

Reasoning Well, Judging Poorly: Theory of Mind Scaffolding Exposes the Limits of LLM Persuasion Judges

Baktash Ansari¹ Mouly Dewan² Afra Mashhadi¹

¹Computing and Software Systems, University of Washington, Bothell, WA, USA

²Information School, University of Washington, Seattle, WA, USA

{baktash, mdewan, mashhadi}@uw.edu

Abstract

LLM-as-judge frameworks are widely used for subjective NLP evaluation, yet their reliability on social-cognition tasks remains unclear. We investigate whether Theory-of-Mind (ToM) scaffolding can help LLM judges evaluate persuasion on Reddit ChangeMyView (CMV). We test four judges (GPT-5, GPT-4.1, Claude Sonnet 4.6, Qwen3-30B-A3B-Instruct-2507) under three conditions that share the same ToM annotations but vary how they are presented: No-ToM, Full-ToM (flat), and our proposed PAToM (Persuasion-Aware ToM), which structures reasoning across five sequential belief-stack turns. Our results show that better reasoning does not lead to better verdicts. Although PAToM improves reasoning coherence by $1.6\times$, semantic coverage by $1.1\text{--}1.8\times$, and our proposed process metric PRISM by $2\text{--}3\times$, all four judges still struggle to reliably distinguish persuasive comments from high-quality non-persuasive ones. Process-level analysis shows that reasoning quality and verdict accuracy are empirically decoupled under this scaffolding, an open challenge for prompt-based approaches to persuasion judging. We release prompts, ToM annotations, the PRISM metric suite, and a 1,000-item benchmark.

1 Introduction

The rise of LLM-as-judge frameworks (Zheng et al., 2023; Panickssery et al., 2024) has made automated evaluation a routine part of NLP pipelines for subjective tasks. These systems are attractive because they are scalable and correlate reasonably well with human judgments on factual and fluency-oriented tasks (Chiang et al., 2023; Kim et al., 2024). However, far less is known about whether LLM judges can evaluate tasks requiring social cognition, the ability to reason about the mental states, beliefs, and intentions of other agents (Ullman, 2023).

Persuasion is a clear example of this challenge because persuasive success cannot be determined from surface text quality alone. Evaluating persuasion requires reasoning about the target audience’s beliefs, attitudes, and likelihood of attitude change rather than only

linguistic quality or fluency. This makes persuasion fundamentally a social-cognitive evaluation task. Theory of Mind (ToM) is the capacity to attribute mental states to oneself and others (Premack and Woodruff, 1978). Developmental studies link ToM to persuasive behavior (Slaughter et al., 2013; McAlister and Cornwell, 2009; Peterson et al., 2018), suggesting that persuasion depends on humans’ ability to model and anticipate the thoughts, intentions, and reactions of others. Thus, without ToM, a judge evaluating persuasion is effectively rating how convincing an argument sounds rather than how likely it is to change someone’s mind.

Despite growing interest in both LLM evaluation reliability and LLM Theory of Mind (Kosinski, 2024; Le et al., 2019; Chen et al., 2024), little work has studied how these areas connect. We ask: how much do LLM judges use Theory of Mind when evaluating persuasion, and does the structure of ToM annotations measurably affect evaluation quality?

To study this, we use Reddit ChangeMyView (CMV) as a real-world persuasion benchmark and compare three prompting conditions that use the same post, comment, rubric, and ToM annotations, differing only in how the information is structured (Section 3). A ToM profile covers six categories: beliefs, desires, intentions, emotions, knowledge, and perspective-taking. Each category contributes a short keyword list and a one-sentence rationale, giving the twelve fields we refer to throughout. We introduce No-ToM (no mental-state information), Full-ToM (all 12 ToM fields in a single flat block), and our proposed PAToM (Persuasion-Aware ToM), which distributes the same fields across five sequential turns guided by first-, second-, and third-order ToM and Social Judgment Theory (Sherif and Hovland, 1961), a mechanism we call sequential trajectory shaping. We evaluate four judges (GPT-5, GPT-4.1, Claude Sonnet 4.6, and Qwen3-30B-A3B-Instruct-2507) on 1,000 stratified CMV items (334 P + 333 H + 333 E) (Section 4.1 for details), auditing reasoning trajectories using ROSCOE (Golovneva et al., 2023), Reasoning-Flow (Zhou et al., 2026), and our proposed Process Reasoning Index via Structured Measures (PRISM), a reference-free metric combining trajectory arc length, directional consistency, step coherence, and ReCEval chain-quality signals (Section 3.5).

We organize our investigation around three research questions, each isolating a possible explanation for why LLM judges might fail at persuasion:

- **RQ1** (Section 5.1): Can LLM Judges detect persuasion from text alone, without any ToM information?
- **RQ2** (Section 5.2): Does providing explicit ToM annotations improve detection, and does the structure of those annotations matter?
- **RQ3** (Section 5.3): Does the richer reasoning produced by structured ToM prompting translate into more accurate verdicts?

The answer to all three points to the same conclusion: under every condition we test, these judges do not reliably detect persuasion.¹ Without ToM scaffolding, the gap between persuasive comments and well-written non-persuasive ones is half a point across all judges. Providing ToM in a structured flat block does not help. PAToM does reshape reasoning substantially (step coherence $\sim 0.41 \rightarrow 0.69$, PRISM: $0.26 \rightarrow 0.64$, covering $1.1\text{--}1.8\times$ more semantic ground), and it yields the largest discrimination gap in every judge. But absolute accuracy moves only modestly: AUC-ROC gains are $+0.03\text{--}+0.07$, and every reasoning metric that improves under PAToM correlates negatively with correct verdicts. The judges reason harder but not more accurately, suggesting a boundary on what this form of ToM scaffolding can achieve for persuasion judging.

2 Related Work

2.1 LLM Judges and Persuasion Evaluation

LLM-as-judge systems (Zheng et al., 2023) are widely used in NLP evaluation. Known failure modes include positional, verbosity, and self-enhancement biases (Wang et al., 2024; Saito et al., 2023; Panickssery et al., 2024), while fine-tuned judges partially improve calibration (Kim et al., 2024; Zhu et al., 2025). Most prior work targets factual or preference-based tasks rather than socially grounded reasoning. Persuasion evaluation differs because judges must reason about the target’s beliefs and likelihood of attitude change (Sherif and Hovland, 1961). ChangeMyView provides real-world persuasion labels through delta awards (Tan et al., 2016), and prior work has studied argument quality (Habernal and Gurevych, 2016) and linguistic features of persuasive success (Yang et al., 2019). More recent work has explored persuasion capabilities in LLMs: Singh et al. (2025) introduced PersuasionBench for evaluating persuasive generation, Rescala et al. (2024) studied whether LLMs can predict audience appeal, and others explored persuasion strategy reasoning (Labruna et al., 2026), safety evaluation (Liu et al., 2025), and conversational persuasion (Karande et al., 2024; Breum et al., 2024). Automated judgment is applied to socially grounded content beyond persuasion as well, for instance in hybrid human–model moderation of live-stream chat (Ansari et al., 2026).

¹Code, data, prompts, and the PRISM metric suite are available at <https://github.com/baktash81/PaTom>.

2.2 Theory of Mind in Language Models

Theory of Mind benchmarks such as ToMi (Le et al., 2019), HI-ToM (Wu et al., 2023), and OpenToM (Xu et al., 2024) show that higher-order ToM reasoning remains a consistent weakness for LLMs (Chen et al., 2024). Prior work also suggests that chain-of-thought prompting alone does not reliably resolve these limitations (Shapira et al., 2024). Related approaches such as CCoToM (Lin et al., 2025) aim to improve ToM benchmark performance through constrained intermediate reasoning. Other work improves LLM reasoning on socially grounded questions by aligning the model with the topic under discussion (Amirizani et al., 2026).

2.3 Linking Persuasion and Theory of Mind

Premack and Woodruff introduced Theory of Mind (ToM) as the capacity to attribute mental states to oneself and others (Premack and Woodruff, 1978). Slaughter et al. (2013) showed that stronger ToM abilities correlate with more sophisticated persuasive arguments, and subsequent work connected persuasion to perspective-taking and mental-state reasoning (McAlister and Cornwell, 2009; Peterson et al., 2018), suggesting that persuasion evaluation inherently requires modeling another agent’s internal state. In language models, however, Amirizani et al. (2024) found persistent gaps between LLM and human ToM capabilities, and little work has examined whether explicit ToM scaffolding improves LLM judge reliability during persuasion evaluation. Our work does not optimize ToM task accuracy itself; instead, we investigate whether structuring identical ToM information changes the reasoning trajectories, calibration, and evaluation behavior of LLM judges in a real-world persuasion setting.

3 Methodology

We treat persuasion evaluation as a scoring and ranking task on Reddit ChangeMyView (CMV). Each item is a (post, comment) pair (x_p, x_c) where the original poster (OP) holds a stated view and the commenter tries to change it. A binary label $y \in \{0, 1\}$ records whether the OP awarded a delta (Δ), the publicly acknowledged signal of view change. Hard-negative comments ($y = 0$) come from the same thread and fall in the top quartile of community scores; they received strong community approval but did not earn a delta, meaning the community found them compelling while the OP did not find them persuasive. Easy-negative comments ($y = 0$) fall in the bottom quartile of community scores and serve as a lower baseline.

An LLM judge $J(x_p, x_c)$ returns a persuasiveness score $s \in \{1, \dots, 10\}$ and a free-form reasoning trace r . We compare three conditions that differ only in how Theory-of-Mind (ToM) information is presented to the judge: No-ToM, Full-ToM, and our proposed Persuasion-Aware Theory-of-Mind (PAToM). All three use the same scoring rubric. No-ToM receives no mental-state information at all; Full-ToM and PAToM

receive the same twelve ToM field-value pairs and differ only in how those fields are presented, which lets us isolate reasoning order as the key variable. Verbatim prompts for all three conditions are provided in Appendix D.

3.1 ToM Profile Extraction

We extract six ToM categories from both the post and the comment separately: beliefs, desires, intentions, emotions, knowledge, and perspective-taking (Premack and Woodruff, 1978; Baron-Cohen et al., 1985). Table 1 presents our working definitions.

Table 1: ToM categories and definitions.

Category	Definition
Beliefs	What the author thinks is true/false
Desires	What the author wants or values
Intentions	What the author plans or advocates
Emotions	Emotional states or appeals
Knowledge	What the author knows or assumes
Perspective-taking	Considering other viewpoints

Each category produces a short keyword list (values) and a one-sentence rationale (content). Together, the twelve fields form the ToM profile $\mathcal{T}(x)$ for each party. Extraction is performed once per text using Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025) (4-bit AWQ) in two passes (Beliefs/Desires/Intentions, then Emotions/Knowledge/Perspective-Taking); full extraction prompts are in Appendix C. The extractor is separate from the three closed judges; Qwen is the self-extraction exception described in Section 4.2. Wherever ToM information is provided it comes from the same $\mathcal{T}$, so Full-ToM and PAToM differ in presentation rather than content; No-ToM receives no profile.

Table 2 shows one extracted pair, for a post arguing that colonization was a net positive and the comment that earned its delta. The contrast that matters for persuasion is in the last row: the OP’s perspective-taking is characterized by dismissal and overgeneralization, the commenter’s by explicit engagement with the viewpoint the OP does not consider. PAToM is built to make a judge reason over exactly this kind of asymmetry rather than over surface argument quality.

3.2 No-ToM: Zero-Shot Chain of Thought

The No-ToM condition follows the standard LLM-as-judge format (Zheng et al., 2023): the model sees only (x_p, x_c) and produces a zero-shot chain-of-thought followed by a 1–10 score. No ToM profiles are provided.

3.3 Full-ToM: Flat Context Block

The Full-ToM condition gives the judge the entire ToM profile $\mathcal{T}(x_p) \cup \mathcal{T}(x_c)$ all at once in a single block prepended to the scoring instruction, producing a score and reasoning in one generation step. Because the model is not required to work through the belief stack before giving a verdict, the profiles act as extra context rather

Table 2: Extracted ToM values for one (post, comment) pair (“CMV: Colonisation overall was a net positive for the world”), abbreviated. Each field also carries a one-sentence content rationale, omitted here for space; the full profile format appears in Appendix C.

Category	Post author (OP)	Comment author
Beliefs	Progress, civilization, order	Justice, accuracy, historical truth
Desires	Gratitude, recognition	Equity, recognition, truth
Intentions	Persuasion, reorientation	Education, correction, accountability
Emotions	Conviction, gratitude, confidence	Anger, disgust, empathy
Knowledge	Certainty, authority, expertise	Evidence-based, critical, informed
Perspective-taking	Dismissal, assumption, overgeneralization	Empathy, rejection of superiority, historical inclusion

than structured reasoning guides, producing trajectories similar to No-ToM (Section 5).

3.4 PAToM: Sequential Belief-Stack Prompting

PAToM (Persuasion-Aware Theory of Mind) uses the same twelve ToM fields as Full-ToM but spreads them across five sequential turns. Each turn introduces a relevant subset of the profiles and asks the model a step-specific question. The model’s answer is added to the conversation context before the next turn, requiring the judge to build an explicit intermediate understanding at each level of the belief stack before moving on. The step order follows the persuasion-theoretic sequence target $\rightarrow$ persuader $\rightarrow$ gap $\rightarrow$ social management $\rightarrow$ verdict:

1. **1st-order OP beliefs** (post beliefs, desires, intentions, knowledge, perspective-taking). The judge sees the post and the OP’s full profile (except emotions) and identifies the OP’s position and key assumptions. The comment is withheld so the judge forms an independent picture of the target’s mental state (Wellman, 1990; Baron-Cohen et al., 1985).
2. **1st-order commenter beliefs** (comment beliefs, desires, intentions). The comment text is introduced with the commenter’s own stance. Perspective-taking and knowledge are held back for Step 3, since they represent the commenter’s model of the OP rather than the commenter’s own position.
3. **2nd-order targeting: latitude of acceptance** (comment perspective-taking, knowledge). The commenter’s perspective-taking and knowledge are introduced as evidence of how well the commenter understood the OP. The judge evaluates whether the argument falls within the OP’s latitude of acceptance (Sherif and Hovland, 1961), the zone close enough to seem credible but far enough to shift the OP’s view.
4. **3rd-order social management** (post emotions, com-

ment emotions). The judge assesses whether the commenter signals good faith, manages credibility, and matches the OP’s emotional tone. This captures social and emotional conditions for persuasion that are distinct from argument quality (Tan et al., 2016).

5. **Verdict** (no new fields). The judge integrates all four intermediate representations to rate the comment as a persuasive act toward this specific OP (1–10 with brief reasoning).

All three conditions receive identical text inputs, and the two that are given ToM information receive the same twelve fields; only the presentation structure varies. PAToM is the only condition that enforces a reasoning order, builds intermediate representations, and uses ToM annotations as step controllers rather than as passive context.

3.5 Metrics

We evaluate judges along two complementary axes: score-level metrics assess whether the final numeric verdict separates comment types; reasoning-level metrics audit the internal reasoning trajectory regardless of the final score.

Score-level metrics. Judges receive a fixed 1–10 scoring rubric in the system message with qualitative anchor bands: 1–2 (irrelevant), 3–4 (no new framing), 5–6 (partial engagement), 7–8 (well-targeted), 9–10 (exceptionally persuasive). We treat scores of ≥ 7 as the delta threshold, corresponding to arguments strong enough to plausibly earn a delta; this threshold anchors all calibration analyses below. We chose 7 because the rubric anchors 7–8 as “well-targeted”, so a score of at least 7 marks an argument plausibly strong enough to earn a delta. Other choices, including a binary formulation, were tried during development and led to the same conclusions. The threshold affects only the FPR and FNR columns; our primary measures, the discrimination gap, AUC-ROC, and Cohen’s d , are threshold-free. We report:

- Mean predicted score per comment type (Persuasive, Hard-Negative, Easy-Negative) per condition and model.
- Discrimination gap: $\text{mean}(P) - \text{mean}(H+E)$, the primary measure of whether the judge separates persuasive from non-persuasive comments.
- Score ordering: whether $\bar{s}_E < \bar{s}_H < \bar{s}_P$ holds, the necessary condition for a calibrated judge.
- AUC-ROC for the P-vs-H binary task (the hardest cut, since hard negatives are community-approved and surface-quality comments).
- Cohen’s d with two-sided Mann–Whitney U for pairwise condition comparisons.
- False-positive rate (negatives scored ≥ 7) and false-negative rate (persuasives scored < 7) at the delta threshold.

Reasoning-level metrics. We propose a metric suite that audits the reasoning trajectory produced by each

judge condition. Let $\mathcal{C} = (r_1, \dots, r_T)$ be the sequence of reasoning steps, $\mathbf{e}_t = \text{enc}(r_t) \in \mathbb{R}^d$ their sentence embeddings (Reimers and Gurevych, 2019), and $\mathbf{e}_x$ the embedding of the source text (post and comment concatenated). Steps are defined as the five PAToM controller outputs for the chain condition, and sentence-split segments of the free-form response for flat conditions.

ROSCOE. (Golovneva et al., 2023) is a reference-free framework for evaluating multi-step reasoning chains at the step level. It produces scores that capture whether consecutive steps are locally coherent, whether each step stays grounded in the source text, and whether the chain avoids redundant repetition across steps. We adopt three of its scores:

$$\text{SC} = \frac{1}{T-1} \sum_{t=2}^T \cos(\mathbf{e}_{t-1}, \mathbf{e}_t) \quad (1)$$

$$\text{SG} = \frac{1}{T} \sum_{t=1}^T \cos(\mathbf{e}_t, \mathbf{e}_x) \quad (2)$$

$$\text{SR} = \binom{T}{2}^{-1} \sum_{i < j} \cos(\mathbf{e}_i, \mathbf{e}_j) \quad (3)$$

Step Coherence (SC) captures how smoothly the chain transitions between consecutive steps; grounding SG measures per-step alignment with the source text; redundancy SR measures average pairwise similarity (high values indicate uniform, less exploratory chains).

ReCEval. (Prasad et al., 2023) is a reference-free framework that evaluates the quality of a reasoning chain’s final conclusion. It measures how well the final step is grounded in the source text and how consistent it is with the rest of the chain, both without requiring a gold reference. We use two of its signals:

$$\text{CG} = \cos(\mathbf{e}_T, \mathbf{e}_x) \quad (4)$$

$$\text{CC} = \cos\left(\mathbf{e}_T, \frac{1}{T} \sum_{t=1}^T \mathbf{e}_t\right) \quad (5)$$

Conclusion grounding CG checks whether the final verdict is anchored in the source text; chain consistency CC checks whether the verdict follows naturally from the preceding chain. We additionally track informativeness (per-step novelty relative to all prior steps) as a diagnostic comparator but exclude it from our combined index, since it inverts for multi-step chains where each conditioned step is penalized as redundant relative to prior ones (Appendix B).

Trajectory geometry. Reasoning Flow (Zhou et al., 2026) models a reasoning chain as a geometric trajectory in embedding space, characterizing inference by path-level properties rather than step-level content.² It distinguishes genuine multi-step reasoning from verbose repetition by measuring how far the chain travels, how directionally consistent its moves are, and how well the final step departs from the prior context. We adopt

²Reasoning Flow is concurrent work; we adopt only its standard geometric measures (L , ρ , δ_f), which are well-defined independently of the original paper.

three measures:

$$L = \sum_{t=2}^T \|\mathbf{e}_t - \mathbf{e}_{t-1}\|_2 \quad (6)$$

$$\rho = \frac{1}{T-2} \sum_{t=2}^{T-1} \cos(\mathbf{d}_t, \mathbf{d}_{t+1}), \mathbf{d}_t = \mathbf{e}_t - \mathbf{e}_{t-1} \quad (7)$$

$$\delta_f = \left\| \mathbf{e}_T - \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{e}_t \right\|_2 \quad (8)$$

Arc length L measures total semantic distance covered; step-displacement cosine ρ captures directional consistency between successive moves (negative values indicate zig-zagging rather than coherent progress); final deviation δ_f measures how far the final verdict sits from the centroid of earlier steps.

PRISM. As the ablation shows (Appendix B), no single metric family is sufficient on its own: ROSCOE captures step quality but not path shape; trajectory geometry captures exploration but not verdict quality; ReCEval captures conclusion grounding but inverts for informativeness. We therefore introduce the Process Reasoning Index via Structured Measures (PRISM), which combines signals from all three frameworks:

$$\text{PRISM} = L \cdot (-\rho) \cdot \text{SC} \cdot \sqrt{\text{CG} \cdot \text{CC}}, \quad (9)$$

PRISM rewards (i) trajectory length L (genuine semantic exploration), (ii) directional deliberation $(-\rho)$ (penalizing random walks), (iii) local step coherence SC , and (iv) verdict quality $\sqrt{\text{CG} \cdot \text{CC}}$, the geometric mean of the two ReCEval signals that consistently improve under structured prompting. The geometric mean ensures neither signal alone can dominate and collapses the index whenever the final verdict is ungrounded or off-chain. PRISM is reference-free and applies to all three conditions.

By design, PRISM measures reasoning process structure: whether the chain is coherent, grounded, exploratory, and internally consistent. It is not designed to predict whether the final verdict is correct; that is a separate empirical question we address in Appendix A. The distinction matters because, as we show, reasoning quality and verdict accuracy are empirically decoupled on this task.

Verdict correctness. To correlate reasoning metrics with verdict accuracy, we define a continuous correctness signal: $c = s$ for persuasive comments (higher score is more correct) and $c = 10 - s$ for negatives (lower score is more correct), where s is the judge’s 1–10 score. This encoding maps all comment types onto a common scale where higher values indicate more accurate verdicts.

4 Experimental Setup

4.1 Datasets

Reddit ChangeMyView (CMV). Our dataset is the ChangeMyView subreddit, a widely used benchmark for persuasion research (Tan et al., 2016). We collected all posts marked with a “Delta(s) from OP” label between 2013 and 2024. After removing deleted or moderated content and threads with broken comment

trees, we extract three comment types for each post: Persuasive (P), comments for which the OP explicitly awarded a delta; Hard Negative (H), comments from the same thread in the top quartile of community score but without a delta (in 91% of posts, at least one hard negative has a higher community score than the persuasive comment, meaning community approval does not reliably predict persuasive success); and Easy Negative (E), bottom-quartile comments that serve as an easy baseline. From posts that have at least one P comment and balanced H/E coverage, we sample a stratified evaluation set of 1,000 (post, comment) pairs (334 P + 333 H + 333 E), stratified by comment type rather than by post: the 1,000 items are drawn from 794 distinct posts, and only 8 posts contribute one comment of each type, so P, H, and E items are not matched within a thread. Subset statistics (median comment lengths: P 137 words, H 89 words, E 56 words; mean community scores: P 18.8, H 40.1, E −0.2) confirm that hard negatives are objectively well-argued.

4.2 Models

We evaluate four LLM judges spanning three model families and two deployment regimes (closed and open source): GPT-5 (OpenAI, 2025a) (gpt-5-reasoning, OpenAI Chat Completions API, 2025-08 snapshot), GPT-4.1 (OpenAI, 2025b) (same API, same snapshot), Claude Sonnet 4.6 (Anthropic, 2026) (Anthropic Messages API, claude-sonnet-4-6 snapshot), and Qwen3-30B-A3B-Instruct-2507 (Qwen Team, 2025) (open-source MoE, served locally via vLLM (Kwon et al., 2023); abbreviated as Qwen in tables). All four judges share the same system prompt and scoring rubric across conditions; only the user turns differ in ToM exposure (Section 3). Decoding is temperature 0.2 with `max_new_tokens` = 768 per turn. All four models are evaluated on the same 1,000-item stratified sample.

ToM human annotation. ToM profiles are extracted as described in Section 3.1, with truncation flags passed to downstream evaluators.

One caveat follows from that design: Qwen3-30B-A3B-Instruct-2507 is both the extraction model and one of the four judges, so for that judge alone the scaffolding is self-produced. The other three judges never see profiles they wrote, and every finding holds for them, with PAToM discrimination gaps of +1.00, +1.14, and +1.13 against Qwen’s +1.03. Extraction is also a single pass, not the iterative self-refinement in which bias amplification has been reported.

To validate extraction quality, two annotators labeled a 30-pair sample on delta validity and faithfulness of each ToM category. On delta validity, at least one annotator found the delta comment persuasive for 93% of items and both did for 70%, with the two giving the same answer on 77%. Mean per-category inter-annotator agreement on the ToM fields is 82.2%, matching prior reports on automatic ToM extraction (Xu et al.,

2024). Per-category breakdowns are in Table 9 (Appendix C); the full annotation procedure and question definitions are described in Appendix F.

We also asked whether the profiles depend on which model wrote them. Using the same prompts, we extracted ToM again for the same 30 comments with a different model, Claude Sonnet 4.6. For 83% of fields, the Claude profile for a comment is closer to the Qwen profile for that same comment than to any other comment in the sample, against 3.3% by chance, and the two models rank the six categories in the same order our annotators did (Appendix C). This shows only that the profiles do not depend on one model’s habits, not that they are correct: two models can share a mistake a person would catch.

Reasoning-step human annotation. The study above checks the ToM profiles the judge is given. A second study checks the reasoning the judge produces from them. Three annotators scored every step of 50 PAToM chains (29 GPT-5, 21 Claude Sonnet 4.6) as correct, partly correct, or wrong, agreeing with each other on 80% of Step 3 ratings and up to 96% on Step 2. Scores are near perfect for the early steps, then fall at Step 3, where the judge has to decide whether the argument lands close enough to the OP’s current view to be persuasive: 14% of Step-3 ratings are below full marks against 2–3% at Steps 1, 2, and 4, and the judge picks the right category on 86% of chains. The most common complaint is that the judge will not choose: it names two or three possible answers and moves on, and Step 5 carries that hesitation into the verdict (Appendix G).

4.3 Conditions

We evaluate three conditions that isolate the contribution of access to ToM information from structured reasoning over it:

NOTOM-COT No-ToM with chain-of-thought instruction; standard LLM-as-judge recipe (Section 3.2).

FULLTOM-COT Full-ToM with chain-of-thought instruction; all twelve ToM fields presented simultaneously, single generation step (Section 3.3).

PAToM Sequential five-step belief-stack prompting (Section 3.4).

FULLTOM-COT is the direct comparison for PAToM: both include explicit step-by-step reasoning, so any difference in reasoning quality cannot be explained by chain-of-thought alone. The key variable is whether ToM fields are given one at a time across turns (PAToM) or all at once (FULLTOM-COT).

4.4 Evaluation

Full definitions of score-level and reasoning-level metrics are given in Section 3.5. Here we summarize the statistical conventions used throughout the paper.

Reasoning texts are split into steps (one per PAToM controller turn; sentence-split for flat conditions)

and embedded with allMiniLML6v2 (Reimers and Gurevych, 2019). PCA projections for trajectory plots are fit jointly across all conditions to make the geometries directly comparable.

Statistics. All effect sizes use sample variance (ddof = 1); all significance tests use two-sided Mann–Whitney U . We mark $^+p < 0.10$, $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. For the multi-metric tables we report uncorrected p and rely on the consistency of effect-size directions across metrics for interpretation, following standard ROSCOE practice. For the key claim that PRISM correlates negatively with verdict correctness, we additionally verified significance under Bonferroni correction across all ten metrics ($\alpha=0.005$); the correlation remains significant in 6 of 12 model–condition cells.

5 Results

We report findings for four judges (GPT-5, GPT-4.1, Claude Sonnet 4.6, Qwen3-30B-A3B-Instruct-2507) on the 1,000-item stratified CMV sample (334 P + 333 H + 333 E), organized by research question.

5.1 RQ1: Without Theory of Mind, LLMs Cannot Reliably Detect Persuasion

Table 3 shows mean scores by comment type. Under NOTOM-COT, the persuasive-vs-hard-negative gap is uniformly small: +0.47 (GPT-5), +0.52 (GPT-4.1), +0.30 (Claude), −0.00 (Qwen). All judges score hard negatives within half a point of genuinely persuasive comments. The required ordering $\bar{s}_E < \bar{s}_H < \bar{s}_P$ holds for all closed models; Qwen under NOTOM-COT is the sole exception (6.69 vs. 6.69). Without ToM scaffolding, no judge can reliably tell apart comments that actually changed the OP’s mind from well-written comments that did not. Surface text quality alone is not enough to detect persuasion.

5.2 RQ2: Theory of Mind Information Improves Discrimination but Calibration Remains Low

The weak discrimination in RQ1 could stem from two causes: the judge either lacks access to the relevant mental-state information, or it has the information but cannot reason over it effectively. FULLTOM-COT tests the first explanation and PAToM tests the second.

Adding the full ToM profile in a flat block produces discrimination changes of −0.07 (GPT-5), −0.02 (GPT-4.1), +0.09 (Claude), +0.22 (Qwen) relative to No-ToM, showing that access to ToM information alone has little effect. PAToM, by contrast, scores hard negatives significantly lower than FULLTOM-COT in every judge ($p < 0.01$ or better; Table 3), while persuasive scores stay roughly flat, yielding the largest discrimination gap in every judge (+1.00, +1.14, +1.13, +1.03). This gain comes entirely from increased skepticism (hard-negative suppression) rather than from better recognition of persuasive comments.

Table 3: Mean judge scores and calibration metrics ($n = 1,000$ items per model). $\text{DISC} = \text{mean}(P) - \text{mean}(H+E)$. $\text{AUC} = \text{AUC-ROC}$ for P -vs- H ; $d = \text{Cohen's } d$ (P vs. H); $\text{FPR} = \text{negatives scored } \geq 7$; $\text{FNR} = \text{persuasives scored } < 7$. Superscripts on PAToM cells: significance vs. FULLTOM-COT (Mann–Whitney U). $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Best value per column per model bolded.

Model	Condition	Score-level				Calibration (≥ 7)			
		P	H	E	Disc	AUC	d	FPR	FNR
GPT-5	NOTOM-COT	5.18	4.71	3.75	+0.95	0.576	0.265	0.126	0.722
	FULLTOM-COT	5.16	4.76	3.79	+0.88	0.567	0.234	0.111	0.731
	PAToM	5.03	4.40 **	3.66	+1.00	0.609	0.397	0.066	0.835
GPT-4.1	NOTOM-COT	5.39	4.87	3.90	+1.02	0.581	0.289	0.156	0.680
	FULLTOM-COT	5.49	4.90	4.08	+1.00	0.588	0.323	0.173	0.626
	PAToM	5.28	4.41 ***	3.87	+1.14	0.634	0.485	0.147	0.710
Claude 4.6	NOTOM-COT	4.47	4.17	3.13	+0.82	0.550	0.162	0.060	0.862
	FULLTOM-COT	4.74	4.30	3.35	+0.91	0.572	0.239	0.084	0.796
	PAToM	4.41**	3.69 ***	2.88 ***	+1.13	0.617	0.398	0.054	0.859
Qwen	NOTOM-COT	6.69	6.69	5.60	+0.54	0.504	−0.001	0.530	0.401
	FULLTOM-COT	7.99	7.69	6.78	+0.76	0.516	0.137	0.736	0.150
	PAToM	6.76	6.10 ***	5.35	+1.03	0.566	0.234	0.461	0.380

However, absolute calibration remains low (Table 3, right panel). AUC-ROC ranges from 0.50 to 0.63; Cohen’s d improves under PAToM for all models but stays small-to-moderate (0.23–0.49). FNR under PAToM is 84% (GPT-5), 71% (GPT-4.1), 86% (Claude), 38% (Qwen): most genuinely persuasive comments remain scored below 7. In short, PAToM makes judges more skeptical, which widens the discrimination gap, but absolute detection accuracy remains limited. A worked case study in Appendix E illustrates this pattern on a single post: all judges correctly rank $P > H > E$, yet even on this well-separated item PAToM lowers persuasive scores rather than raising them.

5.3 RQ3: Structured ToM Reshapes Reasoning Without Proportional Verdict Gains

If PAToM only modestly improves discrimination (RQ2), what does the sequential belief-stack actually change? Table 4 reports reasoning metrics across all four judges. Against FULLTOM-COT, PAToM leads on every metric but redundancy and informativeness, which both move the wrong way (Appendix B): step coherence rises from 0.42–0.48 to 0.68–0.73, arc length from 2.0–3.0 to 2.9–3.2, and PRISM from 0.22–0.33 to 0.58–0.66 (2.0–2.9 $\times$). Yet despite these 2–3 $\times$ reasoning gains, AUC-ROC increases by only +0.03–+0.07.

The reasoning trajectory reveals a clear disconnect between process quality and verdict accuracy. Figure 1 isolates items where both FULLTOM-COT and PAToM gave the same wrong score (< 7). Across all 176 agreed-wrong items, PAToM covers significantly more semantic ground ($p < 10^{-24}$, Mann–Whitney U), confirming that reasoning depth is a property of the prompting structure, not of verdict correctness. Appendix I presents two of these chains in full. Table 5 makes this explicit: Spear-

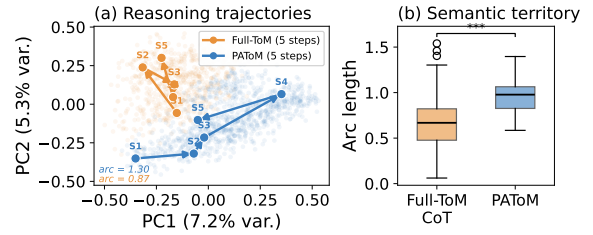


Figure 1: Score agreement masks reasoning divergence. Agreed-wrong GPT-4.1 items (persuasive, scored < 7 by both conditions). (a) PCA trajectories for a single item; both conditions produce five steps and the same score (5), yet PAToM (blue) covers more ground than FULLTOM-COT (orange). (b) Arc-length distributions across all 176 agreed-wrong items; $p < 10^{-24}$ (Mann–Whitney U).

man correlations between reasoning metrics and verdict correctness under PAToM show that arc length, final deviation, and informativeness are the only positive predictors, while all ROSCOE and remaining ReCEval signals are consistently negative. PRISM itself is negative across all four judges (−0.05 to −0.13). A per-family ablation (Appendix B) confirms that no single metric family is sufficient on its own; only PRISM remains stable across subsets and model families ($d = +2.55$ *** on all persuasive items, $d = +2.51$ *** on agreed-wrong items; four-model means).

In summary, PAToM improves PRISM by 2.0–2.9 $\times$ yet AUC-ROC gains only +0.03–+0.07, and PRISM correlates negatively with correctness. The judges reason better but do not judge better; we interpret this disconnect in Section 6.

Table 4: Mean reasoning metrics across four judges. N/F/P denote NOTOM-COT/FULLTOM-COT/PAToM; arrows indicate direction and bold marks the best value. Means exclude undefined values (Qwen PRISM $n=518/869/1000$).

Family	Metric	GPT-5			GPT-4.1			Claude Sonnet			Qwen		
		N	F	P	N	F	P	N	F	P	N	F	P
ROSCOE	step_coh. $\uparrow$	0.409	0.419	0.682	0.436	0.446	0.711	0.415	0.437	0.690	0.468	0.476	0.725
	grounding $\uparrow$	0.465	0.472	0.626	0.440	0.453	0.622	0.438	0.458	0.587	0.551	0.529	0.667
	redundancy $\downarrow$	0.407	0.409	0.634	0.425	0.441	0.645	0.416	0.440	0.657	0.466	0.476	0.667
Trajectory	arc_length $\uparrow$	2.497	2.124	3.150	2.643	2.679	3.001	2.429	3.031	3.120	1.615	1.980	2.928
	disp. cos. $\uparrow$	-0.489	-0.469	-0.445	-0.486	-0.505	-0.402	-0.500	-0.507	-0.467	-0.484	-0.497	-0.417
	final dev. $\downarrow$	0.954	0.977	0.667	0.935	0.905	0.667	0.942	0.894	0.640	0.965	0.919	0.628
ReCEval	informative. $\uparrow$	0.698	0.711	0.436	0.664	0.653	0.417	0.687	0.639	0.429	0.714	0.671	0.410
	concl. grnd. $\uparrow$	0.365	0.350	0.537	0.265	0.291	0.538	0.331	0.359	0.511	0.486	0.421	0.594
	chain_cons. $\uparrow$	0.756	0.764	0.841	0.744	0.759	0.840	0.759	0.752	0.856	0.815	0.798	0.861
Combined	PRISM $\uparrow$	0.257	0.218	0.638	0.235	0.269	0.577	0.246	0.331	0.658	0.261	0.275	0.634

Table 5: Spearman ρ between each reasoning metric and verdict correctness under PAToM, per judge. Full correlation analysis including flat conditions is in Appendix A. $^+p<0.10$, $^*p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

Family	Metric	GPT-5	GPT-4.1	Claude	Qwen
ROSCOE	step_coh.	-0.119***	-0.152***	-0.139***	-0.124***
	grounding	-0.086**	-0.155***	-0.060 $^+$	-0.120***
	redundancy	-0.149***	-0.206***	-0.157***	-0.139***
ReCEval	informative.	+0.132***	+0.167***	+0.148***	+0.122***
	concl. grnd.	-0.108***	-0.157***	-0.092**	-0.088**
	chain_cons.	-0.125***	-0.175***	-0.130***	-0.082**
Trajectory	arc_length	+0.118***	+0.149***	+0.137***	+0.123***
	disp. cos.	-0.017	+0.064*	-0.020	+0.034
	final_dev.	+0.124***	+0.174***	+0.130***	+0.083**
Combined	PRISM	-0.087**	-0.125***	-0.074*	-0.048

5.4 Generalization to a Second Dataset

On CMV, we know an argument worked because the OP said so. To see whether our findings hold when success is measured by what people actually did, we replicated the core comparison on Persuasion for Good (Wang et al., 2019), a set of conversations in which one person asks another to donate to charity and the record shows how much money changed hands. We built 2,180 pairs (545 P + 1,090 H + 545 E) the same way we built the CMV set and ran all three conditions on Qwen, GPT-4.1, and GPT-5 (Appendix H).

The main finding carries over. PAToM again improves the reasoning far more than the flat conditions do, lifting PRISM to 0.54–0.63 from 0.14–0.22 and step coherence from about 0.37 to 0.67. The verdicts again barely move: AUC stays between 0.56 and 0.62 in every condition, and judges that score higher on PRISM are again slightly more likely to be wrong (ρ from -0.073 to -0.219, all $p < 0.001$). Holding for one open model and two frontier models on a behaviorally grounded label, this is our strongest evidence that better reasoning and better verdicts come apart for reasons not particular to one model or setting.

The two ToM conditions rank differently across the two datasets. On CMV, PAToM gives the best persuasive-versus-hard-negative AUC for every judge; on Persuasion for Good, FULLTOM-COT does, and

PAToM mostly just lowers scores. Which of the two ToM conditions produces better verdicts is therefore not stable across datasets, which is why we describe PAToM as changing how a judge reasons rather than as a reliable way to make it more accurate.

6 Discussion

The results show a consistent pattern across four judges and three model families: structured Theory-of-Mind prompting reshapes how LLM judges reason about persuasion without making them judge it accurately.

6.1 Surface Quality as the Default Evaluation Dimension

Without ToM scaffolding, all four judges score hard negatives within ~ 0.5 points of persuasive comments (Section 5.1). Hard negatives are, by construction, community-approved arguments that the OP did not find persuasive: in 91% of posts at least one outscored the delta-earning comment. A judge that treats “well-argued” as “persuasive” will tie or reverse the P-vs-H ranking, which is what we observe. Adding the full profile in a flat block does not shift this default: discrimination changes by -0.07 for GPT-5, -0.02 for GPT-4.1, +0.09 for Claude, and +0.22 for Qwen. The judges’ default dimension is surface argument quality, and access to ToM information without structured reasoning collapses to the No-ToM baseline.

6.2 Why Structured ToM Reasoning Fails

In human development, persuasive ability depends on ToM capacity (Slaughter et al., 2013; McAlister and Cornwell, 2009; Peterson et al., 2018), and the relationship is constructive: a child who can represent another’s beliefs uses that representation to choose arguments within the target’s latitude of acceptance (Sherif and Hovland, 1961; Baron-Cohen et al., 1985). PAToM was designed to reproduce that chain, and the reasoning metrics confirm it succeeds at structuring it ($d=1.1$ – 3.9 across the ROSCOE and ReCEval metrics; Table 7).

Verdict accuracy does not follow. The step-level signals PRISM rewards each correlate negatively with cor-

rect verdicts ($\rho = -0.08$ to -0.18), while informativeness ($\rho \approx +0.14$) decreases under PAToM (Appendix A). The models appear to follow ToM-flavored instructions, producing text that reads like belief-stack reasoning, without maintaining a separate representation of another agent’s mental state. Wellman (1990) showed that children begin describing others’ beliefs around age three but cannot use them to predict actions until four or five, and Wu et al. (2023) report an analogous gap in LLMs. The agreed-wrong analysis shows this directly: on the 176 items where both conditions give the same wrong verdict, PAToM covers significantly more semantic ground ($p < 10^{-24}$) and still arrives at the same wrong answer. Appendix I works through two of these chains in full: in both, Step 3 correctly places the argument inside the OP’s latitude of acceptance, and Step 5 then marks the comment down for tone and lack of concession, discarding the judgment the scaffold was built to produce.

6.3 Hard-Negative Suppression and the Latitude of Acceptance

PAToM’s discrimination gain comes from suppressing hard negatives rather than from recognizing persuasive comments: hard-negative scores drop in every judge ($p < 0.01$) while persuasive scores stay flat or fall as well, and 71–86% of genuinely persuasive comments from the closed models remain scored below 7. The belief-stack makes judges more skeptical, not more accurate.

This asymmetry fits Social Judgment Theory. Step 3 asks the judge whether the argument falls within the OP’s latitude of acceptance; forced to articulate a mismatch, it correctly lowers hard-negative scores. The symmetrical task, identifying arguments that fall within the zone and succeed, appears to require a deeper model of attitude change. Our step-level annotation supports this: Step 3 is where errors concentrate, and the judge answers “within latitude” in 111 of 150 cases, rarely reaching for the categories that would mark an argument as likely to fail (Appendix G).

6.4 Implications for LLM-as-Judge Evaluation

Because the judges’ default mode evaluates argument quality rather than persuasive fit, reasoning coherently about the wrong dimension produces more confidently wrong verdicts. Process quality therefore cannot be assumed to track outcome accuracy on social-cognition tasks, and score-level metrics alone cannot separate a judge that reasons well and judges correctly from one that reasons well and judges wrongly. We therefore recommend reporting both score-level and reasoning-level metrics for any LLM-as-judge pipeline on subjective tasks. Chain-of-thought (Wei et al., 2022) does not by itself elicit social-cognitive reasoning even when ToM information is available (Shapira et al., 2024): both flat conditions produce trajectories much like No-ToM. That the miscalibration persists across four judges, three model families, and two deployment regimes suggests the limitation is not model-specific, and that progress

may require fine-tuning on delta-labeled data, retrieval of similar outcomes, or hybrid systems combining LLM reasoning with behavioral signals.

7 Conclusion

We evaluate four judges under three ToM conditions on 1,000 CMV items and replicate the core comparison on Persuasion for Good. On CMV, PAToM improves reasoning quality by 2–3 $\times$ and yields the largest discrimination gap in every judge, mainly by suppressing hard-negative scores; its AUC remains 0.57–0.63. Across both datasets, prompt restructuring changes reasoning trajectories without reliably improving verdicts, and the structural signals rewarded by PRISM correlate negatively with correctness (Appendix A). Improving verdicts will likely require supervision or behavioral signals beyond prompt structure; subjective LLM evaluation should report process and outcome metrics together.

Limitations

We study two persuasion settings, online argument and charity dialogue; other settings such as debate or negotiation remain untested. For closed-source judges we can only analyze the reasoning they surface, not their internal states, though every finding also holds for the open-weight Qwen judge.

ToM profiles come from one extraction model and are held fixed across conditions, so they cannot account for any difference between conditions. A second model reproduces them closely (Appendix C), but agreement between two models is not the same as validation by people: our annotators accepted or rejected the model’s output rather than writing their own, which cannot detect a mistake both would make, so the agreement figures in Section 4.2 are best read as an upper bound.

The delta is a binary award and does not capture partial changes of mind. It also reflects more than the argument itself, though our design fixes one such factor, since every negative comment was posted before the delta-winning one and was therefore already visible to the OP. Finally, we compare judges with one another rather than with people: we have no study of humans predicting which comment an OP will reward, so we cannot say how much of the remaining gap is model limitation and how much is task difficulty. For reference, aggregate human approval does not track deltas either, ranking persuasive comments against hard negatives at AUC=0.27, below chance. We therefore claim only that these judges do not exceed a surface-quality heuristic under the conditions we test; the reasoning–verdict result at the center of the paper is measured within model and across conditions, and does not depend on a human baseline. Collecting one would calibrate the absolute numbers and is the clearest next step.

References

- Maryam Amirizani, Baktash Ansari, Chirag Shah, and Afra Mashhadi. 2026. [TopicTune: A topic alignment approach to improve LLMs reasoning on social questions](#). In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 1–11. ACM.
- Maryam Amirizani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. [Can LLMs reason like humans? assessing theory of mind reasoning in LLMs for open-ended questions](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 34–44. ACM.
- Baktash Ansari, Elias Martin, and Afra Mashhadi. 2026. [ToxiTwitch: Toward emote-aware hybrid moderation for live streaming platforms](#). In *Companion Proceedings of the ACM Web Conference 2026*, pages 425–434. ACM.
- Anthropic. 2026. System card: Claude Sonnet 4.6. <https://www.anthropic.com/claude-sonnet-4-6-system-card>. Accessed 2026-08-29.
- Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. 1985. [Does the autistic child have a “theory of mind”?](#) *Cognition*, 21(1):37–46.
- Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. [The persuasive power of large language models](#). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. [ToMBench: Benchmarking theory of mind in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. <https://lmsys.org/blog/2023-03-30-vicuna/>. Blog post.
- Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. [ROSCOE: A suite of metrics for scoring step-by-step reasoning](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Ivan Habernal and Iryna Gurevych. 2016. [Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.
- Shirish Karande, Santhosh V, and Yash Bhatia. 2024. [Persuasion games with large language models](#). In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 576–582, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLP AI).
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Michal Kosinski. 2024. [Evaluating large language models in theory of mind tasks](#). *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, pages 611–626. ACM.
- Tiziano Labruna, Arkadiusz Modzelewski, Giorgio Satta, and Giovanni Da San Martino. 2026. [Detecting winning arguments with large language models and persuasion strategies](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1888–1915, Rabat, Morocco. Association for Computational Linguistics.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. [Revisiting the evaluation of theory of mind through question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.
- Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2025. [Constrained reasoning chains for enhancing theory-of-mind in large language models](#). In *PRICAI 2024: Trends in Artificial Intelligence*, volume 15282 of *Lecture Notes in Computer Science*, pages 354–360, Singapore. Springer Nature Singapore.
- Minqian Liu, Zhiyang Xu, Xinyi Zhang, Heajun An, Sarvech Qadir, Qi Zhang, Pamela J. Wisniewski, Jin-Hee Cho, Sang Won Lee, Ruoxi Jia, and Lifu Huang. 2025. [LLM can be a dangerous persuader: Empirical study of persuasion safety in large language models](#). *arXiv preprint arXiv:2504.10430*.
- Anna R. McAlister and T. Bettina Cornwell. 2009. [Preschool children’s persuasion knowledge: The contribution of theory of mind](#). *Journal of Public Policy & Marketing*, 28(2):175–185.

- OpenAI. 2025a. GPT-5 system card. <https://openai.com/index/gpt-5-system-card/>. Accessed 2026-08-29.
- OpenAI. 2025b. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>. Accessed 2026-08-29.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802. Curran Associates, Inc.
- Candida C. Peterson, Virginia Slaughter, and Henry M. Wellman. 2018. Nimble negotiators: How theory of mind (ToM) interconnects with persuasion skills in children with and without ToM delay. *Developmental Psychology*, 54(3):494–509.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. ReCEval: Evaluating reasoning chains via correctness and informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.
- David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526.
- Qwen Team. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388. Model checkpoint: Qwen/Qwen3-30B-A3B-Instruct-2507, <https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507>.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.
- Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian’s, Malta. Association for Computational Linguistics.
- Muzafer Sherif and Carl I. Hovland. 1961. *Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change*. Yale University Press, New Haven, CT.
- Somesh Singh, Yaman K Singla, Harini SI, and Balaji Krishnamurthy. 2025. Measuring and improving persuasiveness of large language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Virginia Slaughter, Candida C. Peterson, and Chris Moore. 2013. I can talk you into it: Theory of mind and persuasion behavior in young children. *Developmental Psychology*, 49(2):227–231.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 613–624. ACM.
- Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghui Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Henry M. Wellman. 1990. *The Child’s Theory of Mind*. MIT Press, Cambridge, MA.
- Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Hainiu Xu, Runcong Zhao, Lixing Zhu, Jinhua Du, and Yulan He. 2024. [OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8593–8623, Bangkok, Thailand. Association for Computational Linguistics.

Diya Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. [Let’s make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru R. Zhang. 2026. [The geometry of reasoning: Flowing logics in representation space](#). In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. [JudgeLM: Fine-tuned large language models are scalable judges](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.

A Metric–Verdict Correlation Analysis

The results in Section 5 show that PRISM reliably distinguishes PAToM from flat conditions even when scores are identical. A natural follow-up question is whether any reasoning metric, individually or in combination, predicts whether the judge reaches a correct verdict. We report Spearman rank correlations between each reasoning metric and verdict correctness across all four judges and three conditions.

Correlation target. For each item we define a continuous correctness signal: $c = s$ for persuasive comments (higher score = more correct) and $c = 10 - s$ for negatives (lower score = more correct), where s is the judge’s 1–10 score. Spearman’s ρ is computed between each reasoning metric and c .

Sign consistency. Because a single pooled correlation can mask model- or condition-specific reversals, we also compute ρ in each of the $4 \times 3 = 12$ model–condition cells and report the fraction of cells where the sign matches the dominant direction (**Sign cons.**), together with the number of individually significant ($p < 0.05$) cells.

Results. Table 6 shows a consistent pattern. The two metrics with the highest sign consistency and strongest per-condition correlations point in opposite directions: **informativeness** is the only metric that consistently increases with correct verdicts ($\rho = +0.11$ to $+0.14$, 92% sign consistency, 11/12 significant cells), while **grounding** consistently decreases with correctness ($\rho = -0.12$ to -0.16 , 100% consistency, 10/12 significant). In other words, judges that bring novel information per step tend to score more accurately, while judges that stay close to the source text tend to err.

Step coherence, redundancy, conclusion grounding, and chain consistency all negatively correlate with correctness. Because PRISM aggregates these same components, its negative correlation with correctness ($\rho = -0.07$ to -0.14 , 100% consistency) follows directly from the individual-metric pattern rather than constituting an independent finding. We therefore emphasize that **the scientific content lies in the per-metric correlations**: each structural signal that PAToM improves is individually anti-correlated with correct verdicts. PRISM is best understood as a convenient process-structure index that separates PAToM from flat conditions ($d > 2.2$), not as a verdict predictor.

Among trajectory metrics, arc length and final deviation show positive correlations under PAToM ($+0.13$, $+0.12$) but are near zero under flat conditions, consistent with the interpretation that PAToM makes longer chains semantically meaningful rather than merely verbose.

Can a combined metric do both? A natural question is whether any combination of reasoning metrics can simultaneously (a) distinguish PAToM from flat conditions (large positive condition effect, like PRISM) and (b) positively correlate with verdict correctness. We performed an exhaustive search over multiplicative and ratio combinations of all nine component metrics, evaluating each candidate on both objectives across all four models and three conditions (48 cells).

The search reveals a fundamental trade-off. Metrics that increase under PAToM (step coherence, grounding, conclusion grounding, chain consistency) all anti-correlate with correctness. The only metric that consistently predicts correct verdicts, informativeness ($\rho = +0.14$, 92% sign consistency), decreases under PAToM ($d = -4.04^{***}$). Combining the two directions, *e.g.* $SC^2 \cdot Inf/SG$, yields a metric that satisfies both criteria but with weak effect sizes on each (correctness $\rho = +0.05$, condition $d = +0.99$), far below PRISM’s $d = +2.54$ or informativeness’s $\rho = +0.14$ individually.

This confirms that **no single reasoning metric can serve as both a process auditor and a verdict predictor** on this task. PRISM remains the strongest process-structure index ($d \in [2.3, 3.2]^{***}$ across models). Its negative correlation with correctness inherits from its components (Table 6) and reflects the central empirical observation: the individual coherence, grounding, and consistency signals that PAToM improves each anti-correlate with correct verdicts. The judges reason more

Table 6: Spearman correlation (ρ) between each reasoning metric and verdict correctness. Columns show per-condition correlations (all four models pooled) and the overall pooled ρ . **Sign** is the dominant direction across 12 model-condition cells; **Sign cons.** is the fraction of cells sharing that sign; **Sig. cells** counts cells with $p < 0.05$. ⁺ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Family	Metric	Per-condition ρ			Overall ρ	Sign	Sign cons.	Sig. cells
		NOTOM	FULLTOM	PAToM				
ROSCOE	step_coherence	-0.074***	-0.083***	-0.135***	-0.017 ⁺	-	100%	8/12
	grounding	-0.156***	-0.142***	-0.124***	-0.079***	-	100%	10/12
	redundancy	-0.091***	-0.080***	-0.156***	-0.029**	-	83%	9/12
ReCEval	informativeness	+0.107***	+0.102***	+0.143***	+0.028**	+	92%	11/12
	concl. grounding	-0.101***	-0.084***	-0.121***	-0.068***	-	92%	8/12
	chain_consistency	-0.009	+0.008	-0.120***	-0.012	-	58%	6/12
Trajectory	arc_length	-0.007	+0.001	+0.133***	+0.021*	+	58%	9/12
	step_disp_cos.	+0.058***	+0.005	-0.003	+0.031**	+	67%	3/12
	final_deviation	+0.036*	+0.015	+0.120***	-0.001	+	75%	8/12
Combined	PRISM	-0.135***	-0.101***	-0.072***	-0.036***	-	100%	9/12

carefully by every structural measure, yet that careful reasoning does not translate into more accurate verdicts.

B Analysis and Ablation

Failure modes. Persuasive items where all three conditions score below the delta threshold (agreed-wrong) share three patterns: (i) **OP-as-arbiter mismatch** (~50%): the chain correctly places the argument within the latitude of acceptance at Step 3, but the final verdict falls back on a generic argument-quality assessment; (ii) **Latitude collapse on cover-and-reframe** (~25%): Step 3 misclassifies a successful reframe as simple agreement with the OP (Tan et al., 2016); (iii) **Step-4 emotional override** (~25%): sharp-toned comments are penalized for missing a good-faith signal that the OP did not actually require. In all three cases the geometric trajectory reaches the correct ToM assessment; only process-level auditing reveals where the final verdict goes wrong.

B.1 Full Per-Model Ablation

We evaluate two nested subsets of persuasive items for each model. **S1** contains all persuasive items ($n=334$ per model); scores may differ across conditions, so S1 tests whether reasoning quality improves regardless of verdict. **S3** contains the triple agreed-wrong subset: items where all three conditions assign the same score below the delta threshold (GPT-5 $n=64$, GPT-4.1 $n=33$, Claude $n=55$, Qwen $n=12$). Because the score is identical by construction, any metric difference on S3 is purely a reasoning-process effect.

All effect sizes below compare PAToM against FULLTOM-COT, the strongest flat-prompt baseline. Both conditions receive the same twelve ToM fields, so any difference reflects the sequential belief-stack structure rather than information access.

Table 7 reports Cohen’s d on S1. Every ROSCOE and ReCEval conclusion metric except redundancy fa-

Table 7: Cohen’s d (PAToM vs. FULLTOM-COT) on **all persuasive items** (S1, $n=334$ per model). Tests omit undefined metric values; Qwen PRISM uses $n=291$ items with values defined in both conditions. Positive d means PAToM is higher. Two-sided Mann–Whitney U . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Family	Metric	GPT-5	GPT-4.1	Claude	Qwen
ROSCOE	step_coh. $\uparrow$	+3.20***	+3.91***	+3.18***	+3.21***
	grounding $\uparrow$	+1.71***	+2.14***	+1.49***	+1.74***
	redundancy $\downarrow$	+2.72***	+3.08***	+2.89***	+2.47***
Trajectory	arc_len. $\uparrow$	+2.21***	+0.38***	-0.04*	+2.45***
	disp. cos. $\uparrow$	+0.27**	+1.00***	+0.48***	+0.80***
	final_dev. $\downarrow$	-3.02***	-2.65***	-2.38***	-2.81***
ReCEval	inform. $\uparrow$	-5.08***	-4.67***	-3.78***	-5.06***
	concl. grnd. $\uparrow$	+1.17***	+1.97***	+1.05***	+1.30***
	chain_cons. $\uparrow$	+1.25***	+1.60***	+1.52***	+1.19***
Combined	PRISM $\uparrow$	+3.16***	+2.29***	+2.21***	+2.55***

vors PAToM with large effect sizes ($d > 1.0$) in every model. Redundancy moves the other way: the positive d means PAToM chains repeat more, which Table 4 counts as a cost. Informativeness also inverts ($d = -3.8$ to -5.1) because PAToM’s conditioning steps are penalized as redundant relative to earlier steps. Arc length and displacement cosine show moderate positive effects, confirming that the chain covers more semantic ground and moves more consistently. PRISM is stable across all four models ($d = +2.2$ to $+3.2$).

Table 8 reports the same analysis on S3. The pattern holds: ROSCOE step coherence ranges from $d = +2.69$ to $+3.66$ even on items where the judge gives the same wrong answer under all three conditions. This confirms that PAToM reasoning is more locally coherent regardless of verdict accuracy. PRISM remains stable ($d = +2.2$ to $+3.2$). On S3, displacement cosine is not significant for GPT-5, and chain consistency is not significant for Qwen. These two cells are consistent with the smaller model-specific subsets.

Table 8: Cohen’s d (PAToM vs. FULLTOM-COT) on **agreed-wrong items** (S3: all three conditions assign the same score <7). Sample sizes: GPT-5 $n=64$, GPT-4.1 $n=33$, Claude $n=55$, Qwen $n=12$. Tests omit undefined metric values; Qwen PRISM uses $n=9$ items with values defined in both conditions. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Family	Metric	GPT-5	GPT-4.1	Claude	Qwen
ROSCOE	step_coh. $\uparrow$	+3.66***	+3.27***	+2.69***	+3.17***
	grounding $\uparrow$	+1.86***	+2.09***	+1.41***	+1.57***
	redundancy $\downarrow$	+3.01***	+2.66***	+2.47***	+2.01***
Trajectory	arc_len. $\uparrow$	+2.41***	+0.82***	+0.49**	+2.43***
	disp. cos. $\uparrow$	+0.23	+0.72*	+0.33*	+1.89***
	final_dev. $\downarrow$	-2.95***	-2.38***	-2.33***	-2.03***
ReCEval	inform. $\uparrow$	-4.84***	-3.97***	-3.49***	-3.78***
	concl. grnd. $\uparrow$	+1.30***	+1.88***	+0.99***	+1.27*
	chain_cons. $\uparrow$	+1.23***	+1.24***	+1.33***	+0.59
Combined	PRISM $\uparrow$	+3.19***	+2.36***	+2.32***	+2.18**

C ToM profile extraction

Table 9: Inter-annotator agreement on ToM extraction quality ($n = 30$ items). “A1 yes” and “A2 yes” are the number of items each annotator marked as faithfully extracted. **Delta validity** checks whether the ground-truth delta label (awarded by the OP) is consistent with the comment content, *i.e.* whether the labeled persuasive comment plausibly deserved its delta; it validates the dataset, not the extraction. The remaining rows evaluate extraction faithfulness per ToM category.

Category	A1 yes	A2 yes	Agree	Agree %
Delta validity	28	21	23	77%
Beliefs	25	29	26	87%
Desires	25	30	25	83%
Intentions	26	30	26	87%
Emotions	23	29	22	73%
Knowledge	27	30	27	90%
Perspective-Taking	22	22	22	73%
Mean (ToM)				82.2%

Cross-extractor agreement. The validation in Table 9 asks annotators to accept or reject a single model’s output. As a complementary check on whether the profiles depend on the choice of extractor, we re-extracted ToM for the same 30 comments with Claude Sonnet 4.6, using the same two-pass prompts, the same truncation, and the same parser; only the model differs. All 30 items parsed with all six categories present.

We embed each extractor’s values and content text with allMiniLML6v2 and compare the two profiles for the same comment (**matched**) against profiles for different comments (**mismatched**). The mismatched column is the important control: ToM descriptions share a great deal of generic phrasing, so raw similarity overstates agreement, and only the gap between the two shows that the extractors agree on content specific to each comment. **Top-1** is the fraction of items for which the matching profile is the nearest of all 30 candidates (chance $=1/30=0.033$).

Table 10: Cross-extractor agreement between Qwen3-30B-A3B-Instruct-2507 and Claude Sonnet 4.6 on the same $n=30$ comments. All categories differ from the mismatched baseline at $p<10^{-8}$ (Wilcoxon signed-rank on matched vs. mean mismatched similarity per item). Category ordering tracks the human agreement rates in Table 9 (Spearman $\rho=0.85$, $p=0.032$).

Category	Matched	Mismatched	Gap	Top-1
Beliefs	0.726	0.180	0.546	1.00
Desires	0.619	0.250	0.368	0.83
Intentions	0.641	0.306	0.335	0.77
Emotions	0.658	0.380	0.278	0.63
Knowledge	0.704	0.230	0.474	1.00
Perspective-Taking	0.594	0.318	0.277	0.73
Mean	0.657	0.277	0.380	0.83

Two extractors from different families therefore recover substantially the same profiles, so the results in Section 5 do not rest on a Qwen-specific artifact. The agreement is also structured in an interpretable way: beliefs and knowledge, which are close to propositional content stated in the text, are recovered almost perfectly, while emotions and perspective-taking, which require inferring attitude and stance, agree least. Our annotators reached the same conclusion independently, marking emotion values as inferred rather than explicit and perspective-taking as generously read. As noted in the Limitations, agreement between two models is weaker evidence than independent extraction by humans, which remains future work.

CMV post — extraction Pass 1

You are analyzing a post from r/ChangeMyView (CMV) through Theory of Mind (ToM).

<post>
Title: {title}

Body: {body}
</post>

Analyze ONLY these three ToM categories (keep output brief):

1. **BELIEFS:** What the author thinks is true/false. Values (1-4 words each, comma-separated) and Content (brief).
2. **DESIRES:** What the author wants/values. Values and Content.
3. **INTENTIONS:** What the author plans or wants others to do. Values and Content.

Output Format:

```
<analysis>
2-3 relatively short sentences of
analysis and reasoning.
</analysis>
<beliefs>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</beliefs>
```

```

<desires>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</desires>
<intentions>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</intentions>
IMPORTANT: Output only these four
tags. Be concise.

```

CMV post — extraction Pass 2

You are analyzing a post from
r/ChangeMyView (CMV) through Theory
of Mind (TOM).

```

<post>
Title: {title}

```

```

Body: {body}
</post>

```

Analyze ONLY these three TOM
categories (keep output brief):

4. **EMOTIONS**: What the author feels
or attributes to others. Values and
Content.
5. **KNOWLEDGE**: What the author
knows/doesn't know. Values and
Content.
6. **PERSPECTIVE-TAKING**: How the
author considers other viewpoints.
Values and Content.

Output Format:

```

<emotions>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</emotions>
<knowledge>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</knowledge>
<perspective-taking>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</perspective-taking>
IMPORTANT: Output only these three
tags. Be concise.

```

CMV comment — extraction Pass 1

You are analyzing a comment from
r/ChangeMyView (CMV) through Theory
of Mind (TOM).

```

<context>
Original Post Title: {title}
</context>
<comment>
{body}
</comment>

```

Analyze ONLY these three TOM

categories (keep output brief):

1. **BELIEFS**: What the author thinks
is true/false. Values and Content.
2. **DESIRES**: What the author
wants/values. Values and Content.
3. **INTENTIONS**: What the author
plans or wants others to do. Values
and Content.

Output Format:

```

<analysis>
2-3 relatively short sentences of
analysis and reasoning.
</analysis>
<beliefs>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</beliefs>
<desires>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</desires>
<intentions>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</intentions>
IMPORTANT: Output only these four
tags. Be concise.

```

CMV comment — extraction Pass 2

You are analyzing a comment from
r/ChangeMyView (CMV) through Theory
of Mind (TOM).

```

<context>
Original Post Title: {title}
</context>
<comment>
{body}
</comment>

```

Analyze ONLY these three TOM
categories (keep output brief):

4. **EMOTIONS**: What the author feels
or attributes to others. Values and
Content.
5. **KNOWLEDGE**: What the author
knows/doesn't know. Values and
Content.
6. **PERSPECTIVE-TAKING**: How the
author considers other viewpoints.
Values and Content.

Output Format:

```

<emotions>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</emotions>
<knowledge>
Values: Value1, Value2 (or
''none'')
Content: Brief description
</knowledge>
<perspective-taking>

```

```

Values: Value1, Value2 (or
`none`)
Content: Brief description
</perspective-taking>
IMPORTANT: Output only these three
tags. Be concise.

```

D Prompts

D.1 Rubric, anchor, and response format.

All judges see a shared 1–10 persuasiveness rubric in the system message (qualitative bands from non-persuasive to exceptionally persuasive). We treat scores of 7–8 and above as the region anchored to plausible delta-worthy arguments, matching the threshold used when comparing model scores to ground-truth Δ s in the main text. Every judge must answer with Reasoning: (short justification) and Score: (integer 1–10). No-ToM and Full-ToM use one generation with the system prompt below; Full-ToM additionally inserts the twelve ToM fields into the user message as structured profiles. PAToM uses its own system prompt (opening subsection below) and four analytical turns before the same Reasoning: / Score: verdict format on turn five.

D.2 No-ToM and Full-ToM (single-turn scoring)

System prompt — No-ToM and Full-ToM

You are an expert evaluator of persuasive arguments. You analyze comments from r/ChangeMyView (CMV), a Reddit community where users post a view they hold but are open to changing. Commenters try to change the original poster's (OP)'s mind. A comment that succeeds earns a ``delta'' (Δ) from the OP as explicit acknowledgment that it changed their view.

A persuasive comment typically:

- Directly engages with the OP's core beliefs and stated concerns.
- Introduces new framing, evidence, or perspectives the OP had not considered.
- Acknowledges the OP's emotional state before challenging their reasoning.
- Reframes the OP's goals or plans toward a different conclusion.
- Does NOT simply agree with or echo the OP's existing arguments.
- Does NOT parrot the OP's own knowledge claims back at them without adding new insight.

You will rate each comment on a persuasiveness scale from 1 to 10:

1--2 : Irrelevant, dismissive, or purely agreeing --- no persuasive value.

3--4 : Engages with the topic but adds no new framing or evidence.

5--6 : Offers some new angle but does not squarely address the OP's core concerns.

7--8 : Well-targeted argument that addresses key concerns and introduces new framing.

9--10 : Exceptionally persuasive --- directly dismantles the OP's position with compelling logic.

User message — No-ToM (Full-ToM adds ToM blocks)

```

### Post
Title: {post.title}

{post.body truncated to ~1500 chars}
[Full-ToM only: Post Author ---
Theory of Mind Profile:
  six categories with values and/or
content lines]

### Comment to Evaluate
{comment.body truncated to ~1000
chars}

[Full-ToM only: Comment Author ---
Theory of Mind Profile:
  six categories with values and/or
content lines]

[Full-ToM only: instruction to
use profiles to assess engagement
with beliefs/desires, reframing
intentions/emotions, and avoiding
parroting knowledge claims]

Provide a brief reasoning (2--3
sentences) and then a score.
Respond in exactly this format:
Reasoning: <your reasoning>
Score: <integer 1--10>

```

D.3 PAToM (five-turn chain)

System prompt (shared across Steps 1–5):

System Prompt — PAToM

You are an expert evaluator of persuasive arguments in r/ChangeMyView (CMV), a Reddit community where users post a view they hold but are open to changing. Commenters try to change the original poster's (OP's) mind; a successful comment earns a ``delta'' (Δ) from the OP.

Persuasion in CMV is not about a generically good argument --- it is about an argument *calibrated to this specific OP's* epistemic commitments and the gaps in their stated reasoning. You will reason step-by-step through a nested Theory of Mind analysis. Each step builds on the last; do not jump to a verdict until asked.

Step 1 (1st-order OP: OP's belief state). ToM fields shown: beliefs, desires, intentions, knowledge, perspective-taking.

Step 1 of 5 — 1st-order OP: OP's belief state

```
### Step 1 of 5 --- 1st-order OP:
OP's belief state
**Post**
Title: {post.title}
{post.body[:1500]}
{op.tom.profile}
Answer in 3--5 sentences:
1. What does the OP actually
believe, and what position are they
defending?
2. What are the load-bearing
assumptions underneath their
stated view --- the premises
they treat as self-evident that a
challenger would need to undermine?
Do not mention the comment yet ---
it has not been shown.
```

Step 2 (1st-order Commenter: argument and epistemic stance). Introduces comment text. ToM fields shown: beliefs, desires, intentions.

Step 2 of 5 — 1st-order Commenter: argument & epistemic stance

```
### Step 2 of 5 --- 1st-order
Commenter: commenter's argument
and epistemic stance
{commenter.tom: beliefs, desires,
intentions}
Comment being evaluated
{comment.body[:1200]}
Answer in 3--5 sentences:
1. What is the commenter's core
argument or counter-position?
2. What is their own epistemic
stance --- what do they believe,
and from what perspective are they
arguing?
```

Step 3 (Latitude of acceptance, Social Judgment Theory). ToM fields shown: perspective.taking, knowledge (commenter's model of OP).

Step 3 of 5 — Latitude of Acceptance (Social Judgment Theory)

```
### Step 3 of 5 --- Latitude of
acceptance (Social Judgment Theory)
{commenter.tom: perspective.taking,
knowledge}
The OP's anchor position and
load-bearing assumptions are
established in Step 1. The
commenter's argument is established
in Step 2.
Social Judgment Theory (Sherif &
Hovland, 1961) holds that persuasion
succeeds only when an argument lands
within the target's latitude
of acceptance --- the range of
positions they would consider.
Answer in 3--5 sentences:
1. What is the OP's likely
latitude of acceptance --- how
far from their anchor would they be
```

willing to move, given their stated beliefs and reasoning style?
2. Does the commenter's argument land **within** that latitude --- far enough to move the OP but not so discrepant it triggers rejection?
3. Or does it **assimilate** (essentially restates what the OP already believes) or **contrast** (positions itself so far from the anchor it would be dismissed)?

Step 4 (3rd-order ToM: credibility and intent management). ToM fields shown: emotions for both OP and commenter.

Step 4 of 5 — 3rd-order ToM: Credibility & Intent Management

```
### Step 4 of 5 --- 3rd-order
(Commenter → OP → Commenter):
credibility and intent management
{op.tom: emotions}
{commenter.tom: emotions}
Answer in 2--4 sentences:
1. Does the commenter's emotional
stance acknowledge or match the
OP's emotional register --- showing
genuine engagement rather than
dismissal?
2. Do they signal good faith --- intellectual respect, shared
values, or intellectual humility
before challenging the OP?
3. Does the commenter preemptively
manage how the OP will perceive
their credibility (e.g., conceding
minor points, acknowledging the OP's
strongest concern)?
```

Step 5 (Verdict). No new ToM fields. Conditioned on the full recursive belief-stack from Steps 1–4.

Step 5 of 5 — Verdict

```
### Step 5 of 5 --- Verdict
Using your full nested analysis:
• Step 1 --- OP's beliefs and
load-bearing assumptions
• Step 2 --- Commenter's argument
and epistemic stance
• Step 3 --- Whether the argument
lands within the OP's latitude of
acceptance (SJT)
• Step 4 --- Whether the commenter
managed OP's perception of their
credibility/intent
Rate whether this comment is a
well-calibrated persuasive act
toward this specific OP.
Respond in exactly this format:
Reasoning: {2--3 sentences grounded
in your chain-of-analysis above}
Score: {integer 1--10}
```

E Case Study

We present a single CMV item that illustrates the three comment types evaluated in this paper, together with the extracted ToM profiles and judge scores across conditions. The post, comment texts, and ToM profiles below are shown verbatim (lightly truncated for space).

E.1 Post

CMV post: "I don't see why people like flowers as a gift"

Of course, they should be appreciated like all gifts. But really? They suck, they die fast WHILE being expensive as hell to buy. ``Just get a vase'' --- What's the point of a gift if I have to buy another thing for it myself? ``The colors are nice'' --- Pretty colors for like a week then bam: ugly colors taking up space. ``Just throw them out'' --- I don't like throwing out gifts, I wanna keep them forever so I can remember the thought. You can't do that with flowers because once they're dead you gotta get rid of them. ``Flowers are supposed to be short-term gifts'' --- What is the point of a short-term gift? Who likes a gift you're made to get rid of? ``What about fake flowers'' --- They're made of harmful plastics! ``What if the flowers are potted?'' --- Your gift is responsibility? Now I have to love and care for this houseplant or I'll feel horrible?

E.2 Three Comment Types

Persuasive (P): delta awarded.

Persuasive comment (truncated)

I'm going to focus on the idea you raised ``who wants a short term gift?'' The answer is plenty of people. Think about ``experience'' gifts: a trip, a show, a special meal. These are based on how the experience feels in the moment, and the memories you retain. The same is the case for chocolates and alcohol. The pleasure is fleeting but still meaningful. Of course some people prefer long-term souvenirs, but others prefer the sensory, temporary experience. I'm a visual person and I get genuine pleasure every time I look at a vase of flowers, for the 5 or so days I can enjoy them. If the flowers are a gift then every time I look at them it reminds me of the person who gave them to me. I even enjoy seeing how they change over those days. In fact, the temporary nature actually

Table 11: Extracted ToM profile for the OP (post). Values are short keyword lists; Content is a one-sentence rationale.

Category	Values / Content
Beliefs	V: Transient, wasteful, burdensome C: Flowers are a poor gift because they die quickly, require extra effort (vases, water), and cause distress when discarded.
Desires	V: Lasting, meaningful, low-maintenance C: The author desires gifts that can be kept long-term and evoke positive memory without decay or disposal.
Intentions	V: Recipient autonomy, emotional preservation C: The author wants others to stop giving flowers and instead offer enduring, mindful alternatives.
Emotions	V: Frustration, disappointment, resentment C: Frustration over the impermanence and disposability of flowers, disappointment in the expectation to maintain them.
Knowledge	V: Practicality, environmental awareness C: The author knows flowers are expensive, short-lived, and require maintenance; even fake ones are environmentally harmful.
Perspective	V: Empathy, receptivity C: Acknowledges counterarguments (vases, composting, potted plants) but rejects them as inadequate.

makes me appreciate them more. So for me the temporary nature is a factor in the plus column, not the minus column!

Hard Negative (H): no delta, high community score.

Hard-negative comment

I agree when it's about bought flowers. I've been visiting rose plantations in Ecuador and talked to people working there, and the working conditions are really bad and it's really bad for the environment. There are alternatives though. My grandmother had a beautiful garden with lots of flowers. When she was visiting people she sometimes picked some flowers in her garden to give to people. People really appreciated it. If you have a place to pick wild flowers it's also really nice and special.

Easy Negative (E): no delta, low community score.

Easy-negative comment

Walmart sells bouquets for \$5. I get them for my girl randomly all the time and she loves them.

Table 12: Judge scores for the three comments across conditions. The ordering $\bar{s}_P \geq \bar{s}_H > \bar{s}_E$ holds in every cell; under PAToM, GPT-4.1 ties the persuasive and hard-negative comments at 5.

Model	Condition	P	H	E
GPT-5	NOTOM-COT	7	4	3
	PAToM	6	4	3
GPT-4.1	NOTOM-COT	7	5	2
	PAToM	5	5	3
Claude 4.6	NOTOM-COT	7	3	1
	PAToM	6	4	2

E.3 Judge Scores

E.4 Discussion

This item is one of the clearest successes in the dataset: all judges correctly rank $P > H > E$ under both conditions. The persuasive comment directly targets the OP’s load-bearing assumption (“who wants a short-term gift?”) by reframing impermanence as a feature rather than a flaw, drawing on analogies to experience gifts (trips, meals). The hard negative partially agrees with the OP and offers an alternative (garden-picked flowers) but sidesteps the core premise rather than challenging it. The easy negative is dismissive and anecdotal, offering no engagement with the OP’s reasoning.

The extracted ToM profiles (Table 11) capture these distinctions: the OP’s profile highlights frustration, a desire for permanence, and rejection of maintenance-based solutions; the persuasive comment’s profile shows empathy, perspective-shifting, and an understanding that impermanence can enhance value. Under PAToM, the judge’s Step 3 (latitude of acceptance) correctly identifies that the persuasive comment lands within the OP’s acceptance range by validating their preference for meaningful gifts while reframing what “meaningful” can include.

Yet even on this well-separated item, no judge scores the persuasive comment above 7, and PAToM actually lowers the persuasive score from 7 to 5–6 in all three judges. The structured reasoning improves hard-negative suppression but does not boost persuasive-comment recognition, consistent with the aggregate pattern reported in Section 5.2.

F Human Annotation Procedure

To validate ToM extraction quality, two annotators—a computer-science Master’s student and an Information School PhD student—labeled a stratified sample of 30 CMV items. Each item consisted of the original post, the persuasive comment (the one that earned a delta), and the OP’s delta reply. For each item, annotators answered seven binary (Yes/No) questions:

1. **Persuasiveness (Q1):** Does the persuasive comment give a clear, substantive reason for the OP to change their view?

2. **Beliefs (Q2):** Do the extracted belief-values match what the commenter seems to believe or assert?
3. **Desires (Q3):** Do the desire-values match what the commenter wants or prioritizes?
4. **Intentions (Q4):** Do the intention-values match what the commenter is trying to do or achieve?
5. **Emotions (Q5):** Do the emotion-values match the emotional tone or appeals in the comment?
6. **Knowledge (Q6):** Do the knowledge-values match what the commenter claims to know or how they use evidence?
7. **Perspective Taking (Q7):** Do the perspective-taking values match how the commenter considers other viewpoints?

Annotators were provided with category definitions from Table 1 and instructed to judge each ToM category independently against the comment text. If a category had no clear content in the comment, annotators were told that the extraction saying “none” or being minimal still counts as correct. Annotators could also flag unusable samples. Progress was saved after each submission, and annotators could revisit and revise earlier answers. On Q1, at least one annotator found the delta comment persuasive for 93% of items and both did for 70%; the two gave the same answer on 77%. Mean per-category agreement on Q2–Q7 is 82.2% (Table 9). The annotation interface is released with the paper.

Sampling and topical coverage. The 30 items were drawn uniformly at random from the cleaned delta-post pool with a fixed seed, and come from 30 distinct posts. The sample is topically broad rather than concentrated in one debate area, spanning criminal justice, party politics, colonialism, religion, gun policy, education, parenting, gender and LGBTQ questions, the Ukraine conflict, labor and economics, and metaphysics. Post titles, lightly truncated, are:

1. Louis CK doesn’t deserve to be punished
2. Stray cats should be euthanized and it should be illegal to let your...
3. I believe the attorney-client privilege should be abolished CMV
4. Democrats need to move to the right in order to win in the future
5. there is a disturbing tendency of teenage girls in the US claiming to...
6. There’s never been a better time to be rich and powerful: 54% can’t r...
7. Colonisation overall was a net positive for the world.
8. European countries should not extradite military-age Ukrainian men fo...
9. Banning extremist political groups from Reddit is often a bad idea
10. Sharepoint sucks. It Reduces Productivity Than It Actually Helps.
11. We need to solve the important stuff first. Everything else can wait.
12. There’s no real issue with the National Guard being deployed
13. Two State solution is dead, no one knows what’s next
14. Tips are a practical and effective solution for encouraging good cust...
15. School dress codes aren’t sexist

16. Accepting the theory of binary multi-universes necessitates immortali...
17. Before we see any major change in American capitalist practices, we m...
18. "Free Will" is sufficient complexity in a decision-making process. It...
19. most people definitely smelled terrible in medieval times
20. It is not immoral for a parent of an autistic child to make their chi...
21. if gay incest is wrong, then so is homosexuality between non-siblings.
22. Religion is archaic
23. For couples with children, not wanting a single income lifestyle with...
24. I am against an assault weapons ban in the US.
25. Altruism is a hoax, power is a virtue, happiness is not an end
26. Vodka is useless as an ingredient to cocktails because it has no flavor
27. Unnecessary cosmetic procedures do more harm than good
28. Being a PoC is harder than being queer
29. It sucks, but an Armistice is the only realistic end to the Ukraine c...
30. School are teaching Maths the wrong way

G Reasoning-Step Annotation

Three annotators scored every step of 50 PAToM chains (29 GPT-5, 21 Claude Sonnet 4.6) drawn from the main evaluation set. Each step was marked 0 (wrong or unsupported), 1 (partly correct), or 2 (correct), following written guidelines that we release with the code. Two steps came with extra questions. For Step 3, annotators noted which of three answers the judge gave about how the argument relates to the OP’s view (“within latitude”, “assimilation”, or “contrast”) and whether that answer was right. For Step 5, they noted whether the verdict follows from the four steps before it and whether it brings in anything those steps did not say.

Table 13: Reasoning-step annotation (50 chains, 3 annotators, 150 ratings per step). **Rated 2** is the share of ratings at the maximum; **Agree** is the range of exact pairwise agreement across the three annotator pairs. Errors concentrate at Step 3, the latitude-of-acceptance judgment, and propagate into the Step 5 verdict.

	Step	Mean	Rated 2	Agree
1	OP’s belief state	1.97	96.7%	92–94%
2	Commenter’s argument	1.98	98.0%	96%
3	Latitude of acceptance	1.85	86.0%	80%
4	Credibility & tone	1.97	97.3%	94–98%
5	Final verdict	1.91	92.0%	84–90%

Step 3 is the weakest point by every measure: the lowest average score, the fewest full marks, and the lowest agreement between annotators. That last number matters, because it says the judgment is hard for people too, not only for the model. The judge gives the right answer on 43 of 50 chains (86%), and it answers “within latitude” 111 times out of 150. In other words, it seldom reaches for the two categories that would mark an argument as unlikely to land, which are exactly the ones that would separate a hard negative from a persuasive

comment.

The annotators wrote comments on 45 steps they marked down. The most common complaint, about a quarter of them, is that the judge will not choose: it lists two or three possible answers and moves on. One annotator writes that a step “can’t decide between three different outcomes instead of picking one”; another, that the reasoning “supports both Within Latitude and Assimilation, but does not clearly establish which.” Step 5 then carries that hesitation into the verdict, as one annotator puts it, “the same indecision as Step 3.” The chains still hold together: the verdict follows from the earlier steps in 94.7% of ratings and brings in new reasoning in only 3.3%. So the reasoning stays orderly while failing to tell the two kinds of comment apart, which is the split described in Section 5 showing up one step at a time.

H Generalization to Persuasion for Good

The corpus and its label. Persuasion for Good (Wang et al., 2019) contains 1,017 conversations between two people, in which one tries to convince the other to give part of their task payment to Save the Children. Each conversation ends with a record of how much the second person gave. We use that amount, and ignore the field recording how much they said they would give, which leaves 545 conversations where money was donated and 472 where none was. This is what makes the dataset a useful second test: on CMV the label is an award the OP chose to give, while here it is a record of what someone actually did.

Fitting it to our setup. These are back-and-forth conversations, not a post with replies, so we had to map them onto the (post, comment) format our judges expect. We use the donor’s first three messages as the “post”, since these show what they believed and objected to before being talked round. Later messages already reflect the persuader’s influence and would give away the outcome, so we leave them out. Everything the persuader said, joined in order, becomes the “comment”. We cut both to the same lengths we used for CMV (1,500 and 1,200 characters). Persuaders took 10.4 turns on average when they succeeded and 10.7 when they failed, so how much someone talked does not by itself reveal the answer.

Building the pairs. On CMV, the negative comments come from the same thread as the persuasive one. Persuasion for Good has no equivalent, since each conversation involves just one pair of people, so we match across conversations instead. Every successful conversation supplies one persuasive item. We sort the failed conversations by how many turns the persuader took, as a rough measure of how hard they tried: the top two-thirds (317) become hard negatives and the bottom third (155) easy negatives. Drawing two hard and one easy negative for each successful conversation (seed 42) gives 2,180 pairs (545 P + 1,090 H + 545 E). ToM profiles come from Qwen3-30B-A3B-Instruct-2507 (Qwen

Team, 2025), the same extractor used for CMV, using the six categories of Appendix C; all 2,180 parsed without error.

Table 14: PFG results across three judges ($n=2,180$). $\text{DISC} = \text{mean}(P) - \text{mean}(H+E)$; AUC on the P-vs-H cut. ρ_{PRISM} is the Spearman correlation between PRISM and verdict correctness under PAToM. Best AUC per judge in **bold**. On every judge FULLTOM attains the best AUC, and PRISM rises sharply under PAToM while correlating negatively with correctness.

Judge	Cond.	Disc	AUC	PRISM	ρ_{PRISM}
Qwen	NOTOM	+0.82	0.573	0.138	
	FULLTOM	+1.28	0.601	0.208	
	PAToM	+0.82	0.561	0.538	-0.171***
GPT-4.1	NOTOM	+0.27	0.562	0.173	
	FULLTOM	+0.66	0.617	0.215	
	PAToM	+0.68	0.581	0.550	-0.219***
GPT-5	NOTOM	+0.30	0.557	0.190	
	FULLTOM	+0.62	0.621	0.140	
	PAToM	+0.54	0.606	0.625	-0.073***

*** $p < 0.001$. Step coherence rises analogously (NOTOM $\rightarrow$ PAToM: Qwen 0.37 $\rightarrow$ 0.67, GPT-4.1 0.39 $\rightarrow$ 0.68, GPT-5 0.37 $\rightarrow$ 0.66).

What holds, and what does not. The main result holds. PAToM again improves the reasoning by a wide margin on all three judges, raising PRISM to between 2.6 and 4.5 times the flat conditions and step coherence from about 0.37 to 0.67. The verdicts again stay close to chance in every condition (AUC 0.56–0.62), and judges scoring higher on PRISM are again slightly more likely to be wrong. This holds for one open model and two frontier models whose raw scores differ by a factor of three or four, which is the best evidence we have that the pattern is not a quirk of any one system.

The two ToM conditions rank differently on the two datasets. On CMV, PAToM gave the best persuasive-versus-hard-negative AUC for every judge. Here FULLTOM does (0.621, 0.617, 0.601), and PAToM mainly pushes scores down instead of separating the two groups more sharply. So which ToM condition yields better verdicts changes from one dataset to the other. This fits our main claim, that reasoning quality and verdict accuracy move independently, but it rules out reading the CMV numbers as evidence that PAToM reliably improves verdicts. We take the cautious view: structured ToM prompting changes how a judge reasons much more than whether it is right, and its small effects on verdicts do not carry from one setting to the next.

I Reasoned Better, Judged Wrong

Section 5 reports that on the 176 persuasive items where FULLTOM-COT and PAToM both score below the delta threshold, PAToM still covers significantly more semantic ground ($p < 10^{-24}$). This appendix shows two of those items in full, so the disconnect can be read at the level of a single chain rather than a distribution. Both comments earned a delta from the OP. Both were scored

6 by PAToM, one point short of the threshold. Steps are lightly truncated.

Example 1: insurance as a Ponzi scheme. The OP argues that all insurance is “literally a Ponzi scheme”. The delta-winning comment explains that a Ponzi scheme pays early investors out of new investors’ money, whereas an insurer pays claims out of pooled premiums and profits on the margin. FULLTOM-COT scores this 5, PAToM 6.

PAToM chain — cwn2zj/eyd2hq1

Step 1 (OP’s belief state). The OP believes all forms of insurance are fundamentally fraudulent and unnecessary, likening them to Ponzi schemes... individuals would be better off saving their own money for emergencies.

Step 2 (commenter’s argument). The commenter’s core argument is that insurance differs from a Ponzi scheme because insurers profit on the margin between premiums collected and claims paid, not by using new customers’ payments to pay earlier ones.

Step 3 (latitude of acceptance). The OP’s latitude of acceptance is likely very narrow... they would only consider arguments that acknowledge some flaws in insurance but show it is not fraudulent, or that demonstrate a clear, practical distinction between insurance and Ponzi schemes.

Step 4 (credibility and tone). The commenter’s emotional stance does not directly acknowledge the OP’s frustration and cynicism; it maintains a tone of confident correction... which may come across as dismissive.

Step 5 (verdict). The comment accurately distinguishes insurance from a Ponzi scheme and provides a clear, factual correction, **which lands within the OP’s narrow latitude of acceptance.** However, it does not address the OP’s deeper skepticism... nor does it manage credibility by conceding any points or showing empathy. **Score: 6.**

Example 2: race and gender in police violence. The OP argues that police violence is as much a gender issue as a race issue. The delta-winning comment clarifies that it never denied a gender aspect, only that the gender aspect did not warrant the same level of response as the racial one. FULLTOM-COT scores this 2, PAToM 6.

PAToM chain — gse90x/fs4s701

Step 1 (OP’s belief state). The OP believes police violence is not

solely a racial issue but equally a gender issue... both race and gender must be addressed.

Step 2 (commenter's argument). The commenter is not denying gender bias, but rejects the idea that gender and race deserve equal emphasis in public discourse about police violence.

Step 3 (latitude of acceptance). Given the OP's nuanced, intersectional reasoning... their latitude of acceptance likely includes arguments that recognize both factors but differ on the degree of emphasis. They are open to arguments that prioritize race somewhat more than gender, as long as gender is not dismissed.

Step 4 (credibility and tone). The commenter's emotional stance is primarily defensive... they do not explicitly engage with the OP's emotional urgency.

Step 5 (verdict). The comment **lands within the OP's latitude of acceptance** by acknowledging the gender aspect but prioritizing the racial component... However, it does not directly engage with the OP's core emotional concerns or manage credibility through explicit empathy... it lacks the depth and emotional resonance needed for a higher score. **Score: 6.**

What the two have in common. In both chains the belief-stack works as designed through Step 3. The judge reconstructs the OP's position, states the latitude of acceptance in specific terms, and then in Step 5 explicitly concludes that the argument *falls inside* that latitude, which is the judgment the protocol was built to elicit and which the delta confirms is correct. The verdict nonetheless comes out below threshold, and in both cases for the same reason: Step 4 supplies a deficit of empathy, concession, or "emotional resonance", and Step 5 treats that deficit as decisive, marking the comment down from the placement Step 3 had earned.

This is the decoupling in miniature. The scaffold succeeds at the social-cognitive sub-task and then discards the result, because the final score is dominated by the same tone and argument-quality dimensions that drive the No-ToM baseline (Section 6.1). It also matches what our annotators found at Step 5, where errors concentrate second only to Step 3 (Appendix G): the chain stays coherent, and the verdict does not follow the part of the chain that carries the persuasion signal.

J Use of AI Assistants

AI assistants (Claude, ChatGPT) were used during this work for code debugging and revising some text of the paper.