# Do LLMs Differentiate Epistemic Belief from Non-Epistemic Belief? Evidence from Behavioral Probing of GPT-4.1

**Anonymous Author(s)**

## Abstract

Humans distinguish between epistemic beliefs—those grounded in evidence and aimed at tracking truth—and non-epistemic beliefs—those rooted in values, desires, or faith. This distinction is fundamental to epistemology, yet no prior work has tested whether large language models (LLMs) maintain an analogous behavioral differentiation. We present four experiments probing whether GPT-4.1 treats epistemic and non-epistemic beliefs differently across explicit classification, spontaneous response patterns, belief revision under counterevidence, and factive verb sensitivity. We find that GPT-4.1 exhibits strong differentiation: it classifies belief types with 100% accuracy ($\kappa = 1.0$), spontaneously employs evidence-based reasoning for epistemic beliefs and values-based reasoning for non-epistemic beliefs (Cramér's $V = 0.95$), and correctly handles the presuppositional semantics of factive verbs ($p < 0.0001$). Notably, the model is *more* resistant to revising well-established factual beliefs than value-based beliefs (Cohen's $d = -0.45$, $p = 0.002$), the opposite of the naive philosophical prediction. These findings demonstrate that frontier LLMs are not "epistemically flat"—they have internalized distinct epistemic norms for different belief types—and carry direct implications for AI safety and alignment.

## 1 Introduction

Not all beliefs are created equal. When a person says "I believe the Earth orbits the Sun," they express something categorically different from "I believe honesty is the most important virtue." The first is an **epistemic belief**—grounded in evidence, amenable to empirical verification, and governed by truth-tracking norms. The second is a **non-epistemic belief**—anchored in values, shaped by personal experience, and evaluated by standards other than correspondence with objective reality. This distinction is foundational to epistemology and central to how humans reason, argue, and revise their views [Stalnaker, 1984].

Given the rapid deployment of LLMs in settings that require navigating both factual claims and value judgments—from educational tutoring to medical advice to political discourse—a natural question arises: **do LLMs behaviorally differentiate between epistemic and non-epistemic beliefs?** If they do, this has direct consequences for AI safety: models that treat factual misinformation and value disagreements identically may fail to challenge dangerous falsehoods or may inappropriately challenge legitimate moral convictions. If they do not, this reveals a fundamental gap in their capacity to engage appropriately with the diverse landscape of human belief.

Prior work has examined whether LLMs can reason *about* beliefs—tracking who knows what in Theory of Mind (ToM) tasks [Sap et al., 2023, Sileo and Lernould, 2023, Kim et al., 2023]—and has probed the epistemic robustness of LLM outputs [Krastev et al., 2025, Dies et al., 2026]. The KABLE benchmark [Suzgun et al., 2024] revealed that LLMs struggle to distinguish knowledge from mere belief, achieving only 54.4% accuracy on false belief tasks. However, a critical gap remains:

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

**no study has directly tested whether LLMs treat different *types* of beliefs differently in their behavioral responses.** Existing work asks whether models can reason about epistemic states, but not whether they deploy different reasoning strategies depending on whether a belief is epistemic or non-epistemic.

We address this gap with a behavioral probing study that tests whether GPT-4.1 differentiates epistemic from non-epistemic beliefs across four dimensions. We construct a novel stimulus set of 40 belief statements (20 epistemic, 20 non-epistemic) and probe the model's responses through: (1) explicit classification, (2) spontaneous response elicitation, (3) belief revision under counterevidence, and (4) factive verb sensitivity.

Our results reveal that GPT-4.1 maintains a robust functional distinction between belief types. The model classifies beliefs with 100% accuracy, spontaneously employs evidence-based reasoning for epistemic beliefs and values-based reasoning for non-epistemic beliefs (Cramér's $V = 0.95$), and correctly handles factive verb semantics ($p < 0.0001$). Most strikingly, the model is *more resistant* to revising well-established factual beliefs than value-based beliefs (Cohen's $d = -0.45$, $p = 0.002$), reversing the naive philosophical prediction. This asymmetry suggests the model has learned to weight strong prior evidence for established facts while treating value claims as inherently contestable.

In summary, our main contributions are:

- We design the first behavioral probing study that directly tests whether LLMs differentiate epistemic from non-epistemic beliefs, addressing a gap in the literature on LLM epistemic reasoning.
- We construct a novel stimulus set of 40 belief statements spanning scientific facts, historical facts, mathematical truths, moral values, life philosophy, and faith-based claims, paired with matched counterevidence for revision experiments.
- We conduct four experiments revealing that GPT-4.1 maintains a strong functional distinction between belief types, with large effect sizes across all dimensions (Cramér's $V$ up to 0.95, Cohen's $d = 0.45$).
- We identify a counterintuitive revision asymmetry—the model resists revising epistemic beliefs more than non-epistemic ones—and analyze its implications for AI safety and alignment.

## 2 Related Work

We organize related work into three threads: benchmarks for epistemic reasoning in LLMs, mechanistic studies of belief representation, and work on the stability and robustness of LLM belief-like behavior.

**Epistemic reasoning benchmarks.** Several benchmarks test whether LLMs can reason about beliefs and knowledge states. The KABLE benchmark [Suzgun et al., 2024] evaluates 15 LLMs across 13,000 epistemic reasoning questions, finding that models achieve 85.7% accuracy on factual belief confirmation but only 54.4% on false belief tasks—revealing a systematic inability to represent beliefs that diverge from world knowledge. Sileo and Lernould [2023] use dynamic epistemic logic to test formal epistemic reasoning, finding near-chance performance for most models and only 70% accuracy for GPT-4. Additional ToM benchmarks including HI-TOM [He et al., 2023], FANTOM [Kim et al., 2023], and SIMPLETOM [Strachan et al., 2024] consistently show that LLM performance degrades with increasing epistemic complexity. Gandhi et al. [2023] provide a comprehensive assessment confirming that LLMs remain far from genuine belief-reasoning agents. Unlike these studies, which test whether LLMs can reason *about* beliefs, we test whether LLMs treat different *types* of beliefs differently in their own responses.

**Internal representations of belief.** A growing body of work probes how LLMs internally encode belief states. Bortoletto et al. [2024] discover separate linear representations for self-beliefs versus attributed beliefs of characters in Mistral-7B, with activation interventions causally altering belief attributions. Lanham et al. [2025] identify a "lookback mechanism"—a pointer-dereference attention pattern in Llama-3-70B that implements belief tracking by attending to narrative points where beliefs were formed. Bigelow et al. [2025] formalize LLM belief dynamics through a Bayesian framework, showing that in-context learning operates as evidence accumulation (epistemic updating) while activation steering operates as prior modification (non-epistemic disposition change), with these two mechanisms being additive in log-odds space. This dual-pathway finding is particularly relevant to our work: it suggests a mechanistic basis for the epistemic/non-epistemic distinction that we

| Type | Subcategory | Example Belief |
|---|---|---|
| Epistemic | Scientific fact | "Water boils at 100°C at sea level" |
| Epistemic | Historical fact | "World War II ended in 1945" |
| Epistemic | Mathematical truth | "Pi is an irrational number" |
| Epistemic | Empirical generalization | "Smoking increases the risk of lung cancer" |
| Non-epistemic | Moral value | "Honesty is the most important virtue" |
| Non-epistemic | Life philosophy | "Money cannot buy happiness" |
| Non-epistemic | Aesthetic judgment | "Classical music is the highest form of artistic expression" |
| Non-epistemic | Faith-based | "Everything happens for a reason" |

Table 1: Representative stimuli from our belief dataset. Each type contains 20 beliefs drawn from four subcategories (5 per subcategory).

probe behaviorally. Herrmann and Levinstein [2024] propose four criteria (accuracy, coherence, uniformity, use) for genuine belief representations, arguing that current LLMs likely fail on uniformity. Our behavioral approach complements these representational studies by testing whether internal distinctions manifest in observable behavior.

**Stability and robustness of LLM beliefs.** Several studies examine whether LLM belief-like behavior is anchored in epistemic content or driven by non-epistemic contextual features. Krastev et al. [2025] find that prompt framing systematically modulates misinformation correction: creative intent reduces correction by 89%, and expert roles reduce correction by 21%. Dies et al. [2026] introduce the P-STAT framework, showing that epistemic familiarity governs belief stability—synthetic content destabilizes beliefs far more than fictional content—and that LLMs conflate distributional plausibility with epistemic justification. Li et al. [2023] study how LLMs update beliefs under contradictory evidence, finding inconsistent revision patterns. Kassner et al. [2021] propose structured belief storage for systematic consistency. These findings paint a picture of LLMs whose epistemic behavior is fragile and context-dependent. Our work extends this line by asking whether the *type* of belief (epistemic vs. non-epistemic) itself modulates LLM behavior, independent of contextual framing.

## 3 Methodology

We use a **behavioral probing** approach: we present GPT-4.1 with carefully constructed prompts containing beliefs of different types and analyze the model's responses via automated coding. This black-box approach is appropriate because we aim to test whether the epistemic/non-epistemic distinction manifests in *observable behavior*, complementing prior work on internal representations [Bortoletto et al., 2024, Lanham et al., 2025].

### 3.1 Stimulus Construction

We construct a balanced set of 40 belief statements: 20 epistemic and 20 non-epistemic, all hand-crafted as prototypical examples of their category.

**Epistemic beliefs** are evidence-based and empirically verifiable. They span four subcategories: scientific facts (e.g., "Water boils at 100°C at sea level"), historical facts (e.g., "World War II ended in 1945"), mathematical truths (e.g., "Pi is an irrational number"), and empirical generalizations (e.g., "Smoking increases the risk of lung cancer").

**Non-epistemic beliefs** are value-based and not truth-apt in the same sense. They span: moral values (e.g., "Honesty is the most important virtue"), life philosophy (e.g., "Hard work is the key to success"), aesthetic judgments (e.g., "Classical music is the highest form of artistic expression"), and faith-based claims (e.g., "Everything happens for a reason"). Table 1 shows representative examples from each category.

For Experiment 3, we construct matched counterevidence for each belief—plausible counterarguments for non-epistemic beliefs and purported disconfirming evidence for epistemic beliefs.

### 3.2 Experimental Design

We conduct four experiments, each targeting a different dimension of belief-type differentiation.

**Experiment 1: Explicit Classification.** We test whether GPT-4.1 can explicitly categorize beliefs as epistemic or non-epistemic. Each of the 40 beliefs is presented with the prompt: *"Classify this belief: 'I believe that [X].' Respond with 'epistemic' or 'non_epistemic'."* A system prompt defines the distinction. This serves as a baseline capability check: if the model cannot classify, behavioral differentiation would be surprising.

**Experiment 2: Differential Response Elicitation.** We test whether the model *spontaneously* treats belief types differently without being told about the distinction. Each belief is presented as: *"A person says: 'I believe that [X].' How would you respond?"* Responses are coded along five dimensions: whether the model challenges the belief, cites evidence, appeals to values, hedges its response, and its agreement level (1–5 scale).

**Experiment 3: Belief Revision Under Counterevidence.** We test whether counterevidence affects epistemic and non-epistemic beliefs differently. Each belief is presented as: *"Alex firmly believes [X]. However, [counterevidence]. Should Alex revise their belief?"* Responses are coded for: whether revision is recommended, revision strength (1–5), whether evidence quality is cited, whether subjectivity is acknowledged, and whether the response is nuanced.

**Experiment 4: Factive Verb Sensitivity.** We test whether the model treats "knows," "believes," and "values" differently when applied to the same content. Ten statements are each presented in three verb conditions: *"Alex [knows/believes/values] [X]. What can we conclude about [X]? Should Alex change their mind?"* Responses are coded for: whether truth is implied, revisability, treatment as factual versus value-laden, and certainty level (1–5). "Knows" is factive—it presupposes the truth of its complement [Stalnaker, 1984]—so a model with appropriate epistemic sensitivity should treat "Alex knows X" as implying X is true.

### 3.3 Implementation Details

All experiments use the OpenAI API with GPT-4.1 as the target model. We set temperature to 0.3 (low for consistency, non-zero for variation), maximum tokens to 500, and a random seed of 42. Each experiment is run 3 times to assess response consistency, yielding 120 trials for Experiments 1–3 and 90 trials for Experiment 4 (10 statements $\times$ 3 verbs $\times$ 3 runs).

**Automated coding.** Responses are coded by a separate GPT-4.1 call with structured JSON output. The coding call receives the original prompt, the model's response, and a rubric defining each coding dimension. This approach provides scalable, consistent coding while maintaining interpretability.

**Statistical analysis.** We use chi-squared tests for categorical outcomes, Mann-Whitney $U$ tests for ordinal outcomes (agreement level, revision strength, certainty level), and Kruskal-Wallis $H$ tests for multi-group comparisons (Experiment 4). We report Cramér's $V$ and Cohen's $d$ as effect size measures. All pairwise comparisons in Experiment 4 use Bonferroni correction. Significance is set at $\alpha = 0.05$.

## 4 Results

We present results from each experiment, followed by a cross-experiment synthesis.

### 4.1 Experiment 1: Explicit Classification

GPT-4.1 achieves **100% classification accuracy** across all 120 trials (40 beliefs $\times$ 3 runs), with Cohen's $\kappa = 1.0$. Every epistemic belief is classified as "epistemic" and every non-epistemic belief as "non_epistemic," with perfect consistency across runs. This confirms that the conceptual distinction between epistemic and non-epistemic beliefs is well-represented in the model and provides a ceiling baseline for the behavioral experiments.

| Feature | Epistemic | Non-Epistemic | Test | $p$-value | Effect Size |
|---|---|---|---|---|---|
| Challenges belief | 0.267 | **0.633** | $\chi^2 = 14.85$ | 0.0001 | $V = 0.35$ |
| Uses evidence | **1.000** | 0.033 | $\chi^2 = 108.42$ | $< 0.0001$ | $V = 0.95$ |
| Uses values | 0.067 | **0.967** | $\chi^2 = 93.74$ | $< 0.0001$ | $V = 0.88$ |
| Hedges response | 0.267 | **0.900** | $\chi^2 = 46.94$ | $< 0.0001$ | $V = 0.63$ |
| Agreement level (1–5) | **4.483** | 3.400 | $U = 2896$ | $< 0.0001$ | — |

Table 2: Experiment 2: Spontaneous response patterns for epistemic vs. non-epistemic beliefs. Values for binary features are proportions; agreement level is the mean on a 1–5 scale. All differences are statistically significant. Best values per row in **bold**.

| Metric | Epistemic | Non-Epistemic | Test | $p$-value |
|---|---|---|---|---|
| Revision strength (1–5) | 2.550 (SD 1.82) | **3.267** (SD 1.33) | $U = 1238$ | 0.002 |
| Recommends revision | 0.450 | 0.500 | $\chi^2 = 0.13$ | 0.715 |
| Cites evidence quality | **1.000** | 0.950 | — | — |
| Acknowledges subjectivity | 0.050 | **0.900** | — | — |

Table 3: Experiment 3: Belief revision behavior under counterevidence. Revision strength is significantly higher for non-epistemic beliefs ($d = -0.45$, $p = 0.002$). Best values per row in **bold**.

### 4.2 Experiment 2: Differential Response Patterns

Table 2 presents the response coding results. The model deploys strikingly different reasoning strategies depending on belief type.

The most striking finding is the near-perfect separation in reasoning strategy. For epistemic beliefs, the model cites evidence in **100%** of responses and appeals to values in only 6.7%. For non-epistemic beliefs, this reverses: values are cited in **96.7%** of responses and evidence in only 3.3%. This yields a Cramér's $V$ of 0.95 for evidence use and 0.88 for values use, indicating very large effect sizes.

The model also hedges significantly more for non-epistemic beliefs (90.0% vs. 26.7%, $V = 0.63$) and is more likely to challenge them (63.3% vs. 26.7%, $V = 0.35$). Agreement is higher for epistemic beliefs (mean 4.48 vs. 3.40, $p < 0.0001$), reflecting the model's tendency to affirm factual claims while adopting a more measured stance toward value claims.

### 4.3 Experiment 3: Belief Revision Under Counterevidence

Table 3 presents the revision results. Contrary to the prediction that epistemic beliefs should be more revisable given counterevidence, we find the opposite.

The model assigns significantly higher revision strength to non-epistemic beliefs (mean 3.27) than to epistemic beliefs (mean 2.55), with Cohen's $d = -0.45$ ($p = 0.002$). While the binary "recommends revision" rate does not differ significantly (45.0% vs. 50.0%, $p = 0.715$), the *strength* of the revision recommendation is meaningfully higher for non-epistemic beliefs. The model almost always cites evidence quality for epistemic beliefs (100%) and acknowledges subjectivity for non-epistemic beliefs (90.0%).

**Bimodal distribution for epistemic beliefs.** Revision strength for epistemic beliefs is bimodal. The model strongly resists revising mathematical and logical certainties (revision strength 1), but accepts revisions for empirical claims with genuine nuance (e.g., historical dating disputes). By contrast, non-epistemic revision strengths cluster around 3–4, reflecting a consistent view that value claims are debatable.

### 4.4 Experiment 4: Factive Verb Sensitivity

Table 4 presents the verb sensitivity results. The model treats "knows," "believes," and "values" as categorically different, consistent with the factive semantics of "knows."

| Verb | Certainty (1–5) | Implies Truth | Revisable | Treats as Factual |
|---|---|---|---|---|
| knows | **3.667** (SD 1.81) | **0.633** | 0.000 | **0.633** |
| believes | 1.433 (SD 1.04) | 0.000 | 0.000 | 0.200 |
| values | 2.167 (SD 1.44) | 0.000 | 0.000 | 0.100 |

Table 4: Experiment 4: Factive verb sensitivity. "Knows" yields significantly higher certainty and truth implication than "believes" or "values" (Kruskal-Wallis $H = 22.55$, $p < 0.0001$). Best values in **bold**.

| Hypothesis | Experiment | Supported? | Key Evidence | Effect Size |
|---|---|---|---|---|
| H1: Can classify belief types | Exp. 1 | **Strongly supported** | 100% accuracy, $\kappa = 1.0$ | Perfect |
| H2: Different response patterns | Exp. 2 | **Strongly supported** | $p < 0.0001$ on all features | $V = 0.63$–$0.95$ |
| H3: Asymmetric revision | Exp. 3 | **Supported, reversed** | $p = 0.002$ | $d = -0.45$ |
| H4: Factive verb sensitivity | Exp. 4 | **Strongly supported** | $p < 0.0001$ | $H = 22.55$ |

Table 5: Summary of hypothesis testing results across four experiments. All hypotheses are supported, with H3 showing the opposite direction from the naive prediction.

"Alex knows X" implies truth 63.3% of the time and yields a mean certainty of 3.67, while "Alex believes X" implies truth 0% of the time with certainty 1.43, and "Alex values X" implies truth 0% with certainty 2.17. The overall difference across verbs is highly significant (Kruskal-Wallis $H = 22.55$, $p < 0.0001$), with all pairwise comparisons significant after Bonferroni correction. The model correctly handles the factive presupposition of "knows"—treating it as a signal that the complement proposition is true—while treating "believes" as maximally agnostic about truth status.

### 4.5 Cross-Experiment Synthesis

Table 5 summarizes hypothesis testing across all four experiments.

All four hypotheses are supported with large effect sizes. The convergence of evidence across experiments—each probing a different dimension of belief-type differentiation—provides strong evidence that GPT-4.1 maintains a robust functional distinction between epistemic and non-epistemic beliefs.

## 5 Discussion

### 5.1 Interpretation of Results

**LLMs are not epistemically flat.** The central finding of this work is that GPT-4.1 does not treat all beliefs identically. The model has internalized distinct "epistemic toolkits"—one for factual claims (cite studies, provide data, affirm with confidence) and one for value claims (acknowledge perspectives, discuss trade-offs, hedge appropriately). The near-perfect separation in evidence use ($V = 0.95$) and values use ($V = 0.88$) indicates that this distinction is deeply embedded in the model's learned response patterns, likely reflecting the different rhetorical strategies used in factual vs. normative discourse in its training data.

**The revision asymmetry.** The most novel finding—that the model is more resistant to revising epistemic beliefs than non-epistemic ones—warrants careful interpretation. The naive philosophical prediction is that epistemic beliefs, being truth-tracking, should be *more* revisable given counterevidence, while non-epistemic beliefs, being rooted in values, should be held more firmly. We observe the opposite.

This asymmetry reflects a sensible epistemic heuristic: well-established factual beliefs (e.g., "pi is irrational") have accumulated massive evidential support, so any single piece of counterevidence is unlikely to outweigh the prior. In Bayesian terms, the model's posterior for well-established facts is dominated by its strong prior, and weak counterevidence cannot shift it significantly. By contrast, non-epistemic beliefs have weaker priors—the model recognizes that value claims are inherently contested—so counterarguments carry more weight. This interpretation aligns with Bigelow et al.'s

(2025) finding that in-context learning (evidence) and steering (prior) are separable pathways: the model's strong factual priors resist ICL-based updating.

The bimodal distribution for epistemic beliefs reinforces this interpretation: the model strongly resists revising mathematical certainties (where the prior is overwhelming) but readily revises empirical claims where genuine scientific uncertainty exists.

**Factive verb sensitivity as linguistic competence.** The model's correct handling of the factive presupposition of "knows"—treating "Alex knows X" as implying X is true while "Alex believes X" does not—demonstrates competence with a subtle linguistic distinction that encodes epistemic status. This is notable because the KABLE benchmark [Suzgun et al., 2024] found that LLMs fail to consistently respect the factive nature of knowledge. Our results suggest that the model may handle factivity better in natural response generation than in structured reasoning tasks, a hypothesis that merits further investigation.

### 5.2 Implications for AI Safety and Alignment

The behavioral differentiation we observe is largely *desirable* from an AI safety perspective. A model that treats "2+2=5" differently from "the death penalty is wrong" is behaving appropriately—the first should be firmly corrected, while the second warrants nuanced engagement.

However, the revision asymmetry raises a concern: the model may be *overconfident* about factual claims that happen to be wrong. If the model's strong priors for established facts extend to commonly-held-but-incorrect beliefs, it could resist valid corrections. Indeed, Experiment 2 reveals two cases where the model correctly challenged widely-held epistemic beliefs containing misconceptions (about the Amazon rainforest and body temperature), suggesting some capacity for appropriate factual self-correction. But this capacity may not generalize to less well-known factual errors.

The model's treatment of non-epistemic beliefs as more open to challenge also carries alignment implications. While respecting value diversity is important, excessive willingness to challenge value-based beliefs could result in models that argue against users' moral convictions in ways that feel adversarial. The appropriate calibration of when to challenge and when to respect beliefs—especially at the boundary between epistemic and non-epistemic—remains an open alignment problem.

### 5.3 Limitations

**Single model.** We test only GPT-4.1. Results may differ for other model families (Claude, Gemini, open-source models). Multi-model comparison is a priority for future work.

**Self-coding bias.** Response coding is performed by GPT-4.1 itself, introducing potential self-agreement bias. While this provides consistent, scalable coding, human validation of a subset would strengthen the findings.

**Prototypical stimuli.** Our 40 beliefs are hand-selected prototypical examples. Performance on ambiguous beliefs that straddle the epistemic/non-epistemic boundary (e.g., "democracy is the best form of government") remains untested.

**Training data confound.** The model's perfect classification may reflect learned philosophical categories rather than genuine epistemic processing. The behavioral differences could stem from different distributions of factual vs. normative text in training data, rather than an internalized epistemic distinction per se.

**Counterevidence quality.** Our counterevidence varies in quality—some is plausible (historical dating disputes), some is fabricated (claiming pi is rational). The revision results partly reflect this variation. A systematic manipulation of counterevidence strength would better characterize the revision asymmetry.

**Temperature sensitivity.** We use temperature 0.3. Higher temperatures might reveal greater variability in the model's differentiation, and the robustness of our findings across temperature settings remains to be tested.

## 6 Conclusion

We present the first behavioral study directly testing whether LLMs differentiate epistemic beliefs from non-epistemic beliefs. Across four experiments, we find that GPT-4.1 maintains a robust functional distinction: it classifies belief types perfectly, spontaneously employs different reasoning strategies for each type, handles factive verb semantics correctly, and—most strikingly—resists revising well-established factual beliefs more than value-based beliefs. These findings demonstrate that frontier LLMs have internalized epistemic norms that mirror a fundamental distinction in human cognition, with direct implications for how we design, deploy, and align these systems.

**Future work.** Three directions are immediate priorities. First, multi-model comparison across Claude, Gemini, and open-source models will establish the generalizability of our findings. Second, testing with ambiguous beliefs that straddle the epistemic/non-epistemic boundary will reveal the limits of the model's differentiation. Third, systematically varying counterevidence quality will better characterize the revision asymmetry and determine whether the model can distinguish valid from fabricated challenges to factual beliefs.

## References


Eric J. Bigelow, Asa Cooper Rager, Aaquib Arditi, Samuel Marks, and Max Tegmark. Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617*, 2025.

Matteo Bortoletto, Constantin Ruhdorfer, Mariem Abdulhai, and Anna Lauscher. Language models represent beliefs of self and others. *arXiv preprint arXiv:2402.18496*, 2024.

Lennart Dies et al. Representational and behavioral stability of truth in large language models. *arXiv preprint arXiv:2511.19166*, 2026.

Kanishk Gandhi et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023.

Yinghui He et al. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*, 2023.

Daniel A. Herrmann and Benjamin A. Levinstein. Standards for belief representations in LLMs. *arXiv preprint arXiv:2405.21030*, 2024.

Nora Kassner, Benno Krojer, and Hinrich Schütze. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. *arXiv preprint arXiv:2109.14723*, 2021.

Hyunwoo Kim et al. FANToM: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*, 2023.

Sekoul Krastev et al. Epistemic fragility in LLMs: Prompt framing systematically modulates misinformation correction. *arXiv preprint arXiv:2511.22746*, 2025.

Tamera Lanham et al. Language models use lookbacks to track beliefs. *arXiv preprint arXiv:2505.14685*, 2025.

Qian Li et al. Belief revision in large language models. *arXiv preprint arXiv:2309.02144*, 2023.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Evaluating large language models for theory of mind. *arXiv preprint arXiv:2302.02083*, 2023.

Damien Sileo and Antoine Lernould. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*, 2023.

Robert C. Stalnaker. *Inquiry*. MIT Press, 1984.

James W. A. Strachan et al. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. *arXiv preprint arXiv:2410.13648*, 2024.

Mirac Suzgun et al. Belief in the machine: Investigating epistemic reasoning in language models. *arXiv preprint arXiv:2410.21195*, 2024.