CPD: CONTRASTIVE PERSPECTIVE DECODING FOR FIRST-ORDER FALSE-BELIEF ANSWER SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In a false-belief story, a character who misses an object's move should still answer from their earlier information. On 174 narratives, three language models correctly answer both complementary Yes/No location questions on only 1–11% of stories. We study whether decoding can select an answer less influenced by information unavailable to the character. Contrastive Perspective Decoding (CPD) scores each candidate twice with the same frozen model: once with the full narrative and once with attention to an inaccessible sentence blocked. It favors candidates whose support rises under the mask. The relevant sentence and the character's access to it are supplied. Across three location datasets and three models, CPD reaches 0.81–0.95 macro-F1, improving on masking alone by 0.22–0.48; however, the supplied access and question type determine the binary label. On 928 stories paired with a constructed informed counterpart, CPD selects both actions on 0.915–0.952 of pairs. Yet a diagnostic classifier trained on 219 separate stories reaches 0.975 using only action wording and supplied access. This supervised control is not a like-for-like decoding baseline; it shows that the action-choice test also permits a shortcut. Our results therefore characterize conditional answer selection, not autonomous inference of what the character knows or shortcut-resistant transfer.

1 INTRODUCTION

Theory of mind (ToM) is the capacity to attribute mental states to other agents Wimmer & Perner (1983). In a first-order false-belief task, a character observes an object at one location, misses a later move, and should therefore still believe that the object is at the original location. Answering correctly requires using the character's outdated information rather than the current world state.

The answer format can obscure this distinction. On the same 400 false-belief narratives from BigToM Gandhi et al. (2023), one model reaches 96% accuracy when choosing between two named locations but only 10% when answering whether the object remains at its original location. This gap motivates a controlled test, but does not by itself isolate the cause because both the question and the answer candidates change.

We therefore ask two complementary Yes/No questions about every narrative: according to the character, is the object at its original location, and is it at its current location? The two questions encode the same underlying location belief, but change the queried predicate and reverse the correct answer. Across three model families, fewer than 11% of the narratives receive both required answers. Qwen and Llama favor the current-location formulation, whereas Mistral favors the original-location formulation. This test therefore diagnoses sensitivity to question formulation and answer polarity, not pure token identity. A Yes-minus-No score is a candidate score for one question, not a direct measure of the character's believed location.

Figure 1 illustrates the conflict between the character's outdated belief and the current state. Inhibitory-control accounts of false-belief errors describe the need to select a belief-consistent answer despite a salient reality-consistent one Leslie et al. (2005); Birch & Bloom (2007). This account motivates the contrast in our decoder; behavioral results do not establish that a language model implements the same cognitive process.

Contrastive Perspective Decoding (CPD) implements this idea with two scoring passes through one frozen language model. The full pass retains the relevant state change. The masked pass blocks

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

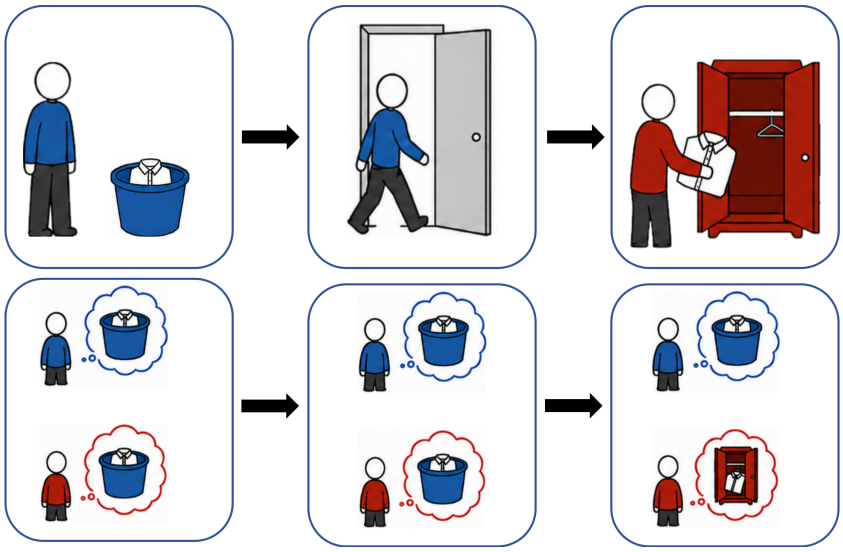

Figure 1: The character misses the shirt's move from basket to wardrobe. Across 174 analogous stories, the three models answer both complementary location questions correctly only 1–11% of the time (Table 1).

attention to that change when the character did not witness it. CPD then favors an answer when its score rises after the inaccessible change is blocked. The text, answer candidates, and model parameters remain identical, so the difference measures how answer support responds to the selected span rather than to a rewritten narrative.

Our experiments hold the supplied perspective information fixed. We compare CPD with masking alone and with a one-pass prompt receiving the same change and witnessing status; complementary questions and matched random masks test whether the result can be attributed to a fixed answer preference or arbitrary token removal. The information supplied to these methods is unusually strong: in the original-location binary protocol, witnessing status and question type determine the label. Accordingly, we study a conditional answer-selection intervention, not an end-to-end ToM system.

The paper contributes a paired-question diagnostic of formulation sensitivity and a controlled test of full-versus-masked answer-score contrast. We also examine variable action choices under paired information-access conditions. An option-only control exposes why high action-choice accuracy cannot, on this benchmark, establish transfer of belief reasoning.

2 Related Work

ToM prompting methods provide demonstrations, decompose the problem, or explicitly reconstruct a character's observations. SymbolicToM Sclar et al. (2023) tracks belief states, while SimToM Wilf et al. (2024), PercepToM Jung et al. (2024), and Decompose-ToM Sarangi et al. (2025) retain or restate information available to the character. PICTURE Jeon & Lee (2026) asks the model to describe what the character knows before answering. SimpleToM Gu et al. (2026) tests whether models use an inferred information state to predict a character's next action; we use its action candidates for a separate transfer experiment. Discrete World Models supply structured descriptions of environmental changes and interactions Huang et al. (2024); DEL-ToM scores explicit belief-update traces at inference time Wu et al. (2025). These approaches change the textual context or generate intermediate reasoning. CPD instead keeps the text fixed, blocks direct attention to selected event tokens, and contrasts answer scores from the two attention conditions. Our two one-pass notes are narrow equal-information controls, not implementations of these structured methods.

Contrastive decoding methods combine two output distributions. Contrastive Decoding Li et al. (2023b) compares model sizes, CAD Shi et al. (2024) compares the same model with and without

context, and DoLa Chuang et al. (2024) compares transformer layers. DeCoRe Gema et al. (2025), MaskCD Deng & Yang (2025), and PH3 Jin et al. (2024) mask branches or heads to reduce hallucination and knowledge conflict. CPD also compares two scores from one model and uses a CAD-like linear contrast. Its contribution is the question-specific full/masked attention pair, the exact empty-mask identity, and an evaluation that reverses the correct answer token while holding the belief fixed. Attention knockout has been used to trace factual recall Geva et al. (2023), while causal tracing localizes mediating activations Meng et al. (2022); CPD places the mask intervention inside decoding rather than using it only for post-hoc analysis.

Representation-based studies show that belief-related information can be decoded from language-model activations. RepBelief Zhu et al. (2024) and *Brittle Minds* Bortoletto et al. (2025) learn belief directions and use the learned directions for intervention; ITI Li et al. (2023a) similarly selects and edits activation directions. These methods require labeled belief examples, probes, or learned steering vectors. CPD does not learn a belief direction. It needs a change span and a visibility decision, then intervenes only through the attention mask and decoding rule. This distinction removes belief-probe training but retains the need for white-box model access and reliable event localization.

3 Contrastive Perspective Decoding

CPD receives a narrative, a question, candidate answers, and two pieces of perspective information: a relevant text span and whether the target character had access to its content. In location tasks, this is the sentence containing an object move and whether the character witnessed it. CPD first converts this information into a question-specific attention mask. It then scores the same candidates with the same frozen model under complete and masked attention. Finally, it amplifies the score of a candidate when blocking the inaccessible span increases that candidate's support.

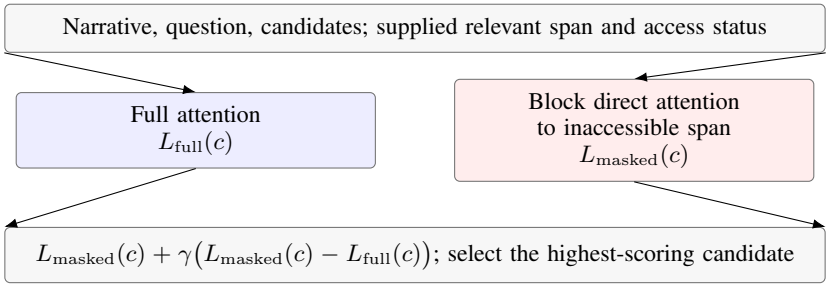

Figure 2: CPD holds the prompt, candidates, and model fixed while blocking direct attention to a supplied inaccessible span.

3.1 Events, Text Spans, and Visibility

Let $S = (s_1, \ldots, s_n)$ be the narrative sentences. An *event* e is a change in the queried world state, such as moving an object from a drawer to a box. In our implementation, an *event sentence* is the complete narrative sentence containing one or more such changes, and the *event span* $r(e) = [a_e, b_e)$ is the interval of input-token positions covered by that sentence. The mask blocks direct attention to that whole sentence as a key. If a sentence contains multiple changes, they share one span; we do not claim clause-level localization.

Let Q be the question, x the target character, and C the answers. E_Q contains the events that can change the state queried by Q; for a location question, these are moves of the queried object. $R_Q = \{r(e) : e \in E_Q\}$ contains their token spans. A visibility flag $w_e^x \in \{0, 1\}$ records whether character x perceived or later learned about event e. Visibility describes access to a state change, not access to the sentence that expresses it.

The same attention contrast can be applied to an inaccessible fact rather than a state change. In the action-choice transfer experiment, the selected span is the sentence stating a fact unknown to the

actor. We do not call this sentence an event or interpret action-choice accuracy as location-belief accuracy.

In the binary original-location protocol used here, an unwitnessed move requires "Yes" and a witnessed move requires "No"; the complementary current-location question reverses the mapping. A rule supplied with the relevant transition, witnessing status, and question type recovers every label in the three location datasets without CPD. Our experiments isolate how a frozen model's candidate scores respond to a mask under this strong information condition. They do not measure the model's ability to infer the transition or its visibility.

3.2 Masking Inaccessible Narrative Information

Let $P \in \{\text{full}, \text{masked}\}$ denote the same tokenized prompt under the corresponding attention mask, and let $L_P(c)$ denote the score of candidate c. For $c \in \{\texttt{Yes}, \texttt{No}\}$, $V(c)$ contains the accepted first-token IDs of candidate c, including leading-space and capitalization variants. We use $L_P(c) = \log \sum_{t \in V(c)} p(t \mid P)$. For a multi-token candidate $c = (c_1, \ldots, c_m)$, we use the length-normalized autoregressive score $L_P(c) = m^{-1} \sum_{j=1}^{m} \log p(c_j \mid P, c_{<j})$. These are decoding scores, not probabilities that the character holds a belief.

The full-context pass uses the model's standard causal attention and produces $L_{\text{full}}(c)$. For the masked pass, we collect the token positions $U_x = \bigcup_{e \in E_Q : w_e^x = 0} r(e)$. This pass is question specific: it masks relevant state changes that the character did not perceive rather than reconstructing every sentence available to the character. For action choices, U_x instead contains the supplied key-fact sentence when the actor is unaware of it, and is empty after the actor is explicitly informed.

We construct an additive attention mask $M \in \mathbb{R}^{1 \times 1 \times L \times L}$, where L is the scored sequence length at that step. Entries whose key position lies in U_x receive the minimum bf16 value; all other entries retain the causal mask.

Later tokens cannot attend directly to those selected-span key positions. This does not guarantee removal of information paraphrased elsewhere in the narrative. Input tokens, token positions, and model parameters remain unchanged. Applying M to every attention layer and head yields $L_{\text{masked}}(c)$; both scores come from the same frozen language model.

3.3 Contrastive Decoding Rule

Figure 2 summarizes the two scoring passes and their contrast. CPD predicts the candidate with the largest masked-pass score plus a scaled contrast against the full-context score:

$$\hat{y} = \arg\max_{c \in C} \left[L_{\text{masked}}(c) + \gamma \big(L_{\text{masked}}(c) - L_{\text{full}}(c) \big) \right]. \tag{1}$$

The first term is the candidate score after masking the selected span. The contrast is positive when the mask raises a candidate's score and negative when it lowers that score. It measures how each answer responds to suppressing the selected span. Increasing $\gamma \geq 0$ gives this contrast greater influence. In the inhibitory-control analogy, the contrast upweights support that survives suppression of the inaccessible information.

The construction has an exact no-change case. When the actor has access to all selected information, U_x is empty, M changes no attention edge, and $L_{\text{masked}} = L_{\text{full}}$. Equation 1 then reduces to the base-model decision for every value of γ. This identity is a property of the decoder and always holds when U_x is empty. The access decision determines whether a particular item should receive such an empty mask.

4 Experiments

We first measure sensitivity to complementary questions, then compare decoding methods under the same supplied perspective information and test transfer across narrative and answer formats. All reported CPD results use supplied perspective-relevant spans and access conditions.

4.1 Experimental Setup

Datasets and models. The primary test contains 300 OpenToM questions Xu et al. (2024): 174 false-belief (FB) cases, where the character misses a relocation, and 126 true-belief (TB) cases, where the character witnesses it. BigToM Gandhi et al. (2023) contributes 800 balanced narratives in a different template, and ExploreToM Sclar et al. (2025) contributes 180 multi-event narratives. We test Qwen2.5-7B-Instruct, Llama-3-8B-Instruct, and Mistral-7B-Instruct-v0.3. For action-choice transfer, SimpleToM Gu et al. (2026) supplies 928 evaluation stories from 557 story cores, disjoint from 219 development stories. Each story has two different natural-language actions: one judged appropriate when the actor lacks a key fact, and one for the counterfactual informed actor. We add an explicit disclosure to construct the informed version; the added sentence itself was not separately human-annotated.

Perspective information and selection. CPD receives the relevant state-change span and witnessing status, holding perspective inference fixed so that the experiment isolates Equation 1. OpenToM spans are matched from object, mover, and destination metadata with an answer-independent lexical fallback. BigToM uses the state-changing sentence in its generated template, and ExploreToM uses the generator's event and witness traces. A disjoint 296-item OpenToM development split selects $\gamma \in \{1, 2, 4, 8\}$ by macro-F1 subject to true-belief accuracy ≥ 0.90; the selected $\gamma = 8$ is fixed for every reported test. The supplied information contains no explicit answer, although, as discussed in Section 3, it can determine the answer in this binary protocol. For SimpleToM, the supplied span is its key-information sentence; it is masked only when the actor is unaware. Neither the span nor the access condition is inferred by CPD.

Metrics and conditions. For Yes/No questions, macro-F1 is primary because OpenToM's FB answer is always Yes; we also report FB and TB accuracy. The complementary test asks both whether the object is at its original location and whether it is at its current location, and counts a narrative as jointly correct only when both opposite answers are correct. For SimpleToM, we score the two action strings in shuffled order and count a story as jointly correct only when both its unaware and informed action choices are correct. We use the same length-normalized candidate scoring for all conditions; these values are not directly comparable to published SimpleToM multiple-choice generation scores. The evaluation split was defined before this experiment but is not an untouched prospective holdout. For a diagnostic of answer-writing cues, we also train an L2-regularized logistic classifier on the 219 development stories. It uses word and character n-grams and the lengths of the two options; at test time, supplied access selects its unaware prediction or the opposite option. It does not read the test story. Its penalty is chosen within development. This control uses development answer labels and was introduced after inspection of initial test results, so it is neither label-free nor a prospective baseline for CPD. *Mask only* applies the supplied attention mask without score contrast. The *one-pass note* states the same change and witnessing status in natural language but never states the answer or preferred candidate. On OpenToM, a second one-pass control explicitly asks the model to track the character's location belief, updating it only after witnessed changes. It receives the same span and status but no answer label. A deterministic rule derives the label from the supplied variables in this binary format; it is a protocol diagnostic, not a learned baseline.

Baselines and statistics. We include zero-shot, four-shot, chain-of-thought, CAD Shi et al. (2024), DoLa Chuang et al. (2024), PercepToM Jung et al. (2024), PICTURE Jeon & Lee (2026), Decompose-ToM Sarangi et al. (2025), and SimToM Wilf et al. (2024). Published prompts are retained. Deterministic methods use greedy decoding; PICTURE uses temperature 0.3 and four generation seeds. Models use bf16 and eager attention. Invalid generated answers are counted as incorrect in class accuracy and as neither label in macro-F1; for example, chain-of-thought produces 14 and 50 invalid outputs for Llama and Mistral on the 300 OpenToM questions. Exact paired McNemar tests are computed on item-level correctness and Holm-corrected within each prespecified family. Macro-F1 differences use 10,000 paired bootstrap resamples and report unadjusted 95% intervals. Decoder cost is measured relative to the corresponding single-pass Base run; hardware is matched within each ratio.

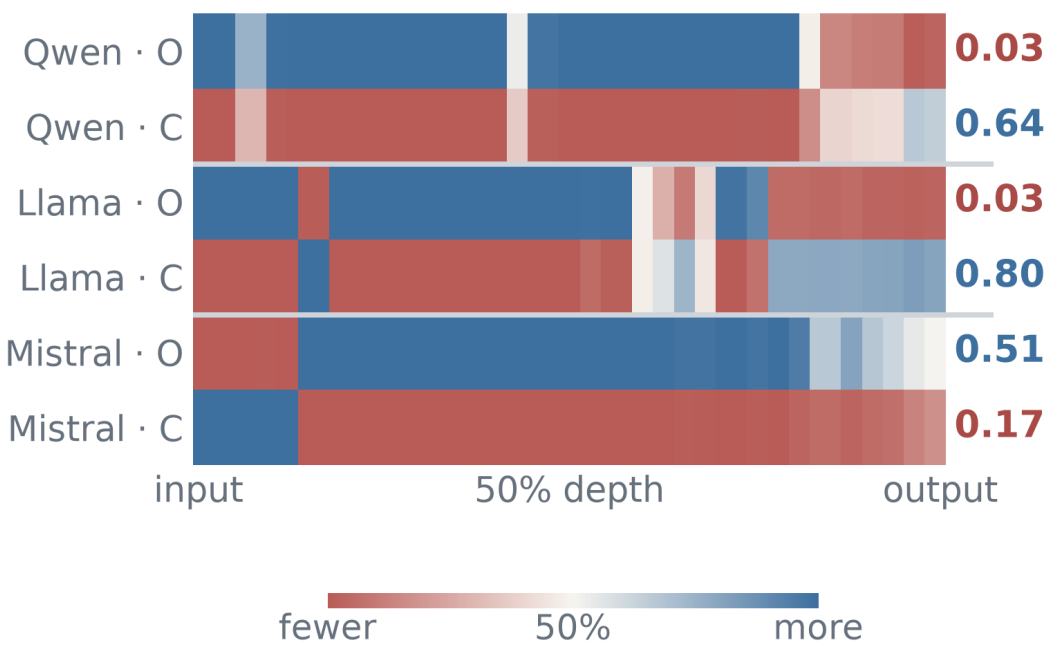

Figure 3: Fraction of false-belief items whose correct Yes/No token has the higher layerwise score. O/C denote original/current-location questions; right labels are output-layer values.

4.2 Complementary Questions and Answer Support

For each false-belief narrative, we ask whether the object is at its original location and whether it is at its current location from the character's perspective. The believed location is unchanged, but the correct token reverses.

Table 1: Joint accuracy on 174 complementary false-belief pairs. Both opposite answers must be correct; random-mask values average 20 draws.

Condition	Qwen2.5	Llama	Mistral
Base	0.011	0.017	0.109
Mask only	0.063	0.080	<u>0.454</u>
Matched random	<u>0.102</u>	<u>0.136</u>	0.107
CPD	**0.500**	**0.236**	**0.523**

Table 1 shows that Base answers both correctly on only 0.011, 0.017, and 0.109 of narratives. CPD raises joint accuracy above Mask only and the mean of 20 matched random masks for every model. It is higher than every random draw; after a global 60-test correction, all Qwen and Mistral draws but only one Llama draw remain significant. We therefore claim a consistent numerical effect, not uniform draw-level significance. The individual questions remain asymmetric: for Llama, CPD raises the original-location question from 0.121 to 0.598 but lowers the current-location question from 0.810 to 0.362. The test supports improved joint consistency, not invariance to question formulation.

Figure 3 traces the score of the correct token through the network for the two complementary questions. Qwen and Llama end near 0.03 on the original-location question but at 0.64 and 0.80 on the current-location question; Mistral shows the reverse preference (0.51 versus 0.17). This is a layerwise answer-score diagnostic, not a measure of believed location. Its reversal across two questions about the same belief is precisely why the decoder contrasts candidate scores rather than interpreting a Yes-minus-No margin as a belief variable.

Table 2: OpenToM macro-F1 on 300 questions and relative latency in a separate 40-item A6000 benchmark. Lower rows give false-/true-belief accuracy.

Method	Qwen2.5	Llama-3	Mistral	Latency ×
Without supplied perspective information				
Zero-shot	.302	.309	.336	1.00
Few-shot ($k = 4$)	.296	.363	.345	1.44–1.64
Chain-of-thought	.296	.434	.341	1.43–42.71
CAD	.322	.330	.342	1.38–1.44
DoLa	.302	.296	.315	1.06–1.19
PercepToM	.381	.317	.440	.92–.94
PICTURE	.375	.487	.576	19.89–45.91
Decompose-ToM	**.461**	.454	.352	203.09–256.01
SimToM	.413	**.556**	**.672**	43.10–71.94
Same supplied change span and witnessing status				
Mask only	.333	.351	.557	1.00
One-pass note	.296	.445	.427	–
State-tracking note	.296	.417	.607	–
CPD	**.810**	**.806**	**.870**	1.41–1.43
Controlled comparison: FB/TB accuracy				
Mask only	.034/1.000	.052/1.000	.293/.984	–
One-pass note	.000/1.000	.149/1.000	.132/.992	–
State-tracking note	.000/1.000	.121/.992	.603/.619	–
CPD	**.672**/1.000	**.667**/1.000	**.787**/.984	–

Bold/underline mark the top two macro-F1 scores within each information block. The two note prompts were not measured in the separate latency benchmark.

4.3 Controlled Decoder Comparison on OpenToM

Table 2 distinguishes ordinary methods from controlled conditions that receive the correct change span and witnessing status. The first block provides context, not an equal-information ranking against CPD. Within the controlled block, CPD reaches macro-F1 0.810, 0.806, and 0.870. Relative to Mask only, the gains are 0.476, 0.455, and 0.313; relative to the one-pass note, they are 0.514, 0.361, and 0.443. All six paired comparisons remain significant after correction (largest adjusted $p = 3.23\times10^{-20}$). The contrast adds to masking and to this specific one-pass natural-language note. An explicit state-tracking note reaches 0.296, 0.417, and 0.607 macro-F1 on the same OpenToM items; CPD remains higher for all three models. These prompts do not exhaust the structured methods that could use the same information.

CPD preserves TB accuracy at 1.000, 1.000, and 0.984. This is expected from the empty-mask identity: a correctly identified witnessed change is not blocked, so the two scores are equal.

The implementation reuses the full scores for an empty mask; all $1{,}944$ true-belief model–items across the three datasets have identical stored full and masked scores. On OpenToM, 57, 58, and 37 false-belief items remain incorrect for Qwen, Llama, and Mistral; in 33, 34, and 27 of these errors, masking does not increase the correct answer's margin over the alternative. The remaining errors have a favorable but insufficient contrast. In the separate 300-item runs, CPD costs 1.57–$1.59\times$ Base and the one-pass note 1.05–$1.10\times$; across all three datasets, CPD costs 1.30–$1.59\times$.

4.4 Transfer Across Location Narratives

Figure 4 holds the answer format, supplied information, and $\gamma = 8$ fixed while changing the narratives. CPD exceeds Mask only and the one-pass note in all nine model–dataset combinations, by 0.22–0.48 and 0.14–0.52 macro-F1, respectively. All 18 paired bootstrap intervals exclude zero; both nine-test families remain significant after Holm correction (largest adjusted p: 0.040).

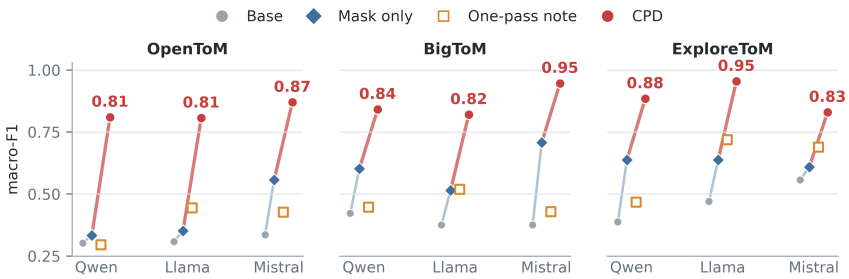

Figure 4: Macro-F1 across three datasets with matched perspective inputs.

4.5 Action Choices and Option-Writing Cues

The fixed Yes/No mapping in the location tasks leaves open whether score contrast helps with other answers. SimpleToM asks what a character will do when an important fact is not known to them; its options are different actions, such as buying a sealed item or reporting a defect. For each story, we also test a version in which a reliable source tells the character that fact before the next action. The dataset's counterfactual action becomes the target in this constructed informed version; it was not independently annotated. We keep $\gamma = 8$, the same three models, and the same two candidate strings. Alongside Mask only, we compare a one-pass perspective instruction, deletion of the first key-fact sentence, and a rewrite from the unaware actor's viewpoint. All use identical candidate scoring. Two no-story diagnostics test option cues: an access-conditioned length rule identified post hoc, and the development-trained option-only classifier described above.

Table 3: Paired SimpleToM action accuracy (928 stories).

	Joint accuracy							
Model	Base	Mask	Delete	Perspective prompt	Rewrite	Access +length	Option only†	CPD
Qwen2.5	.117	.818	.829	.472	.807	.915	**.975**	.934
Llama-3	.117	.637	.661	.468	.645	.915	**.975**	.915
Mistral	.074	.859	.867	.759	.866	.915	**.975**	.952

CPD exceeds the model-based Mask-only and rewritten-story controls. However, the access-plus-length rule reaches 0.915, and the option-only classifier reaches 0.975, above CPD in all three models. The classifier uses labeled development data (†); it is a diagnostic of the test rather than a fair label-free competitor. Development and test story cores are disjoint. On the 895 test pairs where neither action string appears in development, it still scores 0.974; with only length features it scores 0.903. The paired metric is vulnerable because the informed target is defined as the other option: learning which option is typical for an unaware actor, then using supplied access to reverse the choice, suffices. We therefore cannot interpret the action-choice result as shortcut-resistant ToM transfer.

4.6 Sensitivity and Event Specificity

Increasing γ raises FB accuracy while TB is unchanged by the empty mask. At $\gamma = 8$, FB accuracy is $0.672/0.667/0.787$ and TB accuracy is $1.000/1.000/0.984$ for Qwen/Llama/Mistral. An unrelated-span mask does not reproduce CPD (Table 1), supporting contrast with a correctly specified event rather than robustness to mask errors.

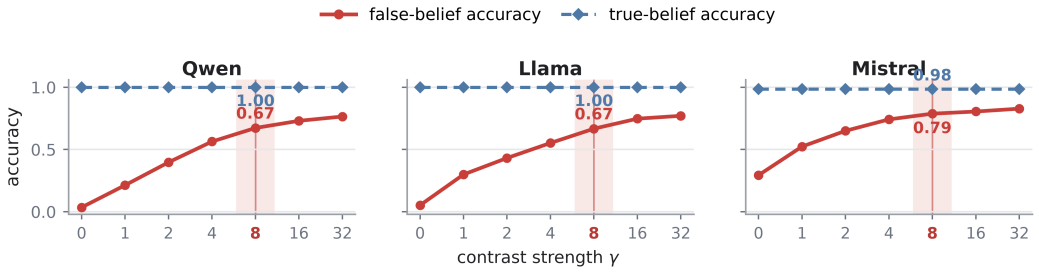

Figure 5: OpenToM accuracy versus contrast strength with supplied perspective information. Shading marks the selected $\gamma = 8$.

5 Limitations

The experiments assume that the relevant event or fact span and the actor's access status are available. They do not establish an autonomous solution to first-order false-belief reasoning. In the evaluated binary location protocol, the supplied variables determine the correct label: the deterministic rule scores 1.000 on every binary-location model–dataset condition and outperforms CPD there. Masking the event sentence also blocks collateral text in that sentence and does not erase correlated evidence elsewhere. CPD still makes errors under supplied information, and its performance varies with question formulation. Forced Yes/No selection is not interchangeable with a native location-choice or free-form answer. The action-choice transfer tests varied answers and paired access states, but its key-fact spans come from dataset metadata, and its informed narratives are constructed disclosures rather than newly annotated natural stories. The supplied access state combined with answer length alone gets 0.909–0.915 joint accuracy. A supervised, development-trained option-only diagnostic scores 0.975 without reading test stories. Variable action strings therefore do not eliminate answer leakage. We test one model size per family. The method requires white-box attention-mask access. The inhibitory-control account motivates the intervention but is not established as a mechanism of the model.

6 Conclusion

Complementary questions reveal substantial formulation sensitivity in first-order false-belief answer selection. When the relevant change and its visibility are supplied, CPD improves over masking alone and a one-pass same-information note across three models and three datasets. The improvement in paired action choice over model-based masking is outweighed by an option-only diagnostic trained on development examples. These are conditional decoding results: a direct rule solves the binary location protocol, and the action test does not establish transfer beyond option-writing cues.

AI use statement

Generative AI tools assisted with manuscript drafting and editing, figure design, code development, and the design and interpretation of the exploratory answer-shortcut audits. The authors are responsible for verifying all AI-assisted analyses and for the final manuscript and claims.

References

Susan A. J. Birch and Paul Bloom. The curse of knowledge in reasoning about false beliefs. *Psychological Science*, 18(5):382–386, 2007. doi: 10.1111/j.1467-9280.2007.01909.x.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Brittle minds, fixable activations: Understanding belief representations in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22521–22543, Suzhou, China,

November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1226.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Jingyuan Deng and Yujiu Yang. MaskCD: Mitigating LVLM hallucinations by image head masked contrastive decoding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18854–18866, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1025.

Kanishk Gandhi, Jan-Philipp Fraenken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 13518–13529. Curran Associates, Inc., 2023. doi: 10.52202/075280-0595.

Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10003–10039, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.531.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023. doi: 10.18653/v1/2023.emnlp-main.751.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/524ff06d5375d76e93d8490e654a81bf-Abstract-Conference.html.

X. Angelo Huang, Emanuele La Malfa, Samuele Marro, Andrea Asperti, Anthony G. Cohn, and Michael J. Wooldridge. A notion of complexity for theory of mind via discrete world models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2964–2983, 2024. doi: 10.18653/v1/2024.findings-emnlp.167.

Eojin Jeon and SangKeun Lee. PICTURE: Enhancing theory-of-mind in large language models by revealing, not hiding, characters' lack of knowledge. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36165–36192, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1674.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70.

Chani Jung, Dongkwan Kim, Jiho Jin, Jiseon Kim, Yeon Seonwoo, Yejin Choi, Alice Oh, and Hyunwoo Kim. Perceptions to beliefs: Exploring precursory inferences for theory of mind in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19794–19809, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1105.

Alan M. Leslie, Tim P. German, and Pamela Polizzi. Belief-desire reasoning as a process of selection. *Cognitive Psychology*, 50(1):45–85, 2005. ISSN 0010-0285. doi: 10.1016/j.cogpsych.2004.06.002.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023a. doi: 10.52202/075280-1797.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10228–10241, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.682/.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13960–13980, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.780.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore theory of mind: program-guided adversarial data generation for theory of mind reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=246rHKUnnf.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.69.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8292–8308, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.451.

Heinz Wimmer and Josef Perner. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128, 1983. ISSN 0010-0277. doi: 10.1016/0010-0277(83)90004-5.

Yuheng Wu, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. DEL-ToM: Inference-time scaling for theory-of-mind reasoning via dynamic epistemic logic. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11383–11397, 2025. doi: 10.18653/v1/2025.emnlp-main.573.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8593–8623, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.466.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62638–62681. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhu24o.html.