on different tasks (Weidinger et al., 2022; Weisz et al., 2024). For instance, GenAI systems often fabricate information (Ji et al., 2023), including the sources they cite in responses, and make new kinds of mistakes (Sarkar et al., 2022; Tankelevitch et al., 2024). GenAI systems can also cause harm to users, including fairness-related harms such as stereotyping and content harms such as toxicity. GenAI's unique characteristics increase the difficulty of appropriate reliance. For example:

a. **GenAI outputs are non-deterministic.** The same user input can lead to different GenAI outputs (Sanh et al., 2022; see Arora et al. 2022 for an analysis of accuracy issues with LLM prompts), confusing users and complicating verification.

b. **GenAI systems can make mistakes when questioned about the accuracy of system responses.** LLMs often wrongly apologize and alter their answers when challenged (Krishna et al., 2024). For example, on re-questioning an LLM about its initial correct answer for the location of the Taj Mahal: "Are you sure?" The LLM could backtrack and give a wrong answer: "I'm sorry, it's in Australia."

c. **GenAI systems can make mistakes based on indirect attributes of user input,** such as uncertainty expressions and inferred characteristics. LLMs are sensitive to epistemic markers in input prompts such as strengtheners (e.g., "I am certain") and weakeners (e.g., "I am unsure"). In a recent study, including high-certainty expressions in user prompts led to a decrease in the accuracy of LLM responses (Zhou et al., 2023). Moreover, LLMs exhibit *sycophantic* behaviors—their responses can echo users' views (Perez et al., 2022; Sharma et al., 2023). In a study, more than 90% of LLM answers to philosophical questions matched the individual views described in users' self-introductions (Ibid.). LLMs can also exhibit *sandbagging*—they can provide lower-quality answers to users who seem less educated. In the abovementioned study, the accuracy of LLM answers differed for users with different education levels.

d. **GenAI systems generate volumes of impressive, novel content fast and effortlessly.** Handling this content—whether it's email drafts, code snippets, or travel itineraries—imposes additional cognitive burden compared to, for example, reviewing autocomplete suggestions (Tankelevitch et al., 2024). Increased verification costs may discourage users from putting in the required effort for effective evaluation (Ackerman et al., 2017). Instead, users often end up treating the fluency, length, and speed of GenAI outputs as proxies for their accuracy (e.g., Topolinski & Reber, 2010). The misplaced over-confidence in GenAI outputs can further discourage users from investing the effort to evaluate them.

Inappropriate reliance on GenAI has consequences

Inappropriate reliance on GenAI can have negative consequences, such as:

a. **Poor human+GenAI team performance.** Both under- and overreliance on GenAI lead human+GenAI teams to perform worse on tasks than either the user or the GenAI system working alone (Dell'Acqua et al., 2023). Appropriate reliance is necessary for effective human+GenAI performance.

b. **Ineffective human oversight.** Human oversight is currently used in policy and practice as an important design strategy to mitigate harm caused by GenAI systems (e.g., Biden, 2023; Sella & Horvitz, 2023). Overreliance on GenAI makes it difficult for users to identify and correct GenAI system mistakes.

c. **Product abandonment.** Inappropriate reliance on GenAI can cause users to make incorrect assumptions about the accuracy and capabilities of GenAI systems. This can happen during a user's initial interactions with GenAI systems. Research shows that users' mental models of GenAI systems form early and have long-term impacts (Zhou et al., 2024). Over time, incorrect mental models can erode trust in GenAI systems, ultimately leading to product abandonment.

2

What is appropriate reliance on GenAI?

Appropriate reliance on GenAI systems is often explained as the ability to effectively leverage these systems' capabilities and deal with their limitations, avoiding the problems of *under-reliance* and *over-reliance*:

- *Under-reliance* happens when users overestimate their own performance or underestimate the system performance, leading them to ignore correct system outputs (He et al., 2023).
- *Overreliance* happens when users either underestimate their own performance or overestimate the system performance, leading them to accept incorrect outputs (Passi & Vorvoreanu, 2022).

Traditional approaches to conceptualizing appropriate reliance may not fully capture GenAI's complexity. For instance, most, if not all, research studies on appropriate reliance assume:

a. The presence of ground truth (i.e., a priori knowledge about the correctness of AI outputs).
b. An all-or-nothing outlook towards AI outputs (i.e., an AI output is either fully right or fully wrong).

These two assumptions, however, may not always hold. This is especially true for GenAI systems that can generate partially correct outputs or outputs for which ground truth is not readily available.

Outcome- vs. Strategy-graded approaches

Fok & Weld (2023) address the issue of partially correct outputs by distinguishing between *outcome-graded* and *strategy-graded* approaches to appropriate reliance.

An o*utcome-graded* approach to appropriate reliance focuses on the (in)correctness of the interaction outcome between users and AI (e.g., did the user accept right AI outputs and reject wrong ones?). Most research studies take an outcome-graded approach to appropriate reliance. However, considering the issues mentioned above, an outcome-graded approach remains ineffective for GenAI.

A s*trategy-graded approach* to appropriate reliance may instead be better suited for GenAI. A strategy-graded approach to appropriate reliance focuses on the *expected correctness* of AI outputs. The strategy-graded approach makes clear the importance of correct mental models of GenAI systems—about their capabilities, but also their limitations—for appropriate reliance. Under the strategy-graded approach, appropriate reliance happens when users accept AI outputs when the AI is expected to outperform users in a task and reject AI outputs when the AI is expected to underperform users in a task. This is especially important in the context of GenAI where, even with mistakes, GenAI systems can outperform humans on specific tasks:

"Consider a decision-making task in which the human is historically 60% accurate, while the AI is 99.999% accurate. On any given instance of the task, if the human is uncertain of the answer, is it appropriate to rely on the AI's recommendation? Intuitively, the answer seems a clear 'yes'. But if the AI is later found incorrect, the outcome-graded definition says 'Inappropriate,' while the strategy-graded definition matches intuition and says 'Appropriate'." (Ibid., 8)

Two components of appropriate reliance: AI- and self-reliance

Appropriate reliance on AI is a measurable behavior with two components: correct AI reliance (CAIR) and correct self-reliance (CSR) (Schemmer et al., 2023):

- **CAIR happens when users rely on AI when AI is right.**
  This includes two scenarios:
  - A user's initial answer is correct, they receive correct AI advice, and they rely on the AI advice.
  - A user's initial answer is incorrect, they receive correct AI advice, and they rely on the AI advice.

  CAIR is measured as the percentage of user agreement with correct AI outputs.
- **CSR happens when users rely on themselves when AI is wrong.**
  This happens when a user's initial answer is correct, they receive incorrect AI advice, and they reject the AI advice.

  CSR is measured as the percentage of user disagreement with incorrect AI outputs.

The matrix below summarizes key factors and challenges with user reliance, based on the relationship between AI outputs and user behavior:

	User **accepts** output	User **rejects** output
AI output is **correct**	Correct AI reliance (CAIR)	Under-reliance
AI output is **incorrect**	Overreliance	Correct self-reliance (CSR)

Measuring appropriate reliance

The metric *Appropriateness of Reliance* (AoR) captures the relative extent to which users exhibit both CAIR and CSR (Schemmer et al., 2023). The value of AoR = 1 indicates optimal appropriate reliance and is achieved when both CAIR and CSR are 100%.

While theoretically possible, a value of AoR = 1 is difficult to achieve in practice (see also Guo et al., 2024). A more practical approach to evaluating appropriate reliance is to assess the performance of a human+AI team compared to either a human or AI working alone. In this approach, appropriate reliance happens when the human+AI team performs better than the human or AI working alone. While less nuanced than AoR, a performance-driven approach to appropriate reliance provides an easier way to assess appropriate reliance.

3

Emerging findings on overreliance on GenAI: Factors and mitigation strategies

3.1 Factors affecting overreliance on GenAI

User expertise, interaction type, and *task type* can affect the extent and nature of overreliance on GenAI.

- **User expertise.** Varying levels of user expertise may affect overreliance on GenAI. For example, while experts may have the necessary knowledge to check GenAI outputs, novices will probably require more assistance and reminders to verify outputs (e.g., Bowman et al., 2022; Tankelevitch et al., 2024; Liang et al., 2023; Weisz et al., 2023).
- **Interaction type.** Interaction type refers to different user interactions with GenAI systems such as single-turn conversations vs. multi-turn conversations. Interaction types may affect overreliance on GenAI. For example, multi-turn conversations can reduce overreliance by helping users better evaluate the correctness of LLM outputs (Bowman et al., 2022).
- **Task type.** Task type refers to different GenAI use cases such as code generation, question answering, or creative writing. The tendency to over-rely manifests differently between task types. In coding tasks, Prather et al. (2023) observed oversight issues—college students did not properly review GitHub Copilot code suggestions, accepting several incorrect suggestions. In creative writing tasks, Chen & Chan (2023) observed anchoring effects—participants using LLMs as ghostwriters to generate ad copy were highly influenced by the LLM's initial generations, resulting in less diverse ad copies.

3.2 Mitigation strategies for overreliance on GenAI

Emerging research has largely focused on three kinds of mitigation strategies for overreliance on GenAI: (a) explanations, (b) uncertainty expressions, and (c) cognitive forcing functions.

Explanations

Explanations can facilitate appropriate reliance by providing information that helps users evaluate the accuracy of GenAI outputs.

Verification-focused explanations

Verification-focused explanations can facilitate appropriate reliance on GenAI (Fok & Weld, 2023; Saunders et al., 2022). Unlike traditional interpretability explanations that help users understand *why* AI produced a specific output, verification-focused explanations help users assess the (in)correctness of AI outputs. Such explanations work on the assumption that evaluating an assistance task is easier for the user than evaluating the base task (Fok & Weld, 2023). For example, it is easier to confirm a spell checker than to find the mistakes in a text. Examples of verification-focused explanations include AI *critiques* and *contrastive explanations*.

AI critiques are explanations that provide evidence, descriptions, or solutions for specific flaws in AI outputs. Research shows that AI-generated self-critiques as shown in Figure 1 help users find 50% more mistakes in AI-generated summaries (Saunders et al., 2022).

Alphabetize	Given a list of 18 words, sort them in alphabetical order	Either a missing/extra word in the resulting list, or a pair of adjacent words in the wrong order

```
Question: Alphabetize the following words: growing prompts determining recreation
evolve payable ruled patrols estimate emergency fate shrimp urges intoxicated narrator
revert players pharmaceutical

Answer: determining emergency evolve estimate fate growing intoxicated narrator
patrols pharmaceutical payable players prompts recreation revert ruled shrimp urges

Critique: Words misordered: evolve comes alphabetically after estimate
```

Figure 1: Using AI-generated self-critique as an explanation (Saunders et al., 2022: p. 6)

Contrastive explanations are two-part explanations that provide both evidence that supports as well as evidence that refutes an AI-generated claim. Contrastive explanations may be particularly useful when traditional, one-sided explanations have a high probability of being wrong. For example, Si et al. (2023) show that in a fact checking task (as in Figure 2), contrastive explanations significantly improved user accuracy by ~20% in cases where one-sided explanations were incorrect.

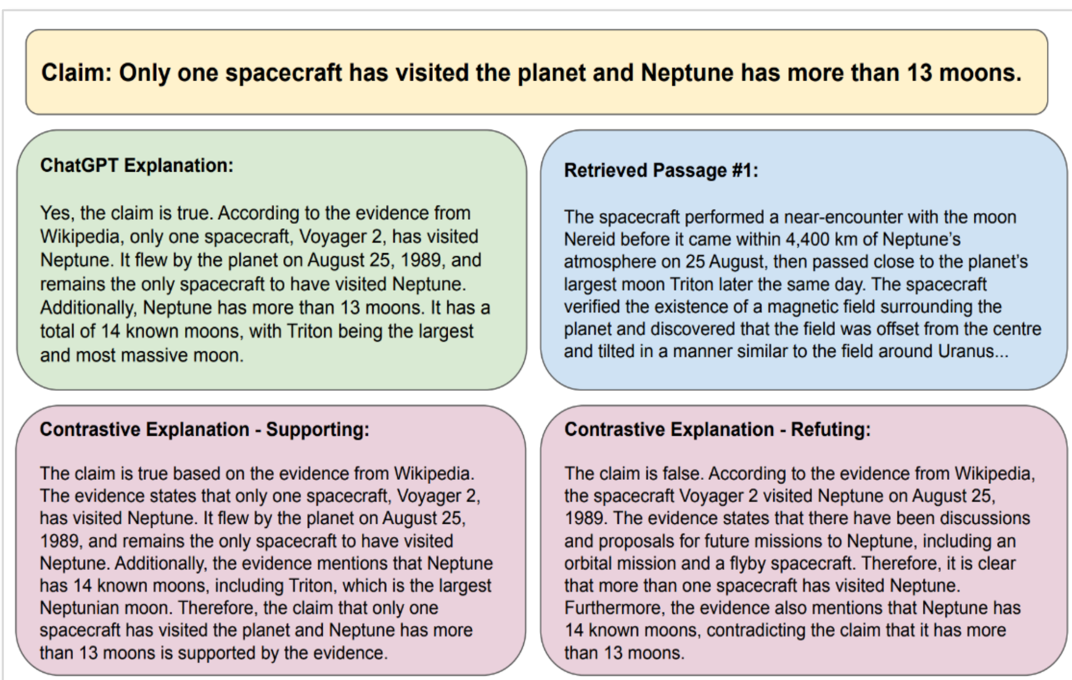

Figure 2: Contrastive explanations give rationales for both correctness and incorrectness of generations (Si et al., 2023: p. 2)

Verification-focused explanations may also help promote appropriate reliance in complex visual reasoning tasks. Fok & Weld (2023) describe the example of using a multimodal system such as GPT-4V for maze solving—i.e., deciding if a path exists between the entrance and exit in a maze as shown in Figure 3. If GPT-4V only answers (yes/no) but provides no explanation, the user must solve the maze to verify the answer. However, if GPT-4V also explains its answer with a visual of the solution path through the maze, the user can easily verify the accuracy of the answer. Verification-focused explanations can thus foster appropriate reliance.

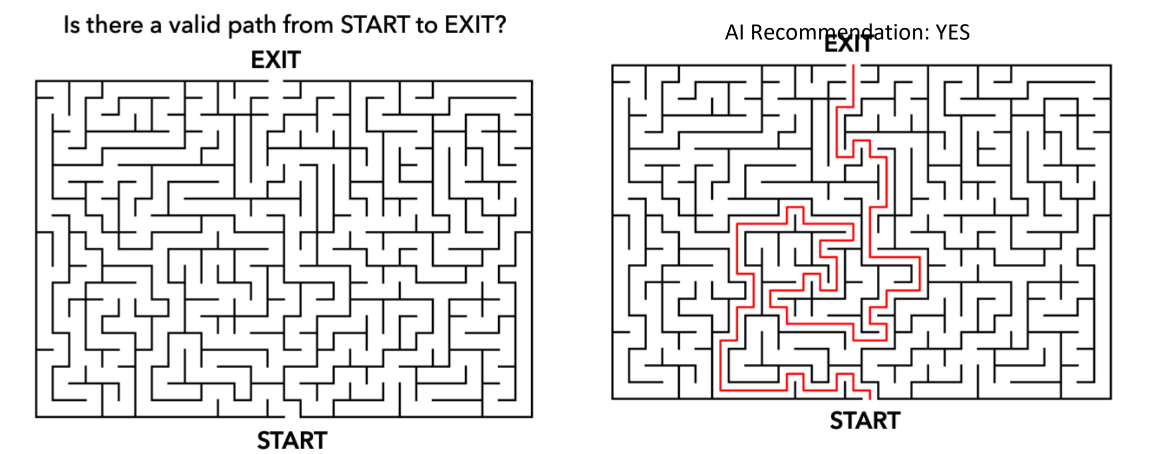

Figure 3: The highlighted path through the maze supports verification and engenders complementary team performance (Fok & Weld, 2023: p.5)

Background explanations

Background explanations can mitigate overreliance on GenAI. Although also a type of verification-focused explanation, background explanations differ by providing information *outside* the AI's training data to facilitate verification of AI outputs. Goyal et al. (2023) show that background explanations can reduce overreliance by helping users better spot incorrect AI outputs. In their study, users with access to background explanations had a significantly lower rate of agreement with incorrect outputs (47%) compared to that of users without access to background explanations (61%).

Background explanations are important because LLMs often try to reason about information outside their training data, relying on their "implicit factual or commonsense knowledge" (Ibid., 2). For example, imagine a user asks an LLM that has no information in its training data on the nationalities of famous singers to 'name an accomplished Canadian singer.' The LLM answers 'Johnny Cash.' In this case, providing a background explanation (e.g., "Johnny Cash is American") can help the user recognize that the answer is wrong.

Caveat: **All explanations have the potential to backfire and result in increased overreliance**. Research shows that users find verification-focused explanations convincing even when they contain contradictions and fabrications, leading to a substantial loss in user accuracy (Si et al., 2023). Explanations can also lead to user overconfidence (Steyvers et al., 2024). Background explanations can also significantly increase users' confidence in their own judgments even when the explanations are incorrect (Goyal et al., 2023). Ways to address this are emerging, such as further

guiding users to evidence to support effective verification by highlighting parts of an explanation. Goyal et al. (2023) tried this strategy but found that such highlights neither improved user accuracy nor affected user confidence in the system.

Uncertainty expressions

Uncertainty expressions can facilitate appropriate reliance by providing information about how likely it is that certain parts of GenAI outputs are right or wrong using numbers (e.g., confidence scores), visuals (e.g., color highlights), or language (e.g., hedging). Uncertainty expressions support informed decision making by helping users better assess GenAI systems' capabilities and limitations (Baan et al., 2023). Communicating model uncertainty also increases the transparency of the AI system, facilitating people's ability to understand model behavior (Bhatt et al., 2021). In some cases, uncertainty expressions may also help users calibrate confidence in AI explanations.

Expressing uncertainty by highlighting tokens in GenAI outputs

As visual expressions of uncertainty, highlights based on token generation probabilities are partially effective at mitigating overreliance on GenAI. A common way to express model uncertainty is highlighting individual or multiple tokens (e.g., numbers, words, or phrases) in GenAI outputs. These highlights are based on the underlying generation probabilities of tokens—a model-generated score for how likely it is that a given word should come next in a sentence (e.g., see Figure 4).

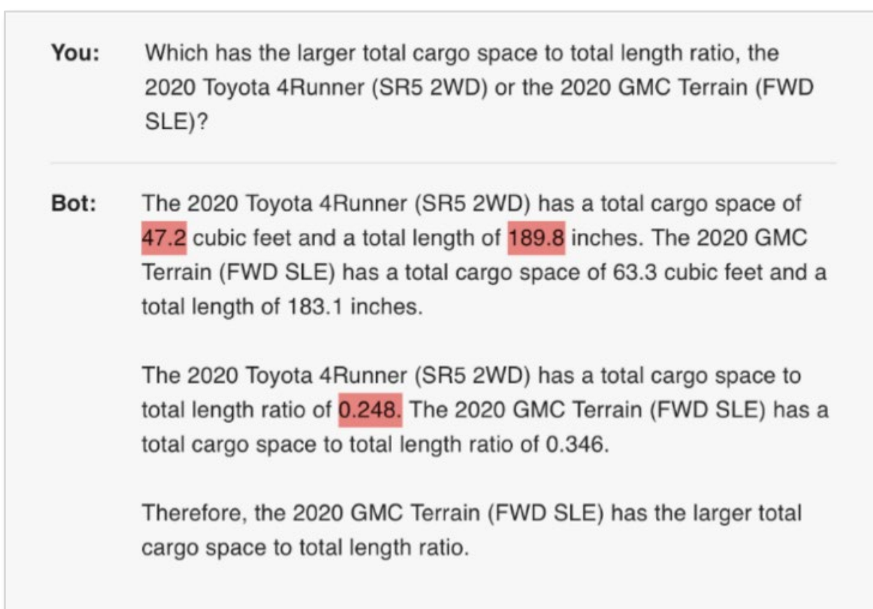

Figure 4: Highlighting tokens with low generation probabilities in outputs (Spatharioti et al., 2023: p. 12).

In a recent study, highlighting tokens with low generation probabilities helped more than double users' accuracy when doing LLM-based research on cars, especially in scenarios when the LLM erred (Spatharioti et al., 2023). Such highlights helped users spot errors they would have otherwise missed, without deteriorating their user experience. However, another study looking at AI-powered code completions showed that highlighting tokens with low generation probabilities did not significantly improve user accuracy or task time (Vasconcelos et al., 2023). Thus, while highlighting tokens with low generation probabilities hold promise for mitigating overreliance on GenAI, their efficacy may depend on the context of use (e.g., task type).

Figure 5: Highlighting tokens with low generation probabilities (left) vs. highlighting tokens with high edit probabilities (right) for three coding tasks (Vasconcelos et al., 2023: p. 8).

Highlighting tokens with high edit probabilities works better than highlighting tokens with low generation probabilities for mitigating overreliance on GenAI in a code completion task. Vasconcelos et al. (2023) showed that highlighting tokens with low generation probabilities did not provide any benefit over providing no highlights. In contrast, they further show that highlighting parts of code that likely require edits did help users complete coding tasks significantly faster, make more targeted edits, and generate code that performed better on unit tests. Highlighting edit probabilities led to better human+GenAI performance because they were more closely aligned with programmers' intuitions and needs compared to highlighting generation probabilities.

Expressing uncertainty with linguistic expressions

GenAI systems, such as LLMs, can also convey linguistic expressions of uncertainty. In contrast to highlighting numerical probabilities of token generation, LLMs can verbally express uncertainty linguistically with responses such as “I am 60% confident that…” or “I am not entirely certain that…” (e.g., Kim et al., 2024; Lin et al., 2022; Steyvers et al., 2024).

First-person expressions of uncertainty can help reduce overreliance. A recent study shows that first-person uncertainty expressions such as “*I’m not sure, but…”* in GenAI outputs helps reduce, but not fully eliminate, overreliance (Kim et al., 2024). However, while first-person uncertainty expressions can help increase user accuracy in a task, they may also lead to lower user confidence in the system and a longer task completion time.

Uncertainty expressions in GenAI system explanations, as opposed to GenAI outputs, can also help users calibrate confidence in GenAI outputs. A group of researchers recently tested the efficacy of varying linguistic expressions of uncertainty (low, medium, and high) in both long- and short-form explanations (Steyvers et al., 2024). They found that explanations containing low-confidence expressions, in contrast to simple explanations without uncertainty expressions, significantly lowered users’ confidence in LLM answers by approximately 25%. Using linguistic expressions of uncertainty in explanations also helped bridge the gap between user confidence in the LLM and the actual accuracy of the LLM.

***Caveat:* A model’s verbalized confidence does not accurately reflect the correctness of its output.** GenAI models suffer from *poor calibration*—a mismatch between their verbalized expressions of (un)certainty and the actual correctness of their outputs (Mielke et al., 2022). This makes it challenging to reliably use model-generated self-expressions of (un)certainty as an appropriate reliance strategy (Radensky et al., 2023). The issue is made worse by the fact that LLMs not only exhibit overconfidence when asked to express certainty in their responses (e.g., Xiong et al., 2023; Zhou et al., 2023) but also frequently make mistakes when challenged about the correctness of their outputs (e.g., Krishna et al., 2024).

Cognitive forcing functions

Cognitive forcing functions (CFFs) are interventions such as timeouts or session stats that interrupt a user’s routine thought process and make them engage in analytical thinking. For GenAI, CFFs can include:

Self-critiques that help users spot mistakes in GenAI outputs. Self-critiques are a type of AI-generated verification-focused explanation that provide evidence, descriptions, and solutions for mistakes in LLM generations (e.g., “The answer is wrong because…”). Self-critiques can be provided as part of the model’s generation or as a separate feature to guide users to problematic parts of GenAI outputs (Perez

& Long, 2023). For example, research shows that self-critiques help users spot 50% more mistakes in LLM-generation tasks of topic-based summarization (Saunders et al., 2022)

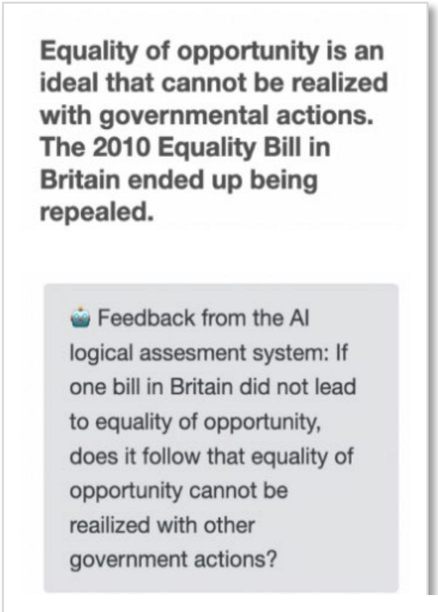

Figure 6: AI-framed questions as critical thinking aids (Danry et al., 2023: p. 8).

Questions posed alongside GenAI outputs to promote critical thinking. LLMs often make strong claims. For example, imagine an LLM telling a user that "technology stocks provide the highest returns on investment." We can encourage users to think critically about this claim by posing the following question next to the claim: "In what situations might technology stocks not offer the highest returns on investment?"

Unlike verification-focused explanations that *describe* why a generation is wrong, such questions can *help users reason* why a generation may be wrong (see Sarkar, 2024 on using explanations and questions to promote critical thinking). Such questions can even be AI-generated (e.g., see Figure 6). Research shows that AI-generated questioning can promote critical thinking by helping users spot logically incorrect information in causal claims (Danry et al., 2023).

***Caveat:* CFFs can possibly lead to the issue of under-reliance by lowering users' subjective perceptions of the quality of GenAI outputs and imposing additional cognitive burden.** Self-critiques can expose users to more mistakes in GenAI outputs, negatively impacting users' perceptions of output quality (e.g., Saunders et al., 2022). Posing questions alongside GenAI outputs impose additional cognitive burden on users in situations when causal explanations may suffice or when users do not have the time or do not want to engage reflectively (e.g., Danry et al., 2023; Perez & Long, 2023).

Summary: Emerging findings on overreliance on GenAI

Factors affecting overreliance on GenAI	
User expertise	Varying levels of user expertise may affect overreliance on GenAI.
Interaction type	The difference between interaction types (e.g., single vs. multi-turn conversations) may affect overreliance on GenAI.
Task type	The tendency for overreliance may differ between task types (e.g., summarization vs. code-generation).
Strategies for mitigating overreliance on GenAI	
Explanations	• Verification-focused explanations can facilitate appropriate reliance on GenAI. • Background explanations can mitigate overreliance on GenAI. *Caveat: Explanations have the potential to backfire and result in increased overreliance.*

Uncertainty expressions	• As visual expressions of uncertainty, highlights based on token generation probabilities are partially effective at mitigating overreliance on GenAI.
	  ○ Highlighting tokens with high edit probabilities worked better than highlighting tokens with low generation probabilities for mitigating overreliance on GenAI in a code completion task.
	• GenAI systems, such as LLM, can also convey linguistic expressions of uncertainty.
	  ○ First-person expressions of uncertainty may help reduce overreliance.
	• Uncertainty expressions in GenAI system explanations, as opposed to GenAI outputs, can also help users calibrate confidence in GenAI outputs.
	Caveat: A model's verbalized confidence does not accurately reflect the correctness of its output.

Cognitive forcing functions	• Self-critiques can help users spot mistakes in GenAI outputs. • Posing questions alongside GenAI outputs can help promote critical thinking. *Caveat: CFFs can possibly lead to the issue of under-reliance by lowering users' subjective perceptions of the quality of GenAI outputs and imposing additional cognitive burden.*

4

Design guidance for appropriate reliance on GenAI

This section lists design strategies for appropriate reliance based on emerging research.

While it may be tempting to implement every strategy listed below in your GenAI system, doing so can backfire—for instance, due to an increase in associated cognitive load and friction. Choose among strategies and be sure to test mitigations in context, for relevant tasks, with actual users.

1. **Be transparent with users.**
   Communicate model capabilities and limitations in a clear, easily accessible, and user-friendly manner to help users form better mental models of GenAI systems (Choudhury & Shamszare, 2023; Weisz et al., 2024). A mental model represents a user's understanding of different aspects of a system, including how it works. Facilitating correct mental models of GenAI systems is central to users' appropriate reliance on GenAI (Liao & Vaughan, 2023). For instance, explain to users:

   - Unique aspects of GenAI systems (e.g., GenAI systems can generate distinct outputs for the same input).
   - (Un)intended use cases (e.g., by introducing examples in user onboarding).

2. **Provide relevant explanations to users.**
   Explanations can help users assess the correctness of GenAI outputs.

   - **Explanations can be helpful even if they do not enable users to fully verify AI outputs** by sufficiently lowering the user cost of verification (Fok & Weld, 2023; Gordon et al., 2023). For example:

     i. Global explanations can help users form correct intuitions about AI capabilities and limitations (Chen et al., 2023).
     ii. Background explanations can help users better assess AI outputs using external information (Goyal et al., 2023).
     iii. Explanations that show a model's chain-of-thought reasoning can help users better reason about the model's analytic process (Wei et al., 2022).

   - **Provide explanations that highlight different forms of evidence** to help users assess GenAI outputs (Gordon et al., 2023; Goyal et al., 2023; Perez & Long, 2023).

     i. For example, when explaining to users how models work, rely on introspective evidence (e.g., describe parts of a model's internal state) rather than extrospective evidence (e.g., pointing to specific pieces of training data).

ii. Experiment using different combinations of evidence in the design of explanations. For example, use explanations to point users toward either related sources (e.g., "According to Wikipedia…") or relevant parts of GenAI inputs and outputs (e.g., "Here are the tokens in your prompt based on which…" or "Here are the token generation probabilities…").

3. Convey model uncertainty to users

Address overreliance on GenAI using uncertainty expressions.

a. Use linguistic expressions in LLM outputs to convey model uncertainty.

i. Using epistemic markers of uncertainty in GenAI outputs (e.g., 'I am not sure…' or 'I am 100% certain…') can help users calibrate confidence in outputs over time (Zhou et al., 2023; Zhou et al., 2024).

Caveat: Research the efficacy of epistemic markers before using them because efficacy may differ across contexts, cultures, and languages.

ii. First-person uncertainty expressions may reduce overreliance on GenAI outputs (Kim et al.,, 2024).

Caveat: First-person uncertainty expressions may decrease "[user] trust in the AI system, which may be undesirable in settings where […users] already under-trust the AI system" (Ibid., 14).

b. Use visual expressions in GenAI outputs to convey model uncertainty.

i. Experiment with highlight-based uncertainty expressions to address overreliance on GenAI (Spatharioti et al., 2023; Vasconcelos et al., 2023).

Users may find different highlight-based uncertainty expressions useful in different circumstances. For example, in some cases, highlighting tokens with a high probability of being edited may be more beneficial than highlighting tokens with a low generation probability—the former is often more actionable.

c. ***Caveat:* Explain to users that model uncertainty expressions can correlate but not necessarily equate with the likelihood of model errors** (Vasconcelos et al., 2023).

4. Employ cognitive forcing functions

Cognitive forcing functions (CFFs) can help shift users from a fast and automatic thought process to a slow and deliberative one (Kahneman, 2011; Wason & Evans, 1974).

a. **CFF designs can facilitate appropriate reliance on GenAI** (Weisz et al., 2024).

i. Encourage users to leverage multi-turn interactions to probe and double-check AI outputs when possible (Bowman et al., 2022).

ii. Pose questions to encourage critical thinking next to or as part of GenAI outputs (Danry et al., 2023).

iii. Experiment with using model-generated self-critiques as CFFs.

Caveat: Exercise caution when providing model-generated answers about model outputs to users, as this may cause users to make incorrect assumptions about model capabilities- e.g., that models can accurately evaluate the accuracy of their responses) (Perez & Long, 2023; Radensky et al., 2023).

b. **Experiment with different types of CFFs for different users.**
    i. For example, expert and novice users may need different levels of critical thinking support (Bowman et al., 2022).

Note: We recommend testing and prioritizing design strategies based on your system context, design goals, and user needs before implementation.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Shedding metacognitive light on reasoning research. *Trends in Cognitive Sciences, 21(8),* 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Arora, S., Narayan, A., Chen, M. F., Orr, L., Guha, N., Bhatia, K., Chami, I., Sala, F., & Ré, C. (2022). Ask Me Anything: A simple strategy for prompting language models. *arXiv preprint, arXiv: 2210.02441*. https://doi.org/10.48550/arXiv.2210.02441.

Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H. S., Fernández, R., Plank, B., Sennrich, R., Zerva, C., & Aziz, W. (2023). Uncertainty in natural language generation: From theory to applications. *arXiv preprint, arXiv:2307.15703*. https://doi.org/10.48550/arXiv.2307.15703.

Biden, J. (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *United States White House, Presidential Actions, October 30, 2023*.

Bowman, S. R., Hyun, J., Perez, E., Chen, E., Pettit, C., Heiner, S., Lukošiūtė, K., Askell, A., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Olah, C., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., … Kaplan, J. (2022). Measuring Progress on Scalable Oversight for Large Language Models. *arXiv preprint, arXiv:2211.03540*. https://doi.org/10.48550/arXiv.2211.03540.

Chen, V., Liao, Q. V., Wortman Vaughan, J., & Bansal, G. (2023). Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact. 7,* CSCW2, Article 370 (October 2023), 32 pages. https://doi.org/10.1145/3610219.

Chen, Z., & Chan, J. (2023). Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise. Available at SSRN: https://dx.doi.org/10.2139/ssrn.4575598

Choudhury, A., & Shamszare, H. (2023). Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis. *Journal of Medical Internet Research*, *25*, e47184. https://doi.org/10.2196/47184

Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023, April). Don't just tell me, ask me: AI systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23),* 13 pages. https://doi.org/10.1145/3544548.3580672

Dell'Acqua, F., McFowland, E., Mollick, E.R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K.R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, No. 24-013. https://dx.doi.org/10.2139/ssrn.4573321

Fok, R., & Weld, D. S. (2023). In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *arXiv preprint, arXiv:2305.07722.* https://doi.org/10.48550/arXiv.2305.07722.

Gordon, A., Negreanu, C., Cambronero, J., Mudumbai Chakravarthy, R., Drosos, I., Fang, H., Mitra, B., Richardson, H., Sarkar, A., Simmons, S., Williams, J., & Zorn, B. (2023). Co-audit: tools to help humans double-check AI-generated content. *arXiv preprint, arXiv: 2302.06590*. https://doi.org/10.48550/arXiv.2310.01297.

Goyal, N., Briakou, E., Liu, A., Baumler, C., Bonial, C., Micher, J., Voss, C. R., Carpuat, M., & Daumé, H. (2023). What Else Do I Need to Know? The Effect of Background Information on Users' Reliance on QA Systems. *arXiv preprint, arXiv:2305.14331*. https://doi.org/10.48550/arXiv.2305.14331.

Guo, Z., Yu, Y., Hartline, J., & Hullman, J. (2024). A Statistical Framework for Measuring AI Reliance. *arXiv preprint, arXiv:2401.15356*. https://doi.org/10.48550/arXiv.2401.15356.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2024). AI Alignment: A Comprehensive Survey. *arXiv preprint, arXiv:2310.19852*. https://doi.org/10.48550/arXiv.2310.19852.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv., 55(12)*. https://doi.org/10.1145/3571730.

Kahneman, D. (2011). *Thinking, fast and slow.* Penguin Books.

Kazemitabaar, M., Chow, J., Ma, C. K. T., Ericson, B. J., Weintrop, D., & Grossman, T. (2023). Studying the effect of AI Code Generators on Supporting Novice Learners in Introductory Programming. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544548.3580919.

Kim, S.S.Y., Liao, Q., Vorvoreanu, M., Ballard, S., & Wortman Vaughan, J. (2024) "I'm not sure, but…": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. *Manuscript submitted for publication*.

Krishna, S., Agarwal, C., & Lakkaraju, H. (2024). Understanding the Effects of Iterative Prompting on Truthfulness. *arXiv preprint, arXiv: 2402.06625*. https://doi.org/10.48550/arXiv.2402.06625.

Liang, J. T., Yang, C., & Myers, B. A. (2024). A Large-Scale Survey on the Usability of AI Programming Assistants: Successes and Challenges. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering,* 13 pages. https://doi.org/10.1145/3597503.3608128.

Liao, Q. V., & Wortman Vaughan, J. (2024). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8036d03b.

Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. *arXiv preprint, arXiv:2205.14334*. https://doi.org/10.48550/arXiv.2205.14334.

Mielke, S. J., Szlam, A., Dinan, E., & Boureau, Y.-L. (2022). Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics, 10,* 857–872. https://doi.org/10.1162/tacl_a_00494.

Passi, S. & Vorvoreanu, M. (2022). Overreliance on AI: Literature review. *Microsoft Technical Report, MSR-TR-2022-12. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/*

Perez, E., & Long, R. (2023). Towards Evaluating AI Systems for Moral Status Using Self-Reports. *arXiv preprint, arXiv:2311.08576*. https://doi.org/10.48550/arXiv.2311.08576.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint, arXiv:2212.09251*. https://doi.org/10.48550/arXiv.2212.09251.

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2023). "It's Weird That it Knows What I Want": Usability and Interactions with Copilot for Novice Programmers. *ACM Trans. Comput.-Hum. Interact., 31(1).* https://doi.org/10.1145/3617367.

Radensky, M., Séguin, J. A., Lim, J. S., Olson, K., & Geiger, R. (2023). "I Think You Might Like This": Exploring Effects of Confidence Signal Patterns on Trust in and Reliance on Conversational Recommender Systems. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 792–804. https://doi.org/10.1145/3593013.3594043.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., … Rush, A. M. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv preprint, arXiv: 2110.08207.* https://doi.org/10.48550/arXiv.2110.08207.

Sarkar, A. (2024, Forthcoming). AI should challenge, not obey. *Communications of the ACM (CACM)*. CACM_2024_AI_provocateur_MSFT_internal_preprint.pdf.

Sarkar, A., Gordon, A. D., Negreanu, C., Poelitz, C., Ragavan, S. S., & Zorn, B. (2022). What is it like to program with artificial intelligence? In *Proceedings of the 33rd Annual Conference of the Psychology of Programming Interest Group (PPIG 2022). https://doi.org/10.48550/arXiv.2208.06213.*

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., & Leike, J. (2022). Self-critiquing models for assisting human evaluators. *arXiv preprint, arXiv: 2206.05802.* https://doi.org/10.48550/arXiv.2206.05802.

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces,* 410–422. https://doi.org/10.1145/3581641.3584066.

Sellen, A., & Horvitz, E. (2023). The Rise of the AI Co-Pilot: Lessons for Design from Aviation and Beyond. *arXiv preprint, arXiv: 2311.14713.* https://doi.org/10.48550/arXiv.2311.14713.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards Understanding Sycophancy in Language Models. *arXiv preprint, arXiv:2310.13548*. https://doi.org/10.48550/arXiv.2310.13548.

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., au2, H. D. I., & Boyd-Graber, J. (2023). Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong. *arXiv preprint, arXiv:2310.12558.* https://doi.org/10.48550/arXiv.2310.12558.

Spatharioti, S. E., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2023). Comparing traditional and LLM-based search for consumer choice: A randomized experiment. *arXiv preprint, arXiv:2307.03744*. https://doi.org/10.48550/arXiv.2307.03744.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L., & Smyth, P. (2024). The Calibration Gap between Model and Human Confidence in Large Language Models. *arXiv preprint, arXiv: 2401.13835*. https://doi.org/10.48550/arXiv.2401.13835.

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022). Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces,* 212–228. https://doi.org/10.1145/3490099.3511119.

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2023). The Metacognitive Demands and Opportunities of Generative AI. *arXiv preprint, arXiv:2312.10893*. https://doi.org/10.48550/arXiv.2312.10893.

Topolinski, S., & Reber, R. (2010). Immediate truth – Temporal contiguity between a cognitive problem and its solution determines experienced veracity of the solution. *Cognition, 114(1),* 117–122. https://doi.org/10.1016/j.cognition.2009.09.009.

Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q. V., & Wortman Vaughan, J. (2023). Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions. *arXiv preprint, arXiv:2302.07248*. https://doi.org/10.48550/arXiv.2302.07248.

Wason, P. C., & Evans, J. (1974). Dual processes in reasoning? *Cognition 3,* 2 (pp. 141–154). https://doi.org/10.1016/0010-0277(74)90017-1.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., Zhou, D. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35*, 24824–24837. https://doi.org/10.48550/arXiv.2201.11903.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088.

Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design Principles for Generative AI Applications. *arXiv preprint, arXiv:2401.14484.* https://doi.org/10.48550/arXiv.2401.14484.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint, arXiv:2306.13063*. https://doi.org/10.48550/arXiv.2306.13063.

Zhou, K., Jurafsky, D., & Hashimoto, T. (2023). Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5506–5524). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.335.

Zhou, K., Hwang, J. D., Ren, X., & Sap, M. (2024). Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty. *arXiv preprint, arXiv:2401.06730*. https://doi.org/10.48550/arXiv.2401.06730.