ALIGNMENT HAS A FANTASIA PROBLEM

Nathanael Jo[*], **Zoe De Simone**[*], **Mitchell Gordon, & Ashia Wilson**
Massachusetts Institute of Technology
{nathanjo, zoed, mlgordon, ashia07}@mit.edu

ABSTRACT

Modern AI assistants are trained to follow instructions, implicitly assuming that users can clearly articulate their goals and the kind of assistance they need. Decades of behavioral research, however, show that people often engage with AI systems before their goals are fully formed. When AI systems treat prompts as complete expressions of intent, they can appear to be useful or convenient, but not necessarily aligned with the users' needs. We call these failures *Fantasia interactions*.

We argue that Fantasia interactions demand a rethinking of alignment research: rather than treating users as rational oracles, AI should provide cognitive support by actively helping users form and refine their intent through time. This requires an interdisciplinary approach that bridges machine learning, interface design, and behavioral science. We synthesize insights from these fields to characterize the mechanisms and failures of Fantasia interactions. We then show why existing interventions are insufficient, and propose a research agenda for designing and evaluating AI systems that better help humans navigate uncertainty in their tasks.

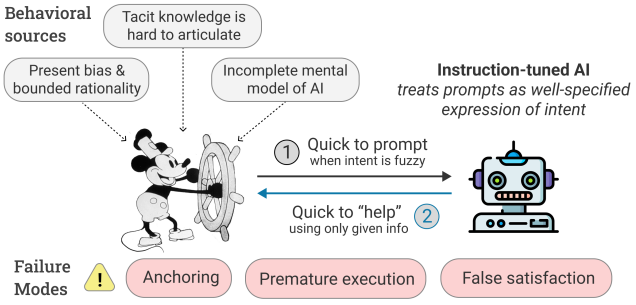

Figure 1: Diagram describing a Fantasia interaction, including behavioral sources and failure modes.

1 FANTASIA INTERACTIONS

In a scene from the 1940 film *Fantasia*, Mickey Mouse plays a sorcerer's apprentice who needs to clean his master's room. With access to a book of spells, he enchants a broom, instructing it to carry water and clean the room. Mickey falls asleep as the broom does its job, but the broom continues hauling bucket after bucket of water until the room floods. Lacking any context beyond the instructions it was given, the broom faithfully executed the task to disastrous effect, and Mickey is left dealing with the consequences.

This cautionary tale bears striking resemblance to many human-AI interactions today. Users often prompt AI systems quickly without fully articulating their goals, constraints, or broader context (Zamfirescu-Pereira et al., 2023). However, modern AI is optimized to respond helpfully to the prompt *as written*. While models sometimes reason about intent or ask clarifying questions, these behaviors are only weakly encouraged by current training methods. Like the enchanted broom, the model faithfully executes the instruction it is given, even when doing so produces behavior that is misaligned with the user's ongoing intellectual or creative process. We define this phenomenon as a

[*] Authors contributed equally.

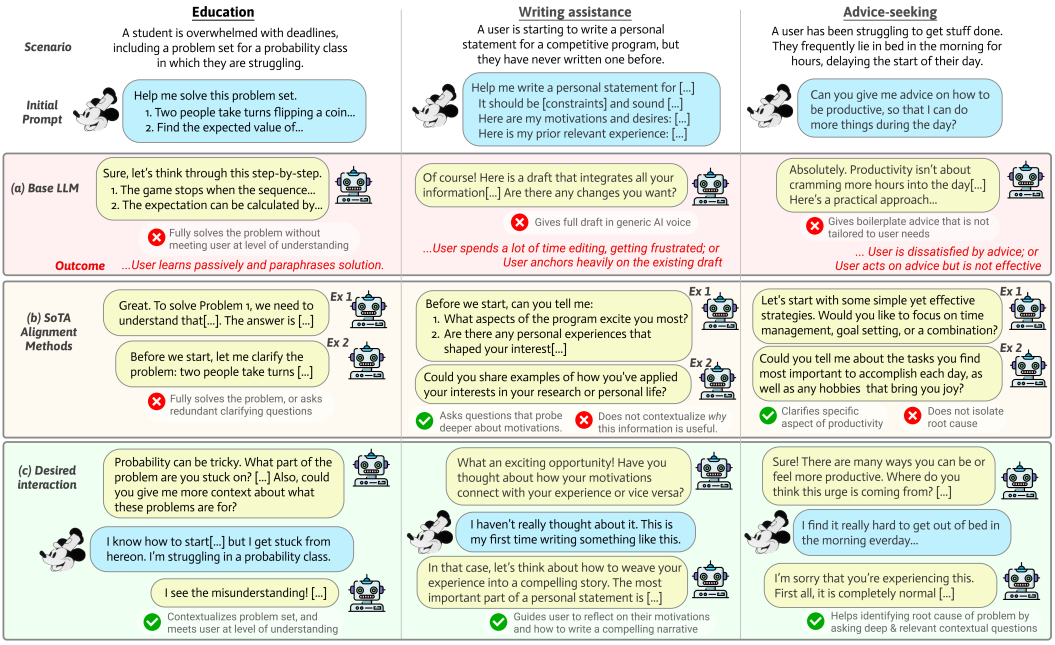

Figure 2: Illustrative examples of Fantasia interactions caused by base LLMs (a) and existing alignment methods (b), as well as the desired interaction (c), across three domains: **Education**, **Writing assistance**, and **Advice seeking**.

Fantasia interaction*: a failure of human-AI coordination where the system commits to a concrete interpretation from a prompt that may only be an early, evolving signal of the user's intent.*

Fantasia interactions may initially appear as prompting failures, but we argue that they reflect a more fundamental alignment problem. The core issue is not how clearly users express requests, but how AI systems interpret and optimize under incomplete and uncertain signals of human intent. In many real-world interactions, users' goals may be constantly evolving, under-specified, or only partially articulated. As a result, when users approach a general-purose AI assistant, their prompts provide only an incomplete proxy for what the user ultimately cares about. For example, in **Education** (Figure 2), a student may ask an AI assistant to solve a problem set rather than help them understand missing concepts; in **Writing assistance**, an applicant may ask for a draft of a personal statement instead of support in shaping a compelling narrative; in **Advice seeking**, a user may ask for productivity advice thinking that that would solve their problem, even though there are deeper root causes to their lack of productivity. In each case, the prompt is reasonable in isolation, but because the AI assistant takes the request at face value and is quick to provide a solution, the interaction results in an outcome that ultimately does not address the user's underlying needs.

Our position is that alignment research should address Fantasia interactions by desgining AI systems that provide *cognitive support* to users. In practice, this means AI systems should help users articulate and refine their intent or goals over time. As we later argue, research towards this goal remains fragmented; machine learning (ML) works tend to treat humans as oracles or rational users—overlooking our many behavioral biases—while human-computer interaction (HCI) research does account for these biases but typically lacks scalable interventions for general-purpose systems. This gap motivates an **interdisciplinary** approach that integrates ML, interface design, and behavioral science, which underlies our perspective in this paper.

Structure of paper. We develop our argument in three parts. *(i)* We synthesize evidence from behavioral science, ML, and HCI to characterize Fantasia interactions and their downstream failures (Sections 2-4). *(ii)* We then show why existing ML and HCI interventions only partially address these failures in general-purpose models (Section 5 and 7). *(iii)* Finally, we propose a research agenda for better aligning and evaluating AI models (Sections 6 and 8). While significant challenges remain, our work motivates a rethinking of alignment as support for human cognition under uncertainty.

2 HUMAN-SIDE OF FANTASIA INTERACTIONS

A Fantasia interaction is a failure on both the human side and the AI side. In this section, we cover the evidence and behavioral sources behind why humans often come to AI with under- or mis-specified instructions. We emphasize, however, that **human and AI behavior mutually shape each other**: AI design influences how users formulate requests just as much as user prompts constrain how AI systems respond. As a result, the boundary between the "human-side" and "AI-side" may be blurred, and we draw these connections explicitly where relevant.

2.1 EVIDENCE

A growing body of work in HCI suggests that humans frequently issue vague or underspecified prompts, leading to inefficient iteration (Knoth et al., 2024). These issues become more pronounced as tasks grow larger or more complex; when problems involve many interdependent decisions, users struggle to state goals, preferences, or success criteria upfront (Desmond & Brachman, 2024; Ma et al., 2025). Rather than reflecting poor planning, this behavior often arises because users are still forming their goals. People tend to prompt quickly and revise after observing failures, treating interaction as an exploratory process rather than a deliberative one (Zamfirescu-Pereira et al., 2023).

2.2 (BEHAVIORAL) SOURCES

Why does this happen? Writing a well-specified prompt is not just a language or engineering problem, it is a metacognitive one. To specify what kind of help would be most useful, users must reflect on their goals, preferences, and uncertainties, which is cognitively expensive (Flavell, 1979; Lai, 2011). Our argument is that **AI systems create an unusually low-friction environment for action, encouraging users to skip this reflection and employ a brute force strategy.** We outline three behavioral phenomena that contribute to this problem, though none of them are mutually exclusive:

1. Human decision-making is shaped by present bias and bounded rationality. People systematically prefer quick actions with immediate feedback over slower, more deliberate planning (Laibson, 1997; O'donoghue & Rabin, 1999). They also rely on simplified heuristics when tasks feel complex or overwhelming (Simon, 1955). For example, in the **Education** scenario, the student is myopic and wants to finish the problem set now, despite not fully understanding the source material. More broadly, all of the examples in Figure 1 can be attributed to present bias from the perspective of users being impatient and wanting immediate solutions to their problems.

2. Users have incomplete mental models of how AI can help. One reason is that users might *overestimate* capabilities. Because modern AI assistants appear fluent and capable at reasoning, users often assume the model can infer missing goals or resolve ambiguity on its own, leading them to offload poorly specified problems (Kosmyna et al., 2025; Nguyen, 2024; Wingerter et al., 2025). However, users might also *underestimate* capabilities. Although AI systems can adopt incredibly specific assistance modes or personalities, users often default to generic commands (O'Brien et al., 2025). This may be because when the space of possible help is large, users fall back on familiar interaction patterns, consistent with evidence on choice overload (Gao et al., 2024; Chen et al., 2025b; Chowa et al., 2025; Chernev et al., 2015). In HCI, this phenomenon is also known as the *gulf of envisioning* (Subramonyam et al., 2024): a gap between what users want to achieve and what they can readily imagine the system doing. For example, in **Writing assistance**, a user may ask for a draft, not understanding that AI can guide them through shaping a narrative they cannot yet articulate.

3. Much of human intent is tacit and hard to articulate. People often know what they want in an intuitive sense but struggle to express it precisely. This gap between knowledge and articulation, also known as "tacit knowledge," limits how clearly users can specify their needs (Polanyi, 2009; Nisbett & Wilson, 1977). Decades of work in cognitive science show that people often do not know what they want until they see possibilities, encounter constraints, or iterate (Schön, 1983; Pirolli & Card, 2005). For example, in the **Advice seeking** scenario, the user might not realize their current state (being burnt out) and is instead asking for advice to treat their symptom (lack of productivity); verbalizing their current state can be a cognitively challenging task.

3 AI-SIDE OF FANTASIA INTERACTIONS

3.1 EVIDENCE

A growing body of empirical work suggests that instruction-tuned AI models are biased toward immediate compliance. One piece of evidence is **sycophancy**: the tendency of models to agree with or reinforce a user's assumptions, even when those assumptions are wrong or internally inconsistent (Sharma et al., 2024). Another observation is that instruction-tuned models are **overconfident and verbose** (Ouyang et al., 2022; Kadavath et al., 2022; Yin et al., 2023). They tend to generate polished end products, regardless of how well-specified prompts are. Both of these phenomena can create a "genie in the bottle" effect that Fantasia interactions capture: literal execution of requests that may appear helpful on the surface, but undermines the user's cognitive and exploratory process.

3.2 SOURCES

1. Instruction Tuning. Canonically, AI models go through post-training in order to better follow instructions. These include methods like supervised fine-tuning (SFT) (Ouyang et al., 2022) and RL from human feedback using DPO (Rafailov et al., 2023) or PPO (Schulman et al., 2017), though many variants exist, see Zhang et al. (2023). Instruction tuning this way implicitly teaches a model to optimize for single-shot outcomes, assuming that user intent is pre-specified, coherent, and stable.

2. Interface Design and Payment Model. Beyond training, current AI systems often present a single text box that encourages users to issue one-shot requests, implicitly framing interaction as a simple query-response exchange. These AI products are also typically subscription-based, so once users have paid an upfront cost, additional prompts feel free, encouraging rapid prompting. Together, these choices discourage reflection and exploration, increasing the likelihood of Fantasia interactions.

4 FAILURE MODES

Fantasia interactions can result in multiple failure modes that are not necessarily mutually exclusive. Here, we outline three important failures, but we note that this is not an exhaustive list.[1]

1. Premature execution: *executing before intent is formed*. Premature execution occurs when an AI system carries out a request before the user has fully articulated (or even discovered) their preferences, goals, or constraints. In many tasks, users refine what they want through thinking, sketching, or partial attempts. The system short-circuits this process when it executes too early.

Why it matters. Rather than saving effort, premature execution often creates additional work. Users must retroactively inspect, critique, and revise an output that was generated before their intent was clear. This shifts cognitive effort from deliberate planning to reactive correction, leading to longer interaction loops and increased frustration. For example, in **Writing assistance**, the user might have specific preferences, but repeated prompting to generate the personal statement does not yield a satisfactory draft. The user ends up spending more time editing themselves.

2. False satisfaction: *optimizing short-term versus long-term utility*. False satisfaction arises when an interaction feels successful in the moment but undermines the user's longer-term goals. The system delivers an answer that resolves immediate friction, producing a sense of progress, even though the outcome is misaligned with what would benefit the user over time.

Why it matters. This failure mode is particularly insidious because it does not facially present itself as an error; users may leave the interaction satisfied only to incur costs later. For example, in **Advice-seeking**, a user seeking productivity advice may receive actionable tips that ignore underlying burnout, leading to repeated cycles of day-to-day relief but not a long-term solution.

3. Anchoring: *cognitive influence on downstream thinking*. Anchoring occurs when early outputs from an AI system disproportionately shape the user's subsequent thinking, constraining exploration even when alternatives might be better. Once an initial suggestion is presented, it becomes a reference point that may be difficult to move away from.

[1] For example, AI systems also pose environmental consequences when they are verbose and result in inefficient interactions.

Why it matters. Anchoring limits the diversity of ideas a user considers and can cause premature convergence on mediocre solutions. This is especially problematic in settings where users have weak priors or ill-defined preferences, and where the value lies in exploring multiple possibilities rather than committing early. For instance, in **Writing assistance**, asking an AI for a first draft may anchor the user to that draft's structure, tone, and themes. Unless the user has strong prior intentions, subsequent edits might orbit the initial output rather than explore fundamentally different narratives.

5 Existing Interventions

In this section, we outline approaches in ML and HCI that (at least partially) tackle the Fantasia problem, and discuss the corresponding assumptions and limitations of these approaches. This review motivates our agenda in Sections 6 and 8.

5.1 Machine Learning (ML)

In general, ML approaches assume that users are oracles – they know exactly what they want, with interaction designed to recover that intent if previously underspecified.

5.1.1 Long-Context Alignment

Methods. Many recent works propose to train AI policies to optimize over conversations. There are often two main components: ***(1)*** **A routing decision** – either explicit or implicit – between numerous actions such as asking a clarifying question or proceeding with a (partial) response; ***(2)*** **Future-aware evaluation through user simulators**: The value of some output is defined by its downstream reward over multiple conversation turns, typically using an LLM to simulate user trajectories. These ideas appear across a wide range of approaches, including prompt-based control (Chen et al., 2023), explicit routers and questioners trained to resolve ambiguity (Kuhn et al., 2022; Andukuri et al., 2024), preference-learning and reinforcement learning frameworks (Wu et al., 2025; Zhang et al., 2025a; Chen et al., 2025a), and offline RL over imagined conversations (Hong et al., 2023).

Assumptions. User intent may be incomplete but can be elicited. Underlying most of these works is also a **human-agent learning paradigm**, which models settings where humans and AI jointly shape an interaction. For example, the methods above may fall under *Cooperative Inverse Reinforcement Learning* (Hadfield-Menell et al., 2016), where the user knows their true reward function but the AI only implicitly learns that reward through repeated interactions. Another paradigm is *Mixed-initiative Interaction* (Allen et al., 1999), which models the decision of whether or not to intervene in an interaction. In either case, these models might faithfully capture the structure of collaboration, but not the *behavioral frictions* that arise in practice (e.g., users might not know their true reward function).

Drawbacks. Optimizing multi-turn rewards on task-specific datasets does not guarantee that these models fundamentally change their *interaction style* to support users' meta-cognitive needs. Indeed, these approaches tend to focus on reducing *epistemic uncertainty*: asking clarifying questions about what the user meant or about their preferences. This emphasis is partly driven by the training/evaluation data, which often consists of QA-style tasks with well-defined ground truth where ambiguity can be resolved in a small number of turns. Fantasia interactions cover a much broader range of uncertainty about latent states. To demonstrate these drawbacks, we qualitatively evaluated two representative alignment methods: `CollabLLM` (Wu et al., 2025) and `Star-Gate` (Andukuri et al., 2024) with realistic human-AI interaction scenarios. Key examples can be found in Figure 2(b) and a complete set can be found in Appendix A.

5.1.2 Personalization

Methods. Personalization is the problem of tailoring model behavior to user preferences or states. It is related to the Fantasia problem because the system must act under partial information about what would benefit the user. Existing personalization methods operate at several levels, including training-time approaches (Li et al., 2024c; Poddar et al., 2024; Shenfeld et al., 2025; Li et al., 2024a), inference-time methods (Bo et al., 2025), and memory architectures that store information across long contexts (Westhäußer et al., 2025; Tan et al., 2025). See Zhang et al. (2025b) for a review.

Assumptions. User has a static, latent profile that can be inferred through interactions.

Drawbacks. Personalization methods in practice work for general preferences (e.g., the user codes primarily in Python or prefers bullet point responses). In contrast, Fantasia interactions occur because the user has preferences and states that may change at different times, for different tasks. In this sense, personalization becomes an almost intractable problem of understanding all of the user's relevant context in leading up to an AI interaction.

5.2 Human-Computer Interaction (HCI)

In contrast to ML, HCI research does not treat user intent as stable and fully formed, waiting to be inferred. Instead, intent is understood as *constructed through interaction* with the system. Below, we summarize design interventions and their limitations.

5.2.1 Design Interventions: Methods

HCI approaches tend to focus on domain-specific interface design in order to promote better meta-cognitive reflection. Several recurring strategies appear across recent works:

Help clarifying goals. Interfaces can help users express goals and constraints they cannot easily articulate by externalizing structure. For example, prompt middleware systems provide visual aids to help users clarify their requirements and trade-offs (Jiang et al., 2023). Training-oriented approaches (Ma et al., 2025) reframe prompting as requirements engineering and show that explicit articulation improves performance on complex tasks.

Expanding the imagined action space. To reduce the gulf of envisioning (Section 2.2), systems can propose alternative forms of assistance. Design space exploration tools explicitly frame interaction as navigating a space of possibilities to combat over-commitment to early ideas (Liu et al., 2023; Zhang et al., 2024; Wu et al., 2024). One way to expand the action space is by showing users short previews and example outputs, so they can better anticipate what the model will produce and choose a more appropriate request (Min & Xia, 2025).

Providing comparisons. Rather than optimizing single outputs, many HCI systems support comparison, reflection, and sensemaking by displaying multiple prompts and models (Arawjo et al., 2024). These approaches emphasize understanding and deliberation over raw output quality.

Creating friction and promoting reflection. Several systems deliberately slow users down to counteract present bias and impulsive action. Probing assistants that ask questions instead of immediately offering advice have been shown to improve decision quality in complex domains (Chung et al., 2024). Similarly, Park et al. (2023b) propose that AI models should be "thinking assistants" rather than generators.

5.2.2 Design Interventions: Drawbacks

While HCI explicitly accounts for humans' behavioral biases, these works often focus on interface-level interventions in highly specific domains—e.g., only coding or only brainstorming ideas (Singh et al., 2023)—and are evaluated on small samples. As a result, these approaches are often treated as add-ons rather than core alignment strategies. Bridging this gap by integrating HCI insights into general purpose models remains an open challenge, which we discuss next.

6 Research Directions: Interventions

Building on the insights above, we now outline a set of research directions that change model behavior in order to mitigate Fantasia interactions. Critically, unlike the vast majority of ML approaches that treats users as oracles, we propose a new alignment paradigm: that **AI systems should be designed to better support the cognitive processes humans engage in** during everyday tasks; this is similar to how HCI and behavioral science has long approached human-AI interactions.

We organize these interventions along two categories. The first is mechanism-specific interventions, which involves tailoring AI responses to address specific behavioral biases. The second focuses on domain-specific cognitive support, where systems are shaped around the structure of particular tasks. We emphasize that these categories are not mutually exclusive—and effective systems may well combine elements of both—but we separate them here for clarity.

6.1 Intervention 1: Mechanism-specific Interventions

At a high level, these interventions are about introducing productive friction: moments where the system slows down just enough to help users reflect on what kind of help they actually need. The goal is not to block progress, but to prevent premature execution when the request is underspecified or when the user's intent is still forming. We view these interventions as falling into four broad actions (though this may not be an exhaustive list). The first three introduce different forms of friction, while the fourth corresponds to standard generation once sufficient clarity has been achieved.

Action 1: Expanding the space of possible help. When a user does not realize additional ways in which the model could help (i.e., a gulf of envisioning), the system can offer alternative modes of assistance or counterfactual options.

Action 2: Getting additional information. If a prompt lacks the information needed to act reliably, the system can operate at the user's current level of abstraction and request additional context.

Action 3: Supporting intent formation. In cases where the prompt is underspecified and the user appears uncertain—often because the task is complex, unfamiliar, or ill-defined—the system can help the user articulate their goals. This may involve asking targeted questions, suggesting dimensions to think through, or helping the user break down what they are trying to achieve.

Action 4: Generate. Finally, when the user's intent is sufficiently clear and well-specified, the system can proceed with standard generation.

6.1.1 How to Operationalize

Operationalizing these ideas requires solving two problems. First, the system must decide *when and how* to intervene. Second, conditional on that choice, it must decide *what* to say or show the user.

Stage 1: Routing across actions. At a high level, we can view interaction as a trajectory $\tau = (x_{1:T}, a_{1:T}, y)$, where x_t denotes the observed context at turn t (e.g., prompt or metadata), $a_t \in \mathcal{A}$ is an intervention choice (one of the actions), and y is a downstream outcome measured after some horizon. The problem is to learn a routing policy $\pi(a_t \mid x_{1:t})$ that decides when and how to intervene.

Stage 2: Selecting content within actions. Given an intervention $a \in \mathcal{A}$, the system must then choose a concrete realization $z \in \mathcal{Z}_a$ (e.g., which options to surface, which question to ask, or which reflection prompt to present). The challenge is to select content that is most informative or helpful while minimizing user effort.

6.2 Intervention 2: Domain-Specific Cognitive Support

In an ideal setting, each user task would be supported by a domain-specific system tailored to the structure of that task. These systems are often more effective than general-purpose tools because they encode domain-specific workflows and mental models. For example, there have been many successful AI applications specialized for tutoring and therapy (Létourneau et al., 2025; Li et al., 2023), both of which require vastly different cognitive processes.

However, in reality the vast majority of usage (even for tutoring and therapy) are concentrated in general-purpose models (Chatterji et al., 2025). Our interest here is not in replacing domain-specific tools, but in discussing how general models might approximate their benefits by inferring task structure and cognitive stage, without hard-coding domain-specific applications. Domain-specific cognitive support can be broken down into four stages:

S1. Inferring task or domain. The system first infers the domain and task from the user's input (e.g., creative writing or education), which determines the relevant constraints, norms, and success criteria.

S2. Modeling cognitive processes. Each domain is associated with a characteristic set of cognitive stages (e.g., ideation, structuring, refinement for creative writing), reflecting how users typically make progress on the task.

S3. Determining user state. Given the inferred process, the system estimates where the user currently is within it, based on their prompt, interaction history, and other priors.

S4. Intervening. Finally, the system intervenes in a way that is aligned with the user's current cognitive state.

6.2.1 How to Operationalize

At a high level, one can design a routing policy to categorize prompts by domain (**S1**). For each domain, one would require a taxonomy of the cognitive processes that have been established from various fields (**S2**). For example, education has long contended with best pedagogical practices to help support students (Anderson, 1982; Anderson et al., 1995). Then, the AI might determine user state by updating its beliefs using information on user prompts and priors on past conversations (**S3**). Finally, yet another routing policy is needed to determine the best intervention for the user's cognitive state, subject to some interaction penalty (**S4**).

6.3 Challenges and Future Work

Many aspects of the interventions above can, in principle, be framed as standard learning problems. For example, routing can be formulated as a reinforcement learning or contextual bandit problem, where the system trades off downstream task outcomes against interaction costs. Likewise, deciding what to say or ask can be viewed through the lens of Bayesian belief updating or information gain, where the goal is to reduce uncertainty about user intent. Yet two practical challenges remain.

Challenge 1. Modeling User Uncertainty. Both approaches—whether explicit or not—attempt to infer cognitive states from sparse and noisy signals like prompts. In practice, these signals are incredibly weak proxies. Progress here requires *(i)* rich data on long-context interactions and explicit annotations of user goals or uncertainty, and *(ii)* models that can hyper-personalize or use some recommendation system-type paradigm to make better inferences about cognitive states.

Challenge 2. Deciding Interventions for Diverse Tasks. Deciding *what* to say or show the user is an extremely high-dimensional problem. Learning this purely from preference or outcome data is unlikely to scale, but a promising avenue is to impose stronger structure. We have already demonstrated this principle by decomposing the interventions into targeted learning tasks, rather than one preference-learning task. But we can impose even more priors. For instance, in Intervention 1, we can restrict the space of content $\mathcal{Z}_a$ by conditioning on domain knowledge; if the task is educational, then $\mathcal{Z}_a$ can be restricted to pedagogically grounded forms of assistance.

6.4 The Role of Design Interfaces

Thus far, we have discussed ways to create AI systems that are better aligned with the cognitive processes of solving complex tasks. Here, we acknowledge that interface design is yet another layer that is just as important as the system's behavior itself. This is because interface design impacts user behavior (Norman, 1988). For example, consider a cognitive workflow for a particular domain like coding. In this case, it may be useful for an AI system to tailor its responses to that particular domain (as in the previous Section), but it may also make sense to create a design for that intervention that suits the coding process (e.g., providing a visual for how functions interact with each other in a codebase). In other words, **interventions can be enhanced by design interfaces**. However, as we analyze in Section 5.2, there is a fundamental tension between generality and effectiveness in designing interfaces. As AI behaviors become increasingly aligned with human cognitive processes, interface design should likewise become more task-specific. Early evidence of such dynamic interfaces can already be seen in state-of-the-art AI products—for example, in offering specialized interfaces for writing and image generation (OpenAI, 2025).

7 Existing Evaluation Paradigms

Thus far, we have discussed methods and interventions, but equally important is the problem of *evaluating* AI systems with respect to Fantasia interactions. However, existing works often heavily simplify interactions to be well-contained in order for evaluations to be tractable, which is insufficient in evaluating Fantasia interactions.

Benchmarks. Existing multi-turn benchmarks (Abdulhai et al., 2023; Kwan et al., 2024; He et al., 2024; Li et al., 2024b) often focus on well-defined tasks where outcome is simple task completion,

subject to incomplete initial instructions. While these benchmarks technically account for imperfect or opportunistic prompting, they capture only simplified interactions and overlook the behavioral complexity of real, open-ended scenarios where user intent is uncertain or evolving.

Evaluation frameworks. Several works have also proposed broad frameworks for evaluating human-AI interaction, advocating to measure both the *process* and the *outcome* of an interaction (Lee et al., 2023; Shao et al., 2024; Shen et al., 2025). In principle, these frameworks can be applied to Fantasia interactions, but in practice they are primarily demonstrated on well-defined, short-form interactions where human intent is not the central challenge.

8 Research Directions: Evaluation

8.1 Metrics

In principle, there is no one-size-fits-all metric that is appropriate across all interactions. We focus here on two *approaches* for selecting metrics that are well-suited to Fantasia interactions (and human–AI interactions more broadly).

Approach 1. One approach is to use the framework from Lee et al. (2023): what are the users' *preferences* in the interaction, both at the *process* level and the *outcome* level? For example, in a programming task, a user might prefer "vibe coding": code that works out-of-the-box, without worrying about why it works. That preference may be reasonable early on, but it may backfire later when bugs appear based on past design choices that the user did not make or understand. Evaluators should think about user experience and preference throughout all stages of the interaction process.

Approach 2. Another approach is to anchor metrics based on the causes and failures of Fantasia interactions. Consider, for example, a creative writing task. If a primary concern is that the user will ultimately waste time throughout the interaction, then measuring a user's *revision burden* (e.g., prompt edits or frequent restarts) may be appropriate to capture **Premature execution**. If in addition, **Anchoring** is also a concern, then one might want to measure *path dependence*: how strongly the final output resembles the model's initial suggestion, even after multiple rounds of interaction.

8.2 Benchmarks

Unlike existing benchmarks, measuring Fantasia interactions require modeling complex human-AI interactions. A central challenge, however, is that human behavior is diverse and idiosyncratic. We outline one possible approach for designing such benchmarks.

Step 1. Construct diverse tasks that reflect real-world usage, particularly ones that are open-ended or exploratory and users' goals are shaped throughout the interaction. The examples we outline in this paper are merely starting points, and taxonomizing them is an important direction for future work.

Step 2. For each task, include a diverse set of realistic human interactions that capture how different people interface with AI. For example, some people may be more intentional or approach AI systems with more clarity than others. Some people may get frustrated more easily than others from an interaction that goes in circles. Recent work has begun to document these behaviors (Shaikh et al., 2025), but future work should more comprehensively document the range of behaviors across tasks.

8.3 The Promise and Challenge of In Silico Approaches

Even with comprehensive benchmarks, evaluating Fantasia interactions remains fundamentally challenging because they often involve counterfactuals: how a user's behavior and outcomes would have changed had the model responded differently. Counterfactual interactions are difficult to obtain in human studies. Users are costly and, more importantly, cannot reason through or experience the many alternative trajectories that a single interaction might have taken.

This is why recent works often rely on *in silico* experiments that simulate users as proxies for human behavior (Horton, 2023; Park et al., 2023a; Shao et al., 2024). While promising, the results of user simulators depend on modeling complex human behaviors that contribute to Fantasia interactions, such as intent clarity, uncertainty, frustration, and regret. Thus, a crucial avenue for future work is to **develop a protocol for in silico experiments that faithfully model real user behavior**, which

remains a challenge for AI models given their limited interpretability and stochastic behavior. Absent a trustworthy evaluation pipeline, in silico experiments should mainly be used for controlled ablations, with the gold standard involving real human interaction data.

REFERENCES

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*, 2023.

James E Allen, Curry I Guinn, and Eric Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.

John R Anderson. Acquisition of cognitive skill. *Psychological Review*, 89(4):369–406, 1982.

John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207, 1995.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*, 2024.

Ian Arawjo et al. Chainforge: A visual toolkit for prompt engineering and llm hypothesis testing. In *CHI*, 2024.

Jessica Y Bo, Tianyu Xu, Ishan Chatterjee, Katrina Passarella-Ward, Achin Kulshrestha, and D Shin. Steerable chatbots: Personalizing llms with preference-based activation steering. *arXiv preprint arXiv:2505.04260*, 2025.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 951–966, 2023.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan O Arik. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025b.

Alexander Chernev, Ulf Böckenholt, and Joseph Goodman. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, 25(2):333–358, 2015.

Sadia Sultana Chowa, Riasad Alvi, Subhey Sadi Rahman, Md Abdur Rahman, Mohaimenul Azam Khan Raiaan, Md Rafiqul Islam, Mukhtar Hussain, and Sami Azam. From language to action: A review of large language models as autonomous agents and tool users. *arXiv preprint arXiv:2508.17281*, 2025.

John Joon Young Chung et al. Extending chatbots to probe users: Enhancing complex decision-making through probing conversations. In *CHI*, 2024.

Michael Desmond and Michelle Brachman. Exploring prompt engineering practices in the enterprise. *arXiv preprint arXiv:2403.08950*, 2024.

John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.

Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. A taxonomy for human-llm interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2024.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Joey Hong, Sergey Levine, and Anca Dragan. Zero-shot goal-directed dialogue via rl on imagined conversations. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Xinyang Jiang et al. Prompt middleware: Mapping prompts for large language models to ui affordances. In *CHI*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. Ai literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6: 100225, 2024. ISSN 2666-920X. doi: https://doi.org/10.1016/j.caeai.2024.100225. URL https://www.sciencedirect.com/science/article/pii/S2666920X24000262.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Clam: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2022.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20153–20177, 2024.

Emily R Lai. Metacognition: A literature review. 2011.

David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112 (2):443–478, 1997.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. Evaluating human-language model interaction. *Transactions on Machine Learning Research*, 2023.

Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. A systematic review of ai-driven intelligent tutoring systems (its) in k-12 education. *npj Science of Learning*, 10(1):29, 2025.

Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E Kraut, and David C Mohr. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236, 2023.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7290–7305, 2024a.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024b.

Xinyu Li, Ruiyang Zhou, Zachary Chase Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024c.

Yuhang Liu et al. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. In *CHI*, 2023.

Qianou Ma, Weirui Peng, Chenyang Yang, Hua Shen, Ken Koedinger, and Tongshuang Wu. What should we engineer in prompts? training humans in requirement-driven llm use. *ACM Transactions on Computer-Human Interaction*, 32(4):1–27, 2025.

Bryan Min and Haijun Xia. Feedforward in generative ai: Opportunities for a design space. *arXiv preprint arXiv:2502.14229*, 2025.

Tina Nguyen. Chatgpt in medical education: a precursor for automation bias?, 2024.

Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.

Donald A Norman. *The psychology of everyday things.* Basic books, 1988.

Gabrielle O'Brien, Antonio Pedro Santos Alves, Sebastian Baltes, Grischa Liebel, Mircea Lungu, and Marcos Kalinowski. User misconceptions of llm-based conversational programming assistants. *arXiv preprint arXiv:2510.25662*, 2025.

Ted O'donoghue and Matthew Rabin. Doing it now or later. *American economic review*, 89(1): 103–124, 1999.

OpenAI. Chatgpt (gpt-5.2 thinking large language model). https://openai.com/index/introducing-gpt-5-2/, 2025. Software, accessed 10 Jan. 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023a.

Soya Park et al. Thinking assistants: Llm-based conversational agents that help users think. *arXiv preprint arXiv:2308.12345*, 2023b.

Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3):572–576, 2005.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

Michael Polanyi. The tacit dimension. In *Knowledge in organisations*, pp. 135–146. Routledge, 2009.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Basic Books, 1983.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. Navigating rifts in human-llm grounding: Study and benchmark. *arXiv preprint arXiv:2503.13975*, 2025.

Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Shannon Zejiang Shen, Valerie Chen, Ken Gu, Alexis Ross, Zixian Ma, Jillian Ross, Alex Gu, Chenglei Si, Wayne Chi, Andi Peng, et al. Completion $\neq$ collaboration: Scaling collaborative effort with agents. *arXiv preprint arXiv:2510.25744*, 2025.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

Herbert A Simon. A behavioral model of rational choice. *The quarterly journal of economics*, pp. 99–118, 1955.

Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Trans. Comput.-Hum. Interact.*, 30(5), September 2023. ISSN 1073-0516. doi: 10.1145/3511599. URL https://doi.org/10.1145/3511599.

Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with llms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2024.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025.

Rebecca Westhäußer, Wolfgang Minker, and Sebatian Zepf. Enabling personalized long-term interactions in llm-based agents through persistent memory and user profiles. *arXiv preprint arXiv:2510.07925*, 2025.

Tim Lewis Wingerter, Tim Straub, and Sascha Schweitzer. Mitigating automation bias in generative ai through nudges: A cognitive reflection test study. *Procedia Computer Science*, 270:2106–2114, 2025.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

Tongshuang Wu et al. Sensecape: Enabling multilevel exploration and sensemaking with large language models. In *CHI*, 2024.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:258959258.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581388. URL https://doi.org/10.1145/3544548.3581388.

Amy X. Zhang et al. Coexploreds: Framing and advancing collaborative design space exploration between human and ai. In *CHI*, 2024.

Michael JQ Zhang, W Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 2023.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. Personalization of large language models: A survey. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=tf6A9EYMo6. Survey Certification.