Life of Research · Published 22 November 2026 · DOI 10.5281/zenodo.24442238

A Belief Probe Is Not a Coordination Test: What Explicit Epistemic State Buys Language-Model Agents, Why Added Theory of Mind Helps in Some Studies and Hurts in Others, and Where a Symbolic Guard Can Stand In

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Language-model agents are now routinely asked to work together, and when they fail the usual diagnosis is that they lack a model of what their partners know. A 2026 proposal answers that diagnosis with architecture: an explicit epistemic layer over a gossip-style record of who has told whom what, and a symbolic guard that permits an agent's proposed action only when the authoritative state satisfies its precondition. Two published literatures pull against each other on whether such machinery is needed. One reports that current models already pass false-belief and higher-order belief tests at or above adult human levels and carry linearly decodable representations of other agents' beliefs. The other reports that the same models fail to coordinate, and that adding explicit theory-of-mind modules or instructions sometimes helps, sometimes does nothing, and sometimes makes coordination worse. In this analysis the author reviews and synthesises published results to show that the two literatures measure different links in one chain. The paper separates four links, tracking who observed what, inferring a belief from that record, converting a belief into an action or message, and calibrating the model of the partner against the partner's actual behaviour, plus a fifth component, a gate that vetoes invalid actions. Most reported gains from explicit modules come from tracking or conversion, the two links that a symbolic component can take away from the language model; inference, the link the theory-of-mind debate argues over, is close to solved once the record is supplied; and calibration, the link a guard cannot reach, appears to be where the largest remaining failures sit. On one benchmark, GPT-4o, GPT-5 and Claude-3.5-Sonnet each answer more than 95 percent of belief questions correctly but predict the resulting behaviour correctly in only 49.5 to 67.0 percent of cases, and reminding them of their own belief answer raises one model's behaviour accuracy from 67.0 to 96.9 percent. The paper consolidates 25 published results in one table, states an attribution procedure for a reported coordination gain, and identifies what a guard over an authoritative state can and cannot supply when no such state exists outside the testbed.

Keywords: theory of mind; multi-agent coordination; language-model agents; epistemic logic; belief tracking; neuro-symbolic agents; symbolic guards; evaluation methodology

1. Introduction

When two language-model agents fail to cooperate, the failure usually looks social. One agent withholds a fact its partner needed; another keeps executing a plan after its partner has visibly changed course; a third answers a hint as if the hinter could see what it sees. Taxonomies of multi-agent failure give these episodes names - information withholding, ignored input, reasoning that does not match the action taken - and the remedy that suggests itself is to give each agent a model of the others' minds (Cemri et al., 2025). A line of work in multi-agent systems has argued for decades that coordination is a matter of reasoning about knowledge: what each participant knows, what each knows the others know, and when a fact has become common ground (Halpern and Moses, 1990). Gossip protocols, in which agents exchange secrets through pairwise calls until everyone knows everything, have become the standard small world in which such reasoning is made exact (Apt et al., 2016; Cooper et al., 2016).

A 2026 proposal brings the two traditions together (Nasiri et al., 2026). It describes Epistemic Probabilistic Language Agents, an architecture in which a language model proposes typed actions, an epistemic logic core maintains the authoritative symbolic state, a conditional-belief engine holds graded uncertainty, and a symbolic guard decides whether a proposed action may execute. The formal work is set in a gossip testbed whose states are call histories, enriched with probability weights over the histories each agent cannot tell apart. Its central technical result is that, for the knowledge fragment the guards use, these weights never change whether a guard's precondition is true. The authors are careful about scope: they describe the learned components and feedback channels as "intended interfaces and empirical hypotheses, not implementation results," and state that their theorems about the gossip abstraction do not imply that an implementation realises it. The proposal is therefore an argument about what coordination requires, and it can be tested against what the published evidence shows.

That evidence points two ways. Models now pass classical false-belief batteries: one June 2023 model solved 75 percent of a 40-task battery in which every task required all 16 of its prompts to be answered correctly (Kosinski, 2023), GPT-4 matched or exceeded a sample of 1,907 human participants on several theory-of-mind measures (Strachan et al., 2024), and GPT-4 scored

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

93 percent on sixth-order belief statements against 82 percent for adult humans (Street et al., 2024). Belief status from different agents' perspectives can be decoded linearly from model activations, and editing those directions changes theory-of-mind answers (Zhu et al., 2024). On the other side, the same generation of models fails embodied coordination tasks that require acting on what partners know (Juneja et al., 2026), fails to pool distributed evidence in hidden-profile tasks (Li, Y. et al., 2025), and lags humans and even simple word-embedding baselines in an interactive code-guessing game (Lupu et al., 2025). Explicit belief-state and theory-of-mind modules are reported to help (Li, H. et al., 2023, whose module is a world-state record; Cross et al., 2024; Zhang, C. et al., 2024). Instructions to reason at a specified theory-of-mind depth are reported to do nothing reliable, although in the same benchmark a reference architecture with an explicit theory-of-mind module, PCM-LLM, succeeds in every condition (Yan et al., 2026). And added theory of mind is reported to hurt when two agents reason to mismatched depths about each other (Mu et al., 2026).

The question this paper addresses is the one the anchor proposal raises without settling: when a language-model agent fails to coordinate, which part of the failure would an explicit epistemic state repair, and which part would a guard over an authoritative state repair? In this analysis the author reviews and synthesises published results to show that "theory of mind" names at least four links in a chain, plus a gate, and that the published contrasts touch them unevenly. Tracking (T) is the bookkeeping of who observed what: which events each agent witnessed, which messages reached it, what its local history contains. Inference (I) is the step from such a record to an attributed belief, including nested beliefs about beliefs. Conversion (C) is the step from an attributed belief to an action or message that serves the joint goal. Calibration (K) is the fit between the agent's model of its partner - the partner's policy, its depth of reasoning, what it will do with a hint - and the partner it actually faces. The gate (G) is any component that vetoes a proposed action that violates a rule or precondition. Classical theory-of-mind tests score I, mostly with T supplied by the story. Coordination benchmarks score the product of all four, often with G present or absent without comment.

The paper argues four things. First, most published gains from adding explicit epistemic machinery to a language model are gains in T or C - a record of perspectives built outside the model, or a rule that forces the action to follow the belief - and these are exactly the links a symbolic component can take over. Second, I, the link the theory-of-mind debate argues over, is close to solved once T is supplied, so that debate predicts little about coordination. Third, K appears to be where the largest residual failures sit, it is the link on which adding more theory of mind can make things worse, and it is the link no guard can reach, because a guard checks actions against a state, not against a partner. Fourth, a guard over an authoritative state can supply T only where such a state exists, which is true of gossip testbeds and game engines and false of most deployments, where each agent's observations are private.

Section 2 sets out the two literatures. Section 3 defines the four links and the gate and maps each published intervention onto them. Sections 4 to 7 take the links in turn. Section 8 treats the gate and the anchor's guard. Section 9 consolidates the measured values in Table 2. Section 10 states an attribution procedure (Algorithm 1). Sections 11 to 13 give the strongest objection, the open questions, and the studies that would settle them.

2. Two Literatures With Opposite Verdicts

2.1 The models already know what others know

The term theory of mind entered the literature as a question about chimpanzees (Premack and Woodruff, 1978), and its standard test, the false-belief task, as a study of what young children understand about deception (Wimmer and Perner, 1983). Language-model research inherited both, along with question-answering versions such as ToMi, which recast the false-belief story as reading comprehension (Le et al., 2019). Through 2023 and 2024 the headline results moved steadily upward. Kosinski's battery, built from false-belief scenarios with matched true-belief controls and reversed versions, was solved at the 20 percent level by two 2022-2023 models and at the 75 percent level by a June 2023 model (Kosinski, 2023). Gandhi et al. (2023) generated a benchmark from causal templates and concluded that GPT-4's inference patterns resemble human ones, though less reliably. Strachan et al. (2024) compared two model families with a large human sample across false belief, indirect requests, irony, misdirection and faux pas; GPT-4 was at or above human level on several, and its weak faux-pas score was traced to an unwillingness to commit to a conclusion rather than to a failure of the underlying inference. Street et al. (2024) wrote a multi-order test suite and found GPT-4 and Flan-PaLM at or near adult level overall.

Representational work added a second line of support. Zhu et al. (2024) decoded belief status from the perspectives of several agents with linear probes and showed that intervening on those directions changes the model's answers to theory-of-mind questions. Sobotka et al. (2026) found in incomplete-information games that probes on internal activations recover the latent game state more accurately than the model's own verbal reports do, a gap they call the observation-belief gap. Okcu et al. (2026) report a related pattern in text-rendered grid worlds: task-relevant state and value information is decodable more reliably than it shows up in what the agent does.

2.2 But they do not coordinate

The coordination literature describes the same models very differently. On the hard split of EnactToM, a set of embodied household tasks with private observations and constrained communication, all seven frontier models

evaluated score 0.0 percent on functional completion when success is required in all three runs, while answering a substantial share of literal belief probes on the same tasks; single-run averages are 63.3 percent literal against 6.7 percent functional for Gemini-Pro and 52.5 against 5.0 for o3 (Juneja et al., 2026). The authors attribute 93 percent of sampled failures to epistemic coordination breakdowns: withheld information, broken epistemic chains, ignored partner constraints, misallocated messages and sabotaged private incentives. HiddenBench, built on the hidden-profile paradigm from social psychology, finds that groups of agents reach 30.1 percent accuracy when the evidence is distributed among them, against 80.7 percent for a single agent given all of it, across 15 models (Li, Y. et al., 2025). LLM-Coordination found agents competent where the right action depends on visible environmental variables and weak where it depends on partners' beliefs and intentions (Agashe et al., 2023). In Decrypto, frontier models trail humans and word-embedding baselines, and newer reasoning models do worse than older ones on tasks adapted from two classic cognitive-science experiments (Lupu et al., 2025). Collab-Overcooked reports strong goal interpretation alongside weak active collaboration and adaptation (Sun et al., 2025). In finitely repeated two-by-two games, GPT-4 does well in self-interested games and poorly in the Battle of the Sexes, which rewards taking turns (Akata et al., 2023).

2.3 Explicit machinery helps, does nothing, and hurts

The intervention literature does not resolve the conflict, because its results also run both ways. On the helping side: a textual belief state about the task world shortened GPT-4 teams' completion time in a search-and-defuse game and raised their accuracy on theory-of-mind questions (Li, H. et al., 2023); a module that generates hypotheses about other agents' strategies and scores them against observed behaviour outperformed language-agent and reinforcement-learning baselines in Melting Pot (Cross et al., 2024); an agent that infers teammates' intentions and revises its beliefs from their behaviour beat population-trained baselines in Overcooked (Zhang, C. et al., 2024); a plug-in multi-character belief tracker raised GPT-3 Davinci's ToMi accuracy by 38 points (Sclar et al., 2023); a two-stage prompt that first filters the story to what a character witnessed raised GPT-4's ToMi false-belief accuracy from 25.5 to 87.75 percent (Wilf et al., 2023). Latent "theory-of-mind tokens" trained to match a partner's actual reasoning improve performance across five multi-agent environments (Sivakumaran et al., 2026).

On the other side: in a vision-language benchmark that varies the depth of reasoning a task demands, instructing models to reason at a specified depth produced no reliable change in behaviour (Yan et al., 2026); a multi-agent architecture combining theory of mind, internal beliefs and a logic solver found the mechanisms "two-edged", improving some models and lowering the lower confidence bound of others (Kostka and Chudziak, 2026); a model fine-tuned on theory-of-mind data played Decrypto worse than its base model in both roles (Lupu et al., 2025); larger reasoning budgets lowered accuracy on theory-of-mind benchmarks (Gong et al., 2026). Most pointedly, when two agents each reason one level deep about the other, coordination in a repeated matrix game collapses, while pairs whose depths differ by one reach the game's maximum score (Mu et al., 2026).

2.4 What the two literatures share

Every result in Sections 2.1 to 2.3 is a measurement of some agent doing something with information about another agent. They differ in which information is supplied, which step is scored, and whether the partner is real. A false-belief story supplies the record of who saw what and scores the attribution. A coordination benchmark supplies nothing and scores joint success. An intervention adds a component and scores whichever of the two its authors chose. The disagreement is not, on this reading, about whether models have theory of mind. It is about which link each study measured, and the next section names the links.

3. Four Links and a Gate

3.1 Definitions

The definitions below are stipulations for this paper.

Tracking (T) maintains, for each agent, a record of the events it observed and the messages it received. In a gossip model this is an agent's view of the call history (Apt et al., 2016); in a story it is the subset of sentences the character witnessed; in Hanabi it is the set of hints each player has received and the cards each can see (Bard et al., 2020). T is bookkeeping. It requires no reasoning about minds, only accurate records of observability.

Inference (I) maps a record to an attributed mental state: that Sally believes the marble is in the basket, that Bob does not know the soup is spoiled, that Alice thinks Bob thinks the bomb is red. Nested inference to order k composes k such attributions. I is what false-belief tests and most theory-of-mind benchmarks score.

Conversion (C) maps an attributed state, together with the joint goal, to an action or message: tell Bob about the soup, stop the current skill because the partner has switched to it, alternate in the Battle of the Sexes because the partner alternates. C is what SimpleToM calls applied theory of mind (Gu et al., 2024) and what Riemer et al. (2025) fold into functional theory of mind.

Calibration (K) is the fit between the agent's model of its partner and the partner. It includes the partner's policy, its depth of reasoning about the agent, what conventions it follows, and how it will read a message. K is updated, if at all, by comparing predictions with the partner's actual behaviour. Riemer et al. (2025) define functional theory of mind as rational in-context adaptation to a partner; that is K and C together.

The gate (G) vetoes a proposed action that violates a rule or precondition. A shield in safe reinforcement learning is a gate synthesised from a temporal-logic specification

(Alshiekh et al., 2018); the anchor's symbolic guard is a gate whose preconditions are epistemic formulas evaluated on the authoritative history (Nasiri et al., 2026); a language-model verification step that rejects an action likely to end a Hanabi game is a gate implemented by the model itself (Agashe et al., 2023).

3.2 The human precedent for separating I from C

The separation of inference from use is not an artefact of language models. Keysar et al. (2003) gave adults a communication game in which a director, who could not see a hidden object, gave instructions about mutually visible objects. Participants knew perfectly well that the director could not mean the hidden object, yet they often treated it as the referent and sometimes reached for it; the effect held even when they believed the director had a false belief about what was hidden. The authors read this as a dissociation between the ability to distinguish one's own beliefs from another's on reflection and the routine deployment of that ability in interpreting action. Apperly and Butterfill (2009) proposed two systems: an efficient but inflexible capacity for tracking belief-like states and a flexible but demanding capacity for full belief reasoning. Neither result transfers mechanistically to language models. Both establish that "has theory of mind" and "uses it to act" were already known to come apart in the one system the tests were designed for.

3.3 Where each kind of evidence sits

A classical false-belief item supplies T in the narrative and scores I. A behaviour-prediction item scores I and C on a third party. A coordination episode with a scripted partner scores T, I, C and, if the partner's policy must be learned, K. A coordination episode with another learning agent adds the problem that the partner's K is changing too. A gate, when present, removes some failed actions before they are scored. The mapping is summarised in Table 1 below, which pairs each published intervention with the link it most directly changes.

The classification is the author's and is coarse: several interventions touch more than one link, and the table records the one the source's own ablation or description isolates. Its use is as a map for the sections that follow.

4. Tracking: Most Published Gains Are Bookkeeping

4.1 Belief trackers work by building the record

SymbolicToM is the cleanest case (Sclar et al., 2023). It constructs, for each character and for each nested perspective, a graph of the world state as that character would believe it, updated sentence by sentence, and then passes the language model only the portion of the story consistent with the queried perspective. The model still answers the question. What has changed is that it answers it over a record from which the events the character did not witness have been removed. With this, GPT-3 Davinci's ToMi accuracy rises by 38 points to 92 percent averaged across question types, and question-type gains reach 62 points for first-order false belief with one model and 78 points for second-order false belief with another. The algorithm does almost all of the work of tracking and leaves the inference to the model.

SimToM makes the same move with a prompt (Wilf et al., 2023). Its first stage asks the model to rewrite the story from what a named character knows; its second asks the question over the rewritten story. GPT-4's ToMi false-belief accuracy goes from 25.5 percent at zero shot to 87.75 percent. The ablations are informative about mechanism. Collapsing the two stages into a single prompt loses much of the gain, and replacing the perspective-taking stage with a generic reasoning stage dropped accuracy below that of single-step chain-of-thought. Most telling is the oracle condition. When human annotators wrote the perspective - the subset of the story a character could know, with nothing added - GPT-3.5-Turbo's false-belief accuracy rose from 81.0 to 96 percent on ToMi and from 70.5 to 96.0 percent on BigToM. Once the record is right, a mid-2023 model answers false-belief questions almost perfectly. The residual difficulty in these benchmarks was tracking.

The higher-order results point the same way. HopToM controls for the shortcut by which a nominally fourth-order question can be answered with second-order reasoning, and under direct prompting four frontier models fall from 73.5 percent at first order to 20.0 percent at fourth (You et al., 2026). HopTrace, the method the same authors propose, encodes each interaction once into per-perspective records and resolves nested queries by a deterministic procedure; it adds 22.0 and 38.5 percentage points at fourth and fifth order over strong prompting baselines while using 20 to 39 times fewer tokens than multi-call methods. The authors frame this as improving perspective reasoning. In the vocabulary here, it moves T out of the model and replaces nested I with a lookup.

DEL-ToM decomposes theory-of-mind problems into a sequence of belief updates in dynamic epistemic logic, trains a process verifier on traces produced by a logic simulator, and uses it to choose among candidate belief traces at inference time (Wu et al., 2025). Thought-tracing, modelled on Bayesian theory of mind (Baker et al., 2017), has the model propose hypotheses about an agent's mental state and reweights them against the agent's perceptions and actions (Kim et al., 2025). Both improve benchmark accuracy, and both do it by imposing a structure on how the record of observations is maintained and updated.

4.2 A world record raised the theory-of-mind score

The most instructive result for coordination comes from a study that did not set out to add theory of mind at all. Li, H. et al. (2023) placed three GPT-4 agents in a text game in which they explore rooms and defuse multi-phase bombs, and added an explicit belief state: a textual description, updated after each observation, of task-relevant facts such as room contents and bomb sequences. The belief state is about the world, not about teammates. Both GPT-4 conditions reached the

Table 1. Published interventions and the link each most directly changes (the author's classification; sources on each row).

Intervention	Source	Link changed	What it adds
Multi-character belief graph, story filtered per character	Sclar et al., 2023	T	External record of perspectives
Perspective-filtering prompt; oracle perspectives	Wilf et al., 2023	T	Record (model- or human-built)
Perspective records resolved deterministically	You et al., 2026	T, I	Record plus rule-based nesting
Verifier trained on epistemic-logic simulator traces	Wu et al., 2025	T, I	Step scoring of belief updates
Textual belief state about the task world	Li, H. et al., 2023	T	World record kept in context
Mental-state reminder before behaviour question	Gu et al., 2024	C	Belief placed at decision time
Predict partner's move, then choose	Akata et al., 2023	C	Prediction placed at decision time
Posterior-gated interruption of current skill	Goel et al., 2026	C	Rule from belief to control
Exchange-then-decide message protocol	Li, Y. et al., 2025	C	Rule for what to say when
Hypotheses scored against observed behaviour	Cross et al., 2024	K	Feedback on partner model
Partner depth estimated from past play	Mu et al., 2026	K	Matching reasoning depth
Latent tokens trained on partner's actual thoughts	Sivakumaran et al., 2026	K	Supervised partner model
LLM verification of rule-following and safety	Agashe et al., 2023	G	Veto on invalid moves
Symbolic guard over authoritative history	Nasiri et al., 2026	G (and T)	Veto against exact state

maximum team score, but with the belief state the team needed 12.3 rounds on average instead of 28.3, the share of model outputs that mapped to legal actions rose from 71.8 to 86.1 percent, and accuracy on theory-of-mind questions asked during the mission rose from 80.0 to 97.2 percent for introspection, from 60.0 to 80.1 percent for first-order questions, and from 64.3 to 69.4 percent for second-order questions. The authors' own diagnosis of the baseline's failures is that agents lost track of long-horizon context and hallucinated the task state.

A record of what the world contains is not a model of anyone's mind, yet it moved the first-order theory-of-mind score by about 20 points. The simplest reading is that the agents could infer what teammates knew once they could remember what had happened; the inference was not the bottleneck. The second-order score moved by about five points, which suggests that nested attribution had a separate limitation the world record did not touch. The ground truth for those questions was computed from the global interaction history, which the evaluators had and the agents did not - a point that returns in Section 8.

4.3 What the stress tests found missing

The adversarial benchmarks that deflated early claims tell the same story from the failure side. Ullman (2023) showed that small changes to a vignette that preserve its theory-of-mind logic turned a reported success into failures. Shapira et al. (2023) found reliance on shallow heuristics across six tasks. FANToM moved the test into information-asymmetric conversations and found large gaps to human performance even with chain-of-thought or fine-tuning (Kim et al., 2023). Hi-ToM found accuracy declining with order (He et al., 2023). MindGames, generated from dynamic epistemic logic, found that scaling did not reliably lift smaller model families above chance, with GPT-4 better but imperfect (Sileo and Lernould, 2023). ExploreToM generated adversarial stories by search over a domain-specific language and reported accuracy as low as 0 percent for Llama-3.1-70B and 9 percent for GPT-4o; its authors name unreliable state tracking among the skills whose absence explains the failures (Sclar et al., 2025). An embodied study that gave ReAct agents structured examples derived from planner solutions found them adequate where the task needed only attentional filtering and inadequate where agents had to reason about occluded spaces or weigh the cost of acting to gain information; its authors conclude that explicit belief tracking is needed (Annese et al., 2025). The stress tests that make inference hard mostly do so by making tracking hard: longer stories, more agents, more moves, more partial observability.

4.4 What tracking requires

A tracker needs to know who observed what. In a story, the narrator says so. In a gossip protocol, the call history determines it, and the anchor proposal's careful choice of view-based rather than call-subsequence-based indistinguishability exists because what an agent learns in a call depends on what its interlocutor had already learned (Nasiri et al., 2026). In Hanabi, the rules make observability common knowledge (Bard et al., 2020). In each of these cases the record is computable by a component with access to the global history. A multimodal construction task built to elicit common ground under epistemic asymmetry, annotated across speech, gesture and action, found that prompted language models had difficulty tracking both task progress and belief state, and compared them with an axiomatic pipeline in dynamic epistemic logic performing the same task (Zhu, Y. et al., 2026). In a deployment where agents read private documents, call private tools and hold private conversations, no component has that access unless the system is built to give it one, and then the record is only as complete as the instrumentation. Tracking is the link that symbolic machinery handles best, and it is also the link whose inputs most depend on the world being instrumented.

5. Inference: Where the Debate Is, and Why It Predicts Little

5.1 The debate in brief

Whether language models "have" theory of mind has been argued from both directions with real evidence. The positive side has the batteries in Section 2.1 and the probing results. The negative side has the stress tests in Section 4.3 and a methodological critique: that tests built for humans assume a consistency of reasoning across questions that models do not show, so that strong literal performance licenses little (Riemer et al., 2025; Ullman, 2023). A 2025 survey organises the field into behavioural evaluation, representational evidence and safety risk (Nguyen, 2025). Reasoning-trained models have reopened the question without closing it. De Haan et al. (2026) find reasoning models more robust to prompt and task perturbations and read the gain as robustness in reaching a correct answer rather than a new social capacity. Gong et al. (2026) find that reasoning models do not consistently beat non-reasoning ones, that accuracy falls as responses lengthen, and that removing multiple-choice options helps, which they attribute to option matching. In Decrypto, state-of-the-art reasoning models were significantly worse than older models on the embedded theory-of-mind tasks (Lupu et al., 2025).

5.2 Why the debate matters less for coordination than it seems

Three observations limit how much the inference debate can say about coordination. First, where T is supplied, I is close to ceiling for recent models. SimpleToM's awareness questions, which ask whether a character knows a hidden fact that the two-sentence story makes clear they could not have observed, are answered correctly more than 95 percent of the time by each of the frontier models the authors name in that comparison, from GPT-4o and Claude-3.5-Sonnet to o1 and DeepSeek-R1 (Gu et al., 2024). The oracle-perspective result in Section 4.1 makes the same point with a controlled manipulation.

Second, where T is not supplied, failures that look like failures of inference are often failures of tracking, as the ExploreToM analysis and the Li, H. et al. belief-state result suggest. A coordination benchmark that withholds the record and scores joint success cannot tell the two apart.

Third, the cross-model correlation between theory-of-mind scores and coordination success, which looks like direct support for the importance of I, is weaker evidence than it appears. LLM-Hanabi extracts first-order statements (what a hint was meant to convey) and second-order statements (how the partner will read a hint) from agents' reasoning during play and finds correlations with game score of $r = 0.76$ and $r = 0.58$ across models (Liang et al., 2025). The correlation is computed across models that differ in general capability as well as in theory of mind, so it does not isolate either. The table also contains pairs that show how loose the coupling is: Llama-3.3-70B scores 81.74 on first-order theory of mind and 7.20 on the game's 100-point scale, while gpt-4.1-mini scores 86.70 and 25.60.

None of this says inference is unimportant. It says that the published coordination failures are not, for the most part, cases where the agent had the right record and drew the wrong conclusion about a mind.

5.3 Where inference still bites

Two exceptions should be stated. Nested inference beyond second order remains weak under direct prompting even with shortcut controls (You et al., 2026), and the Li, H. et al. second-order score barely moved with a world record. Coordination that genuinely depends on third- or fourth-order attribution - conventions about what a partner thinks the agent thinks the partner will do - will stress I directly. Second, inference about what a partner wants, as opposed to what it knows, is not reduced to bookkeeping by any tracker reviewed here; it is part of K and is taken up in Section 7.

6. Conversion: The Gap Between Knowing and Acting

6.1 The explicit-to-applied gap

SimpleToM asks three questions about each story: whether the character is aware of a hidden fact, what the character will do, and whether what the character did was reasonable (Gu et al., 2024). The questions are binary, so chance is 50 percent. GPT-4o answers 95.6 percent of awareness questions correctly, 49.5 percent of behaviour questions and 15.3 percent of judgment questions. Claude-3.5-Sonnet answers 97.9, 67.0 and 24.9 percent; GPT-5 answers 98.5, 64.4 and 40.0 percent; o1-preview answers 95.6, 84.1 and 59.5 percent. Judgment accuracy below 50 percent means the model systematically rates the character's uninformed action as unreasonable, as if the character had known what the reader knows.

The authors then tried several cheap interventions. A mental-state reminder - inserting the model's own answer to the awareness question into the prompt for the behaviour question - was run on eight models and raised behaviour accuracy on every one of them. Seven of the eight gained between 21 and 38 points, the largest gain being Claude-3-Haiku's (23.6 to 61.1 percent). The exception was GPT-3.5, which went only from 7.6 to 12.2 percent and which also fails the awareness question itself (36.5 percent). Among the stronger models, GPT-4o's behaviour accuracy rose from 49.5 to 82.8 percent, Llama-3.1-405B's from 58.2 to 89.5 percent, and Claude-3.5-Sonnet's from 67.0 to 96.9 percent, with Claude's judgment accuracy rising from 24.9 to 84.1 percent. Judgment is where the reminder falls short: most models stayed at or below chance on it, and the authors write that "Apart from Claude-3.5-Sonnet, this is not enough to bring accuracies beyond random on the

judgment question", calling such interventions "generally fragile band-aids with limited scope" (Gu et al., 2024). The other comparisons used three strong models. A generic chain-of-thought prompt gave smaller and uneven gains (GPT-4o's behaviour accuracy rose from 49.5 to 62.8 percent, Llama-3.1-405B's moved from 58.2 to 57.2), and a generic system prompt urging attention to mental states changed little. Variants of both that direct the model to the character's awareness did far better, although the system-prompt variant stayed below the reminder, which the source's Table 6 caption judges "less effective than the mental state reminder", and barely moved judgment accuracy (Gu et al., 2024). The awareness-specific system prompt lifted Claude-3.5-Sonnet's behaviour accuracy from 67.0 to 88.9 percent; the awareness-specific chain-of-thought prompt lifted GPT-4o's from 49.5 to 87.4 percent and its judgment accuracy from 15.3 to 86.7 percent; and that prompt combined with the reminder produced the highest numbers reported, behaviour accuracy between 93.5 and 96.9 percent across the three models. None of these interventions adds information the model did not already produce. Each places the character's awareness where the decision is made. That is a conversion repair.

Saxena et al. (2026) replicate the gap and decompose recovery. On SimpleToM, Llama-3.1-8B scores 51.5 percent overall while recognising mental states at 99.4 percent and predicting behaviour at 28.9 percent. Reminders combined with a test-time training method raise its behaviour prediction by 40.8 percentage points; the authors attribute most of the overall improvement to the reminders, with the training method adding 1.3 to 5.5 points beyond them depending on the model. Temporal belief tracking remained hard even after intervention.

6.2 The same gap in play

Akata et al. (2023) asked why GPT-4 fails to alternate in the Battle of the Sexes against a partner that alternates. They ran two diagnostic simulations. When GPT-4 played and was additionally asked to predict the partner's next move, it predicted the alternation correctly from round 5 onward. When it was shown the same history as a game between two other players, it predicted the pattern from round 3. It could see the convention and did not follow it. Prompting it first to predict the partner's move and then to choose its own - which the authors call social chain-of-thought - improved coordination, including with human partners.

Sobotka et al. (2026) measure the gap quantitatively in incomplete-information games with open-weight models. Choosing the best response to the model's own decoded belief would have raised its payoff in about 95 percent of games, and for Llama 3.1 70B the decoded belief already supports 96.2 percent of the improvement achievable with a perfectly accurate belief. The models' own play realises only a fraction of that. Putting the belief into the prompt changes the model's action distribution substantially but does not reliably raise payoffs. This qualifies the SimpleToM reminder result. A plausible reading, not tested directly by either study, is that externalising a belief helps where the mapping from belief to action is simple and well-practised, as in predicting whether a shopper will pay for a spoiled item; it does not substitute for a decision rule where the mapping is strategic.

Goel et al. (2026) supply the decision rule. Their agent pairs a language-model planner with a Bayesian tracker of the partner's current skill in Overcooked and interrupts its own skill only when the partner's actions contradict the current inference. They define a belief-action gap: the fraction of decisions on which an agent with a correct estimate of its partner nonetheless executes a skill that does not complement it. Their agent and ProAgent reach nearly identical accuracy in predicting the teammate's skill - 0.79 and 0.78 in one layout, 0.71 and 0.71 in another - but the gap is 0.20 against 0.41 in the first and 0.28 against 0.38 in the second. In their ablation, using the posterior only to trigger interruption generally beats using it only as planner context, though the effect is noisy, with standard deviations of several hundred reward points, and in two of nine layout-partner cells the context-only or no-belief condition scores higher than the full agent. The authors describe their contribution as closing a literal-functional gap without improving belief accuracy, which is a precise statement of a conversion repair.

6.3 Conversion in failure taxonomies

MAST, a taxonomy built from annotated traces of seven multi-agent frameworks, groups coordination failures into one category and reports per-mode frequencies (Cemri et al., 2025). The most frequent mode in that category is a mismatch between an agent's reasoning and the action it takes, at 13.2 percent of all annotated failures, the largest of the six coordination modes, ahead of task derailment at 7.40 percent, proceeding on wrong assumptions instead of asking at 6.80 percent, conversation resets at 2.20 percent, ignoring another agent's input at 1.90 percent, and withholding information at 0.85 percent. Reasoning-action mismatch is conversion by definition. The MAST authors draw a different conclusion from the same category: that context or communication-protocol fixes are often insufficient for these failures, which call for deeper social reasoning. The two readings are compatible if "social reasoning" is taken to include C and K, and they conflict only if it is taken to mean I.

EnactToM's failure categories are similar in kind: withheld information, ignored partner constraints and misallocated messages are decisions about what to say and to whom, given what the partner knows (Juneja et al., 2026). The benchmark's authors put it directly: the models report their partners' knowledge but cannot reliably act on it.

6.4 A protocol is a conversion rule

HiddenBench offers the cleanest evidence that a fixed rule can do what a better partner model was supposed to do (Li, Y. et al., 2025). The authors report that agents

integrate distributed information correctly once it is disclosed, and that the failure lies in not surfacing it. They then imposed a two-stage protocol: in an exchange stage each agent shares one or two decision-relevant facts and gives one reason the current favourite may be wrong; in a decide stage the group chooses. The protocol does not tell agents that information is distributed or require them to disclose everything. On 18 tasks, accuracy rose from 0.037 to 0.800 for GPT-4.1, from 0.173 to 0.727 for Gemini-2.5-Flash, and from 0.043 to 0.743 for Gemini-2.5-Flash-Lite. Consilience goes further, replacing the fixed schedule with a controller that chooses, at each turn, an intervention and a speaker from a compact state summary, with a conformal calibration step that bounds the one-step regret of the executed action; on hidden-profile tasks across 12 models it improves accuracy over fixed and unstructured protocols and sometimes exceeds a baseline in which every agent sees all evidence (Babu et al., 2026).

A rule that says "share a fact and challenge the favourite" contains no model of anyone's mind. It encodes a policy that is good for the information structure of the task. Among the contrasts reviewed here, its effect sizes are larger than any reported for a component that adds inference alone.

7. Calibration: Whose Mind, at What Depth

7.1 More theory of mind can break coordination

Mu et al. (2026) built language-model agents with fixed reasoning depths. A depth-zero agent responds to the environment; a depth-one agent models its partner as depth zero; a depth-two agent models its partner as depth one. In a repeated anti-coordination matrix game with one round of memory, every pair whose depths differed by exactly one scored 75.00 points, the maximum, while every other pair, including two depth-one agents and two depth-two agents, scored 0.00. The authors' account of the depth-one pair is that each expects the other to change its choice after a miscoordination, so neither does. In Overcooked, misaligned pairs took 83.50 to 100.00 time units against 43.83 to 51.00 for aligned pairs. Their adaptive agent estimates the partner's depth from past play and responds to it, and the paper's own conclusion is that equipping agents with theory of mind does not necessarily improve coordination; only aligned reasoning does.

Two qualifications in the same paper matter. In a three-action variant of the matrix game, the alignment effect weakened and was reversed in several pairings: an aligned depth-one versus depth-two pair scored 46.50, below four of the five misaligned pairs (among them a depth-zero versus depth-two pair at 64.17), and the aligned depth-two versus depth-one pair, at 61.00, scored below two of them. And in cross-play in Overcooked against a heterogeneous set of partners, the adaptive agent's completion time (48.05) was close to that of a population-trained reinforcement-learning baseline (49.09). Depth alignment matters most when the optimal joint action set is small and agents are close to rational, as the authors say.

7.2 Calibration is feedback against the partner

The interventions that improve K share a structure: they compare predictions about the partner with what the partner does. Hypothetical Minds generates natural-language hypotheses about other agents' strategies, uses the top-valued hypotheses to predict behaviour, and reweights hypotheses by how well those predictions come true (Cross et al., 2024). Its ablation on a repeated rock-paper-scissors substrate found that adding hypothesis evaluation and refinement produced positive returns above 10 in all scenarios and particular strength against a best-responding opponent, and that a variant with evaluation and refinement but no theory-of-mind prompting did well against adaptive opponents and poorly against fixed ones. ProAgent infers teammates' intentions and updates its beliefs against their subsequent actions, reporting an average improvement above 10 percent over the previous best method when paired with human proxy models (Zhang, C. et al., 2024). ToM2Vec trains latent tokens against the partner's actual thoughts and reports that, in Overcooked, the gains survive partner model families not seen in training, and that expressing the same reasoning in text loses much of it (Sivakumaran et al., 2026). In each case the component that helps is a feedback loop or a supervised target from the partner, not a deeper inference procedure. CoELA, a modular cooperative agent with perception, memory, communication and planning components, reported GPT-4-driven agents surpassing planning-based methods and producing effective communication in two embodied environments (Zhang, H. et al., 2024); it describes no loop that checks predictions about the partner against the partner's behaviour, so it is not evidence for this pattern either way.

This was the lesson of the pre-language-model coordination literature as well. Machine theory of mind was first posed as meta-learning a prior over agents from their behaviour (Rabinowitz et al., 2018). The Hanabi challenge was proposed because the game puts reasoning about other players' beliefs and intentions in the foreground (Bard et al., 2020), and other-play addressed the finding that conventions learned in self-play do not carry over to novel partners, by exploiting known symmetries of the game so that arbitrary conventions are not learned (Hu et al., 2020). Cicero reached human-level play in Diplomacy by coupling a language model to planning and reinforcement learning that inferred other players' beliefs and intentions from dialogue (Meta FAIR Diplomacy Team et al., 2022); the partner model there was a dedicated, trained component, not a by-product of the language model. LLM-Coordination found language-model agents more robust to unseen partners than reinforcement-learning agents trained in self-play (Agashe et al., 2023), which is consistent with pretraining supplying a broad prior over human-like

conventions. A broad prior is K at the population level; it is not K for a particular partner.

7.3 Eliciting the partner's state is the hard part

Moore et al. (2026) isolate K more cleanly than any coordination benchmark. A persuader must get a target to choose one of three policy proposals by revealing information strategically; success depends on what the target knows and what it values. In the Revealed condition the persuader is shown both. In the Hidden condition it must ask or infer. Against a target programmed to update rationally, o3 succeeded in 65 percent of Revealed games and 6 percent of Hidden games, against a chance baseline of 10 percent; human persuaders succeeded in 22 and 29 percent. Given the partner's state, the model is a far better persuader than people are. Required to find out the partner's state, it does no better than chance (its 6 percent is below the 10 percent baseline, but not significantly) and far worse than people. The authors' interpretation is that the multi-step planning needed to elicit mental-state information and then use it is what the model lacks, and that its success against human targets in later experiments comes from rhetorical skill that does not require such planning.

The Revealed condition supplies T, I and most of K, leaving C. The Hidden condition requires the agent to build K through interaction. The gap between them - 59 percentage points for o3 and a reversal of 7 points in the opposite direction for humans - is a measure of how much of the task's difficulty for the model lay in calibration.

7.4 Learned theory of mind does not transfer to partners

The Decrypto result in Section 2.3 belongs here. Fine-tuning on program-generated ExploreToM stories improved ToMi accuracy by 27 points (Sclar et al., 2025), yet ExploreToM-8B, a Llama 3.1 8B variant fine-tuned on theory-of-mind data, played Decrypto worse than its base model in both roles (Lupu et al., 2025). Training I on stories with supplied records did not produce K for a partner in an interactive game. Riemer et al. (2025) state the general version: literal theory of mind does not imply functional theory of mind, and long-horizon adaptation to a partner is the hard case.

8. The Gate: What a Symbolic Guard Can and Cannot Supply

8.1 Guards in epistemic gossip and shields in reinforcement learning

The anchor's guard has two ancestors. In distributed epistemic gossip protocols, an agent's call is governed by a guard written in a simple epistemic logic, a condition on what the caller knows, and such guards are what make the protocols analysable (Apt et al., 2016; Apt and Wojtczak, 2017). Apt and Wojtczak (2017) extend the guard language with common knowledge and establish when such protocols remain decidable; Cooper et al. (2016) generalise the gossip goal to arbitrary epistemic depths and show that adding negative goals, such as requiring that some agent not learn some secret, makes deciding whether a successful protocol exists NP-complete. In safe reinforcement learning, a shield synthesised from a temporal-logic specification either offers the learner a list of safe actions or corrects an unsafe one after the fact (Alshiekh et al., 2018), and shields of this kind assume domain knowledge in the form of a possibly partial model of the environment, which becomes the binding constraint under partial observability (Carr et al., 2023).

The anchor proposal sits at the meeting point (Nasiri et al., 2026). Its guard permits a call only if the call is structurally possible and the guard model satisfies the call's epistemic precondition at the current history. Its lottery-transparency result shows that the agent-indexed probability weights do not change which knowledge-fragment guards are true. It transfers decidability from Apt and Wojtczak's logic for a modal-depth-one fragment.

8.2 What the guard evaluates against

The guard's precondition is evaluated at the actual call history. That history is global: it is the sequence of all calls, from which each agent's view is derived. A guard that checks whether agent a knows that agent b knows a secret needs the history, not a's view of it, because whether a knows something is a fact about which histories are consistent with a's view, and the guard must know which history is actual to evaluate it. In the testbed this is unproblematic; the epistemic logic core holds the history by construction. The architecture therefore gives its language model, for free, a perfect tracker - the component Section 4 found responsible for most published gains - and a gate.

Outside the testbed the authoritative state is the problem. Halpern and Moses (1990) showed that, formally speaking, common knowledge cannot be attained in practical distributed systems, and introduced weaker variants that are attainable in many cases of interest. An agent deployment with private tool calls, private retrieval and asynchronous messaging is such a system. A guard can still be built, but its state is then whatever the instrumentation records, and its verdicts are only as authoritative as that record is complete. Shielding under partial observability faces the same constraint and handles it by reasoning over a model of the environment (Carr et al., 2023); a guard over other agents' knowledge would need a model of what they observed, which is T again, now estimated rather than given.

8.3 What a veto buys

LLM-Coordination contains the only published ablation located for this paper that separates a gate from theory-of-mind reasoning in a coordination task (Agashe et al., 2023). In Hanabi with GPT-4-turbo, the base agent averaged 4.33 points with a bomb rate of 1.00 in the source's table. Adding a verification step, in which the

model checks that its chosen action follows the rules and will not end the game immediately, raised the score to 10.33 and the bomb rate to 0.00. Adding an explicit theory-of-mind reasoning step on top raised the score to 13.33. The theory-of-mind step was measured only on top of verification, never without it, so how the gain divides between the two depends on the order in which they were added. In the order tested, the gate, implemented by the model itself, more than doubled the base score and contributed twice as much as the theory-of-mind step. Li, H. et al. (2023) report a related effect without a gate: the belief state raised the share of outputs mapping to legal actions from 71.8 to 86.1 percent, and the authors used environment feedback for error correction when outputs were invalid.

A veto removes actions that are illegal or immediately fatal. It does not choose good actions among legal ones. In Hanabi, after the gate brought the bomb rate to zero, the best configuration still averaged 13.33 points, well short of the game's maximum of 25. In the Battle of the Sexes, every action is legal. In HiddenBench, withholding a fact violates no rule. The coordination failures in Sections 6 and 7 are, almost without exception, legal actions that serve the joint goal badly. A permit is not a recommendation.

8.4 What the guard architecture leaves to the language model

Read against the four links, the anchor's architecture assigns T to the epistemic logic core, G to the guard, graded uncertainty over histories to the belief engine, and everything else to the policy model: which permitted call to make, when, and with what content. That is C and K. The architecture's own interfaces for improving the policy - guard feedback supervising representation edits, and reinforcement learning against temporal-logic objectives - are described by the authors as hypotheses. The published evidence reviewed here suggests that the hypothesis worth testing first is whether guard diagnostics improve conversion, since a denied action tells the policy what it may not do, not what it should do.

The lottery-transparency result has a corresponding reading. Because positive probability weights cannot change the truth of a knowledge-fragment guard, the belief engine cannot affect what is permitted. It can affect only which permitted action the policy chooses, through the summary the policy reads. The probabilistic layer is therefore, in this architecture, a calibration and conversion input, not a safety input. That is a coherent division of labour, and it locates the open empirical question precisely.

9. The Measured Values

Table 2 consolidates the published values the argument rests on. Each row is one result in its source's own metric. The "link" column is the author's classification of which link the contrast varies.

Three patterns stand out. The largest effects in the table (reminders, protocols, oracle perspectives, the belief-action interruption) come from components that supply a record or impose a rule; none of them adds an inference procedure, and where attribution accuracy rises, as with the oracle perspective and the belief state, it rises because the record is supplied. The rows in which more or deeper theory of mind lowers performance (Mu et al.) are all calibration rows. And the one gate ablation shows a veto more than doubling a score that remained far from the maximum.

Figure 1 re-presents a subset of the table. Panel (a) shows, for six model-benchmark pairs, accuracy on reporting a belief against accuracy on using it; panel (b) shows six before-and-after contrasts for a single added component, labelled by the link it touches.

10. Attributing a Coordination Gain

The analysis suggests a procedure for any report that adding an epistemic component improved coordination. Its purpose is to assign the gain to a link, so that a reader knows what has been shown. Algorithm 1 states it.

Every step but one corresponds to a contrast already run somewhere in the literature: step S2 to SimToM's oracle perspectives and Li, H. et al.'s belief state; step S4 to the SimpleToM reminder, which supplies the model's own belief at decision time, Goel et al.'s context-versus-control ablation and Sobotka et al.'s best-response calculation; step S5 to Mu et al.'s cross-depth pairs and Sivakumaran et al.'s unseen-family test; step S6 to Agashe et al.'s verification ablation. Step S3 is the exception: no study located for this paper supplies oracle beliefs and compares them with a supplied record. No single study has run all of them. Step S7's last clause comes from a dining-philosophers study in which one model earned a mean reward of 0.827 at five philosophers while producing no meals, because the reward paid for avoiding deadlock (Hasan and BusiReddyGari, 2026); a coordination gain reported only in reward can be a gain in the reward's loopholes.

For the anchor architecture specifically, the procedure translates into four comparisons: the policy model with and without the epistemic core's record in its context (T); with and without the guard, holding the record (G); with the belief engine's summary replaced by a uniform one (which, by lottery transparency, cannot change what is permitted, and so isolates its effect on choice among permitted actions); and across partners that follow different gossip protocols (K).

11. The Strongest Case Against This Paper

The most serious objection is that the four links are not separable in the systems that matter. A model that tracks badly will infer badly; a model that cannot convert will not reveal whether it inferred correctly; calibration is inference about a particular partner. On this view the decomposition is an analyst's convenience, and the field's single word for the whole chain is the honest one.

Table 2. Published values bearing on the four links and the gate, one row per result, source named on each row.

Source	Setting	Contrast	Value	Link
Gu et al., 2024	SimpleToM, GPT-4o	awareness / behaviour / judgment	95.6 / 49.5 / 15.3%	C
Gu et al., 2024	SimpleToM, Claude-3.5-Sonnet	behaviour, none vs. reminder	67.0 vs. 96.9%	C
Gu et al., 2024	SimpleToM, Claude-3.5-Sonnet	judgment, none vs. reminder	24.9 vs. 84.1%	C
Saxena et al., 2026	SimpleToM, Llama-3.1-8B	awareness vs. behaviour	99.4 vs. 28.9%	C
Akata et al., 2023	Battle of the Sexes, GPT-4	predicts alternation from	round 5 (player), round 3 (observer)	C
Sobotka et al., 2026	Incomplete-info games	best response to own decoded belief gains	about 95% of games	C
Goel et al., 2026	Overcooked, Open layout	skill accuracy; belief-action gap	0.79 vs. 0.78; 0.20 vs. 0.41	C
Li, Y. et al., 2025	HiddenBench, 15 models	distributed vs. single full-info	30.1 vs. 80.7%	C
Li, Y. et al., 2025	HiddenBench, GPT-4.1, 18 tasks	baseline vs. protocol	0.037 vs. 0.800	C
Cemri et al., 2025	MAST, all annotated failures	share that are reasoning-action mismatch (largest coordination mode)	13.2%	C
Juneja et al., 2026	EnactToM hard, Gemini-Pro	literal vs. functional, single run	63.3 vs. 6.7%	C, K
Juneja et al., 2026	EnactToM hard, o3	literal vs. functional, single run	52.5 vs. 5.0%	C, K
Wilf et al., 2023	ToMi FB, GPT-4	zero-shot vs. perspective prompt	25.5 vs. 87.75%	T
Wilf et al., 2023	BigToM FB, GPT-3.5-Turbo	model vs. oracle perspective	70.5 vs. 96.0%	T
Sclar et al., 2023	ToMi, GPT-3 Davinci	gain from belief tracker	+38 points, to 92%	T
You et al., 2026	HopToM, 4 frontier models	1st vs. 4th order, direct prompt	73.5 vs. 20.0%	T, I
You et al., 2026	HopTrace	gain at 4th / 5th order	+22.0 / +38.5 points	T, I
Li, H. et al., 2023	Bomb game, GPT-4	rounds; 1st-order ToM, without vs. with belief state	28.3 vs. 12.3; 60.0 vs. 80.1%	T
Street et al., 2024	6th-order ToM	GPT-4 vs. adult humans	93 vs. 82%	I
Liang et al., 2025	LLM-Hanabi	r with score, 1st vs. 2nd order	0.76 vs. 0.58	I, K
Mu et al., 2026	Matrix game, memory 1	aligned vs. misaligned depths	75.00 vs. 0.00 points	K
Mu et al., 2026	3-action matrix game	aligned (1 vs. 2) vs. misaligned (0 vs. 2)	46.50 vs. 64.17 points	K
Moore et al., 2026	Persuasion, o3	target state Revealed vs. Hidden	65 vs. 6% (chance 10%)	K
Moore et al., 2026	Persuasion, humans	Revealed vs. Hidden	22 vs. 29%	K
Agashe et al., 2023	Hanabi, GPT-4-turbo	base / + verification / + ToM step	4.33 / 10.33 / 13.33	G, I

There is force in this. The links interact, and several interventions in Table 1 touch more than one. HopTrace replaces nested inference with a lookup over tracked records, so its gain is T and I together. Mu et al.'s depth-k agents differ in inference depth and in calibration at once. EnactToM's failures are conversion failures that may begin as tracking failures in long embodied episodes. The decomposition does not claim the links are independent. It claims that published contrasts vary them unequally and that the variation can be read off the design: an oracle record holds T fixed; an awareness question isolates I given T; a reminder changes C without changing the information; a partner swap varies K. Where a study runs one of these contrasts, it says something about one link. Where it does not, its gain is a sum whose terms are unknown.

A second objection is that the evidence base is skewed toward small, controlled tasks - two-sentence stories, two-by-two games, gridworld cooking - and that real deployments are messy enough that everything is tracking. That may be right, and if so the paper's conclusion strengthens rather than weakens: tracking is the link symbolic components serve best and the one whose inputs depend most on instrumentation.

A third objection concerns sources. Many key results are 2025-2026 preprints from single groups, several use the authors' own benchmarks, and the largest effects (reminders, protocols) were measured on few models or few tasks: the HiddenBench protocol on 18 tasks and three models, the SimpleToM reminder on eight models (three in the comparison with other interventions). The PCM-LLM reference architecture that succeeds where models fail was also used to generate the benchmark's fine-tuning demonstrations (Yan et al., 2026). The paper reports these limits on each result and places its weight on patterns that recur across groups - the awareness-behaviour gap in SimpleToM and its replication, the decoded-belief gap in two independent groups, and the depth-mismatch failure, which comes from one study but whose mechanism the authors spell out and qualify themselves.

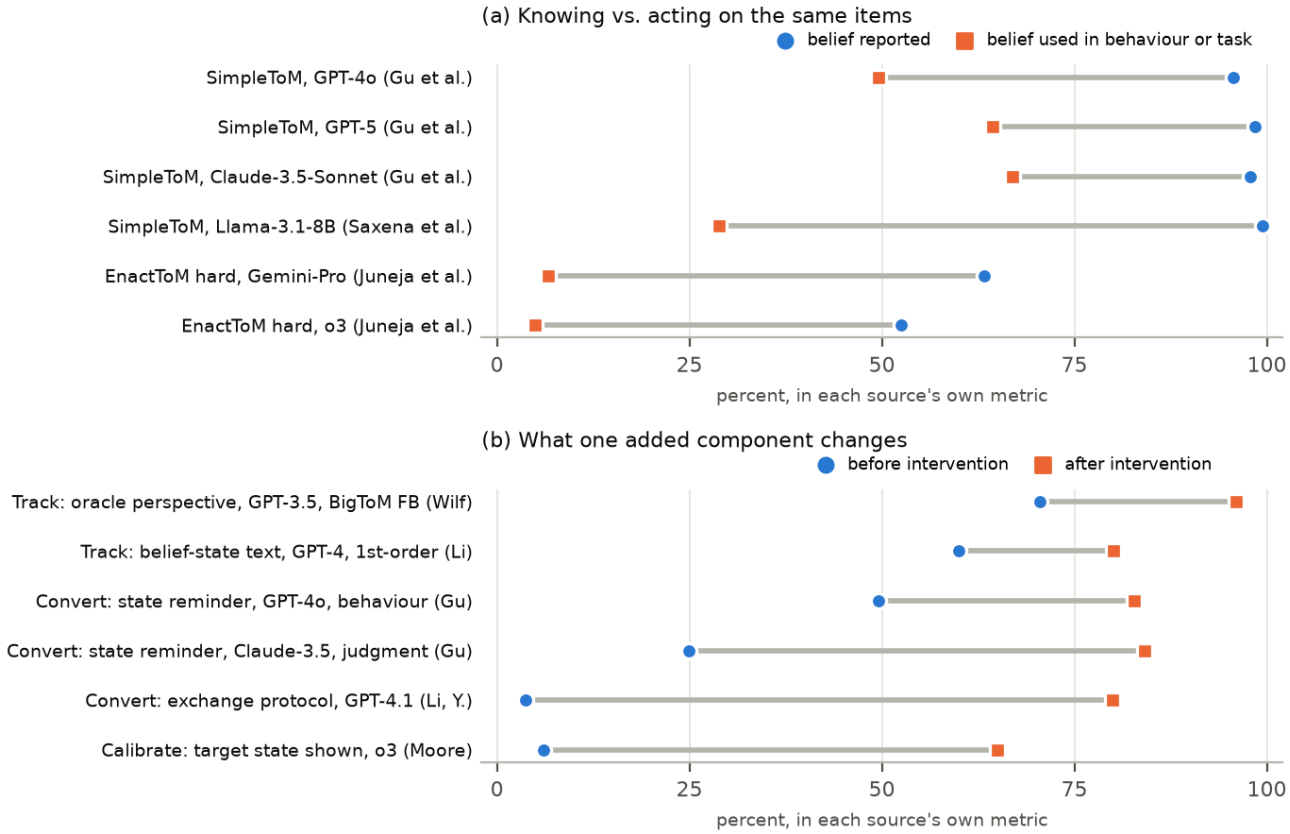


Figure 1. (a) Accuracy on belief questions (circles) against accuracy on behaviour prediction or functional task completion for the same items (squares): SimpleToM rows from Gu et al. (2024) Table 2 and Saxena et al. (2026); EnactToM hard-split single-run averages from Juneja et al. (2026). (b) One added component, before (circles) and after (squares): oracle perspective, GPT-3.5-Turbo, BigToM false belief (Wilf et al., 2023); textual belief state, GPT-4, first-order ToM questions (Li, H. et al., 2023); mental-state reminder, GPT-4o behaviour and Claude-3.5-Sonnet judgment (Gu et al., 2024 Table 3); exchange-then-decide protocol, GPT-4.1, HiddenBench accuracy 0.037 and 0.800 shown as percent (Li, Y. et al., 2025 Table 7); target's state Hidden versus Revealed, o3 persuasion success (Moore et al., 2026, Experiment 1). Every value is as published, each in its source's own metric; the link labels are the author's classification.

Algorithm 1. Attributing a reported coordination gain to a link

```
Input: base agent A; agent A+M with added component M;
       task family F; partner set P (scripted and learning)
step S1  score A and A+M on F with each partner in P
step S2  TRACKING - give A the record M builds (who
         observed what), nothing else changed; if
         A+record recovers most of the gain, credit T
step S3  INFERENCE - give A correct attributed beliefs
         (oracle); if A+oracle beats A+record, the
         remainder is credited to I
step S4  CONVERSION - with oracle beliefs supplied,
         compare beliefs-as-context against a fixed
         belief-to-action rule; that difference, and the
         gap from A+oracle to the best response given
         those beliefs, is credited to C
step S5  CALIBRATION - swap partners across P (other
         depths, other policies, unseen model families);
         a gain that holds for one partner only is K
step S6  GATE - disable the veto, keep all else; record
         the change in score and in invalid-action rate
step S7  report each share together with the partner
         set, and report reward and behaviour apart
Output: gain decomposed into T, I, C, K, G shares
```

12. What Is Not Known

No study located for this paper runs the full attribution in Algorithm 1 on one task. The shares of T, I, C, K and G in any reported coordination gain are therefore unknown.

No implementation of the anchor architecture has been evaluated. Whether its guard diagnostics improve the policy's choices among permitted actions is untested.

It is not known how large the tracking share is in open-ended deployments, because no benchmark reviewed here instruments private observations in a realistic multi-agent system and then compares an oracle record with an agent-maintained one.

It is not known whether conversion failures shrink with scale. SimpleToM's best behaviour and judgment accuracies belong to reasoning models, and Sobotka et al. report the belief-action gap, measured on verbal beliefs, narrowing but not disappearing in frontier API models; Gong et al. and Lupu et al. report reasoning models doing worse on other theory-of-mind tasks. These findings have not been reconciled.

It is not known whether calibration can be learned in a way that transfers across partner populations. ToM2Vec reports transfer to unseen model families in Overcooked; ExploreToM-8B's Decrypto result shows a learned skill failing to transfer to play. The conditions that separate the two outcomes have not been identified.

No published gate over other agents' knowledge has been tested with agents whose observations are private and only partly instrumented.

13. Studies That Would Settle It

Four studies, each within reach of existing benchmarks, would replace the classification in Table 1 with measurements.

First, run Algorithm 1 on EnactToM's hard split, where literal and functional scores are already measured on the same tasks. Supply the formally verified task record as an oracle (step S2), supply correct belief statements (step S3), and supply a scripted conversion rule for the five failure categories the authors identify (step S4). The benchmark's verification of required epistemic depth makes the oracle constructible.

Second, implement the anchor architecture on its own gossip testbed and on a variant in which the guard does not hold the global call history but only an instrumented, lossy log of it. Compare guarded and unguarded policies, and policies with and without the core's record in context. This measures what the guard supplies when its state stops being authoritative.

Third, extend Mu et al.'s depth-mismatch design to tasks in which the optimal joint action set is large, and to partners that are not language models, to measure how often calibration failures of this kind arise outside small games. Their own three-action results suggest the effect is conditional.

Fourth, run the SimpleToM reminder, the HiddenBench protocol and the Goel et al. interruption rule on the same set of models, to test whether a single conversion repair transfers across settings or whether each setting needs its own rule. If one repair transfers, conversion is a general capacity the model lacks; if none does, it is a property of the task's information structure, and the case for symbolic rules strengthens.

14. Conclusion

The debate over whether language models have theory of mind has been conducted mostly on tests that supply the record of who saw what and score the attribution. On those tests, recent models do well, and with an oracle record they do nearly perfectly. Coordination asks for more: a record the agent must keep, an attribution, an action that uses it, and a model of the partner that fits the partner. The published gains from explicit epistemic machinery come mostly from keeping the record and from forcing the action to follow the belief - two links that a symbolic component can take away from the language model. The clearest published loss from added theory of mind is a calibration loss, where reasoning more deeply about a partner who reasons differently makes things worse.

A guard over an authoritative state, the component the 2026 proposal puts at the centre, supplies a perfect record and a veto where such a state exists. It cannot supply the choice among permitted actions or the fit to a particular partner, and in most deployments it does not have the state. The proposal's own scope statements are accurate: its formal results concern the abstraction, and its learned components are hypotheses. The evidence reviewed here says which hypothesis to test first. It is not whether agents can be made to know what their partners know. It is whether they can be made to act on it, with the partner they actually have.

References

Agashe, S., Fan, Y., Reyna, A., & Wang, X. E. (2023). LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. arXiv:2310.03903. doi:10.48550/arXiv.2310.03903.

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). Playing repeated games with Large Language Models. arXiv:2305.16867. doi:10.48550/arXiv.2305.16867.

Alshiekh, M., Bloem, R., Ehlers, R., Konighofer, B., Niekum, S., & Topcu, U. (2018). Safe Reinforcement Learning via Shielding. arXiv:1708.08611. doi:10.48550/arXiv.1708.08611.

Annese, L., Patania, S., Serino, S., Foulsham, T., Rossi, S., Ruggeri, A., & Ognibene, D. (2025). Who Sees What? Structured Thought-Action Sequences for Epistemic Reasoning in LLMs. arXiv:2508.14564. doi:10.48550/arXiv.2508.14564.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? Psychological Review, 116, 953-970. doi:10.1037/a0016923.

Apt, K. R., Grossi, D., & van der Hoek, W. (2016). Epistemic Protocols for Distributed Gossiping. arXiv:1606.07516. doi:10.48550/arXiv.1606.07516.

Apt, K. R., & Wojtczak, D. (2017). Common Knowledge in a Logic of Gossips. arXiv:1707.08734. doi:10.48550/arXiv.1707.08734.

Babu, A., Kaur, R., Pramanik, V., Kotevska, O., Bastian, N. D., Jha, S., Raj, S., Wu, Y., Jha, S. K., & Roy, A. (2026). Consilience: Conformally Calibrated Communication Control for Hidden-Profile Multi-Agent Reasoning. arXiv:2608.20564. doi:10.48550/arXiv.2608.20564.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nature Human Behaviour, 1. doi:10.1038/s41562-017-0064.

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E.,

Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., & Bowling, M. (2020). The Hanabi Challenge: A New Frontier for AI Research. arXiv:1902.00506. doi:10.48550/arXiv.1902.00506.

Carr, S., Jansen, N., Junges, S., & Topcu, U. (2023). Safe Reinforcement Learning via Shielding under Partial Observability. arXiv:2204.00755. doi:10.48550/arXiv.2204.00755.

Cemri, M., Pan, M. Z., Yang, S., Agrawal, L. A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Zaharia, M., Gonzalez, J. E., & Stoica, I. (2025). Why Do Multi-Agent LLM Systems Fail? arXiv:2503.13657. doi:10.48550/arXiv.2503.13657.

Cooper, M. C., Herzig, A., Maffre, F., Maris, F., & Regnier, P. (2016). Simple epistemic planning: generalised gossiping. arXiv:1606.03244. doi:10.48550/arXiv.1606.03244.

Cross, L., Xiang, V., Bhatia, A., Yamins, D. L. K., & Haber, N. (2024). Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. arXiv:2407.07086. doi:10.48550/arXiv.2407.07086.

de Haan, I. B., van der Putten, P., & van Duijn, M. (2026). Evaluating Theory of Mind in Reasoning Models: Robustness over Reasoning. arXiv:2608.04646. doi:10.48550/arXiv.2608.04646.

Gandhi, K., Franken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). Understanding Social Reasoning in Language Models with Language Models. arXiv:2306.15448. doi:10.48550/arXiv.2306.15448.

Goel, H., Ellendula, A. S., Tadiparthi, V., Moradi Pari, E., Nourkhiz Mahjoub, H., & Chinchali, S. P. (2026). Bayesian Partner Modelling enables Adaptive Replanning for LLM Coordination. arXiv:2608.18490. doi:10.48550/arXiv.2608.18490.

Gong, N., Li, H., Dong, S., Lian, J., Fu, Y., & Xie, X. (2026). To Think or Not To Think, That is The Question for Large Reasoning Models in Theory of Mind Tasks. arXiv:2602.10625. doi:10.48550/arXiv.2602.10625.

Gu, Y., Tafjord, O., Kim, H., Moore, J., Le Bras, R., Clark, P., & Choi, Y. (2024). SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs. arXiv:2410.13648. doi:10.48550/arXiv.2410.13648.

Halpern, J. Y., & Moses, Y. (1990). Knowledge and common knowledge in a distributed environment. Journal of the ACM, 37, 549-587. doi:10.1145/79147.79161.

Hasan, N., & BusiReddyGari, P. (2026). Reward and Behavioral Measures of LLM Coordination in Dining Philosophers. arXiv:2606.07845. doi:10.48550/arXiv.2606.07845.

He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023). HI-TOM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. arXiv:2310.16755. doi:10.48550/arXiv.2310.16755.

Hu, H., Lerer, A., Peysakhovich, A., & Foerster, J. (2020). "Other-Play" for Zero-Shot Coordination. arXiv:2003.02979. doi:10.48550/arXiv.2003.02979.

Juneja, G., Lu, D., Agashe, S., Diwane, P., Gunn, E., Srinivasa, J., Liu, G., Wang, W. Y., Du, Y., & Wang, X. E. (2026). EnactToM: An Evolving Benchmark for Functional Theory of Mind in Embodied Agents. arXiv:2605.09826. doi:10.48550/arXiv.2605.09826.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. Cognition, 89, 25-41. doi:10.1016/S0010-0277(03)00064-7.

Kim, H., Sclar, M., Zhou, X., Le Bras, R., Kim, G., Choi, Y., & Sap, M. (2023). FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. arXiv:2310.15421. doi:10.48550/arXiv.2310.15421.

Kim, H., Sclar, M., Zhi-Xuan, T., Ying, L., Levine, S., Liu, Y., Tenenbaum, J. B., & Choi, Y. (2025). Hypothesis-Driven Theory-of-Mind Reasoning for Large Language Models. arXiv:2502.11881. doi:10.48550/arXiv.2502.11881.

Kosinski, M. (2023). Evaluating Large Language Models in Theory of Mind Tasks. arXiv:2302.02083. doi:10.48550/arXiv.2302.02083.

Kostka, A., & Chudziak, J. A. (2026). Evaluating Theory of Mind and Internal Beliefs in LLM-Based Multi-Agent Systems. arXiv:2603.00142. doi:10.48550/arXiv.2603.00142.

Le, M., Boureau, Y.-L., & Nickel, M. (2019). Revisiting the Evaluation of Theory of Mind through Question Answering. Proceedings of EMNLP-IJCNLP 2019, 5871-5876. doi:10.18653/v1/D19-1598.

Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., & Sycara, K. (2023). Theory of Mind for Multi-Agent Collaboration via Large Language Models. arXiv:2310.10701. doi:10.48550/arXiv.2310.10701.

Li, Y., Naito, A., & Shirado, H. (2025). Systematic Failures in Collective Reasoning under Distributed Information in Multi-Agent LLMs. arXiv:2505.11556. doi:10.48550/arXiv.2505.11556.

Liang, F., Zheng, T., Chan, C., Yim, Y., & Song, Y. (2025). LLM-Hanabi: Evaluating Multi-Agent Gameplays with Theory-of-Mind and Rationale Inference in Imperfect Information Collaboration Game. arXiv:2510.04980. doi:10.48550/arXiv.2510.04980.

Lupu, A., Willi, T., & Foerster, J. (2025). The Decrypto Benchmark for Multi-Agent Reasoning and Theory of Mind. arXiv:2506.20664. doi:10.48550/arXiv.2506.20664.

Meta Fundamental AI Research Diplomacy Team (FAIR), Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., et al. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. Science, 378, 1067-1074. doi:10.1126/science.ade9097.

Moore, J., Overmark, R., Cooper, N., Cibralic, B., Haber, N., & Jones, C. R. (2026). Large Language Models Persuade Without Planning Theory of Mind. arXiv:2602.17045. doi:10.48550/arXiv.2602.17045.

Mu, C., Zeng, Y., Zhang, Q., Shao, K., Chu, C., Guo, H., Jia, D., Wang, Z., & Hu, S. (2026). Adaptive Theory of Mind for LLM-based Multi-Agent Coordination. arXiv:2603.16264. doi:10.48550/arXiv.2603.16264.

Nasiri, M., Arvenaghi, M. S., Vaezi, S., & Ardeshir-Larijani, E. (2026). Epistemic-Probabilistic Model for Guarded Multi-Agent LLM Coordination. arXiv:2609.29366. doi:10.48550/arXiv.2609.29366.

Nguyen, H. M. (2025). A Survey of Theory of Mind in Large Language Models: Evaluations, Representations, and Safety Risks. arXiv:2502.06470. doi:10.48550/arXiv.2502.06470.

Okcu, A. S., Wang, W., Xie, T., Dalton, N., Nalmpantis, A., Arghal, R., & Giulianelli, M. (2026). The belief-action gap in language model agents. arXiv:2610.17444. doi:10.48550/arXiv.2610.17444.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1, 515-526. doi:10.1017/S0140525X00076512.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine Theory of Mind. arXiv:1802.07740. doi:10.48550/arXiv.1802.07740.

Riemer, M., Ashktorab, Z., Bouneffouf, D., Das, P., Liu, M., Weisz, J. D., & Campbell, M. (2025). Position: Theory of Mind Benchmarks are Broken for Large Language Models. arXiv:2412.19726. doi:10.48550/arXiv.2412.19726.

Saxena, P., Janzen, S., & Maass, W. (2026). Diagnosing the Explicit-to-Applied Theory of Mind Gap in LLMs. arXiv:2610.31060. doi:10.48550/arXiv.2610.31060.

Sclar, M., Kumar, S., West, P., Suhr, A., Choi, Y., & Tsvetkov, Y. (2023). Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. arXiv:2306.00924. doi:10.48550/arXiv.2306.00924.

Sclar, M., Yu, J., Fazel-Zarandi, M., Tsvetkov, Y., Bisk, Y., Choi, Y., & Celikyilmaz, A. (2025). Explore Theory of Mind: Program-guided adversarial data generation for theory of mind reasoning. arXiv:2412.12175. doi:10.48550/arXiv.2412.12175.

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. arXiv:2305.14763. doi:10.48550/arXiv.2305.14763.

Sileo, D., & Lernould, A. (2023). MindGames: Targeting Theory of Mind in Large Language Models with Dynamic Epistemic Modal Logic. arXiv:2305.03353. doi:10.48550/arXiv.2305.03353.

Sivakumaran, N., Payani, A., Lee, H., Bansal, M., & Stengel-Eskin, E. (2026). ToM2Vec: Continuous Theory-of-Mind Tokens for Multi-Agent Coordination. arXiv:2610.29934. doi:10.48550/arXiv.2610.29934.

Sobotka, J., Karabag, M. O., & Topcu, U. (2026). Why Do LLMs Struggle in Strategic Play? Broken Links Between Observations, Beliefs, and Actions. arXiv:2605.00226. doi:10.48550/arXiv.2605.00226.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 8, 1285-1295. doi:10.1038/s41562-024-01882-z.

Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Aguera y Arcas, B., & Dunbar, R. I. M. (2024). LLMs achieve adult human performance on higher-order theory of mind tasks. arXiv:2405.18870. doi:10.48550/arXiv.2405.18870.

Sun, H., Zhang, S., Niu, L., Ren, L., Xu, H., Fu, H., Zhao, F., Yuan, C., & Wang, X. (2025). Collab-Overcooked: Benchmarking and Evaluating Large Language Models as Collaborative Agents. arXiv:2502.20073. doi:10.48550/arXiv.2502.20073.

Ullman, T. (2023). Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv:2302.08399. doi:10.48550/arXiv.2302.08399.

Wilf, A., Lee, S. S., Liang, P. P., & Morency, L.-P. (2023). Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. arXiv:2311.10227. doi:10.48550/arXiv.2311.10227.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13, 103-128. doi:10.1016/0010-0277(83)90004-5.

Wu, Y., Xie, J., Zhang, D., & Xu, Z. (2025). DEL-ToM: Inference-Time Scaling for Theory-of-Mind Reasoning via Dynamic Epistemic Logic. arXiv:2505.17348. doi:10.48550/arXiv.2505.17348.

Yan, T., Sergeant-Perthuis, G., & Rudrauf, D. (2026). Belief Without Behavior: Measuring the Translation of Theory of Mind into Coordinated Social Action in Vision-Language Models. arXiv:2608.20975. doi:10.48550/arXiv.2608.20975.

You, S., Wang, Z., Liu, H., Liu, Y., He, Y., Hao, X., & Wang, L. (2026). Beyond Belief Nesting: Rethinking Higher-Order Theory of Mind in Large Language Models. arXiv:2610.17777. doi:10.48550/arXiv.2610.17777.

Zhang, C., Yang, K., Hu, S., Wang, Z., Li, G., Sun, Y., Zhang, C., Zhang, Z., Liu, A., Zhu, S.-C., Chang, X., Zhang, J., Yin, F., Liang, Y., & Yang, Y. (2024). ProAgent: Building Proactive Cooperative Agents with Large Language Models. arXiv:2308.11339. doi:10.48550/arXiv.2308.11339.

Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., Shu, T., & Gan, C. (2024). Building Cooperative Embodied Agents Modularly with Large Language Models. arXiv:2307.02485. doi:10.48550/arXiv.2307.02485.

Zhu, W., Zhang, Z., & Wang, Y. (2024). Language Models Represent Beliefs of Self and Others. arXiv:2402.18496. doi:10.48550/arXiv.2402.18496.

Zhu, Y., Bradford, M., Lai, K., Obiso, T., Venkatesha, V., Pustejovsky, J., & Krishnaswamy, N. (2026). Distributed Partial Information Puzzles: Examining Common Ground Construction Under Epistemic Asymmetry. arXiv:2603.05450. doi:10.48550/arXiv.2603.05450.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24442238