HUMAN MODELING MEETS AI METACOGNITION: TOWARD COUPLED COGNITIVE ALIGNMENT FOR INTERACTIVE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI agents increasingly participate in complex human workflows, where failures arise not only from incorrect outputs but also from inappropriate interactions. An agent may answer when it should ask for clarification, personalize despite high uncertainty about the user, or express confidence in ways that distort human trust and reliance. We argue that these failures stem from inaccurate *human modeling* that aims to infer humans' latent cognitive states, and unreliable *AI metacognition* which estimates the agent's own uncertainty, capability boundaries, and risk, and more importantly, misalignment between the two in the interaction policy. A primary bottleneck is that the latent states of humans and AI have to be inferred from partial observations. This position paper advocates *Coupled Cognitive Alignment* (CCA), a framework that unifies *human modeling* and *AI metacognition* as a coupled problem of interaction-level decision making under uncertainty. In CCA, agents maintain uncertainty-aware dynamic models of human states while simultaneously reasoning about their own uncertainty and capability limits. These coupled estimates determine not only what the agent outputs, but whether it should answer, ask, explain, retrieve, decompose tasks, defer, or escalate. We formalize it as a partially observable interactive system, identify its key research challenges, and discuss potential strategies for its learning, evaluation, and deployment.

1 INTRODUCTION

AI agents are increasingly embedded in human workflows, where their role is no longer limited to producing isolated responses. In these settings, agents must decide not only what answer to produce, but also what kind of interaction should happen next. This decision depends on two forms of uncertainty that are rarely resolved by the surface prompt alone: what the human actually needs, and what the agent can reliably do. On the human side, latent goals, constraints, preferences, uncertainty, cognitive limitations, and contextual tradeoffs are not directly observable to the agent, even when they shape what kind of assistance would be appropriate. At the same time, agents may be uncertain about their capability boundaries, failure modes, costs, and risks. When either side is modeled poorly, fluent responses can still produce inappropriate reliance, overconfident action, or ineffective assistance (Lee & See, 2004; Parasuraman & Riley, 1997; Bansal et al., 2021; Vasconcelos et al., 2023).

Figure 1 illustrates this challenge in tutoring. Consider a student new to calculus who is trying to solve a problem independently. The student does not want the final answer, but wants to know whether their current solution path is missing a key step. If the student asks an AI tutor, "Can you help me solve this calculus problem?", a capable language model may produce a correct derivation. Yet in this situation, a complete solution can be mathematically correct while pedagogically misaligned: it satisfies the surface request, but fails to model the student's latent learning need. An aligned tutor would also need to assess whether it has enough evidence to diagnose the student's own solution path; if not, it should ask to see the student's steps rather than defaulting to a canonical derivation.

This example highlights two capabilities that are usually studied separately: *human modeling* and *AI metacognition*. In the tutoring case, human modeling means inferring that the student is seeking diagnosis of their own reasoning rather than a complete solution. AI metacognition means estimating whether the agent has enough evidence and capability to perform that diagnosis, or whether it should

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

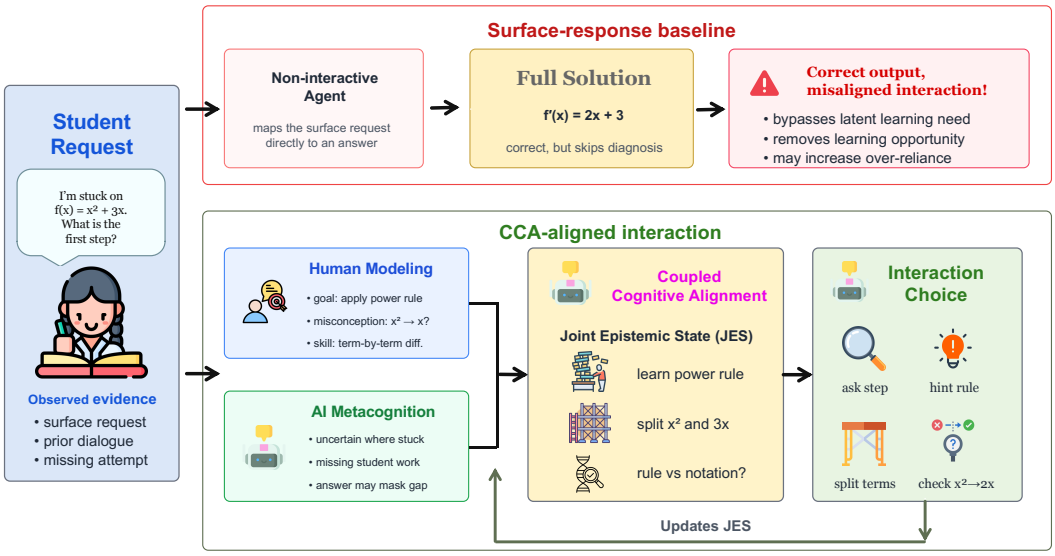

Figure 1: The tutoring challenge: a correct solution may bypass a student's need to diagnose their own reasoning. CCA couples estimated learning needs with the tutor's diagnostic uncertainty to choose between requesting the student's steps and offering a targeted hint.

first ask for the student's intermediate steps. More generally, human modeling refers to estimating latent human cognitive states (Anderson & Lebiere, 2014; Baker et al., 2017; Sap et al., 2022; Kim et al., 2023; Shapira et al., 2024; Chen et al., 2024), including beliefs, goals, preferences, knowledge, trust, uncertainty, cognitive load, and evolving intentions. AI metacognition refers to estimating the agent's own uncertainty, reliability, capability boundaries, expected success rates, computational cost, and operational risks (Flavell, 1979; Koriat, 1997). A primary challenge shared by both tasks is that their states are latent and have to be inferred from partial observations or limited history.

However, each capability in isolation is insufficient for satisfying interactions. Recent work makes this gap more apparent rather than narrowing it. On the human side, personalization and user simulation now move beyond static profiles toward implicit cost and performance preferences, personality anchored behavior, recursive theory of mind reasoning, and human aligned user simulation for complex agent tasks (Zeng et al., 2026; Javadi et al., 2026; Lei et al., 2026; Wang et al., 2026; Li et al., 2026c). These advances improve inference about human states, but they can still personalize too strongly when the agent is uncertain about the user. On the agent side, recent work improves verbal uncertainty alignment, faithful confidence expression, dynamic abstention, uncertainty aware clarification or tool use decisions, and episode level analyses of reasoning traces (Shi et al., 2026; Gani et al., 2026; Zhao et al., 2026; Edwards & Schuster, 2026; Zhou et al., 2026c; Li et al., 2025; 2026d). Such estimates are useful, but they do not determine how uncertainty should be communicated or acted upon for a particular human. The same uncertainty may call for a brief warning to an expert, an explanation for a novice, a diagnostic question for a student, or deferral in a high stakes setting.

Our position is that future interactive agents should couple these two forms of reasoning. We call this direction **Coupled Cognitive Alignment (CCA)**: **the study of agents that maintain uncertainty aware estimates of both evolving human cognitive states and their own epistemic and capability states, and use these estimates jointly to select interaction policies** (Zhou et al., 2024b; Liu et al., 2024). CCA is not a single architecture or a claim that complete human mental states are identifiable. Rather, it is a policy level framing: human modeling should constrain how metacognitive uncertainty is communicated, and metacognition should constrain when personalization, direct action, delegation, or abstention is appropriate.

Under this view, alignment for agentic language systems is not only a property of outputs, but a property of interaction level policy calibration under uncertainty about both the human and the AI. This paper develops CCA as a research framing for modeling, decision making, and evaluation in extended interaction between humans and AI.

Table 1: Positioning of Coupled Cognitive Alignment (CCA) relative to related research categories. CCA differs from prior lines of work by explicitly coupling uncertainty aware human state inference with AI side metacognitive self modeling, and by using the resulting Joint Epistemic State (JES) to select interaction level actions over time.

Research Category	State Modeling		Coupling	Interaction Policy
	Human	AI		
Personalization	✓	✗	✗	✗
Human Simulation	✓	✗	✗	✗
Metacognition	✗	✓	✗	✗
Uncertainty Calibration	✗	✓	✗	✗
Deferral and Abstention	✗	✓	✗	✓
Human-AI Teaming	✓	✓	✗	✓
Planning with Tool Use	✗	✓	✗	✓
CCA (Ours)	✓	✓	✓	✓

2 Missing Pieces of Interactive Agents?

2.1 From Response to Interaction

The dominant evaluation paradigm for language models remains largely response centered. This view remains useful, but it is increasingly incomplete for modern agentic systems. In extended human workflows, the agent's decision problem has shifted from generating a response to selecting an appropriate interaction act within an ongoing policy.

Interaction quality goes beyond response quality. This changes the unit of analysis from response generation to interaction policy. An agent may need to answer directly, ask for missing context, retrieve evidence (Lewis et al., 2020; Nakano et al., 2021), use a tool (Schick et al., 2023; Karpas et al., 2022; Qin et al., 2024; Li et al., 2023b; Patil et al., 2024), decompose the task (Shen et al., 2023; Wang et al., 2023), provide a hint, communicate uncertainty, defer to a human expert (Mozannar & Sontag, 2020), or escalate the decision (Wu et al., 2024). These choices go beyond wording to determine how the human and the system coordinate, how uncertainty is handled, and how the task unfolds over time.

A response can therefore be locally high quality while still being interactionally inappropriate. A tutoring agent may provide a correct derivation when the student needs a hint. A decision support agent may give a plausible recommendation when it should first ask about constraints. A coding assistant may implement a requested change (Jimenez et al., 2024) when it should first inspect dependencies and testing requirements. In these cases, the problem is not only whether the generated content is good, but whether the agent selected the right interaction act.

Interaction requires modeling both sides. We use *interaction level requirements* to refer to the demands that arise in such settings: adapting assistance to the user's state, asking clarifying questions when the task is underspecified, communicating uncertainty in a useful form, avoiding inappropriate automation, supporting calibrated reliance (Lee & See, 2004; Parasuraman & Riley, 1997; Lai et al., 2023), and selecting feasible subtasks when the original request exceeds the agent's reliable capability. Meeting these requirements requires inference about both sides of the interaction: what the human currently needs, understands, and act upon, and what the agent itself can reliably know, do, or risk.

Partially Observable Interactive Process. We formalize this setting minimally as a partially observable interactive process (Kaelbling et al., 1998; Gmytrasiewicz & Doshi, 2005). At each time step t, the agent observes a task context x_t, an interaction observation o_t, and internal evidence e_t about its own uncertainty and reliability. The agent chooses an action a_t^{AI}, and the human responds with a_t^H. The interaction history is

$$\tau_t = (x_{1:t}, o_{1:t}, e_{1:t}, a_{1:t-1}^{AI}, a_{1:t-1}^{H}). \tag{1}$$

We use τ_t broadly to include available interaction history and, when applicable, compact memory summaries maintained by the agent. The central assumption is **double partial observability**: the human cognitive state and the agent's epistemic state are both inferred rather than directly observed.

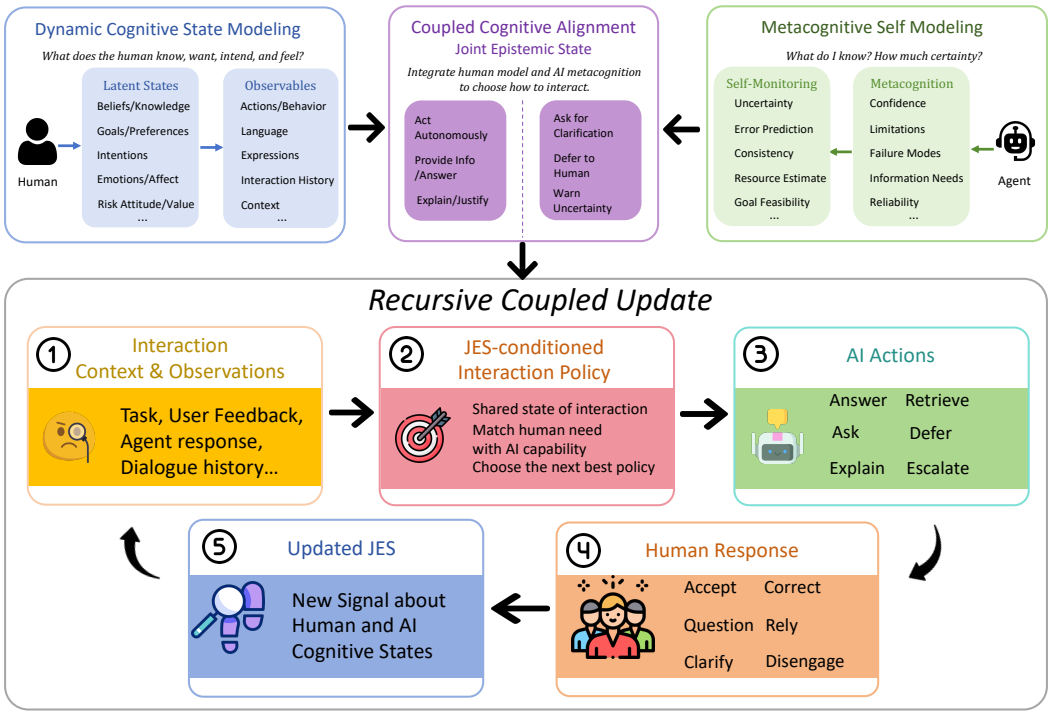

Figure 2: Coupled Cognitive Alignment uses the Joint Epistemic State (JES) as the interface for policy selection. The agent infers a belief about the human state and a metacognitive belief about its own uncertainty, capability, and risk, couples them into a JES, and uses the JES to select the next interaction action. Human responses and new evidence recursively update the coupled posterior and recompute the JES for the next turn.

2.2 Human modeling: Inferring humans' latent states and goals

We use *human modeling* to refer to the agent's attempt to infer task relevant latent cognitive states of the human from observable signals (Baker et al., 2017). These states may include what the user knows, misunderstands, wants, prefers, trusts, doubts, can currently manage, or is likely to do next. The observable evidence may come from language, behavior, feedback, interaction history, task context, or multimodal cues (Sap et al., 2022; Kim et al., 2023; Koh et al., 2024). The key point is that the agent does not directly observe the human state that matters for interaction. It observes traces that must be interpreted under uncertainty.

Dynamic Cognitive State Modeling. Let h_t denote the latent human cognitive state which includes task relevant variables:

$$h_t = (b_t^H, g_t^H, \theta_t^H, k_t^H, u_t^H, \ell_t^H, r_t^H, \sigma_t^H, \ldots). \tag{2}$$

Here b_t^H denotes task beliefs, g_t^H goals, θ_t^H preferences, k_t^H knowledge, u_t^H uncertainty, ℓ_t^H cognitive load, r_t^H reliance or trust, and σ_t^H interaction strategy.

The agent maintains a belief distribution over this state:

$$\mu_t^H(h) = p(h_t = h \mid \tau_t). \tag{3}$$

We call this belief update *Dynamic Cognitive State Modeling.* (DCSM). DCSM is not a claim that the agent can recover a true psychological state with certainty. It is an inference problem for policy selection: maintaining an uncertainty aware distribution over cognitive variables that matter for prediction, intervention, and action selection. The term "dynamic" means that this belief is revised as the interaction history changes (Shin et al., 2021; Abdelrahman et al., 2023; Song et al., 2022).

Human modeling guides adaptive assistance. Human modeling extends beyond static profiles of preferences or demographics (Ricci et al., 2011; 2021; Salemi et al., 2024; Geng et al., 2022) to

transient, context sensitive states such as confusion, cognitive overload (Hart & Staveland, 1988), overconfidence, uncertainty, disengagement, and overreliance (Lee & See, 2004; Buçinca et al., 2021). Beyond predicting behavior, it asks which latent states make particular interventions appropriate. In agentic interaction, this supports three functions: interpreting underspecified requests by inferring hidden tradeoffs among accuracy, cost, time, and other goals; tailoring assistance to the user's state, from concise answers for experts to explanations for novices or warnings against overreliance; and adapting as new evidence reveals how the user understood, questioned, or used previous responses.

Remaining Limitations of Human Modeling. Current approaches often use explicit preferences, brief conversational context, or coarse profiles (Ricci et al., 2011; Tan & Jiang, 2023; Salemi et al., 2024; Geng et al., 2022). Inferred implicit preferences and social or cognitive states remain uncertain and context dependent (Zeng et al., 2026; Javadi et al., 2026). Inferring preferences from behavior (Ng et al., 2000) or predicting actions does not ensure accurate modeling of underlying beliefs, uncertainties, and constraints. Language model simulations produce plausible responses (Aher et al., 2023; Argyle et al., 2023; Park et al., 2023; Wang et al., 2024b), while benchmarks assess state alignment, recursive perspectives, user behavior, and student difficulty (Wu et al., 2026; Lei et al., 2026; Wang et al., 2026; Li et al., 2026c). Yet fluent student simulations can mask unrealistic learner states unless epistemic states and competence boundaries are specified (Yuan et al., 2026). These advances do not establish that models capture latent states needed for intervention: agentic interaction requires modeling users as cognitive agents whose uncertain states evolve with context and interaction.

2.3 Metacognition: Estimating uncertainty, capability, and risk of AI

While human modeling captures the latent states and handles the uncertainty of the user, agentic interaction also requires considering the capability boundary of the AI system itself. We use ***AI metacognition*** to refer to the agent's estimation and regulation of its own epistemic and operational state (Flavell, 1979; Nelson, 1990). This includes uncertainty about an answer, reliability of evidence, capability boundaries, expected success of a planned action, likelihood of failure, cost of acting incorrectly, need for retrieval or tool use, and uncertainty about the agent's own human state inference.

Metacognitive Self Modeling. Let m_t denote the agent's metacognitive state:

$$m_t = (u_t^{AI}, q_t^{AI}, \kappa_t^{AI}, \rho_t^{AI}, s_t^{AI}, \eta_t^{H|AI}, \ldots). \tag{4}$$

Here u_t^{AI} denotes task uncertainty, q_t^{AI} confidence or expected quality, κ_t^{AI} capability boundary information, ρ_t^{AI} risk, s_t^{AI} expected success, and $\eta_t^{H|AI}$ uncertainty over the human state estimate.

The agent maintains a belief distribution over this state:

$$\mu_t^M(m) = p(m_t = m \mid \tau_t). \tag{5}$$

We call this *Metacognitive Self Modeling* (MSM): the agent forms a decision relevant model of its own uncertainty, limits, and risk, rather than relying on a standalone confidence score. Like human modeling, this estimate can draw on interaction history, memory, tool outcomes, user corrections, and prior evidence of failure or success, all represented through τ_t.

AI metacognition guides calibrated action. Metacognition assesses not only output confidence but also action feasibility, capability boundaries, and error costs. These judgments matter in open ended tasks involving ambiguous goals, incomplete evidence, multistep reasoning, tool calls, and human interpretation. Metacognition can prevent infeasible attempts, unsupported claims, and continuation when error costs are too high. It can guide agents to retrieve evidence (Lewis et al., 2020) before answering, decompose difficult requests into safer subtasks (Yao et al., 2022), communicate uncertainty (Lin et al., 2022a; Steyvers & Peters, 2025), or defer to human experts (Mozannar & Sontag, 2020). The central question is how uncertainty should constrain the agent's next action.

Remaining Limitations of Metacognition. Current approaches emphasize calibration and uncertainty estimation (Guo et al., 2017; Lakshminarayanan et al., 2017; Shorinwa et al., 2025), hallucination detection (Manakul et al., 2023; Farquhar et al., 2024; Ji et al., 2023; Min et al., 2023; Li et al., 2023a; Kalai & Vempala, 2024), selective prediction and abstention (Geifman & El-Yaniv, 2017; Mozannar & Sontag, 2020), self consistency (Wang et al., 2022), critique (Saunders et al., 2022), and reflection (Shinn et al., 2023). Recent work adds verbal uncertainty, faithful confidence

expression, dynamic abstention, uncertainty guided tool use, and interpretable reasoning decomposition (Shi et al., 2026; Gani et al., 2026; Davidov et al., 2026; Zhou et al., 2026c; Li et al., 2025; 2026d). Although useful, these methods often focus on individual outputs, reasoning traces, or isolated actions. Such estimates alone do not determine whether to ask, warn, retrieve, simplify, or defer. Agentic interaction therefore requires self assessment that connects uncertainty to capability boundaries, action feasibility, and interaction level risk.

3 Coupled Cognitive Alignment

3.1 Misalignment between Human Modeling and AI Metacognition

Alignment objective. The preceding section introduced human modeling and AI metacognition as two uncertainty aware estimates, μ_t^H and μ_t^M, over latent states. We argue that the central problem is not only that each inference problem is challenging, but that the two are often decoupled in agentic decision making and planning. We use *misalignment* to refer to a policy level mismatch between human needs and the agent's reliability, risk, and effort profile of selected actions or plans. CCA aims to optimize an alignment objective defined on action a_t:

$$a_t^{AI} \in \arg\min_a \mathcal{L}_{\mathrm{CCA}}(a \mid x_t, \mu_t^H, \mu_t^M). \tag{6}$$

Here $\mathcal{L}_{\mathrm{CCA}}$ is not a standard response quality loss. It is a policy level mismatch loss induced by the coupled state of the human and agent. It is low when the action satisfies inferred human objectives, respects hidden human constraints, and uses a level of agent capability, risk, and interaction effort appropriate to what the user cares about. It is high when important human needs are handled by weak or uncertain skills, or when low importance needs trigger unnecessarily costly or complex strategies.

Human modeling must respect agent limitations. This objective clarifies several decoupling failures. Human modeling without metacognitive constraint can route a high priority human utility to a low competence agentic skill. For example, if the user strongly cares about legal correctness but the agent's legal reliability is low, directly answering from memory is misaligned; a better act may be retrieval, citation, narrowing the task, or escalation. More generally, when the human side cost of error is high, metacognitive evidence should shift the policy toward the strongest available reliable route rather than unsupported direct action.

Metacognition must account for human needs. The reverse failure occurs when metacognition is not conditioned on the human's state. The agent may underuse its capability by issuing generic caveats, abstaining, or deferring even when it can safely provide bounded help for an important user need. It may also overuse its capability by applying a complex, costly, or attention demanding strategy to a need that the user does not strongly care about. For instance, a long tool using workflow or extensive evidential scaffolding may be unnecessary when the user only wants a quick formatting suggestion. In both cases, the agent may be capable in isolation, but still interactionally misaligned.

Jointly conditioned action. Therefore, CCA does not seek plans that merely maximize inferred human utility, nor plans that merely minimize agent side risk. It seeks interaction acts that minimize the mismatch between human needs and agent capabilities. Human state estimates determine what matters, how much it matters, and what form of help is usable; agent side estimates determine which actions are feasible, reliable, risky, or unnecessarily costly. The next interaction act should therefore be jointly conditioned on both estimates rather than selected from either one alone.

3.2 Joint Epistemic State

Coupled beliefs. DCSM and MSM provide complementary beliefs (Figure 2). The former captures what the agent believes about the human; the latter captures what the agent believes about itself. For interaction, however, the agent must reason about their relation. Let

$$z_t = (h_t, m_t) \tag{7}$$

denote the coupled state of the human and agent, and let

$$\widehat{p}_t(z) = p(z_t = z \mid \tau_t) \tag{8}$$

denote the coupled posterior. Its marginals are the DCSM and MSM beliefs μ_t^H and μ_t^M.

Joint Epistemic State. We define the *Joint Epistemic State* (JES) as a summary for policy selection derived from this coupled posterior:

$$\widehat{J}_t = \Phi(\widehat{p}_t, \tau_t). \tag{9}$$

The mapping Φ is model dependent and need not be scalar; it may be any representation that preserves decision relevant uncertainty. The JES is not assumed to be a complete model of mutual belief (Clark & Brennan, 1991; Stalnaker, 2002). It is an operational summary of whether the human and the agent are sufficiently coordinated for the next interaction act. This makes explicit that JES is not an additional score placed beside DCSM and MSM; it is the interface through which the two belief states are coupled for policy selection.

3.3 Interaction Policy Conditioned on JES

Structured interaction actions. CCA defines the agent policy over structured interaction actions. Given the task context and the current JES:

$$a_t^{AI} = (y_t, \delta_t, c_t), \tag{10}$$

$$a_t^{AI} \sim \pi^{AI}(\cdot \mid x_t, \widehat{J}_t). \tag{11}$$

Here y_t denotes the task level content, δ_t the interaction mode, and c_t the communication act. The mode δ_t specifies whether the agent should proceed directly or shift to a support action such as clarification, evidence gathering (Deng et al., 2023; Li et al., 2023b), deferral, or escalation. The communication act c_t specifies how the agent presents its epistemic state, including confidence, uncertainty, rationale, limitations, or risk.

Reciprocal constraints. Conditioning on $\widehat{J}_t$ does not discard DCSM or MSM. Rather, it requires the policy to use them through an estimate of shared epistemic ground (Clark & Brennan, 1991; Stalnaker, 2002). The human model determines how metacognitive information should be used and communicated; the metacognitive model determines whether and how strongly the agent should rely on its human model. For example, high AI risk or a violated capability boundary should shift δ_t away from direct action. Evidence that the user is confused, overloaded, or likely to rely too strongly should shift c_t toward more scaffolded communication, stronger limitation statements, or deferral.

3.4 Recursive Coupled Update

CCA is dynamic: the joint epistemic state is not computed once and then held fixed. Each interaction act changes the human state, reveals new evidence about the user's needs and understanding, and may also expose new information about the agent's own reliability. The agent therefore updates its coupled belief after each turn.

Interaction feedback. Let

$$f_{t+1} = (a_t^H, o_{t+1}, e_{t+1}) \tag{12}$$

denote the feedback observed after the agent's action, where a_t^H is the human response, o_{t+1} is the next interaction observation, and e_{t+1} is internal evidence about the agent's uncertainty, capability, or reliability. The interaction history is updated as

$$\tau_{t+1} = \tau_t \oplus (a_t^{AI}, f_{t+1}), \tag{13}$$

where $\oplus$ denotes appending new interaction evidence to the history.

Coupled belief update. Rather than specifying a particular transition or likelihood model, we write the coupled belief update abstractly as

$$\widehat{p}_{t+1} = \mathcal{U}(\widehat{p}_t, a_t^{AI}, f_{t+1}, \tau_t), \tag{14}$$

where $\mathcal{U}$ is a model dependent update operator. This operator may be implemented by Bayesian filtering, learned state tracking, memory update, tool based verification, self evaluation, or other mechanisms. The joint epistemic state is then

$$\widehat{J}_{t+1} = \Phi(\widehat{p}_{t+1}, \tau_{t+1}). \tag{15}$$

Feedback loop.

$$\widehat{J}_t \rightarrow a_t^{AI} \rightarrow f_{t+1} \rightarrow \widehat{p}_{t+1} \rightarrow \widehat{J}_{t+1}. \tag{16}$$

The key point is not the specific form of the update, but the dependency structure: the agent's action affects the human, the human's response provides evidence about whether the assumed shared epistemic ground was adequate, and this evidence revises both the human side and agent side components of the coupled state. A concrete case is given to illustrate this feedback loop in Appendix B.

3.5 What the Formulation Rules Out

This formulation rules out insufficient designs. First, personalization without metacognitive gating is incomplete because the agent may adapt strongly to an uncertain user model or to a task outside its capability boundary. Second, abstention or uncertainty communication without human state modeling is incomplete because epistemic signals must be translated into actions appropriate for the user. Third, static output level evaluation is incomplete because the agent's action changes the human's beliefs, trust, cognitive load, and future reliance. Under CCA, neither μ_t^H nor μ_t^M is sufficient for action alone, and $\widehat{J}_t$ must be revised rather than treated as a compatibility score. Human modeling gates how metacognition is communicated; metacognition gates when personalization is appropriate.

4 High Stakes Applications

Education and tutoring. A tutor responding to a student's incorrect derivation should assess whether the difficulty is conceptual or procedural (Corbett & Anderson, 1994; Piech et al., 2015). Without enough student work, CCA would favor asking for an intermediate step before choosing a hint or explanation (Li et al., 2026c).

Clinical and healthcare decision support. When explaining a medical report, an assistant should tailor explanation depth to the patient's health literacy and concerns (Griot et al., 2025). If clinical context is incomplete, CCA would favor clarifying questions, stating evidence limits, and preparing questions for a clinician.

AI companions and mental health support. Repeated requests for reassurance may reflect distress, uncertainty, or ordinary conversation. A companion should preserve that uncertainty and its non-clinical role; CCA would guide bounded support, boundary reminders, or human involvement as appropriate to the situation (Ren et al., 2026; Reani et al., 2026).

Legal, financial, and organizational decision support. For contract review or resource allocation, an assistant should elicit the user's goals and risk constraints while recognizing its own incomplete context or lack of decision authority. CCA would guide option comparison, explicit assumptions, or referral to accountable experts.

Software engineering and scientific assistance. Before modifying code or planning an experiment, an agent must check the user's intended scope and whether it can verify the proposed steps. CCA would guide clarification, dependency checks, small tests, or handoff before execution.

Human and AI teaming. In emergency response planning, collaborators differ in expertise and workload. CCA would combine estimates of those constraints with the agent's reliability to suggest task assignments, request confirmation, or return control as conditions change.

These scenarios illustrate proposed uses of CCA; Appendix E provides fuller discussions.

5 Challenges

Partial observability. Human beliefs, confusion, trust, and workload are rarely labeled and must be inferred from ambiguous interaction traces. A pause, for example, may reflect confusion, distraction, or careful verification. Such observations should constrain uncertain beliefs rather than serve as direct labels for mental states. Agent side signals are similarly indirect: confidence or retrieval scores do not directly reveal whether an action is feasible or safe.

Evaluation over interaction. Answer accuracy alone cannot establish whether the agent chose an appropriate interaction act. Evaluation should test how policies respond to changes in user expertise, cognitive load, capability boundaries, and error costs. Useful protocols include counterfactual user state evaluation, metacognitive stress testing, and episode level analysis of reasoning and monitoring (Li et al., 2025; 2026d). Longitudinal evaluation can track learning, error recovery, and calibrated reliance, capturing effects that individual responses do not reveal.

Deployment boundaries. Richer human models introduce privacy and manipulation risks and should not automatically justify greater autonomy. Systems should infer only task relevant states, preserve uncertainty, and avoid treating transient signals as stable attributes. In high stakes settings, narrowing the task or handing off a decision may be the appropriate action. Deferral should be treated as a legitimate interaction act rather than a service failure (Mozannar & Sontag, 2020). Appendix F discusses these challenges in detail.

6 Opportunities

Interaction level benchmarks. Benchmarks can specify task context, possible user states, agent uncertainty, and expected interaction modes. This would shift evaluation toward whether the agent should answer, clarify, retrieve evidence, communicate limitations, or defer, rather than final answer quality alone. Recent benchmarks for user simulation, collaboration, companion safety, and clarification illustrate this shift (Wang et al., 2026; Qian et al., 2026b; Ren et al., 2026; Edwards & Schuster, 2026).

Human modeling under uncertainty. User models can represent uncertainty over cognitive states alongside preferences and behavior. Work on implicit preferences, personality anchoring, and recursive perspectives motivates richer user models, but their inferences remain dependent on uncertain latent assumptions (Zeng et al., 2026; Javadi et al., 2026; Lei et al., 2026). When available evidence does not distinguish expertise, confusion, or overload, diagnostic questions and explanations robust to several plausible states may be more appropriate than stronger personalization.

Adaptive uncertainty communication. Metacognitive signals can be translated into forms that match the user's expertise, trust, workload, and decision context. The same uncertainty may call for a caveat, supporting rationale, request for information, safer subtask, or escalation. Uncertainty communication itself must be evaluated, not merely appended to answers (Shi et al., 2026; Gani et al., 2026; Li et al., 2025).

Learning from interventions. Training and evaluation can examine how interaction acts change understanding, trust, workload, reliance, and decisions. These outcomes would connect policy selection to the effects of assistance on users, extending the target beyond predicting their next response or completing the immediate task. Appendix G develops these directions.

7 Conclusion

This paper has argued that AI systems should move beyond response level alignment toward interaction level policy calibration. In extended human workflows, the key question is not only whether an output is correct or preferred, but whether the agent chooses an appropriate next act under uncertainty. Coupled Cognitive Alignment frames this choice as jointly constrained by human modeling and AI metacognition: what the user needs, understands, and can act upon, and what the agent can reliably know, do, and risk. This reframing shifts the research target from better isolated responses to better interaction policies under coupled uncertainty about humans and AI.

Limitations

As a position paper, we present CCA through an illustrative tutoring dialogue and a controlled simulation; no general system is evaluated with real users. Representing and coupling uncertain human and agent states remains underspecified. Future work should test CCA across domains and determine when its benefits justify added complexity and risks to privacy and accountability.

ETHICAL STATEMENTS

CCA involves modeling latent human states such as goals, uncertainty, trust, cognitive load, and reliance. While such modeling may improve interaction, it also raises risks of privacy invasion, manipulation, and unfair treatment. Systems should model only task relevant states, represent uncertainty over inferred states, and avoid treating temporary signals as fixed personal attributes.

In high stakes domains, richer user modeling should not justify greater automation without reliable capability or authority. CCA should support bounded assistance, uncertainty communication, clarification, and appropriate handoff, while preserving privacy, consent, fairness, and calibrated human reliance.

AI USE STATEMENT

During preparation of this submission, AI is used to assist in editing and formatting the text. The authors have reviewed and edited the content as needed and take full responsibility for the submission's content.

REFERENCES

Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

Chris Adams, Arjun Singh Banga, Parveen Bansal, Souvik Bhattacharya, Rujin Cao, Pedro Canahuati, Nate Cook, Brian Ellis, Prabhakar Goyal, Gurinder Grewal, et al. Automating low-risk code review at meta: Radar, risk calibration, and review efficiency. *arXiv preprint arXiv:2605.30208*, 2026.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

John R Anderson and Christian J Lebiere. *The atomic components of thought*. Psychology Press, 2014.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Or Bachar, Or Levi, Sardhendu Mishra, Adi Levi, Manpreet Singh Minhas, Justin Miller, Omer Ben-Porat, Eilon Sheetrit, and Jonathan Morra. Llm performance predictors: Learning when to escalate in hybrid human-ai moderation systems. *arXiv preprint arXiv:2601.07006*, 2026.

Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 0064, 2017.

Safia Baloch and Rahemeen Khan. Toward human-centered multi-agent systems: Integrating cognition, culture, values, and cooperation in ai agents. *arXiv preprint arXiv:2606.08274*, 2026.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16, 2021.

Adam Bawatneh, Sagar Sapkota, Amrit Singh Bedi, Santu Karmaker, and Mubarak Shah. Omnitom: Benchmarking theory of mind in llms via explicit belief modeling. *arXiv preprint arXiv:2605.26322*, 2026.

Suhas BN, Andrew M Sherrill, Rosa I Arriaga, Chris W Wiese, and Saeed Abdullah. Ai safety training can be clinically harmful. *arXiv preprint arXiv:2604.23445*, 2026.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287.

Yancheng Cao, Yishu Ji, Yue Fu, Sahiti Dharmavaram, Meghan Turchioe, Natalie C Benda, Lena Mamykina, Yuling Sun, and Xuhai Xu. More than decision support: Exploring patients' longitudinal usage of large language models in real-world healthcare-seeking journeys. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2026.

Han Chen, Ming Li, Chenguang Wang, Yijun Liang, Dawei Zhou, Hong Jiao, and Tianyi Zhou. LLMs struggle to measure what distinguishes students of different proficiency levels: A study of item discrimination in reading comprehension assessment. *arXiv preprint arXiv:2606.18709*, 2026a.

Jiaju Chen, Yuxuan Lu, Jiayi Su, Chaoran Chen, Songlin Xiao, Zheng Zhang, Yun Wang, Yunyao Li, Jian Zhao, Tongshuang Wu, et al. Humans' almanac: A human collaboration dataset of action-level mental model annotations for agent collaboration. *arXiv preprint arXiv:2606.06388*, 2026b.

Jiaju Chen, Bo Sun, Yuxuan Lu, Yun Wang, Dakuo Wang, and Bingsheng Yao. Collabsim: A cscw-grounded methodology for investigating collaborative competence of llm agents through controlled multi-agent experiments. *arXiv preprint arXiv:2606.06399*, 2026c.

Xinghe Chen, Naiming Liu, and Shashank Sonkar. Malrulelib: large-scale executable misconception reasoning with step traces for modeling student thinking in mathematics. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15112–15138, 2026d.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15959–15983, 2024.

Herbert H Clark and Susan E Brennan. Grounding in communication. 1991.

Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

Hen Davidov, Nachshon Cohen, Oren Kalinsky, Yaron Fairstein, Guy Kushilevitz, Ram Yazdi, and Patrick Rebeschini. Knowing when to quit: A principled framework for dynamic abstention in llm reasoning. *arXiv preprint arXiv:2604.18419*, 2026.

Mengyi Deng, Zhiwei Li, Xin Li, Tingyu Zhu, Ying Zhao, Zhijiang Guo, and Wei Wang. Uncertainty-aware clarification in llm agents with information gain. *arXiv preprint arXiv:2606.03135*, 2026.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, Léo Boisvert, Megh Thakkar, Quentin Cappart, David Vazquez, et al. Workarena: How capable are web agents at solving common knowledge work tasks? *arXiv preprint arXiv:2403.07718*, 2024.

Zhangqi Duan, Shuyan Huang, Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. Who am i? history-aware profiles for student simulation in tutoring dialogues. *arXiv preprint arXiv:2605.30051*, 2026.

Nicholas Edwards and Sebastian Schuster. Ask or assume? uncertainty-aware clarification-seeking in coding agents. *arXiv preprint arXiv:2603.26233*, 2026.

Iizalaarab Elhaimeur and Nikos Chrisochoides. Itas: A multi-agent architecture for llm-based intelligent tutoring. *arXiv preprint arXiv:2604.24808*, 2026.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, 2023.

Xiaoyang Fan, Yufan Cai, Zhe Hou, and Jin Song Dong. Uncertainty reasoning with large language models for explainable disease diagnosis. *arXiv preprint arXiv:2605.25566*, 2026.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Robinson Ferrer, Damla Turgut, Zhongzhou Chen, and Shashank Sonkar. When can we trust llm graders? calibrating confidence for automated assessment. *arXiv preprint arXiv:2603.29559*, 2026.

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

Areeb Gani, Asal Meskin, Gabrielle Kaili-May Liu, and Arman Cohan. Quantifying faithful confidence expression in large reasoning models. *arXiv preprint arXiv:2606.03969*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems*, pp. 299–315, 2022.

Piotr J Gmytrasiewicz and Prashant Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005.

Maharshi Gor, Yoo Yeon Sung, Yu Hou, Eve Fleisig, Irene Ying, Tianyi Zhou, and Jordan Boyd-Graber. Ai, take the wheel: What drives delegation and trust in human-computer cooperative question answering? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL) Findings*, 2026.

Rahul Gorijavolu, Kaushik Madapati, Pritika Vig, Rawan Abulibdeh, Nikhil Jaiswal, Mahri Kadyrova, Zeamanuel Hailu Tesfaye, Charles Senteio, Paula Maurutto, and Leo Anthony Celi. Testing the black box: Structural barriers to independent evaluation of consumer-facing health llms. *arXiv preprint arXiv:2606.08483*, 2026.

Abinitha Gourabathina, Inkit Padhi, Manish Nagireddy, Subhajit Chaudhury, and Prasanna Sattigeri. Answering the wrong question: Reasoning trace inversion for abstention in llms. *arXiv preprint arXiv:2604.02230*, 2026.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. Towards realistic personalization: Evaluating long-horizon preference following in personalized user-llm interactions. *arXiv preprint arXiv:2603.04191*, 2026.

Peixuan Han, Hongyi Du, Jiayu Liu, Yihang Sun, Yutong Liu, and Jiaxuan You. Ψ-Bench: Evaluating persona-sensitive influencing in persuasive dialogues. *arXiv preprint arXiv:2606.02754*, 2026a.

Yi Han, Lingfei Qian, Yan Wang, Yueru He, Xueqing Peng, Dongji Feng, Yankai Chen, Haohang Li, Yupeng Cao, Jimin Huang, et al. Can llm agents be cfos? a benchmark for resource allocation in dynamic enterprise environments. *arXiv preprint arXiv:2603.23638*, 2026b.

Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pp. 139–183. Elsevier, 1988.

Johannes Härtel. Agentic repository mining: A multi-task evaluation. *arXiv preprint arXiv:2605.04845*, 2026.

Rima Hazra, Bikram Ghuku, Ilona Marchenko, Yaroslava Tokarieva, Sayan Layek, Somnath Banerjee, Julia Stoyanovich, and Mykola Pechenizkiy. Safetutors: Benchmarking pedagogical safety in ai tutoring systems. *arXiv preprint arXiv:2603.17373*, 2026.

Lo Gullstrand Heander, Agnia Sergeyuk, Ilya Zakharov, Emma Söderberg, and Nikita Mukhortov. Trust-calibrated code review: A participatory design study of review workflows for llm-generated multi-file changes. *arXiv preprint arXiv:2606.01969*, 2026.

Fangrui Huang, Souhad Chbeir, Arpandeep Khatua, Sheng Wang, Sijun Tan, Kenan Ye, Lily Bailey, Merryn Daniel, Ryan Louie, Sanmi Koyejo, et al. Therapygym: Evaluating and aligning clinical fidelity and safety in therapy chatbots. *arXiv preprint arXiv:2603.18008*, 2026a.

Shuyan Huang, Alexander Scarlatos, Jaewook Lee, and Andrew Lan. Interpretable difficulty-aware knowledge tracing in tutor-student dialogues. *arXiv preprint arXiv:2605.01097*, 2026b.

Shim Jaechang and Unggi Lee. The tutoring effectiveness index: Predicting llm math tutor quality from four conversation signals. *arXiv preprint arXiv:2605.30666*, 2026.

Vahid Sadiri Javadi, Aksa Aksa, Fryderyk Róg, Lucie Flek, and Johanne R. Trippas. Personality anchoring for social simulation: Linking personality, social behavior, and interaction success with llm agents. *arXiv preprint arXiv:2606.06936*, 2026.

Eugene Yu Ji, Igor Grossmann, and Amir-Hossein Karimi. Metacognition should be the scientific framework for bounded and effective self-governance in generative ai. *arXiv preprint arXiv:2605.23981*, 2026.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, pp. 54107–54157, 2024.

Clark Mingxuan Ju, Yuwei Qiu, Tong Zhao, and Neil Shah. Self-supervised user profile generation for personalization. *arXiv preprint arXiv:2606.05336*, 2026.

Prerna Juneja and Lika Lomidze. Persona-grounded safety evaluation of ai companions in multi-turn conversations. *arXiv preprint arXiv:2605.00227*, 2026.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Adam Tauman Kalai and Santosh S Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pp. 160–171, 2024.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*, 2022.

Yusuf Kesmen, Fay Elhassan, Jiayi Ma, Julien Stalhandske, Yena Chang, David Sasu, Alexandra Kulinkina, Akhil Arora, Lars Klein, and Mary-Anne Hartley. Mobayes: A modular bayesian framework for separating reasoning from language in conversational clinical decision support. *arXiv preprint arXiv:2604.20022*, 2026.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14397–14413, 2023.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, 2024.

Rishik Kondadadi and John E Ortega. L2d-clinical: Learning to defer for adaptive model selection in clinical text classification. *arXiv preprint arXiv:2604.13285*, 2026.

Fanqi Kong, Jiayi Zhang, Mingyi Deng, Chenglin Wu, Yuyu Luo, and Bang Liu. Infopo: Information-driven policy optimization for user-centric agents. *arXiv preprint arXiv:2603.00656*, 2026.

George Koomullil. Proof-carrying certificates for llm pipelines: A trust-boundary architecture. *arXiv preprint arXiv:2605.16407*, 2026.

Asher Koriat. Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of experimental psychology: General*, 126(4):349, 1997.

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1369–1385, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Gregory Kang Ruey Lau, Hieu Dao, Nicole Kan Hui Lin, and Bryan Kian Hsiang Low. Uncertainty quantification for multimodal large language models with incoherence-adjusted semantic volume. *arXiv preprint arXiv:2602.24195*, 2026.

John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

Chao Lei, Guang Hu, Meng Yang, Yanbei Jiang, and Nir Lipovetzky. Mind the perspective: Let's reason recursively for theory of mind. *arXiv preprint arXiv:2606.11724*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Changcheng Li, Jiancan Wu, Hengheng Zhang, Zhengsu Chen, Guo An, Junxiang Qiu, Xiang Wang, and Qi Tian. Confidence before answering: A paradigm shift for efficient llm uncertainty estimation. *arXiv preprint arXiv:2603.05881*, 2026a.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023a.

Lingyao Li, Deyi Li, Chen Chen, Renkai Ma, Runlong Yu, Mingquan Lin, Rui Yin, Lizhou Fan, Cathy Shyr, Siyuan Ma, et al. Llm-as-a-judge in healthcare: A scoping analysis of applications, methods, and human alignment. *arXiv preprint arXiv:2605.25273*, 2026b.

Ming Li, Nan Zhang, Chenrui Fan, Hong Jiao, Yanbin Fu, Sydney Peters, Qingshu Xu, Robert Lissitz, and Tianyi Zhou. Understanding the thinking process of reasoning models: A perspective from Schoenfeld's episode theory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18267–18288, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.922. URL https://aclanthology.org/2025.emnlp-main.922/.

Ming Li, Han Chen, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou. Can LLMs estimate student struggles? human-AI difficulty alignment with proficiency simulation for item difficulty prediction. *arXiv preprint arXiv:2512.18880*, 2026c.

Ming Li, Chenrui Fan, Yize Cheng, Soheil Feizi, and Tianyi Zhou. Schoenfeld's anatomy of mathematical reasoning by language models. *arXiv preprint arXiv:2512.19995*, 2026d.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. Api-bank: A comprehensive benchmark for tool-augmented llms. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 3102–3116, 2023b.

Weitao Li, Hao Zhou, Xuanyu Lei, Fandong Meng, Yuanhang Liu, Jingyi Ren, Ante Wang, Xiaolong Wang, Yuanchi Zhang, Fuwen Luo, et al. Enhancing llm metacognition via cognitive pairwise training. *arXiv preprint arXiv:2606.00869*, 2026e.

Jiacheng Lin, Kun Qian, Arvind Srinivasan, Tian Wang, Fang Han, Changran Hu, Junze Liu, Ziyi Wang, Hanwen Xu, Mengmeng Xue, et al. Llm agents enable user-governed personalization beyond platform boundaries. *arXiv preprint arXiv:2605.09794*, 2026.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022b.

Jiahao Liu, Mingzhe Han, Guanming Liu, Weihang Wang, Dongsheng Li, Hansu Gu, Peng Zhang, Tun Lu, and Ning Gu. From hidden profiles to governable personalization: Recommender systems in the age of llm agents. *arXiv preprint arXiv:2604.20065*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations*, volume 2024, pp. 52989–53046, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Chunjiang Mu, Ya Zeng, Qiaosheng Zhang, Kun Shao, Chen Chu, Hao Guo, Danyang Jia, Zhen Wang, and Shuyue Hu. Adaptive theory of mind for llm-based multi-agent coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29608–29616, 2026.

Philip Müller, Nicholas Popovič, Michael Färber, and Peter Steinbach. Benchmarking uncertainty calibration in large language model long-form question answering. *arXiv preprint arXiv:2602.00279*, 2026.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Thomas O Nelson. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, pp. 125–173. Elsevier, 1990.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Thuy Ngoc Nguyen, Duy Nhat Phan, and Cleotilde Gonzalez. Dynamic theory of mind as a temporal memory problem: Evidence from large language models. *arXiv preprint arXiv:2603.14646*, 2026.

Luke Nicholls, Robert Hutto, Zephrah Soto, Hamilton Morrin, Thomas Pollak, Raj Korpan, and Cheryl Carmichael. " ai psychosis" in context: How conversation history shapes llm responses to delusional beliefs. *arXiv preprint arXiv:2604.13860*, 2026.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

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, 2023.

Shishir G Patil, Tianjun Zhang, Xin Wang, and Joseph E Gonzalez. Gorilla: Large language model connected with massive apis. *Advances in Neural Information Processing Systems*, 37: 126544–126565, 2024.

Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, Tianyi Zhou, and Robert Lissitz. Text-based approaches to item difficulty modeling in large-scale assessments: A systematic review. *arXiv preprint arXiv:2509.23486*, 2025.

Dzianis Piatrashyn, Nikita Kotelevskii, Kirill Grishchenkov, Nikita Glazkov, Ivan Nasonov, Ilya Makarov, Timothy Baldwin, Preslav Nakov, Roman Vashurin, and Maxim Panov. Redact: Uncertainty-aware deferral for llm agents. *arXiv preprint arXiv:2604.07036*, 2026.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J. Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper/5654-deep-knowledge-tracing.

Christian Poelitz, Finale Doshi-Velez, and Siân Lindley. A benchmark to assess common ground in human-ai collaboration. *arXiv preprint arXiv:2602.21337*, 2026.

Cheng Qian, Jiayu Liu, and Heng Ji. Userharness: Harnessing user minds for stronger agent theory-of-mind. *arXiv preprint arXiv:2605.27721*, 2026a.

Hong Qian, Yuanhao Liu, Zihan Zhou, Zongbao Zhang, Hanjie Ge, Haotian Shi, Liang Dou, Xiangfeng Wang, Jingwen Yang, and Aimin Zhou. Collabbench: Benchmarking and unleashing collaborative ability of llms with diverse players via proactive engagement. *arXiv preprint arXiv:2606.05793*, 2026b.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Manuele Reani, Hongjian Zhang, and Hongyu Tian. The governance of human-llm interaction: Safety gating, civility steering, and affective default lock-in. *arXiv preprint arXiv:2606.08172*, 2026.

Yanjing Ren, Reza Ebrahimi, and TengTeng Ma. Aicompanionbench: Benchmarking llms-as-judges for ai companion safety. *arXiv preprint arXiv:2606.04867*, 2026.

Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to Recommender Systems Handbook*, pp. 1–35. Springer US, Boston, MA, 2011. ISBN 978-0-387-85820-3. doi: 10.1007/978-0-387-85820-3_1. URL https://doi.org/10.1007/978-0-387-85820-3_1.

Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pp. 1–35, 2021.

Jillian Ross and Andrew W Lo. One size fits none: Heuristic collapse in llm investment advice. *arXiv preprint arXiv:2604.23837*, 2026.

Minyuan Ruan, Ziyue Wang, Kaiming Liu, Yunghwei Lai, Peng Li, and Yang Liu. Beyond words: Evaluating and bridging epistemic divergence in user-agent interaction via theory of mind. *arXiv preprint arXiv:2602.13832*, 2026.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, 2024.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 3762–3780, 2022.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

João Sedoc, Baotong Zhang, and Dean Foster. Trust but verify: Prover-verifier deliberation for selective llm prediction. *arXiv preprint arXiv:2605.25133*, 2026.

Shinas Shaji, Teena Chakkalayil Hassan, Sebastian Houben, and Alex Mitrevski. Evaluating generative models as interactive emergent representations of human-like collaborative behavior. *arXiv preprint arXiv:2605.03855*, 2026.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2257–2273, 2024.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Kaiwen Shi, Zheyuan Zhang, and Yanfang Ye. Sage: Answer-conditioned uncertainty targets for verbal uncertainty alignment. *arXiv preprint arXiv:2606.11512*, 2026.

Dongmin Shin, Yugeun Shim, Hangyeol Yu, Seewoo Lee, Byungsoo Kim, and Youngduck Choi. Saint+: Integrating temporal features for ednet correctness prediction. In *LAK21: 11th international learning analytics and knowledge conference*, pp. 490–496, 2021.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38, 2025.

Xiangyu Song, Jianxin Li, Taotao Cai, Shuiqiao Yang, Tingting Yang, and Chengfei Liu. A survey on deep learning based knowledge tracing. *Knowledge-Based Systems*, 258:110036, 2022.

Ponhvoan Srey, Quang Minh Nguyen, Xiaobao Wu, and Anh Tuan Luu. Towards reliable truth-aligned uncertainty estimation in large language models. *arXiv preprint arXiv:2604.00445*, 2026.

Robert Stalnaker. Common ground. *Linguistics and philosophy*, 25(5/6):701–721, 2002.

Mark Steyvers and Megan AK Peters. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, pp. 09637214251391158, 2025.

Zhaoxuan Tan and Meng Jiang. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*, 2023.

Zhen Tao, Riwei Lai, Chenyun Yu, Weixin Chen, Li Chen, Beibei Kong, Lei Cheng, Chengxiang Zhuo, Zang Li, and Qingqiang Sun. Sager: Self-evolving user policy skills for recommendation agent. *arXiv preprint arXiv:2604.14972*, 2026.

Aditya Taparia, Ransalu Senanayake, Kowshik Thopalli, and Vivek Narayanaswamy. The anatomy of uncertainty in llms. *arXiv preprint arXiv:2603.24967*, 2026.

Virgill van der Meer and Julien Rossi. Legalcheck: Retrieval-and context-augmented generation for drafting municipal legal advice letters. *arXiv preprint arXiv:2605.12012*, 2026.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

Briana Vecchione, Meryl Ye, Livia Garofalo, and Ranjit Singh. Engagement-optimized care: When llms become mental health infrastructure. *arXiv preprint arXiv:2605.23787*, 2026.

Danqing Wang, Akshay Sivaraman, and Lei Li. Crab-bench: Evaluating llm agents under complex task dependencies and human-aligned user simulation. *arXiv preprint arXiv:2606.01815*, 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024a.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. Sotopia-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12912–12940, 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhongyuan Wang and Pratyusha Vemuri. When the tool decides: Llm agents defer blindly to graph neural network tools, and stronger backbones defer more. *arXiv preprint arXiv:2606.14476*, 2026.

Veith Weilnhammer, Kevin YC Hou, Lennart Luettgau, Christopher Summerfield, Raymond Dolan, and Matthew M Nour. Vulnerability-amplifying interaction loops: a systematic failure mode in ai chatbot mental-health interactions. *arXiv preprint arXiv:2602.01347*, 2026.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. Humanlm: Simulating users with state alignment beats response imitation. *arXiv preprint arXiv:2603.03303*, 2026.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

Wenyi Xiao, Xinchi Xu, and Leilei Gan. Vl-calibration: Decoupled confidence calibration for large vision-language models reasoning. *arXiv preprint arXiv:2604.09529*, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Xiaohu Xie, Xiaohu Liu, and Benjamin Yao. Know when you're wrong: Aligning confidence with correctness for llm error detection. *arXiv preprint arXiv:2603.06604*, 2026.

Shuxing Yang, Fujia Chen, Rui Zhao, Junyao Wu, Yize Wang, Haiyao Luo, Ning Han, Qiaolu Chen, Yuze Hu, Wenhao Li, et al. End-to-end autonomous scientific discovery on a real optical platform. *arXiv preprint arXiv:2604.27092*, 2026.

Bingsheng Yao, Chaoran Chen, April Yi Wang, Tongshuang Wu, Toby Jia-Jun Li, and Dakuo Wang. From human-human collaboration to human-agent collaboration: A vision, design philosophy, and an empirical framework for achieving successful partnerships between humans and llm agents. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Zhihao Yuan, Yunze Xiao, Ming Li, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. Towards valid student simulation with large language models. *arXiv preprint arXiv:2601.05473*, 2026.

Jiahao Zeng, Ming Tang, and Ningning Ding. Learning to route llms from implicit cost-performance preferences via meta-learning. *arXiv preprint arXiv:2606.06178*, 2026.

Junli Zha, Jinbo Wang, Chao Zhou, and Xiang Song. Trace2policy: From expert behavior traces to self-evolving decision agents. *arXiv preprint arXiv:2606.10457*, 2026.

Beiwen Zhang, Yongheng Liang, Guowei Zou, Haitao Wang, and Hejun Wu. Distilling llm reasoning into an interpretable policy tree for human-ai collaboration. *arXiv preprint arXiv:2606.08596*, 2026a.

Lechen Zhang, Jiarui Liu, and Tal August. Re-centering humans in llm personalization. *arXiv preprint arXiv:2606.06614*, 2026b.

Xunyi Zhao, Sihao Lin, Gengze Zhou, Zerui Li, Shijie Li, Wei Tao, Jiajun Liu, and Qi Wu. Ask when it pays: Cost-aware open-ended interaction for instance goal navigation. *arXiv preprint arXiv:2606.03175*, 2026.

Jike Zhong, Yuxiang Lai, Ming Li, Yuheng Li, Wuao Liu, Behzad Dariush, Konstantinos Psounis, and Shao-Yuan Lo. From shortcuts to reasoning: Robust post-training of theory of mind with reinforcement learning. *arXiv preprint arXiv:2606.09092*, 2026.

Hongjian Zhou, Xinyu Zou, Jinge Wu, Sean Wu, Junchi Yu, Bradley Max Segal, Tobias Erich Niebuhr, Sara Amro, Michael Petrus, Sheikh Momin, et al. Measuring epistemic resilience of llms under misleading medical context. *arXiv preprint arXiv:2606.12291*, 2026a.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3623–3643, 2024a.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pp. 40975–41019, 2024b.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, et al. Mind the sim2real gap in user simulation for agentic tasks. *arXiv preprint arXiv:2603.11245*, 2026b.

Yijin Zhou, Linqian Zeng, Xiaoya Lu, Wenyuan Xie, Dongrui Liu, Junchi Yan, and Jing Shao. Exploring agentic tool-calling decisions via uncertainty-aligned reinforcement learning. *arXiv preprint arXiv:2606.06976*, 2026c.

Haotian Zong, Binze Li, Yufei Long, Sinyin Chang, Jialong Wu, and Gillian K Hadfield. I-calm: Incentivizing confidence-aware abstention for llm hallucination mitigation. *arXiv preprint arXiv:2604.03904*, 2026.

A What Distinguishes CCA from Existing Research Directions?

Separate research objectives. Existing research directions address important but largely isolated components of interactive intelligence. Personalization, user modeling, theory of mind, and human simulation focus primarily on inferring latent human states such as preferences, beliefs, intentions, knowledge, or likely future behaviors. Their primary objective is to better understand or predict the human, often to improve personalization or behavioral adaptation. In contrast, calibration, uncertainty estimation, hallucination detection, abstention, and learning to defer focus on modeling the agent itself, including confidence, reliability, uncertainty, and capability limitations. Human and AI teaming and shared mental model research recognize the importance of coordination between humans and AI, while agentic planning and tool use systems expand the space of possible actions through retrieval, planning, and external tools. However, these directions generally treat human understanding, AI self assessment, and action selection as separate problems rather than components of a unified interaction policy.

Coupled decision making. CCA differs in both scope and objective. Rather than asking only what the human wants or how reliable the AI is, CCA treats interaction as a coupled decision problem under uncertainty. The central challenge is determining how uncertainty about the human and uncertainty about the AI should jointly influence the next interaction act. An agent that models human needs well but lacks awareness of its own limitations may confidently act beyond its competence. Conversely, an agent with strong self assessment but poor human modeling may defer too often, become too cautious, or fail to provide assistance that is actually useful. CCA therefore explicitly couples Dynamic Cognitive State Modeling (DCSM) and Metacognitive Self Modeling (MSM) through the Joint Epistemic State (JES), creating a representation for decision making that captures uncertainty on both sides of the interaction and their dependencies.

Alignment through interaction policies. The key novelty of CCA is thus not simply combining human modeling and AI metacognition, but reframing alignment as an interaction policy problem. Existing approaches often optimize response quality, personalization quality, calibration quality, or planning quality in isolation. CCA instead focuses on selecting the appropriate interaction strategy: whether the agent should answer, ask a clarifying question, provide a hint, retrieve evidence, explain uncertainty, recommend a safer subtask, defer to a human expert, or escalate the decision. In this view, human state inference and AI self assessment are not end goals themselves; they are coupled prerequisites for adaptive interaction policy selection. This shifts the unit of alignment from individual outputs to the calibration of interaction decisions over time, particularly in settings where human needs, AI capabilities, and associated risks are all uncertain and dynamically evolving.

B Case Study: Coupled Cognitive Alignment in Multi-Turn Tutoring

Consider a four-turn exchange about learning derivatives, as shown in table 2. The HumanState is the tutor's current (and uncertain) estimate of the student's understanding, misconceptions, confidence, and learning needs. The AgentState represents the tutor's self-assessment of its confidence, capabilities, uncertainty, cost, and risk. Together, they form the Joint Epistemic State (JES), which guides interaction policy by matching the student's needs with the tutor's estimated capabilities. After each interaction, both states are updated using the student's feedback, yielding a new JES for the next round. This example shows how the selected action could change as evidence accumulates; it does not measure learning gains.

The example illustrates why the two estimates matter for action selection. The tutor's confidence in the mathematical solution remains high throughout, yet the hypothetical action changes from asking, to hinting, to eliciting the next step, and finally to verifying as HumanState and AgentState evolve.

The estimated HumanState indicates the student's potential learning needs, but not whether the tutor's diagnosis is reliable enough to act upon. The AgentState reflects the tutor's confidence in solving and teaching, but not what the student actually needs. Consequently, the tutor may either explain an unconfirmed misconception or provide the correct answer before identifying the student's difficulty. Even maintaining both estimates independently is insufficient if the interaction policy does not couple and update them jointly. By conditioning the policy on the Joint Epistemic State (JES), the tutor first gathers evidence to refine its estimate of the student's cognitive state, then adapts its tutoring

Table 2: An illustrative four-turn tutoring interaction. HumanState and AgentState are qualitative estimates used to form the JES and select the next interaction act.

Turn and new evidence	**HumanState: tutor's estimate**	**AgentState: tutor's self-assessment**	**JES and next action**
1. Student: "I think the derivative of x^2 is x."	Low knowledge; medium confusion; power-rule error suspected.	Solution confidence: high; diagnosis confidence: low; over-helping risk: high.	Evidence is insufficient for an explanation. Ask: "How did you apply the power rule?"
2. Student: "I thought the exponent just disappears."	Misconception: suspected to supported; confusion: medium to high.	Diagnosis confidence: low to high; solution confidence: high; teaching confidence: medium; over-helping remains a concern.	The new evidence now supports the error diagnosis. Move from asking to a targeted hint: "The exponent becomes a multiplier. What is it for x^2?"
3. Student: "Oh, so I multiply by 2?"	Understanding: improving; confusion: decreasing.	Teaching confidence: medium to high; over-helping risk: high to medium.	The student can continue with limited support. Ask the student to take the next step: "What happens to the exponent next?"
4. Student: "The derivative is $2x$."	Misconception: resolving; knowledge: improved.	Diagnosis confidence: high; interaction risk: low.	Further scaffolding is unnecessary. Verify and briefly summarize the rule.

strategy accordingly, guiding the student toward understanding rather than simply producing the correct answer.

C Controlled Tutoring Simulation: Coupling HumanState and AgentState

Appendix B traces possible decisions in one illustrative interaction. Here, we compare tutoring policies quantitatively in a controlled multi-turn simulation. The student's latent misconceptions and the effectiveness of different teaching actions are known to the simulator but hidden from the tutor.

We construct student profiles by sampling two or three misconceptions from *MalruleLib* (Chen et al., 2026d). For each question associated with a misconception, the student's understanding is represented by one of three latent states (`intact wrong`, `shaken`, or `correct`). Three teaching actions are assigned different transition probabilities between these states, with their effectiveness varying across misconceptions. We generate 40 training episodes with 30 questions each and evaluate on 20 held-out questions.

During training, the tutor must infer both the student's latent state and the effectiveness of different teaching actions from delayed feedback, which is revealed only after five teaching turns. HumanState captures the tutor's current estimate of the student, while AgentState tracks its confidence in these estimates and therefore determines when further exploration is warranted. At test time, these estimates are fixed, allowing us to evaluate whether the learned JES supports effective tutoring decisions. We report success after 5 and 10 turns, AUC over the first 10 turns, mean turns to correction, and the rate of selecting the most effective teaching action.

Method	Success@5	Success@10	AUC@10	Mean Turn	Strong Action Rate
No Modeling	0.576	0.836	0.537	5.469	0.359
HumanState Only	0.551	0.769	0.497	5.794	0.244
HumanState + AgentState	0.656	0.869	0.585	5.016	0.428
Oracle Student Profile	0.720	0.915	0.634	4.576	0.658

Table 3: Outcomes in the controlled tutoring simulation under different forms of cognitive-state modeling.

Table 3 shows that coupling HumanState with AgentState improves tutoring performance across all metrics, whereas HumanState alone can perform worse than no explicit modeling. One possible explanation is premature exploitation: treating an estimated HumanState as reliable by default may lead the tutor to repeat actions it currently believes to be effective, producing a biased interaction history that reinforces incorrect estimates. AgentState provides the missing epistemic signal. When confidence is low, the tutor explores alternative actions to gather informative evidence; as confidence increases, it shifts toward exploiting the actions estimated to be most effective.

Beyond the illustrative case in Appendix B, this controlled simulation provides a minimal quantitative test of CCA's core proposal: choosing teaching actions from both an inferred student state and the tutor's confidence in that inference. The coupled policy achieves higher success, selects the simulator's most effective action more often, and reaches correction sooner than both non-oracle baselines, while remaining below the oracle. These results offer initial support for the coupling mechanism in this simplified setting; tests with real students and broader tasks remain future work.

D Additional Related Work

This section follows the same related work categories used in Tables 1 and 4. The main text uses these categories to position CCA at a high level; here we spell out how each line of work contributes a partial ingredient while leaving open the coupled policy problem that CCA targets.

D.1 Personalization

Personalization and recommender system research studies how systems adapt outputs to users, often through user profiles, preference histories, contextual features, or interaction logs (Ricci et al., 2011; 2021). Recent language model personalization extends this idea to personalized text generation, recommendation, and even model routing under implicit cost and performance preferences (Tan & Jiang, 2023; Geng et al., 2022; Salemi et al., 2024; Zeng et al., 2026). Recent studies also cover human centered personalization, long horizon preference following, user governed personalization across platforms, governable recommender agents, self supervised profile construction, evolving recommendation policies, and persona sensitive influence risks (Zhang et al., 2026b; Guo et al., 2026; Lin et al., 2026; Liu et al., 2026; Ju et al., 2026; Tao et al., 2026; Han et al., 2026a). These methods are important because they show that the same task may require different support for different users.

For CCA, personalization is a useful but incomplete form of human modeling. A profile can help infer stable preferences, prior interests, expected style, or cost tolerance, but interaction policy often depends on transient states: whether the user is confused, uncertain, overloaded, relying too heavily, or asking for a hint rather than a direct answer. Personalization also needs metacognitive gating. If the agent adapts strongly to an uncertain user estimate or acts on an inferred preference in a high risk task, personalization can become overconfident rather than helpful.

D.2 Theory of Mind and Human Simulation

Theory of mind and human simulation work focuses on inferring or reproducing latent human states such as beliefs, goals, intentions, knowledge, and likely reactions. Rational action and inverse planning accounts treat observed behavior as evidence about unobserved mental states (Baker et al., 2017). Recent NLP benchmarks and models test whether language systems can reason about mental states in stories, dialogues, and social situations (Sap et al., 2022; Kim et al., 2023; Shapira et al., 2024; Chen et al., 2024). Recent studies extend this line toward recursive perspective construction, personality anchored social behavior, human aligned user simulation under complex task dependencies, explicit belief modeling, temporal memory for theory of mind, and robust post training of mental state reasoning (Lei et al., 2026; Javadi et al., 2026; Wang et al., 2026; Bawatneh et al., 2026; Nguyen et al., 2026; Zhong et al., 2026).

Another recent direction moves from isolated theory of mind probes toward user and agent interaction and collaboration settings. Work on user harnessing, epistemic divergence, simulation to real gaps in user simulation, and action level mental model annotations studies whether agents can track minds well enough to coordinate with real or simulated partners (Qian et al., 2026a; Ruan et al., 2026; Zhou et al., 2026b; Chen et al., 2026b). This line is especially close to CCA because the relevant target is

Table 4: Representative works in each category of related works.

Research Category	**Representative Works**
Personalization	Ricci et al. (2011; 2021); Tan & Jiang (2023); Geng et al. (2022); Salemi et al. (2024); Zeng et al. (2026); Zhang et al. (2026b); Guo et al. (2026); Lin et al. (2026); Liu et al. (2026); Ju et al. (2026); Tao et al. (2026); Han et al. (2026a)
Theory of Mind and Human Simulation	Baker et al. (2017); Sap et al. (2022); Kim et al. (2023); Shapira et al. (2024); Chen et al. (2024); Aher et al. (2023); Argyle et al. (2023); Park et al. (2023); Wang et al. (2024b); Wu et al. (2026); Lei et al. (2026); Javadi et al. (2026); Wang et al. (2026); Qian et al. (2026a); Chen et al. (2026b); Ruan et al. (2026); Nguyen et al. (2026); Zhou et al. (2026b); Zhong et al. (2026); Bawatneh et al. (2026); Li et al. (2026c)
Metacognition	Flavell (1979); Nelson (1990); Koriat (1997); Saunders et al. (2022); Shinn et al. (2023); Steyvers & Peters (2025); Li et al. (2025; 2026d;e); Ji et al. (2026)
Uncertainty Calibration	Guo et al. (2017); Lakshminarayanan et al. (2017); Mielke et al. (2022); Lin et al. (2022b); Fadeeva et al. (2023); Manakul et al. (2023); Farquhar et al. (2024); Kalai & Vempala (2024); Shorinwa et al. (2025); Shi et al. (2026); Gani et al. (2026); Müller et al. (2026); Xiao et al. (2026); Xie et al. (2026); Li et al. (2026a); Srey et al. (2026); Taparia et al. (2026); Lau et al. (2026); Ferrer et al. (2026)
Deferral and Abstention	Geifman & El-Yaniv (2017); Mozannar & Sontag (2020); Piatrashyn et al. (2026); Davidov et al. (2026); Zhao et al. (2026); Edwards & Schuster (2026); Zong et al. (2026); Bachar et al. (2026); Kondadadi & Ortega (2026); Sedoc et al. (2026); Gourabathina et al. (2026); Deng et al. (2026); Kong et al. (2026)
Human-AI Teaming	Parasuraman & Riley (1997); Lee & See (2004); Bansal et al. (2021); Lai et al. (2023); Vasconcelos et al. (2023); Yao et al. (2026); Qian et al. (2026b); Zhang et al. (2026a); Reani et al. (2026); Ren et al. (2026); Heander et al. (2026); Chen et al. (2026b;c); Poelitz et al. (2026); Shaji et al. (2026); Mu et al. (2026); Baloch & Khan (2026); Gor et al. (2026)
Planning with Tool Use	Yao et al. (2022); Lewis et al. (2020); Karpas et al. (2022); Schick et al. (2023); Shen et al. (2023); Wang et al. (2023); Li et al. (2023b); Qin et al. (2024); Patil et al. (2024); Wu et al. (2024); Edwards & Schuster (2026); Wang & Vemuri (2026); Zhou et al. (2026c)
CCA (Ours)	This work

not only whether an agent can describe another mind, but whether the inferred state changes what the agent should do next.

Human simulation provides a related but distinct route: language models can simulate survey populations, social behavior, or generative agents (Aher et al., 2023; Argyle et al., 2023; Park et al., 2023; Wang et al., 2024b). These systems can be useful for evaluation, stress testing, or data generation, but surface level response imitation does not guarantee that the simulated agent tracks the latent states needed for intervention (Wu et al., 2026). CCA uses the theory of mind motivation, but narrows it to inference for policy selection: the agent need not recover a complete mind; it needs uncertainty aware estimates of the human variables that should change the next interaction act.

D.3 Metacognition

Metacognition originates in cognitive science accounts of monitoring and control: agents monitor their own knowledge, memory, uncertainty, and performance, then use those estimates to regulate future cognition or action (Flavell, 1979; Nelson, 1990; Koriat, 1997). In language agents, related ideas appear in self critique, reflection, process supervision, and verbalized uncertainty, where a system attempts to inspect or regulate its own reasoning and outputs (Saunders et al., 2022; Shinn et al., 2023; Steyvers & Peters, 2025). Recent work also treats metacognition as a trainable or measurable property, including cognitive pairwise training and scientific frameworks for bounded self governance (Li et al., 2026e; Ji et al., 2026).

CCA adopts this regulatory view rather than treating metacognition as a standalone confidence statement. A metacognitive self model should estimate not only whether a final answer is reliable, but also whether a planned action is feasible, whether additional evidence is needed, whether the task exceeds the agent's capability boundary, and whether uncertainty about the human model should limit personalization. This is why CCA links metacognition to interaction policy selection rather than using it only as an output annotation.

D.4 Uncertainty Calibration

Uncertainty calibration asks whether model confidence corresponds to empirical correctness and remains meaningful under shift (Guo et al., 2017; Lakshminarayanan et al., 2017). For language models, this becomes more difficult because outputs are open ended and correctness may depend on evidence, interpretation, or downstream use. Work on truthful answering, verbal uncertainty, and factual reliability studies whether models can express what they know and do not know (Mielke et al., 2022; Lin et al., 2022b; Fadeeva et al., 2023). Recent work sharpens this problem for long form QA, reasoning models, verbal uncertainty targets, multimodal confidence calibration, error detection, confidence estimation before answering, truth aligned uncertainty, multimodal incoherence, and automated assessment (Müller et al., 2026; Shi et al., 2026; Gani et al., 2026; Xiao et al., 2026; Xie et al., 2026; Li et al., 2026a; Srey et al., 2026; Lau et al., 2026; Ferrer et al., 2026).

Other work analyzes uncertainty as a structured object rather than a single confidence number, separating sources such as ambiguity, missing evidence, model instability, and semantic variation (Taparia et al., 2026). This matters for CCA because different uncertainty sources should imply different interaction policies: missing user information may call for clarification, missing external evidence may call for retrieval, and model unreliability may call for abstention or escalation.

Hallucination detection and semantic uncertainty address another part of the same problem: generations may be fluent while unsupported, or lexically diverse while semantically equivalent or conflicting (Manakul et al., 2023; Farquhar et al., 2024; Kalai & Vempala, 2024). Recent surveys emphasize that uncertainty estimation for language agents must cover both prediction level confidence and action level reliability (Shorinwa et al., 2025). CCA treats calibration signals as evidence for Metacognitive Self Modeling, but calibration alone does not choose the interaction act. The same uncertainty may call for a caveat, retrieval, clarification, simplification, or deferral depending on the user's state and the risk of action.

D.5 Deferral and Abstention

Selective prediction and abstention methods allow systems to withhold predictions when confidence is low or risk is high (Geifman & El-Yaniv, 2017). Learning to defer extends this idea by routing difficult or risky cases to a human decision maker rather than forcing the model to answer directly (Mozannar & Sontag, 2020). Recent agent work broadens this space to uncertainty aware deferral, dynamic during generation abstention, asking under cost constraints, confidence aware hallucination mitigation, performance prediction for escalation, adaptive model selection, prover verifier selective prediction, abstention based on reasoning traces, clarification under information gain, and information driven policy optimization (Piatrashyn et al., 2026; Davidov et al., 2026; Zhao et al., 2026; Zong et al., 2026; Bachar et al., 2026; Kondadadi & Ortega, 2026; Sedoc et al., 2026; Gourabathina et al., 2026; Deng et al., 2026; Kong et al., 2026). These approaches are important because they make uncertainty operational: the model's self assessment changes the action it takes.

CCA broadens deferral from a terminal fallback into one possible interaction mode. An agent may abstain, ask a clarifying question, retrieve evidence, narrow the task, hand off to a human expert, or explain why it cannot act safely. Which option is appropriate depends not only on model uncertainty, but also on human expertise, trust, cognitive load, stakes, and the cost of delay. Deferral is therefore not merely a thresholded confidence decision; it is part of the coupled interaction policy.

D.6 Human and AI Teaming

Human and AI teaming and automation research studies how people coordinate with automated systems, how trust and reliance form, and how interface choices affect decisions. Classic work on automation emphasizes appropriate reliance rather than maximal trust (Parasuraman & Riley, 1997; Lee & See, 2004). Recent work on AI assisted decision making and explanation shows that users' reliance can be improved, degraded, or miscalibrated depending on how system capabilities and uncertainty are communicated (Bansal et al., 2021; Lai et al., 2023; Vasconcelos et al., 2023; Gor et al., 2026). Recent studies examine human agent collaboration platforms, diverse player collaboration benchmarks, interpretable collaboration policy trees, companion safety, trust calibrated review workflows, controlled collaboration simulation, common ground benchmarks, emergent collaborative behavior, action level mental models, adaptive theory of mind for multiagent coordination, and human centered multiagent system design (Yao et al., 2026; Qian et al., 2026b; Zhang et al., 2026a; Ren et al., 2026; Heander et al., 2026; Chen et al., 2026c; Poelitz et al., 2026; Shaji et al., 2026; Chen et al., 2026b; Mu et al., 2026; Baloch & Khan, 2026).

CCA treats these findings as evidence that interaction policy must model both sides of the team. The agent needs estimates of the human's expertise, trust, workload, and likely interpretation, while also estimating its own reliability and risk. Human and AI teaming therefore motivates the coupling requirement: the right act is not determined by the human model alone or the AI self model alone, but by how the two jointly affect coordination over time.

D.7 Planning with Tool Use

Agentic planning and tool use systems expand the action space available to language models. Retrieval augmented generation can bring external evidence into the response process (Lewis et al., 2020). Tool use and API use systems let models call external functions, invoke specialized models, or interact with software environments (Karpas et al., 2022; Schick et al., 2023; Li et al., 2023b; Qin et al., 2024; Patil et al., 2024). Planning oriented agents decompose tasks, coordinate subtasks, and operate over longer horizons (Yao et al., 2022; Shen et al., 2023; Wang et al., 2023; Wu et al., 2024). Recent tool use work shows that agents may defer blindly to tools or hallucinate direct responses unless tool calling decisions are explicitly trained and uncertainty aligned (Wang & Vemuri, 2026; Zhou et al., 2026c).

These systems make CCA more necessary rather than less. Once an agent can retrieve, call tools, delegate, browse, plan, or act over many steps, the central problem is not only generating a fluent answer but selecting the right interaction mode under uncertainty. Tool availability does not by itself determine when to ask the user, when to verify, when to proceed, or when to stop. CCA frames these choices as policy decisions conditioned on both Dynamic Cognitive State Modeling and Metacognitive Self Modeling.

E Additional High Stakes Applications

Education and tutoring. In tutoring, the appropriate action often depends on both the student's cognitive state (Corbett & Anderson, 1994; Piech et al., 2015) and the agent's confidence in that state. Recent work makes this interactional nature concrete through text based item difficulty and discrimination modeling, tutoring quality signals, history aware student simulation, difficulty aware knowledge tracing, multi agent tutoring architectures, benchmarks for pedagogical safety, and evidence that LLM estimates of student difficulty or item discrimination can diverge systematically from human calibrated student performance (Peters et al., 2025; Jaechang & Lee, 2026; Duan et al., 2026; Huang et al., 2026b; Elhaimeur & Chrisochoides, 2026; Hazra et al., 2026; Li et al., 2026c; Chen et al., 2026a). A student asking for a solution may need a hint, a diagnostic question, a conceptual explanation, or a worked example. CCA suggests that tutoring agents should not optimize only for task completion or answer correctness. They should reason about whether direct assistance would support or undermine learning, whether the agent has enough evidence to infer the student's misconception, whether an assessment item distinguishes learners at different proficiency levels, and whether the model's own apparent ease on the problem is misleading about the student's difficulty.

Clinical and healthcare decision support. Clinical and health related interactions (Griot et al., 2025) rarely depend on domain evidence alone. Patients may arrive with incomplete information, unclear goals, anxiety, limited medical knowledge, or hidden preferences about risk, cost, quality of life, and professional involvement. Recent work on LLM based clinical reasoning, epistemic resilience under misleading medical context, uncertainty aware diagnosis, healthcare LLM judging, and modular Bayesian decision support shows why an assistant must distinguish evidence gathering from recommendation (Kondadadi & Ortega, 2026; Zhou et al., 2026a; Fan et al., 2026; Li et al., 2026b; Kesmen et al., 2026). Studies of consumer facing health LLM evaluation, longitudinal healthcare seeking, therapy chatbot fidelity, engagement optimized care, and clinically harmful safety training further show that interaction dynamics can shape downstream risk (Gorijavolu et al., 2026; Cao et al., 2026; Huang et al., 2026a; Vecchione et al., 2026; BN et al., 2026). CCA does not imply that agents should make autonomous clinical decisions. Rather, it suggests that they should jointly estimate human state and model uncertainty to decide when to explain, ask additional questions, narrow the task, communicate limitations, or defer to professional support.

AI companions and mental health support. AI companions present a different challenge from traditional assistants because their primary purpose is often social or emotional rather than task oriented. Users may seek encouragement, accountability, companionship, emotional support, or simply someone to talk to. The same response can be helpful in one situation and harmful in another, depending on the user's emotional state, vulnerability, and expectations. Recent work on companion safety and governance of human LLM interaction emphasizes that harmful interaction can arise not only from explicit toxic content, but also from manipulation, affective lock in, unsafe long term dynamics, delusion reinforcing conversation histories, and vulnerability amplifying loops (Ren et al., 2026; Reani et al., 2026; Nicholls et al., 2026; Juneja & Lomidze, 2026; Weilnhammer et al., 2026). Moreover, companion systems must operate within significant capability and ethical limits. They are not therapists, clinicians, or close personal relationships, even if users sometimes interact with them as though they were. This creates a difficult balancing act: the system must be supportive without becoming misleading, emotionally responsive without encouraging unhealthy dependence, and helpful without overstating its understanding. Such settings naturally require long term models of both the user and the system itself.

Legal, financial, and organizational decision support. In domains such as legal advice, financial planning, and enterprise analysis, users often seek concise recommendations under uncertainty. Recent work explores expert trace to policy learning, retrieval augmented municipal legal advice, CFO style resource allocation agents, proof carrying trust boundaries, and failures of one size fits all investment advice (Zha et al., 2026; van der Meer & Rossi, 2026; Han et al., 2026b; Koomullil, 2026; Ross & Lo, 2026). These settings require a separation between assistance, authority, and accountability. The agent may be able to summarize evidence or identify relevant considerations, while lacking the authority or reliability to provide definitive advice. A system designed around CCA would separate high confidence support from speculative judgment, communicate uncertainty in a form appropriate for the user, and route high risk decisions toward human expertise.

Software engineering and scientific assistance. Coding and research agents increasingly perform multistep work: searching, planning, editing, testing (Deng et al., 2023; Koh et al., 2024; Xie et al., 2024; Drouin et al., 2024), and synthesizing (Yao et al., 2022; Schick et al., 2023). In these settings, user intent is often underspecified and the agent's capability varies across subtasks (Wang et al., 2024a; Xi et al., 2025). Recent work on clarification seeking coding agents, trust calibrated review of LLM generated changes, low risk code review automation, agentic repository mining, and autonomous scientific discovery further suggests that the problem is not only whether code or experiments can be generated, but when the agent should ask, expose risk, invite review, or stop for verification (Edwards & Schuster, 2026; Heander et al., 2026; Adams et al., 2026; Härtel, 2026; Yang et al., 2026). CCA suggests that agents should decompose ambitious requests into feasible subgoals, expose uncertainty about dependencies or evidence, and insert verification checkpoints when direct execution is unreliable.

Human and AI agent teaming. Increasingly, AI systems are expected to function as teammates rather than tools. Examples include scientific discovery, software development, emergency response, operations planning, and complex decision support. In these environments, the central challenge is often deciding who should do what. Human collaborators possess different expertise, goals, workload constraints, and situational awareness, while AI agents have different strengths, weaknesses, and uncertainty profiles. Recent collaboration benchmarks, human centered multi agent system analyses, simulation environments, and policy distillation approaches make this coordination problem more concrete by evaluating diverse partners and translating LLM reasoning into interpretable collaboration policies (Qian et al., 2026b; Baloch & Khan, 2026; Chen et al., 2026c;b; Zhang et al., 2026a). Effective collaboration thus depends on maintaining a shared understanding of both human and machine capabilities. An agent may need to decide whether to proceed autonomously, request additional information, explain its reasoning, or hand control back to a human teammate. Poor decisions can lead to overreliance, underuse of expertise, or coordination failures. As AI systems become more capable, these interaction and coordination decisions are likely to become as important as the underlying task performance itself.

Across these applications, the common pattern is not domain specific. The agent must infer what kind of support the human needs while also estimating whether it can provide that support reliably. High stakes use makes the coupling more visible: personalization without self limitation can become unsafe, while uncertainty estimation without communication that is sensitive to the user can become unhelpful.

F Additional Challenges

Data scarcity and partial observability. CCA relies on latent human and agent states that are rarely labeled. Human beliefs, intentions, confusion, trust, workload, and reliance must be inferred from ambiguous traces: the same pause may indicate confusion, skepticism, distraction, overload, or careful verification. Thus, language, history, performance, corrections, latency, feedback, and domain evidence should serve as weak constraints on a belief state, not direct labels for a user's mind. Agent side states are similarly weakly supervised: confidence or retrieval scores do not directly reveal whether an action is feasible or safe.

Evaluation under latent states. CCA cannot be evaluated only by answer accuracy. A system may produce a correct answer while choosing the wrong interaction act, or predict user behavior without identifying the latent cause that requires a different policy. Evaluation should therefore test whether policies change appropriately as user expertise, uncertainty, trust, cognitive load, capability boundaries, and cost of error vary. Useful protocols include counterfactual user state evaluation, metacognitive stress testing, episode level analysis of reasoning and monitoring behaviors (Li et al., 2025; 2026d), and longitudinal measures of learning, error recovery, calibrated reliance (Lee & See, 2004; Bansal et al., 2019; Buçinca et al., 2021), and downstream outcomes.

Deployment boundaries and governance. Modeling confusion, workload, anxiety, trust, or overreliance can improve assistance, but it also creates privacy and manipulation risks. CCA systems should model only task relevant states, represent uncertainty, and avoid treating transient signals as stable attributes. In high stakes settings, richer user models should not justify more autonomy;

the appropriate act may be narrowing the task, explaining uncertainty, preparing questions for a professional, or handing off the decision. Deferral should be treated as a legitimate interaction act rather than a service failure (Mozannar & Sontag, 2020).

G Additional Opportunities

CCA opens several opportunities for studying agents as participants in extended interaction between humans and AI rather than as isolated response generators.

Interaction level benchmarks. A first opportunity is to design benchmarks that move beyond final answer quality by specifying the task context, user state conditions, agent uncertainty, and expected interaction mode. Recent benchmarks for human aligned user simulation, diverse player collaboration, AI companion safety, and agent clarification illustrate this shift toward evaluating interaction policies rather than isolated responses (Wang et al., 2026; Qian et al., 2026b; Ren et al., 2026; Edwards & Schuster, 2026). The target is whether the agent chooses an appropriate next act, such as answering, asking for clarification, retrieving evidence, communicating uncertainty, or deferring.

Human modeling under uncertainty. A second opportunity is to build user models that represent not only profiles, preferences, or likely behaviors, but also uncertainty over possible cognitive states. Work on implicit cost and performance preferences, personality anchoring, and recursive ToM shows that user models are becoming richer, but also more dependent on uncertain latent assumptions (Zeng et al., 2026; Javadi et al., 2026; Lei et al., 2026). When the system is unsure whether the user is a novice, expert, overloaded, or trusting too much, a diagnostic question or robust explanation may be more appropriate than stronger personalization.

Adaptive uncertainty communication. A third opportunity is to translate metacognitive signals into forms that users can interpret and act upon (Lin et al., 2022a; Steyvers & Peters, 2025; Zhou et al., 2024a). Recent work on verbal uncertainty conditioned on the answer, faithful confidence expression, and episode level reasoning analysis highlights that uncertainty communication itself must be evaluated, not merely appended to answers (Shi et al., 2026; Gani et al., 2026; Li et al., 2025; 2026d). The same uncertainty may require a caveat, rationale, request for missing information, safer subtask, or escalation, depending on the user's expertise, trust, workload, and decision context.

Intervention sensitive learning. The final opportunity is to define training and evaluation objectives that measure how interaction acts affect users, not only what users do next. Asking under cost constraints, interpretable collaboration policy distillation, uncertainty aligned tool calling, and trust calibrated review workflows all point toward learning objectives in which the value of an action depends on the human state and downstream coordination, not only task success (Zhao et al., 2026; Zhang et al., 2026a; Zhou et al., 2026c; Heander et al., 2026). Relevant outcomes include understanding, trust, workload, reliance, and decisions, supporting evaluation of adaptation between humans and AI beyond single turn response quality.

H NOTATION

This section summarizes the notation used in the CCA formulation. Table 5 lists the basic interaction variables. Tables 6 and 7 define the human state and agent side metacognitive variables, respectively. Table 8 summarizes the coupled posterior, JES, interaction policy, and recursive update. All latent variables are intended as decision relevant modeling constructs rather than directly observed states.

Table 5: Notation for the interaction process and observable evidence.

Notation	Name	Description
		Time and context
t	Interaction turn	Time step in the interaction between the human and AI.
x_t	Task context	The task context available to the agent at turn t.
o_t	Interaction observation	External observation at turn t, such as user language, behavior, feedback, task state, or multimodal cues.
e_t	Internal evidence	Agent side evidence about uncertainty, reliability, capability boundaries, cost, or risk.
		Actions and history
a_t^{AI}	Agent action	The agent's structured action at turn t. In CCA, $a_t^{AI} = (y_t, \delta_t, c_t)$.
a_t^H	Human response	The human's response or action after the agent acts.
τ_t	Interaction history	History available at turn t: $\tau_t = (x_{1:t}, o_{1:t}, e_{1:t}, a_{1:t-1}^{AI}, a_{1:t-1}^H)$. It may also include compact memory summaries maintained by the agent.

Table 6: Notation for Dynamic Cognitive State Modeling (DCSM).

Notation	Name	Description
		Latent human state
h_t	Human cognitive state	Task relevant latent state of the human at turn t, including beliefs, goals, preferences, knowledge, uncertainty, cognitive load, reliance, and interaction strategy.
b_t^H	Human beliefs	The human's task beliefs or assumptions.
g_t^H	Human goals	The user's intended outcomes or task objectives.
θ_t^H	Human preferences	Preferences, values, constraints, or tradeoffs relevant to the interaction.
k_t^H	Human knowledge	Knowledge state, expertise, or prerequisite understanding.
u_t^H	Human uncertainty	The human's uncertainty about the task, evidence, answer, or decision.
ℓ_t^H	Cognitive load	The user's task relevant cognitive burden or processing difficulty.
r_t^H	Reliance or trust	The user's reliance on, trust in, or dependence on the agent.
σ_t^H	Interaction strategy	The user's interaction strategy, such as seeking a hint, requesting verification, delegating, or testing the agent.
		Human state belief
$\mu_t^H(h)$	Human state belief	The agent's belief distribution over the human state: $\mu_t^H(h) = p(h_t = h \mid \tau_t)$.
DCSM	Dynamic Cognitive State Modeling	The agent's uncertainty aware belief update over latent human cognitive states. It is used for policy selection rather than as a claim of fully recovering the user's true mental state.

Table 7: Notation for Metacognitive Self Modeling (MSM).

Notation	Name	Description
		Latent agent side state
m_t	Metacognitive state	Decision relevant estimate of the agent's own uncertainty, reliability, capability boundaries, risk, expected success, and uncertainty about its human state estimate.
u_t^{AI}	Agent uncertainty	Agent side uncertainty about the task, answer, evidence, or planned action.
q_t^{AI}	Expected quality	Agent confidence or expected output quality.
κ_t^{AI}	Capability boundary	Information about whether the task or subtask lies within the agent's reliable competence.
ρ_t^{AI}	Risk	Estimated risk or cost associated with acting, especially if the action is incorrect or misleading.
s_t^{AI}	Expected success	Expected success of the planned action, subtask, or interaction strategy.
$\eta_t^{H\|AI}$	Human model uncertainty	The agent's uncertainty over its own estimate of the human state.
		Agent side belief
$\mu_t^M(m)$	Metacognitive belief	The agent's belief distribution over its metacognitive state: $\mu_t^M(m) = p(m_t = m \mid \tau_t)$.
MSM	Metacognitive Self Modeling	The agent's decision relevant model of its own uncertainty, limits, reliability, and risk, beyond a standalone confidence score.

Table 8: Notation for Coupled Cognitive Alignment, policy selection, and recursive update.

Notation	Name	Description
	Coupled state and Joint Epistemic State	
z_t	Coupled state	Coupled state of the human and agent, defined as $z_t = (h_t, m_t)$.
$\hat{p}_t(z)$	Coupled posterior	Posterior over the coupled state: $\hat{p}_t(z) = p(z_t = z \mid \tau_t)$. Its marginals correspond to the DCSM belief μ_t^H and the MSM belief μ_t^M.
Φ	JES mapping	Model dependent mapping from the coupled posterior and interaction history to a compact representation for policy selection.
$\hat{J}_t$	Joint Epistemic State	Summary for policy selection derived from the coupled posterior: $\hat{J}_t = \Phi(\hat{p}_t, \tau_t)$. It summarizes decision relevant uncertainty and coordination between the human and the agent.
JES	Joint Epistemic State	The interface through which human modeling and AI metacognition are coupled for policy selection.
	Interaction policy	
π^{AI}	Agent policy	CCA interaction policy conditioned on the task context and current JES: $a_t^{AI} \sim \pi^{AI}(\cdot \mid x_t, \hat{J}_t)$.
y_t	Task level content	Content produced by the agent, such as an answer, explanation, plan, summary, or recommendation.
δ_t	Interaction mode	The selected mode of interaction, such as direct answering, clarification, evidence gathering, decomposition, deferral, or escalation. It is an action component, not an additional latent state.
c_t	Communication act	How the agent communicates confidence, uncertainty, rationale, limitations, or risk. It is an action component, not an additional latent state.
	Recursive coupled update	
f_{t+1}	Feedback	Feedback observed after the agent's action: $f_{t+1} = (a_t^H, o_{t+1}, e_{t+1})$.
$\oplus$	History update	History appending operator: $\tau_{t+1} = \tau_t \oplus (a_t^{AI}, f_{t+1})$.
U	Belief update operator	Model dependent coupled belief update: $\hat{p}_{t+1} = U(\hat{p}_t, a_t^{AI}, f_{t+1}, \tau_t)$. It may be implemented through Bayesian filtering, learned state tracking, memory update, tool based verification, or self evaluation.
$\hat{J}_{t+1}$	Updated JES	Updated Joint Epistemic State after incorporating new feedback: $\hat{J}_{t+1} = \Phi(\hat{p}_{t+1}, \tau_{t+1})$.
CCA loop	Recursive interaction loop	The recursive dependency structure: $\hat{J}_t \to a_t^{AI} \to f_{t+1} \to \hat{p}_{t+1} \to \hat{J}_{t+1}$. The agent's action affects the human, feedback revises the coupled posterior, and the updated posterior determines the next JES.
	Framework	
CCA	Coupled Cognitive Alignment	A policy level framing in which human modeling and AI metacognition jointly constrain interaction decisions under uncertainty.