DialToM: A Theory of Mind Benchmark for Forecasting State-Driven Dialogue Trajectories

Anonymous ACL submission

Abstract

We introduce DialToM, an annotated Theory of Mind (ToM) benchmark built from naturalistic human–human dialogues using a multiple-choice evaluation framework. Concurrent with recent work showing a gap between explicit mental-state inference and applied ToM in synthetic settings (Gu et al., 2024), we establish a stricter *State-Driven Diagnostic Probe* in which models must forecast state-consistent dialogue trajectories solely from isolated mental-state profiles without dialogue context. Our evaluation reveals a systematic reasoning asymmetry—LLMs excel at inferring mental states (Literal ToM) but struggle to leverage them for social forecasting (Functional ToM). Crucially, a domain expert achieves 100% accuracy on this task, proving its validity and establishing a stark human-AI capability gap. Further, a teacher-student reasoning injection probe shows that Gemini 3 Pro—which establishes the leading baseline—possesses robust Functional ToM capabilities for context-free forecasting that are transferable to weaker models. DialToM, its evaluation code, and dataset are publicly available at https://anonymous.4open.science/r/DialToM.

1 Introduction

Theory of Mind (ToM) is a foundation of human cognition that allows us to reason about others' hidden mental states, such as their Beliefs, Desires, and Intentions (the BDI model) (Bratman, 1987; Apperly and Butterfill, 2009). Human social intelligence relies on a dual-stage execution of this capability: (i) inferring latent mental states from context; and (ii) utilizing those mental states to plan downstream actions (Ho et al., 2022; Riemer et al., 2025). Driven by the rapid progress of LLMs, existing research has examined machine ToM in various conditions (Strachan et al., 2024; Shinoda et al., 2025). Existing evaluations, however, remain heavily skewed toward the first stage—tracking states and identifying false beliefs in synthetic narrative domains (Le et al., 2019; Wang et al., 2025)—leaving open whether LLMs can genuinely move from passive inference to functional planning in naturalistic interactions.

A growing body of work suggests this *inference-to-action* gap is a fundamental limitation of current LLMs. Recently, Gu et al. (2024) and Li et al. (2025) demonstrated that while frontier models reliably infer mental states, this competence fails to translate into downstream behavior prediction. We converge on this asymmetry with **DialToM**, a human-verified benchmark over naturalistic conversations that operationalizes the distinction between **Literal ToM** and **Functional ToM**[1] (Riemer et al., 2025). DialToM extends prior investigations along three critical axes: (i) grounding evaluation in high-stakes human dialogues (counseling and persuasion); (ii) adopting a richer six-attribute model featuring a novel *Trust* dimension to capture relational rapport; and (iii) enforcing a bifurcated evaluation that pairs a context-driven **Retrospective task** (Literal ToM: inferring states from dialogue) with a strictly context-free **Prospective Diagnostic Probe** (Functional ToM: forecasting actions from isolated mental states).

The DialToM evaluation pipeline (Figure 1) benchmarks models across both the Retrospective inference task and a context-free Prospective diagnostic forecasting task. Within this framework, models must identify state-consistent dialogue trajectories based solely on an isolated mental-state profile—a setup capturing the literal, high-fidelity internal monologue and communicative intent that directly dictate a speaker's verbal choices. This context-free setup is grounded in standard psychometric principles, which dictate that validating a cognitive construct requires evaluating the

[1] See Appendix A for a detailed theoretical background on the distinction between Literal and Functional ToM.

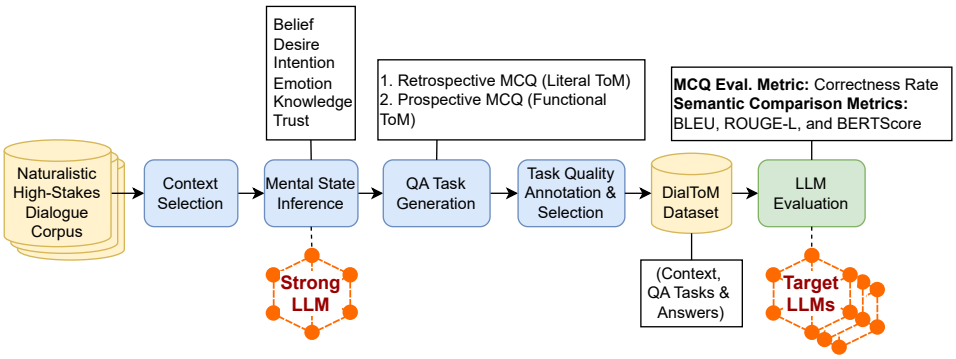

Figure 1: The DialToM Multiple-Choice Question (MCQ) Evaluation Benchmarking Pipeline. The workflow illustrates the transition from Retrospective Inference to Prospective Diagnostic Forecasting through automated mental state inference, distractor generation, and subsequent human verification.

bidirectional mapping between internal states and observed behaviors (Cronbach and Meehl, 1955; Whitely, 1983). The strict design is validated by a professional domain expert who achieved 100% accuracy under these exact conditions, and by the task structure itself[2], where candidate trajectories share high surface-form similarity, forcing a reliance on genuine mental-state reasoning.

Our empirical analysis reveals a systematic *reasoning asymmetry*. While models excel at Retrospective mental state inference, the majority fail at Prospective Diagnostic Forecasting, with several scoring near or below chance. This asymmetry is consistent across domains and model families, replicating and deepening the inference-to-action gap identified by Gu et al. (2024); Li et al. (2025). In addition, we find that LLMs specifically struggle with the Trust attribute, suggesting relational dynamics remain an understudied dimension. A notable exception is Gemini 3 Pro, which achieves ≈83% on the Prospective task—and whose reasoning traces, when applied to weaker models, boost their performance by up to 76 points. This establishes an empirical proof of concept for functional ToM-grounding in LLMs.

2 Related Work

2.1 Benchmarking Theory of Mind in LLMs

Literal ToM & Its Limitations Early ToM benchmarks adapted classic false-belief paradigms (Dennett, 1978; Wimmer, 1983) into rigid, template-based narratives (e.g., ToMi (Le et al., 2019), Hi-ToM (Wu et al., 2023), OpenToM (Xu et al., 2024)). To address this static nature, recent research shifted to multi-agent social simulations and scenarios like FANToM (Kim et al., 2023), SOTOPIA (Zhou et al., 2024), ToMBench (Chen et al., 2024), and ToMATO (Shinoda et al., 2025). While increasing environment dynamism, they remain confined to passive, retrospective state attribution and overlook the broader spectrum of real-world cognition required in social scenarios (Wang et al., 2025; Xiao et al., 2025).

Functional ToM & The Inference-to-Action Gap To bridge the gap, recent benchmarks have pioneered behavioral prediction tasks to measure Functional ToM. While BigToM (Gandhi et al., 2023) explores behavior forecasting in synthetic narrative templates, RecToM (Li et al., 2025) and SimpleToM (Gu et al., 2024) systematically investigate both state attribution and behavior prediction. Crucially, these lines of research independently document a striking inference-to-action gap—termed *cognitive inference vs. behavioral prediction* in RecToM and *Explicit vs. Applied ToM* in SimpleToM—demonstrating that frontier LLMs track mental states proficiently but fail to apply them for behavioral forecasting. Despite this structural progress, existing evaluations rely entirely on context-driven setups where the model retains full conversational or narrative history during prediction. Unlike these works, DialToM uniquely combines context-free prediction with naturalistic human–human dialogue and a Trust dimension, occupying a position in the benchmark landscape that no existing resource addresses (Table 1).

2.2 Distinction of DialToM

DialToM distinguishes itself from current benchmarking methodologies by evaluating ToM over

[2]See Appendix J for concrete examples.

Benchmark	Attributes	Context Type	Functional ToM	Context-Free Prediction	Context Source
ToMi (Le et al., 2019)	B	Goal-oriented stories	✗	✗	Template-based
Hi-ToM (Wu et al., 2023)	B	Goal-oriented stories	✗	✗	Template-based
FANToM (Kim et al., 2023)	B	Synthetic conversations	✗	✗	LLM generated
OpenToM (Xu et al., 2024)	B	Goal-oriented stories	✗	✗	LLM generated
ToMBench (Chen et al., 2024)	BDIEK	Goal-oriented stories	✗	✗	Human-written
ToMATO (Shinoda et al., 2025)	BDIEK	Social simulation	✗	✗	LLM-LLM
DynToM (Xiao et al., 2025)	BIE	Synthetic scenarios	✗	✗	LLM generated
BigToM (Gandhi et al., 2023)	B	Goal-oriented stories	✓	✗	Template-based
SimpleToM (Gu et al., 2024)	B	Synthetic stories	✓	✗	LLM+Human
RecToM (Li et al., 2025)	BDI	Recommendation dialogue	✓	✗	Human-Human
DialToM (Ours)	**BDIEK-T**	**High-stakes multi-turn dialogues**	✓	✓	**Human-Human**

Table 1: Comparison of DialToM with existing Theory of Mind benchmarks. **Functional ToM** indicates whether the benchmark measures applied action/trajectory forecasting based on mental state profiles. **Context-Free Prediction** indicates whether this forecasting task strips out conversational or story context. DialToM uniquely operates at the intersection of both properties while analyzing high-stakes dialogues with Trust modeling.

natural human–human conversations and incorporating a newly proposed Trust dimension between speakers. More importantly, we operationalize *Context-Free Prediction*—the ability to identify state-consistent dialogue trajectories based strictly on an isolated mental state profile. By deliberately withholding prior conversation history, we prevent models from relying on spurious context-to-dialogue correlations.

In Table 1, we compare DialToM with existing benchmarks across five dimensions: the mental state attributes of focus, the type of contexts used, whether the benchmark measures Functional ToM, whether it utilizes Context-Free Prediction, and the source of the contexts.

Ecological Validity & The Trust Dimension Most benchmarks either focus on goal-oriented synthetic stories generated by LLMs or do not follow naturalistic contexts; this makes the findings hard to generalize to real-life human interaction scenarios. To fill this gap, DialToM is constructed from naturalistic, high-stakes human dialogues (mental health support and persuasion).

Existing benchmarks are limited to studying belief as a primary mental state attribute which cannot fully capture the relational complexity of these domains or rapport-building of human interactions. We extend the standard mental state model to include a new component of Trust (T) which is crucial in human interaction.

Context-Free Prediction We introduce context-free prediction (state-driven probing) to rigorously audit the functional utility of ToM. This diagnostic probe evaluates a model's ability to forecast actions without access to scenario or dialogue history. Unlike context-driven forecasting in prior benchmarks, DialToM challenges models to perform strict diagnostic mapping from isolated mental profiles to logical dialogue trajectories. This isolation forces models to demonstrate functionally grounded ToM reasoning rather than exploiting surface-level topic coherence.

Supplementary Semantic Verification While DialToM's core diagnostic relies on multiple-choice evaluation, it is also uniquely supplemented by a gold-standard reference set of human-written mental state rationales. This allows for the semantic comparison of LLM-generated and human-written reasoning traces, providing an additional qualitative lens for evaluating machine ToM alignment.

3 Methodology

3.1 Data Curation & Token Normalization

To ensure ecological validity, DialToM is built upon naturalistic, high-stakes human interactions drawn from three relational domains[3]: Motivational Interviewing via **AnnoMI** (Wu et al., 2022), Emotional Support Conversations via **ESConv** (Liu et al., 2021), and strategic negotiation via **PersuasionForGood** (PFG) (Wang et al., 2019). We preserve a constant dyadic role distribution, mapping participants to an *initiator* (counselor, supporter, persuader) and a *recipient* (client, seeker, persuadee).

To normalize context density and mitigate extreme length variance across domains (ranging from an average of 1,233 words in AnnoMI to 273 words in PFG), we segment full sessions into uniform sliding context windows. We enforce a

[3]Appendix B provides complete corpus descriptions and full pre-/post-segmentation statistics (Table 9).

Statistic	MI	ESC	PFG	Overall
Context Count	865	3,215	1,863	5,943
Word Count per Context	188.2	205.0	210.8	—
Turn Count per Context	5.3	4.5	5.4	—
Word Count per Turn	35.5	45.5	39.1	—
Retrospective Infer. Qns.	5,190	19,290	11,178	35,658
Prospective Forecast. Qns.	865	3,215	1,863	5,943
Human Expert Gold Refs.	300	300	300	900

Table 2: General descriptive statistics of the curated DialToM evaluation suite across its three core dialogue domains.

strict reasoning horizon bounded between 150 and 200 words per window to optimize semantic completeness and token efficiency. Any trailing session fragments falling below the 150-word threshold are dynamically concatenated with their preceding window to maintain interaction continuity.

As summarized in Table 2, this pipeline yields 5,943 evaluation contexts. We generate 6 Retrospective questions (one per BDIEKT attribute) and 1 Prospective question per context, establishing a comprehensive benchmarking suite of 35,658 Literal ToM tasks and 5,943 Functional ToM forecasting problems.

3.2 The Trust Dimension in Mental State Modeling

Previous Theory of Mind (ToM) benchmarks overwhelmingly evaluate five foundational cognitive components: Beliefs (B), Desires (D), Intentions (I), Emotions (E), and Knowledge (K). However, these attributes fail to capture the relational complexity and evolving rapport of naturalistic human interactions. To address this gap, we introduce a novel sixth mental state attribute: **Trust (T)**.

Trust builds incrementally in multi-turn dialogue and serves as the foundation for collaboration and persuasion. While organizational psychology operationalizes trust via three distinct dimensions—ability, benevolence, and integrity (Mayer et al., 1995)—the DialToM framework unifies these factors into a single concept reflecting active interpersonal rapport strength. This integration is grounded in literature establishing trust as a critical metric in communication and high-stakes social interactions (Blacksher et al., 2016; Clark et al., 2019; Hancock et al., 2023). Modeling Trust alongside standard BDIEK attributes allows DialToM to comprehensively audit how LLMs reason about the interpersonal relational dynamics that govern human interaction.

3.3 The Evaluation Tasks

For each dialogue session, the DialToM benchmark consists of two distinct multiple-choice Question-Answering (QA) tasks featuring a single ground-truth answer and three plausible, model-generated distractors:[4]

1. **Retrospective Inference**: The model is provided with the multi-turn conversation history up to turn i ($\mathcal{H}_i$) as its context and must explicitly infer the specific mental state attribute s of the target speaker. This retrospective inference measures the model's capacity for conventional mental state attribution:

$$\mathcal{S}_i = \{B_i, D_i, I_i, E_i, K_i, T_i\}$$
$$\mathcal{O}_{R,i} = \{s_i, d^1_{s,i}, d^2_{s,i}, d^3_{s,i}\}$$
$$\hat{s}_i = \arg\max_{s' \in \mathcal{O}_{R,i}} \mathcal{M}(s' \mid \mathcal{H}_i) \quad \forall s \in \mathcal{S}_i;\ i \in [1, |\mathcal{H}|]$$

   where $\mathcal{S}_i$ represents the set of mental states (Belief, Desire, Intention, Emotion, Knowledge, Trust), $\mathcal{O}_{R,i}$ is the option space containing ground-truth s_i and three distractor labels $d_{s,i}$, and $\hat{s}_i$ is the value predicted by model $\mathcal{M}$.

2. **Prospective Diagnostic Forecasting**: The model is provided strictly with the complete, isolated mental-state profile $\mathcal{S}_i$ and must identify the most state-consistent downstream dialogue trajectory. Vitally, the dialogue history $\mathcal{H}_i$ is entirely withheld—a deliberate design choice that eliminates lexical shortcut exploitation and forces a reliance on Functional ToM.[5] Providing dialogue history allows models to exploit topic coherence rather than engage in true functional-ToM reasoning.[6] Formally, we define the diagnostic probe as:

$$\mathcal{O}_{P,i} = \{\mathcal{H}_{i+1}, \mathcal{D}^1_{i+1}, \mathcal{D}^2_{i+1}, \mathcal{D}^3_{i+1}\}$$
$$\hat{\mathcal{H}}_{i+1} = \arg\max_{\mathcal{H}' \in \mathcal{O}_{P,i}} \mathcal{M}(\mathcal{H}' \mid \mathcal{S}_i) \quad \forall i \in [1, |\mathcal{H}|]$$

   where $\mathcal{O}_{P,i}$ represents the candidate option space consisting of the ground-truth trajectory $\mathcal{H}_{i+1}$ alongside three adversarial distractors $\mathcal{D}$, and $\hat{\mathcal{H}}_{i+1}$ is the trajectory selected by $\mathcal{M}$.

[4] All candidate options are generated by GPT-4o (prompt in Appendix I) and rigorously filtered via our human-verification pipeline.

[5] We provide an extended discussion on the methodological trade-offs, task artificiality, and human-expert validation of this context-free configuration in Appendix A.

[6] Our dialogue context degradation analysis (Section F) confirms this vulnerability directly, demonstrating that model performance increases monotonically as conversational context is restored.

3.4 Human-in-the-Loop Verification

We employ a human-in-the-loop pipeline to audit all model-generated multiple-choice items and establish robust baselines. A group of 22 professional verifiers[7] audited all Retrospective and Prospective task choices to ensure that the ground-truth options were uniquely correct and that the distractors were contextually plausible yet logically incorrect. Separately, a cohort of 9 qualified crowd workers authored open-ended mental state inferences across a subset of 300 contexts to serve as a gold-standard semantic reference.

Human Benchmarking & Expert Ceiling To quantify task difficulty, three crowd workers evaluated a subset of 900 Retrospective items (300 per domain) to establish a baseline. This evaluation was restricted to Retrospective tasks due to the high cognitive load of the Prospective setup. To validate construct validity and establish an expert performance ceiling, a professional counselor and an expert co-author completed a sample of 20 context-free Prospective task items spanning all model-difficulty tiers, stratified by model failure rates.[8]

3.5 Label Aggregation

We finalize true task labels by aggregating annotations using the Dawid-Skene (DS) algorithm (Dawid and Skene, 1979), fitting models independently across annotation batches. To evaluate inter-annotator agreement, we report Gwet's AC1 (Gwet, 2008) due to its stability against the kappa paradox in skewed distributions (Appendix D). The Retrospective task exhibited exceptionally high consensus (AC1 $\in [0.88, 0.92]$), yielding an 83.25% dataset retention rate. Conversely, the Prospective task yielded lower agreement (AC1 $\in [0.45, 0.68]$) and a 54.3% retention rate. More importantly, the lower Prospective retention serves as a vital quality signal. Our adversarial distractor pipeline was intentionally engineered to maintain surface-level semantic similarity to the ground-truth trajectory. This high-difficulty design purposefully filters out borderline contexts where average verifiers disagree.

Despite this high threshold, the filtered subset remains entirely clear and unambiguous to domain specialists, who achieved a near perfect performance ceiling on these retained items.[9] By prioritizing quality over dataset volume, we ensure that every remaining item acts as a strict Functional-ToM probe. While this aggressive filtering yields a smaller final evaluation set, this scale remains robust for zero-shot LLM benchmarking and is sufficiently powered to expose systematic reasoning deficits without ambiguous annotations.

4 Experiments

4.1 Experimental Setup

Models We evaluate ToM capabilities across 14 state-of-the-art LLMs[10] spanning diverse parameter scales (12B to 1T parameters), access tiers, and architectures (Standard vs. Mixture-of-Experts). Our evaluation suite includes: **(a) Closed-source models:** Gemini (3 Pro, 2.5 Pro, 2.5 Flash) (DeepMind, 2025; Comanici et al., 2025), GPT (4.1, 5) (OpenAI, 2025a,b); and **(b) Open-weights models:** Mistral (Nemo, 3.2 24B) (Mistral, 2024, 2025), Moonshot Kimi K2 (Team et al., 2025), Qwen 3 (32B, 235B) (Team, 2025), Llama (3.1 8B, 3.3 70B, 4 Maverick) (Grattafiori et al., 2024; AI, 2025), DeepSeek-V3 (DeepSeek-AI, 2024), and GPT-OSS 120B (Agarwal et al., 2025). Open-weights models are queried via the default OpenRouter API, and proprietary models via official provider APIs.

Prompting Strategy To isolate core diagnostic capacity, the primary benchmark evaluates models using zero-shot prompting without intermediate reasoning chains.[11] This ensures the benchmark measures a model's direct capabilities. Internal reasoning traces are permitted for all reasoning models, e.g., Gemini 3 Pro and GPT-5.

4.2 Evaluation Metrics

We evaluate performance across both tasks using the **Correctness Rate (CR)**, defined as the percentage of correctly answered items:

$$\text{CR} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i] \quad (1)$$

where N is the total number of questions, and $\hat{y}_i, y_i \in \mathcal{O}_i$ represent the model's predicted selection and the ground-truth label from the multiple-choice option space $\mathcal{O}_i = \{A, B, C, D\}$, respectively.

[7] Platforms, compensation, and worker filtering protocols are detailed in Appendix C.

[8] see Appendix C.4 for details.

[9] Details are provided in Appendix C.

[10] Model profiles are summarized in Appendix Table 12.

[11] Complete prompt templates and examples are provided in Appendix I and J, respectively.

5 Empirical Analysis

5.1 Performance Asymmetry & Scale Dynamics

As shown in Table 3, most LLMs achieve high accuracy ($\geq 80\%$) on the Retrospective task, reliably identifying mental state primitives across domains. Conversely, models consistently fail the Prospective Diagnostic Forecasting task; most score near or below the random baseline ($\leq 25\%$). This sharp drop underscores a fundamental reasoning deficit—while models possess the capacity to label underlying mental states, they struggle to predict dialogue trajectories in the absence of local contextual cues. A lone exception is Gemini 3 Pro, which maintains robust Functional ToM capabilities by averaging $\approx 83\%$ accuracy across all three target domains.

Pearson's Chi-Square tests[12] confirm this Retrospective–Prospective gap is universally significant across all 14 evaluated models and 3 domains ($p < 0.01$). Cramér's V effect sizes reveal massive degradation, particularly among frontier models like DeepSeek V3 and GPT-4.1 ($V > 0.80$), proving this is a structural reasoning deficit rather than random variance.

Performance scales consistently with overall model parameters both across and within specific proprietary families. High-capacity frontier models (the Gemini family, Qwen 235B, Kimi K2) serve as the top overall performers. This pattern is mirrored inside isolated families, where Mistral 24B outperforms Mistral Nemo, and GPT-5 outscores GPT-4.1. An isolated anomaly occurs within the Llama family, where the compact Llama 8B surprisingly outperforms its larger counterparts (70B and 4 Maverick).

5.2 Human Baseline and Expert Ceiling

On the Retrospective task, the low-performing Lay Participant baseline highlights that non-experts are significantly outperformed by almost all evaluated LLMs, consistent with prior work (Strachan et al., 2024). This reflects the intensive linguistic overhead of parsing specialized, high-stakes dialogue, where human concentration frequently wavers across lengthy annotations.

On the context-free Prospective task, a professional counselor and an expert dialogue coder achieved 100% and 80% CR scores, respectively, while Gemini 3 Pro matched the expert baseline with an 80% CR on those identical items. The near-perfect expert ceiling demonstrates that the context-free formulation remains objectively unambiguous and solvable for domain specialists. The massive performance gap between the expert ceiling and average LLM accuracy ($\leq 25\%$) points to a profound functional-ToM reasoning deficit, converging with recent independent findings of an inference-to-action gap in LLMs (Gu et al., 2024; Li et al., 2025).

[12]Full statistical reporting is provided in Appendix H.1.

Model	Retrospective			Prospective		
	MI	ESC	PFG	MI	ESC	PFG
Human	32.33	24.33	24	-	-	-
Gemini 3 P	97.71	99.15	98.78	83.09	80.46	86.96
GPT-5	93.46	95.20	95.28	14.71	16.26	16.85
GPT-4.1	93.14	94.07	94.40	5.88	12.81	10.87
GPT-OSS 120B	75.49	88.98	85.25	7.35	15.76	6.52
Gemini 2.5 P	92.48	97.18	94.99	38.24	29.06	19.57
Gemini 2.5 F	94.77	95.48	94.40	26.47	25.12	17.93
Kimi K2	92.16	95.48	93.22	22.06	25.62	13.04
DeepSeek v3	96.73	97.74	98.82	10.29	11.82	10.33
Mistral 24B	61.11	68.64	74.34	5.88	4.93	7.6
Mistral Nemo	42.16	52.26	56.64	6.62	14.29	5.44
Qwen 235B	95.10	97.18	97.94	17.65	29.06	17.93
Llama 70B	77.45	84.75	87.91	18.42	19.70	7.07
Llama 4	87.58	87.57	79.28	9.56	11.82	5.6
Llama 8B	30.39	27.12	34.51	6.62	16.75	10.33

Table 3: Model benchmarking results reporting Correctness Rate (CR) across Retrospective and Prospective tasks alongside a non-expert human baseline.

5.3 Cross-Domain & Cross-Task Generalization

Evaluating cross-domain correlations reveals strong, statistically significant linear connections between all interaction domains across both evaluation paradigms (Retrospective Pearson $r = 0.978$; Prospective Pearson $r = 0.949$; Table 4). This indicates that a model's primary capability to parse a given ToM dimension generalizes exceptionally well across distinct conversational contexts. In contrast, cross-task correlations are weak and statistically non-significant under a linear assumption ($p = 0.35$), though they display a moderate monotonic trend (Spearman $\rho = 0.67$). This suggests that cross-domain proficiency within a task does not imply cross-task synchronization on DialToM.

Domain	Retrospective			Prospective		
	MI	ESC	PFG	MI	ESC	PFG
MI	1.0	0.978*	0.965*	1.0	0.966*	0.941*
ESC	0.978*	1.0	0.987*	0.966*	1.0	0.939*
PFG	0.965*	0.987*	1.0	0.941*	0.939*	1.0

Table 4: Pearson correlations across conversation domains. (*: p<0.001, †: p<0.01, ‡: p<0.05)

5.4 Attribute-Level & Semantic Analysis

Fine-grained attribute mapping (Table 5) demonstrates that nearly all evaluated LLMs perform poorly when isolating the newly introduced *Trust* attribute compared to foundational BDI attributes. For instance, high-performing models like Qwen 235B and Gemini 2.5 Pro experience notable performance drops specifically on Trust items. Conversely, superior tracking metrics on Beliefs and Intentions align with prior literature indicating strong modeling biases toward basic factual perspective-taking (Zhu et al., 2024). These findings validate the inclusion of Trust as an essential but fundamentally underrepresented dimension in current LLM-ToM diagnostics.

Finally, a complete semantic and structural evaluation comparing LLM outputs directly against open-ended human gold references, incorporating standard NLP metrics (such as BLEU, ROUGE-L, and BERTScore) and an LLM-as-a-Judge protocol (evaluating Propositional Agreement, Causal Attribution, and Specificity Alignment), is deferred to Appendix G. Briefly, while current language models show moderate baseline similarity to human references, they consistently exhibit coarse, low-specificity reasoning compared to human baseline.

Model	Retrospective					
	B	D	I	E	K	T
Gemini 3 P	97.65	99.36	98.04	100	98.04	98.34
GPT-5	95.40	97.47	96.10	98.86	97.42	84.15
GPT-4.1	97.13	97.47	96.10	99.43	96.77	78.14
GPT-OSS 120B	88.51	89.87	94.16	91.43	91.61	83.61
Gemini 2.5 P	95.98	94.94	94.81	99.43	96.13	89.07
Gemini 2.5 F	95.98	97.47	98.70	96.57	95.48	86.34
Kimi K2	97.13	90.51	95.45	98.86	94.84	85.79
DeepSeek v3	98.28	98.73	95.45	99.43	98.06	96.72
Mistral 24B	66.67	68.35	74.68	54.86	78.06	68.85
Mistral Nemo	50.57	52.53	63.64	37.14	54.19	48.09
Qwen 235B	97.70	98.73	99.35	99.43	98.71	87.98
Llama 70B	82.76	91.77	87.66	84.57	81.29	74.86
Llama 4	79.17	90.26	89.26	86.55	83.89	81.50
Llama 8B	31.61	30.38	35.06	29.14	32.26	26.23

Table 5: Attribute-level Correctness Rate (CR) for the Retrospective task across six mental state dimensions.

6 Ablations & Diagnostic Probing

To isolate the mechanisms driving the sharp drop in the Prospective Diagnostic Forecasting performance, we run three targeted diagnostic interventions probing on the main benchmark evaluation set (§ 5): (1) model reliance on spurious surface-level shortcuts; (2) robust state-driven reasoning versus static memorization; and (3) structural logic transferability via reasoning injection.

6.1 Distractor Complexity & Out-of-Context Exploitation

A core structural hypothesis for poor model performance on the Prospective task is its rigorous multiple-choice design: the adversarial distractors share deep semantic boundaries with the ground truth, requiring highly precise social reasoning. We contrast this *Hard Set* against a generated *Easy Set* by replacing adversarial distractors with unrelated out-of-context trajectories sampled randomly across dialogue frames—a distractor generation approach employed in Shinoda et al. (2025). This variant introduces distinct topical out-of-distribution shifts, allowing models to identify the correct option purely by leveraging superficial lexical shortcuts or topic coherence without demonstrating genuine functional-ToM reasoning.

As shown in Table 6, lowering the diagnostic barrier boosts model accuracies by 44 to 83 percentage points. Mann-Whitney U tests confirm the Easy–Hard gap is statistically significant across all 13 models and 3 domains ($p < 0.0001$ for 38 of 39 comparisons[13]), with large to very large effect sizes ($r \in [-0.85, -0.40]$). This confirms that the Hard Set successfully eliminates lexical shortcuts, forcing models to rely on state-driven trajectory reasoning alone.

Model	Prospective-Easy		
	MI	ESC	PFG
GPT-5	$94.85_{+80.15}$	$91.63_{+75.37}$	$61.96_{+50.54}$
GPT-4.1	$91.18_{+85.29}$	$90.64_{+77.83}$	$64.13_{+58.70}$
GPT-OSS 120B	$86.76_{+77.21}$	$78.33_{+61.58}$	$55.43_{+48.91}$
Gemini 2.5 P	$94.85_{+56.62}$	$95.07_{+66.01}$	$67.93_{+48.37}$
Gemini 2.5 F	$92.65_{+66.18}$	$91.13_{+66.01}$	$61.41_{+51.63}$
Kimi K2	$93.38_{+71.32}$	$89.66_{+64.04}$	$65.76_{+62.50}$
DeepSeek v3	$93.38_{+83.09}$	$87.68_{+75.86}$	$61.96_{+60.33}$
Mistral 24B	$86.03_{+80.15}$	$82.76_{+77.83}$	$65.76_{+63.59}$
Mistral Nemo	$77.21_{+70.59}$	$66.01_{+51.72}$	$56.52_{+55.43}$
Qwen 235B	$88.24_{+70.59}$	$89.66_{+60.59}$	$57.61_{+51.09}$
Llama 70B	$86.76_{+68.34}$	$88.18_{+68.47}$	$58.15_{+52.72}$
Llama 4	$89.71_{+80.15}$	$90.64_{+78.82}$	$69.02_{+64.67}$
Llama 8B	$75.0_{+68.38}$	$66.5_{+49.75}$	$50.0_{+44.57}$

Table 6: Performance on the Easy Set of the Prospective task. Blue subscripts indicate the performance gains compared directly to the baseline Hard Set.

6.2 Robustness vs. Memorization Verification

Gemini 3 Pro's uniquely high baseline performance on the context-free Prospective task raises a key diagnostic question: is the model executing genuine Functional-ToM forecasting, or is it exploiting benchmark memorization data leaks? To verify its

[13] Full statistics in Appendix H.2

robustness, we engineer an adversarial counterfactual intervention by prompting GPT-4o to replace the ground-truth mental-state profiles $\mathcal{S}$ with contradictory counterfactual mental-state attributes $\mathcal{S}'$ while preserving options (Appendix I). Under this counterfactual environment, the original ground-truth trajectory option is rendered logically invalid, forcing a robust reasoner to select "None of the Above" (NOTA). We test Gemini 3 Pro together with a representative cohort of the four strongest and four weakest models.

As summarized in Table 7, Gemini 3 Pro maintains high accuracy when identifying NOTA selections under counterfactual conditions (CR $\in [80.15\%, 95.11\%]$). This resilience disproves data-leak explanations and establishes its robustness under distributional shift. High counterfactual compliance across most baseline architectures—excluding specific model gaps like Llama 8B—confines the core task barrier to robust relational reasoning rather than simple data memorization. Critically, Gemini 3 Pro serves as an empirical proof of concept that deep social-causal trajectory grounding is attainable.

Model	MI	ESC	PFG
Gemini 3 P	80.15	94.09	95.11
Gemini 2.5 P	70.59	87.68	88.59
Gemini 2.5 F	48.53	76.35	74.46
GPT-5	61.03	86.7	90.22
Kimi K2	80.88	95.07	95.65
Mistral Nemo	59.56	73.89	88.59
Mistral 24B	97.06	99.01	99.46
Llama 4	67.65	78.82	55.98
Llama 8B	37.5	37.93	51.09

Table 7: Robustness vs. Memorization via Counterfactual Intervention. Correctness Rate (CR) for identifying "None of the Above" (NOTA).

6.3 Teacher-Student Reasoning Injection

To test if Gemini 3 Pro's performance stems from its internal logic, we conduct a teacher-student reasoning injection probe targeting weaker student models (Mistral 24B and Mistral Nemo), using GPT-5 as a baseline teacher on the MI subset of Prospective tasks. To ensure that performance shifts are not due to raw label leakage, we manually verify and truncate the final 5 sentences of every reasoning trace, ensuring that all explicit final answer declarations are completely removed. The intermediate steps naturally retain the teacher model's chain-of-elimination text, forcing the student model to parse the reasoning by itself.

The results (Table 8) reveal a striking performance gap. Injecting Gemini 3 Pro's reasoning traces[14] triggers a statistically significant performance surge of ≈ 76 points over original baseline scores (McNemar's test, $p < 0.0001$). In contrast, the negligible 4-point shift yielded by GPT-5's traces is statistically indistinguishable from the un-injected baseline ($p \geq 0.27$). This sharp divergence suggests that Gemini 3 Pro utilizes a highly precise, structured evaluation heuristic that, when externalized, serves as an effective in-context reasoning guide to boost the performance of weaker models.

Teacher Model	Mistral 24B	Mistral Nemo
Gemini 3 Pro	82.35	83.82
GPT-5	9.56	10.29

Table 8: Student model performance when injected with reasoning traces from Gemini 3 Pro vs. GPT-5.

7 Conclusion

We introduced DialToM, a human-verified Theory of Mind (ToM) benchmark designed to evaluate the critical gap between passively inferring mental states and actively utilizing them for behavior prediction in LLMs. Grounded in naturalistic human dialogues and incorporating a novel Trust attribute, DialToM captures the relational dynamics governing human social intelligence.

Evaluation across 14 LLMs exposed a stark *reasoning asymmetry*. While models demonstrate exceptional proficiency at Retrospective mental state inference ($\geq$80% accuracy), they collapse on our context-free Prospective Diagnostic Probe, mostly scoring near random chance ($\leq$25%). Rigorous adversarial ablations demonstrated that standard context-driven evaluation tasks allow models to bypass genuine ToM reasoning by exploiting superficial lexical shortcuts and topical coherence. Conversely, Gemini 3 Pro emerged as an exception, establishing a powerful empirical proof of concept that robust, state-driven trajectory grounding is attainable. Furthermore, our Teacher-Student reasoning injection study verified that Gemini 3 Pro's advanced performance stems from an explicit, transferable internal logic that can be successfully externalized in-context to bridge the Functional ToM capability gap in weaker models without requiring parametric updates.

[14] Excerpts are provided in Appendix J.3.

8 Limitations

Despite the diagnostic rigor of the DialToM framework, several limitations remain. First, while our introduction of the *Trust* attribute represents a critical step toward capturing relational rapport, its operational scope is bounded by the specific counselor-client and persuader-persuadee dynamics of our source corpora. Consequently, it may not fully capture broader, cross-cultural, or highly idiosyncratic definitions of interpersonal trust that manifest in casual, non-professional social settings. Second, our token normalization pipeline enforces a uniform sliding context window (150–200 words) to optimize computing efficiency and maintain semantic consistency. While highly effective for our tasks, this design naturally truncates long-range dependencies.

Third, our data curation pipeline relies on GPT-4o for generating distractors and base mental-state inferences. Although these materials underwent rigorous multi-stage manual filtering via our human-in-the-loop verification framework, we acknowledge the residual risk that subtle, systemic artifacts, or linguistic styles may persist in the candidate option distributions. Fourth, our crowd worker baseline was structurally restricted to the Retrospective Inference task. Because the context-free Prospective task is cognitively demanding and requires highly specialized domain expertise, it proved too functionally complex to serve as an equitable, unbiased baseline for average, untrained crowd workers. Furthermore, our commitment to high-fidelity adversarial distractors required aggressive Dawid-Skene filtering, leading to lower data retention rates on the Prospective task (e.g., 34% in the MI domain). While this limits the sheer volume of our benchmark, we accepted this trade-off to guarantee that the surviving evaluation set is free of semantic ambiguity and rigorously tests Functional-ToM reasoning.

Additionally, while we employ an LLM-as-a-judge protocol for our supplementary semantic alignment analysis (Appendix G), we rely on the established zero-shot evaluative capabilities of frontier models and leave the collection of parallel human meta-judgments for future work. Consequently, these alignment scores are treated as qualitative explorations rather than primary empirical claims.

Our findings are bounded by the current landscape of proprietary and open-weight models available at the time of testing. As scaling laws continue to evolve, the severe inference-to-action performance gaps observed here may shift in response to novel architectures. Finally, Gemini 3 Pro—which serves as the leading baseline in our Prospective Diagnostic evaluation—was discontinued as a preview model on March 9, 2026, and has since been replaced by Google with Gemini 3.1 Pro (Preview) (Google, 2026). While Gemini 2.5 Pro was independently included in our main benchmarking suite, future reproducibility of Gemini 3 Pro's reported results cannot be guaranteed. We have retained all model outputs from our original experiments and will release them upon request to support independent verification.

9 Ethical Considerations

The curation, annotation, and release of DialToM adhered strictly to established ethical guidelines and were formally approved by the Institutional Review Board (IRB) of [Anonymized Institution] under Approval Number [Anonymized Number].

Data Governance and Privacy DialToM is built entirely from public, open-access datasets (AnnoMI, ESConv, and PersuasionForGood). We complied with all original licenses and user agreements. The source transcripts were fully anonymized by their original authors. We conducted a secondary manual audit to ensure that any remaining names, locations, temporal references, or sensitive behavioral disclosures were completely scrubbed or masked to eliminate re-identification risks.

Crowdsourcing Labor and Compensation Professional verifiers and crowd workers from Amazon Mechanical Turk (AMT) and Baidu Crowdsourcing were compensated at an average hourly rate of $12.50 USD, matching or exceeding local living wage standards for digital labor. All participation was strictly voluntary; workers were provided with transparent task descriptions and maintained the right to withdraw at any stage without penalty or loss of accrued compensation.

Dual-Use and Potential Misuse Advanced Functional ToM models present dual-use risks. Malicious actors could leverage automated social-causal reasoning to optimize social engineering campaigns, scale targeted disinformation, or manipulate vulnerable individuals in counseling or negotiation settings. However, concealing evaluation techniques does not prevent these dangers; it

merely blinds the AI safety community to current model vulnerabilities. We release DialToM as a public, non-commercial diagnostic auditing tool to allow researchers to measure, track, and align these social capabilities safely.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Meta AI. 2025. Llama 4.

Ian A. Apperly and Stephen A. Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4):953–970.

Erika Blacksher, Charlene Nelson, Emily Van Dyke, Abigail Echo-Hawk, Deborah Bassett, and Dedra Buchwald. 2016. Conversations about community-based participatory research and trust: "we are explorers together". *Progress in Community Health Partnerships: Research, Education, and Action*, 10(2):305–309.

Michael Bratman. 1987. *Intention, plans, and practical reason*. Stanford Univ Center for the Study.

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.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and Others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20.

Google DeepMind. 2025. Gemini 3.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Daniel C. Dennett. 1978. Beliefs about beliefs [p&w, sr&b]. *Behavioral and Brain Sciences*, 1(4):568–570.

Kanishk Gandhi, Jan-Philipp Fraenken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 13518–13529. Curran Associates, Inc.

Google. 2026. Migrate from Gemini 3 Pro Preview to Gemini 3.1 Pro Preview before March 9, 2026. Google AI Developers Forum. Accessed: May 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and Others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. 2024. SimpleToM: Exposing the gap between explicit tom inference and implicit tom application in llms. *arXiv preprint arXiv:2410.13648*.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

P. A. Hancock, Theresa T. Kessler, Alexandra D. Kaplan, Kimberly Stowers, J. Christopher Brill, Deborah R. Billings, Kristin E. Schaefer, and James L. Szalma. 2023. How and why humans trust: A meta-analysis and elaborated model. *Frontiers in Psychology*, 14.

Mark K. Ho, Rebecca Saxe, and Fiery Cushman. 2022. Planning with theory of mind. *Trends in Cognitive Sciences*, 26(11):959–971.

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

Mengfan Li, Xuanhua Shi, and Yang Deng. 2025. Rectom: A benchmark for evaluating machine theory of mind in llm-based conversational recommender systems. *Preprint*, arXiv:2511.22275.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709.

Mistral. 2024. Mistral nemo.

Mistral. 2025. Mistral small 3.2 24b.

OpenAI. 2025a. Gpt-4.1.

OpenAI. 2025b. Gpt-5.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D Weisz, and Murray Campbell. 2025. Position: Theory of mind benchmarks are broken for large language models. In *International Conference on Machine Learning*, pages 82091–82130. PMLR.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. ToMATO: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. *arXiv preprint arXiv:2501.08838*.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, and 1 others. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, pages 1–11.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and Others. 2025. Kimi K2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Qiaosi Wang, Xuhui Zhou, Maarten Sap, Jodi Forlizzi, and Hong Shen. 2025. Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective. *Preprint*, arXiv:2504.10839.

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.

Susan E Whitely. 1983. Construct validity: Construct representation versus nomothetic span. *Psychological bulletin*, 93(1):179.

H Wimmer. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

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.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-MI: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2025. Towards dynamic theory of mind: Evaluating LLM adaptation to temporal evolution of human states. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24036–24057, Vienna, Austria. 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.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. 2024. Language models represent beliefs of self and others. In *Forty-first International Conference on Machine Learning*.

A Theoretical Background: Literal vs. Functional Theory of Mind

The distinction between *Literal* and *Functional* Theory of Mind (ToM) addresses a critical methodological flaw in machine ToM evaluations, as for-

malized by Riemer et al. (2025). Traditional benchmarks overwhelmingly focus on what the authors define as *Literal Theory of Mind*, which is "the ability to predict the behavior of others." While modern LLMs often achieve high performance on these literal probes, Riemer et al. (2025) caution that such metrics are only informative when models exhibit self-consistent reasoning. In modern LLMs, this consistency cannot be taken for granted. Models can demonstrate strong literal ToM performance that is ultimately illusory, as making accurate predictions about another agent often has "nothing to do with its own chosen course of behavior."

To overcome these limitations, Riemer et al. (2025) introduce the concept of *Functional Theory of Mind*, defined as "the ability to adapt to agents in-context following a rational response to their behavior." Transitioning to a functional framework shifts the benchmarking goal from mere latent state attribution to active forecasting, specifically evaluating "an agent's ability to choose appropriate actions based on the behavior of another agent." DialToM directly operationalizes this theoretical division by structurally partitioning its evaluation workflow into distinct Retrospective and Prospective task suites, enforcing the strict experimental controls necessary to verify that an LLM can actively translate literal inferences into functional behavioral decisions.

A.1 Methodological Trade-offs of the State-Driven Probing Design

Our choice to withhold conversational history during the Prospective task introduces an important methodological trade-off regarding task artificiality versus diagnostic rigor.

Diagnostic Rigor While Riemer et al. (2025) conceptualize Functional ToM through interactive, context-heavy multi-agent environments, context-driven forecasting allows models an unwanted degree of freedom. In a multi-turn setup where the history remains visible, LLMs can bypass the complex structural representation of mental states entirely. Instead, the models can achieve high prediction accuracy by relying on local dialogue heuristics, conversational momentum, or superficial lexical matching with the immediate preceding tokens.

By implementing a *context-free*, *state-driven* configuration, DialToM establishes a strict causal control. Removing the text history breaks the model's reliance on superficial pattern-matching and forces a direct, isolated mapping from the latent mental-state profile to the trajectory choice. If a model successfully completes the task under this setup, it serves as empirically unambiguous proof that the system is treating the explicit mental state primitives as functional, decision-guiding variables rather than passive labels.

Theoretical Basis for Functional ToM Probing From a psychometric perspective, validating a cognitive construct requires testing not only whether an internal state can be inferred from observations, but also whether that state systematically guides downstream behavioral prediction (Cronbach and Meehl, 1955; Whitely, 1983). Existing ToM benchmarks primarily operationalize the cognitive pipeline by mapping contextual evidence ($\mathcal{H}$) to latent mental-state representations ($\mathcal{S}$). DialToM complements this paradigm by establishing a strict structural symmetry, isolating the inverse functional mapping from explicit mental-state configurations ($\mathcal{S}$) to downstream behavioral trajectories ($\mathcal{A}$).

By hiding conversational history, our framework implements a controlled diagnostic setting standard in cognitive evaluation and psychometric testing (e.g., standardized personality and behavioral questionnaires), where explicit traits or internal state profiles are evaluated in isolation to determine their predictive impact on downstream actions. Under standard principles of construct representation (Whitely, 1983) and nomological validity (Cronbach and Meehl, 1955), if mapping contextual evidence to a latent state constitutes operational evidence of state representation, then systematically mapping that same state representation to predictable behavioral outcomes constitutes operational evidence of functional state utilization. Rejecting this context-free alignment as an invalid abstraction would fundamentally undermine the construct validity of standardized psychometric assessments, which routinely isolate internal state profiles to evaluate behavioral trajectories.

Functional Abstraction and Task Validity While this design departs from natural human interaction—where predictions are dynamically intertwined with conversational context rather than derived from a static profile—the concrete examples in Appendix J demonstrate that this abstraction does not undermine construct validity. Across all three conversational domains, candidate trajectories are topically coherent and share deep surface-

form similarities. Consequently, they cannot be distinguished via lexical shortcuts or topic coherence. The correct trajectory is identifiable *only* by reasoning about which continuation remains strictly consistent with the speaker's specific mental-state configuration. This design directly operationalizes functional ToM as defined: evaluating the active mapping from explicit mental-state representations to behavioral predictions, rather than surface-level pattern matching against conversational cues.

Empirical Validation via Human Baseline We counter the potential risk of semantic ambiguity or unfair artificiality through our expert validation protocol. If our context-free isolation rendered the mapping inherently ambiguous or textually disjointed, human performance would naturally degrade. However, as shown in Section 5.2, our professional domain expert achieved a perfect 100% score under these exact context-free conditions. This empirical ceiling proves that the isolated mental profiles contain complete, self-contained, and entirely unambiguous causal information capable of guiding correct functional forecasting. Thus, the trade-off favors diagnostic precision without sacrificing the underlying validity of the task.

B Detailed Breakdown of Conversational Source Corpus

To ground DialToM in naturalistic human interaction, we leverage three distinct, publicly available conversational datasets representing specialized dyadic communication profiles. This section provides the granular background, structural annotations, and linguistic properties of the source material prior to our token normalization and window segmentation pipelines.

AnnoMI Sourced from Wu et al. (2022), this corpus is uniquely designed to study Motivational Interviewing (MI)—a collaborative, goal-oriented style of communication with particular attention to the language of change, widely utilized in psychotherapy and counseling domains. The raw dataset contains 133 full-length counselor-client video transcripts spanning diverse behavioral topics (e.g., smoking cessation, vaccination hesitancy, and dietary modifications). Every utterance turn is expert-annotated with domain-specific behavioral codes detailing counselor strategies (such as *open questions*, *reflections*, or *persuasion with permission*) and client negotiation postures (*change talk* versus *sustain talk*). We utilize all 133 sessions, treating both high-quality and low-quality execution traces as diverse environmental conditions for Retrospective and Prospective tasks.

ESConv The Emotional Support Conversation corpus (Liu et al., 2021) targets the rich cognitive and affective dynamics of empathy and supportive social interaction. It comprises 1,300 multi-turn interaction sessions between human helpers and individuals seeking emotional support. Each interaction is structurally categorized across a taxonomy of help-seeker problems (e.g., academic stress, workplace conflicts, relational distress) and underlying psychological configurations (comprising explicit tags for the seeker's ambient emotion, specific situation, and lived experience). Turns are structurally organized according to progressive emotional support stages: *exploration*, *comforting*, and *action planning*. This progression relies heavily on an agent's ability to recursively model the fluctuating internal distress levels and latent expectations of the speaker.

PersuasionForGood Developed by Wang et al. (2019), this dataset focuses on strategic interpersonal influence and persuasive dialogue. The source corpus captures 1,017 multi-turn natural conversations between human pairs assigned to an asymmetric social task: a *persuader* must convince a *persuadee* to donate a portion of their task compensation to a designated international charity. The data incorporates multi-layered annotations mapping specific communicative strategies used by the persuader (e.g., *logical reasoning*, *emotional appeal*, *credibility framing*) alongside corresponding reactions from the persuadee. Crucially, the dataset provides participant-specific metadata, including self-reported personality profiles (Big Five inventory) and baseline altruism indices, which enrich the cognitive dependencies involved in evaluating strategic interaction trajectories.

C Detailed Data Annotation Process

The annotation study was approved by the Institutional Review Board of [Anonymized Institution] (IRB Approval Number: [Anonymized IRB Number]).

C.1 Crowdsourcing Methodology

Platform Selection We chose separate platforms for our tasks by considering worker-task compatibility. To ensure that Lay Participants who author

Statistic	MI	ESC	PFG	Overall
	Original			
Session Count	133	1300	1017	2450
Word Count per session	1232.97	520.61	272.78	-
Turn Count per session	35.47	11.26	9.39	
Word Count per turn	34.76	46.22	29.07	-
	Final			
Context Count	865	3215	1863	5943
Word Count per context	188.2	205.02	210.8	-
Turn Count per context	5.3	4.48	5.4	-
Word Count/Turn	35.5	45.5	39.1	-
Retrospective Infer. Qns.	5190	19,290	11,178	35,658
Prospective Forecast. Qns.	865	3215	1863	5943
Mental State Inference	100	100	100	300

Table 9: Pre- and post-segmentation statistics of the DialToM source corpora. The **Original** rows report session-level properties of the raw source datasets; the **Final** rows report context-level properties after sliding window segmentation. Each final context yields 6 Retrospective questions (one per BDIEKT attribute) and 1 Prospective question, totalling ≈36,000 evaluation items across the full benchmark suite.

free-form mental state inferences were not exposed to AI-generated inferences before, we recruited Lay Participants for authoring mental state inferences from a different worker pool (AMT) than the verifiers or Lay Participants forming the baseline (Baidu). For the high-volume tasks containing significantly more evaluation samples, we established a contract for mutual benefit with the Baidu Data Crowdsourcing Platform to scale operations efficiently.

Worker Selection To gather high-quality annotations, we filter workers through a selective worker qualification process. In this process, across both platforms AMT and Baidu, we ask the workers to participate in a replication of the Retrospective task of DialToM. Workers were provided with a conversation excerpt, and are needed to answer a multiple-choice question for selecting the appropriate mental state of the recipient. We selectively chose workers that score at least 80% on this test.

C.2 Participant Roles

We categorized crowd workers into two distinct groups based on the nature of the task:

1. **Professional Verifiers (Baidu):** A cohort of 22 individuals who performed the Verification of AI-generated Mental State Inferences and Dialogue Trajectories. These verifiers were compensated at a competitive hourly rate equivalent to $12.50/hour, exceeding fair local wage standards. Their high-quality output was aggregated using the Dawid-Skene algorithm to ensure the final dataset's integrity.

2. **Lay Participants (Baidu/AMT):** These individuals provided the Human Baseline by completing the benchmarking tasks without prior exposure to the correct labels. Their performance serves to quantify the intrinsic difficulty and domain-specific nuance of DialToM.

C.3 Quality Control Methods

Automated We employ the Dawid-Skene (DS) algorithm (Dawid and Skene, 1979) to model annotator skills and verify whether their future annotations remain consistent with previous annotations or not. If the quality diverges significantly, we consider that annotator to be of low-quality and request for a swap of annotations from Baidu.

Manual We also conduct a high-level quality check manually, by going through all samples and verifying if the annotations are as per what we, as researchers with great deal of experience in this field, would expect.

C.4 Human Benchmarking & Expert Ceiling Construction

Human Benchmarking To quantify the intrinsic difficulty of the Retrospective Inference task, three Lay Participants each completed tasks across a stratified subset of 900 evaluation contexts (300 per conversational domain). This data establishes a robust crowd baseline reflecting average human performance. Notably, this crowd benchmarking was restricted exclusively to the Retrospective Inference tasks. This choice was informed by the extreme cognitive load and difficulty of the Prospective Diagnostic Forecasting tasks, which require expert-level social reasoning and long-term logical consistency under a context-free configuration that frequently exceeds the performance bounds of average crowd workers.

Expert Ceiling Validation To validate the construct validity of the context-free Prospective task and establish a definitive performance ceiling, we recruited a professional licensed counselor specializing in Motivational Interviewing (MI) to complete a stratified sample of 20 Prospective items drawn from the MI domain. These items systematically spanned all difficulty tiers, stratified explicitly by baseline LLM failure rates. A co-author

Task	MI				ESC				PFG			
	O1	O2	O3	O4	O1	O2	O3	O4	O1	O2	O3	O4
Retrospective	0.88	0.88	0.9	0.91	0.94	0.95	0.95	0.96	0.87	0.96	0.97	0.97
Prospective	0.37	0.57	0.62	0.67	0.26	0.84	0.84	0.86	0.24	0.82	0.83	0.92

Table 10: Consensus-based verification of task options. Values represent the agreement scores for Professional Verifiers distinguishing the ground-truth (O1) from adversarial distractors (O2–O4).

with extensive MI dialogue assessment experience subsequently completed an independent second-pass evaluation blind to both the model outputs and ground-truth labels. Together, their performance establish a robust expert consensus ceiling.

C.5 Gold Mental-State Inferences Construction

Since our gold mental-state inference task is much more complex than other tasks, it thus requires close to native proficiency of speaking, understanding and writing English. Thus, we only chose 9 annotators from AMT that were not only able to reach an optimal score on our qualification test, but those who also held a Master's qualification and were from predominantly English speaking countries in the West, such as United States and the United Kingdom. These annotators provided subjective, free-form mental state inferences for 300 contexts (100 per conversation domain) to evaluate the semantic alignment of AI-generated mental states.

D Discussions of Verification Results

Metric	Retrospective			Prospective		
	MI	ESC	PFG	MI	ESC	PFG
# Verification	400	400	400	400	300	300
# Retained	306	354	339	136	203	184
% Retention	76.5	88.5	84.75	34	67.67	61.33
Gwet's AC1	0.88	0.92	0.92	0.45	0.68	0.62

Table 11: Verification and label aggregation statistics for the Retrospective and Prospective tasks across all three domains. *# Verification* is the number of samples presented to annotators; *# Retained* is the number passing the Dawid-Skene aggregation threshold; *% Retention* is the corresponding retention rate; and *Gwet's AC1* measures inter-annotator agreement.

In Table 10, we illustrate the average verification statistics averaged over all annotators for both tasks. Following a systematic pattern, all Professional Verifiers found the "ground truth" (O1) much harder (lower score here implies confusion and increased difficulty of the task) to differentiate from the other options.

Table 11 shows the dataset statistics post human verification and Dawid-Skene label aggregation. We were able to retain $\approx 80\%$ of the data for the Retrospective and $\approx 60\%$ for the Prospective task.

Manual error auditing indicates that the lower consensus and data retention observed in the Prospective task is a direct structural consequence of our rigorous diagnostic probing design. The adversarial distractors were deliberately engineered to share high surface-level text and topic similarities with the ground-truth continuation. While this intentionally sets a high-difficulty threshold that challenges average professional verifiers and crowd workers, our expert ceiling check confirms that the items remain completely unambiguous to domain specialists.

E Evaluated Models

Table 12 summarizes the 14 LLMs evaluated in this study. They span parameter scales from 12B to 1T, covering both standard dense and Mixture-of-Experts (MoE) architectures. Closed-source proprietary models (Gemini family, GPT family) are queried via official provider APIs; open-weights models are queried via the OpenRouter API. Parameter sizes and knowledge cut-offs are reported where publicly disclosed by the respective developers; entries marked "-" reflect undisclosed specifications.

Note on Gemini 3 Pro Availability Gemini 3 Pro (`gemini-3-pro-preview`) was available via the Google Gemini API in preview mode at the time of our study. However, Google officially discontinued `gemini-3-pro-preview` on March 9, 2026, replacing it with Gemini 3.1 Pro Preview (`gemini-3.1-pro-preview`) (Google, 2026). As a result, direct replication of our Gemini 3 Pro results using the current Google API is no longer possible. Researchers wishing to verify our findings should note that Gemini 2.5 Pro is independently included in the main benchmarking results (Table 3) and

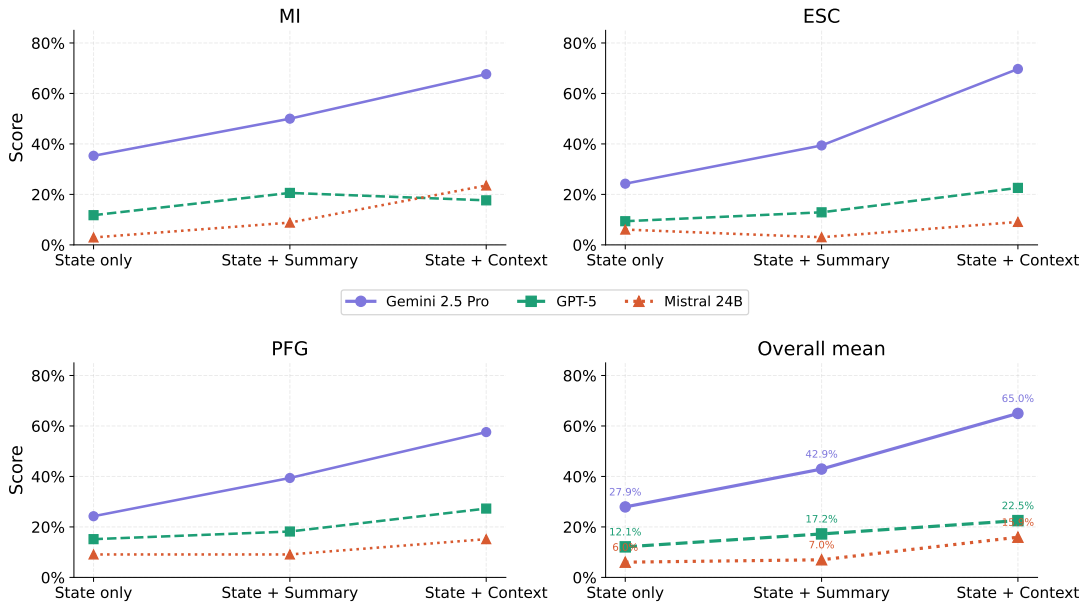

Figure 2: Performance gradient across context conditions

serves as the representative Gemini-family model in the Dialogue Context Degradation Analysis (§ F). All raw model outputs from our original Gemini 3 Pro benchmarking runs have been preserved and are available upon request.

F Dialogue Context Degradation Analysis

To ground DialToM's context-free, state-driven design within a broader evaluation continuum and directly contrast it with context-present frameworks like SimpleToM (Gu et al., 2024) and RecToM (Li et al., 2025), we systematically vary the availability of conversational context. We evaluate these structural configurations across a stratified diagnostic sample of 100 evaluation items (34 MI, 33 ESC, and 33 PFG), maintaining a completely paired experimental design across three representative models: Gemini 2.5 Pro, GPT-5, and Mistral 24B.

We define three specific experimental conditions representing progressive levels of contextual exposure:

- **State-Only**: The canonical DialToM setup. Models are provided exclusively with the latent mental state profile in isolation, establishing the most stringent diagnostic pressure by eliminating all conversational clues.
- **Summary + States**: A dialogue summary generated by Gemini 2.5 Pro[15] is prepended to the core mental state profile, introducing coarse-grained topical contexts while withholding fine-grained lexical patterns.
- **Conversation History + States**: The complete multi-turn conversation history is fully retained and prepended to the profile, closely mirroring conventional context-present evaluation configurations (e.g., (Gu et al., 2024; Li et al., 2025)).

Results are shown in Figure 2. A clear, monotonic performance gradient emerges across all three conditions, with accuracy increasing as contextual exposure scales up. The effect is pronounced for Gemini 2.5 Pro (27.9% → 42.9% → 65.0%), confirming that conversational context provides meaningful signal and that its absence imposes significant performance pressure. However, GPT-5 and Mistral 24B remain near or below the 25% random baseline even under full-dialogue history conditions (22.5% and 15.8%, respectively), demonstrating that their failure on the Prospective task cannot be attributed to the absence of dialogue history alone. These models lack the underlying functional-ToM reasoning capacity regardless of contextual scaffolding. Together, these findings validate DialToM's State-Driven Diagnostic Probe—the context-free condition maximizes diagnostic pressure while remaining objectively solvable for domain experts (§ 5.2), isolating true functional-ToM reasoning from surface-level context exploitation.

[15] See Appendix I for the prompt template.

Model	Parameter Size	Developer	Knowledge Cut-off
Mistral Nemo	12B	Mistral AI	April 2024
Mistral 3.2 24B	24B	Mistral AI	October 2023
Moonshot Kimi K2	1T, 32B active	Moonshot AI	∼October 2024
Qwen3 235B	235B, 22B active	Alibaba	June 2025
Llama 4 Maverick	400B, 17B active	Meta AI	August 2024
Llama 3.3 70B	70B	Meta AI	December 2023
Llama 3.1 8B	8B	Meta AI	December 2023
DeepSeek v3	671B, 37B active	DeepSeek	March 2025
Gemini 3 Pro†	-	Google	-
Gemini 2.5 Flash	-	Google	-
Gemini 2.5 Pro	-	Google	-
GPT-OSS 120B	120B	OpenAI	June 2024
GPT-4.1	-	OpenAI	-
GPT-5	-	OpenAI	-

Table 12: Overview of evaluated LLMs, categorized by architecture and access type. Technical specifications and knowledge cut-offs are provided where publicly available. Models above the horizontal rule are open-weights; models below are closed-source. †`gemini-3-pro-preview` was discontinued by Google on March 9, 2026; see § E for details.

G Semantic Comparison with Gold Reference

NLP Metrics To compare LLM-generated and human-written mental-state inferences, we utilize standard semantic similarity metrics commonly used in NLP evaluations. We adopt both linguistic metrics (**BLEU** and **ROUGE-L**) to measure the linguistic similarity between two texts, and semantic similarities (**BERTScore**) which compare between the generated text of the LLM inference and the gold reference by measuring the cosine similarity between embeddings of two pieces of text. We report the average similarity scores across all references to handle multiple gold references.

In Table 13, we compare the results of LLM-generated inferences with that of human-written gold references. The first row, GPT-4o, shows the similarities between the inferences used in the benchmarking tasks and human annotation.

We find LLMs are only moderately similar to human written gold inferences by semantic similarity (signified by moderate BERTScores) and very low overlap measured by linguistic similarity (signified by low BLEU, ROUGE-L scores). These overall low scores suggest the challenging aspect of generating free-form mental state inferences and the existing lexical diversity in generations while preserving semantic. Nevertheless, the proprietary LLMs (GPT and Gemini families) perform better relative to other models on writing tasks. This overall weak semantic similarity creates a dissonance with respect to the otherwise strong Retrospective task performance.

Model	**Gold Mental-State Reference**		
	BLEU	ROUGE-L	BERTScore
GPT-4o	9.44	0.31	0.44
GPT-5	8.56	0.32	0.41
GPT-4.1	9.71	0.33	0.43
GPT-OSS 120B	8.09	0.32	0.41
Gemini 2.5 P	8.90	0.32	0.44
Gemini 2.5 F	7.98	0.32	0.39
Kimi K2	7.77	0.29	0.38
DeepSeek v3	8.75	0.32	0.42
Mistral 24B	6.98	0.29	0.37
Mistral Nemo	6.21	0.29	0.36
Qwen 235B	8.05	0.30	0.40
Llama 70B	6.67	0.28	0.34
Llama 4	7.32	0.29	0.35
Llama 8B	5.30	0.25	0.31

Table 13: Linguistic alignment between AI and human-written inferences. Low linguistic overlap (BLEU/ROUGE-L) alongside moderate BERTScore values indicate high lexical diversity despite semantic preservation.

LLM-as-a-Judge Evaluation To complement the standard NLP metrics above, we conducted an exploratory LLM-as-a-Judge evaluation[16] using Gemini 3 Pro[17] as a zero-shot judge on LLM-generated inferences to measure how their mental-state inferences align with our human gold references. The LLM judge was provided with the original source dialogue alongside the human and AI inferences, ensuring it evaluated alignment based on the conversational context, rather than lexical overlap. The judge compared these samples across

[16]The evaluation prompt is provided in Appendix I.

[17]Gemini 3 Pro was discontinued on March 9, 2026 (Appendix E); all judge outputs are preserved and available upon request.

three dimensions on a 3-point Likert scale:

- **Propositional Agreement** (PA): Do the AI and human inferences express the same core proposition about the mental state?
- **Causal Attribution** (CA): Do the AI and human anchor the mental state to the same conversational cause or evidence?
- **Specificity Alignment** (SA): Are the AI and human inferences operating at the same level of specificity/granularity?

Model	Gold Mental-State Reference			
	PA (/3)	CA (/3)	SA (/3)	TA (/9)
GPT-5	1.78	2.03	1.53	5.33
GPT-4.1	1.73	2.03	1.61	5.37
GPT-OSS 120B	1.65	1.82	1.45	4.91
Gemini 2.5 P	1.66	1.99	1.55	5.19
Gemini 2.5 F	1.53	1.74	1.37	4.65
Kimi K2	1.58	1.94	1.45	4.97
DeepSeek v3	1.72	2.07	1.60	5.39
Mistral 24B	1.60	1.97	1.48	5.05
Mistral Nemo	1.25	1.40	1.13	3.79
Qwen 235B	**1.82**	**2.17**	**1.65**	**5.63**
Llama 70B	1.62	1.94	1.48	5.04
Llama 4	1.51	1.84	1.42	4.77
Llama 8B	1.16	1.52	1.19	3.87

Table 14: Dialogue-anchored semantic comparison using an LLM-as-a-Judge approach to directly measure causal alignment between written inferences. PA: Propositional Agreement; CA: Causal Attribution Agreement; SA: Specificity Alignment; TA: Total Alignment Score. PA, CA, SA are all scored on a Likert scale of 3, whereas TA = PA + CA + SA, and is scored out of 9.

Table 14 outlines the alignment statistics across all evaluated models (excluding Gemini 3 Pro as the evaluator). Qwen 235B yields the closest alignment to human references across all axes, peaking at a Causal Attribution (CA) score of 2.17/3. On average, models consistently display high Causal Attribution (CA), followed by moderate Propositional Agreement (PA), while exhibiting noticeable divergence on Specificity Alignment (SA).

This uniform pattern indicates that while LLMs successfully isolate and anchor their reasoning to identical conversational triggers as humans, they tend to generate coarse-grained semantic generalizations. In contrast, human annotators formulate highly individualistic, context-specific assertions. These alignment variances indicate that the low SA trends reflect structural differences in abstraction preferences rather than flaws in logical grounding, providing qualitative evidence of how AI-generated mental-state reasoning diverges from human baseline.

H Statistical Analysis

H.1 Inference-to-Action Gap

Table 15 reports the full Pearson's Chi-Square statistics (χ^2), p-values, and Cramér's V effect sizes for the Retrospective–Prospective performance gap across all 14 models and 3 domains. All 42 comparisons are statistically significant at $\alpha = 0.01$, and 41 of 42 at $\alpha = 0.001$, confirming that the inference-to-action gap is a robust, systematic phenomenon rather than an artifact of any particular model or domain.

Effect sizes span a wide range ($V \in [0.11, 0.90]$), reflecting meaningful variation in the magnitude of the gap across models. Gemini 3 Pro yields the smallest effect sizes ($V \in [0.24, 0.33]$), consistent with its uniquely narrow Retrospective–Prospective gap and its strong Functional ToM capability. At the opposite extreme, DeepSeek V3 exhibits the largest effect sizes ($V \in [0.87, 0.90]$), indicating the most pronounced asymmetry of any evaluated model. The single weakest result is Llama 3.1 8B on ESC ($\chi^2 = 7.19$, $p = 0.007$, $V = 0.114$); notably, this small effect does not reflect competence on the Prospective task but rather near-chance performance on both tasks, making the gap itself narrow by floor effect rather than by functional-ToM ability.

Effect sizes are generally consistent across domains within each model, with no systematic domain ordering, suggesting that the inference-to-action gap is a model-level property rather than a domain-specific artifact.

H.2 Distractor Complexity

Table 16 reports the full Mann-Whitney U statistics and rank-biserial effect sizes (r) for the Easy–Hard distractor comparison across all 13 models and 3 domains. All 39 comparisons are statistically significant ($\alpha = 0.05$). Effect sizes are large to very large throughout ($|r| \geq 0.39$ in all 39 cases, and ≥ 0.40 in 38 cases), with stronger effects in MI and ESC ($r \in [-0.853, -0.566]$ and $r \in [-0.788, -0.498]$, respectively) than in PFG ($r \in [-0.587, -0.397]$), indicating that persuasion-domain dialogue trajectories are inherently more topically distinct, modestly reducing the diagnostic pressure of adversarial distractors in that domain. The single weakest result in terms of statistical sig-

Model	MI			ESC			PFG		
	χ^2	p	V	χ^2	p	V	χ^2	p	V
Gemini 3 Pro	29.56	<0.001	0.259	61.78	<0.001	0.333	30.83	<0.001	0.243
Gemini 2.5 Pro	148.08	<0.001	0.579	295.81	<0.001	0.729	310.66	<0.001	0.771
Gemini 2.5 Flash	223.40	<0.001	0.711	299.86	<0.001	0.734	314.34	<0.001	0.775
GPT-5	270.48	<0.001	0.782	356.98	<0.001	0.801	331.87	<0.001	0.797
GPT-4.1	316.89	<0.001	0.847	368.29	<0.001	0.813	362.13	<0.001	0.832
GPT-OSS 120B	173.56	<0.001	0.627	291.37	<0.001	0.723	299.39	<0.001	0.757
Kimi K2	219.48	<0.001	0.705	296.66	<0.001	0.730	333.49	<0.001	0.799
Qwen3 235B	272.05	<0.001	0.785	295.81	<0.001	0.729	358.25	<0.001	0.828
DeepSeek V3	331.74	<0.001	0.866	423.20	<0.001	0.872	422.99	<0.001	0.899
Mistral 3.2 24B	114.26	<0.001	0.508	208.73	<0.001	0.612	209.82	<0.001	0.633
Mistral Nemo	53.74	<0.001	0.349	77.04	<0.001	0.372	129.75	<0.001	0.498
Llama 4 Maverick	244.15	<0.001	0.743	305.20	<0.001	0.740	258.86	<0.001	0.704
Llama 3.3 70B	133.65	<0.001	0.550	226.78	<0.001	0.638	320.01	<0.001	0.782
Llama 3.1 8B	28.66	<0.001	0.255	7.19	0.007	0.114†	35.01	<0.001	0.259

Table 15: Pearson's Chi-Square statistics (χ^2), p-values, and Cramér's V effect sizes for the Retrospective vs. Prospective performance gap, per model and domain. All comparisons are significant at $\alpha = 0.01$. †Llama 3.1 8B on ESC returns the smallest effect ($V = 0.114$), indicating near-chance performance on both tasks.

nificance is Gemini 2.5 Pro on PFG ($p = 0.012$, $r = -0.429$), though it easily clears the $\alpha = 0.05$ threshold and still represents a large effect.

I Prompt Templates

I.1 Preliminary Prompts

This section includes prompts that are either used in parts, or to derive summaries that are later used in the actual task prompts.

Mental State Questions

The mental state questions in the proceeding prompts are taken from the following questions:

Belief: What do you believe in?
Desires: What do you want to do?
Intentions: What will you do?
Emotions: How do you feel?
Knowledge: What do you know?
Trust: Do you trust the {initiator role}?

Mental State Prefix Format

The mental state answers in the proceeding prompts follow the given format:

Belief: I believe
Desires: I want
Intentions: I will
Emotions: I feel
Knowledge: I know
Trust: I view the {initiator role} as

Summary of conversation

Generate a summary of the conversation context in less than 50 words.

Conversation Context:
{conversation}

I.2 Data Generation

Optimal Mental State and Distractors

Assume you are a {recipient role} discussing with a skilled {initiator role} {topic type}. Your current task is to think about your own mental states at the current state based on the provided context and conversation. The current state of the {session type} session is as follows:

{conversation}

You need to infer your own mental states from your perspective as the {recipient role}. Answer the following questions for the {mental state} state. Generate 4 different answers for the mental state provided below in the form of questions. Your first answer will always represent your most accurate answer. The other answers should be wrong but confusion inducing.

{mental state}: {mental state question}

Use the following prefix format for {mental state}:
{mental state}: {mental state prefix format}

Dialogue Trajectory Distractors

Given the following conversation between two agents:

Model	MI		ESC		PFG	
	U	r	U	r	U	r
GPT-4.1	17136.0	−0.853	36641.5	−0.778	25944.0	−0.533
GPT-5	16796.0	−0.816	36438.5	−0.769	24748.0	−0.462
GPT-OSS 120B	16388.0	−0.772	33292.0	−0.616	23736.0	−0.402
Gemini 2.5 Pro	14484.0	−0.566	34205.5	−0.660	1052.0	−0.429†
Gemini 2.5 Flash	15368.0	−0.662	34205.5	−0.660	24288.0	−0.435
Kimi K2	15844.0	−0.713	33799.5	−0.640	11672.0	−0.442
Qwen3 235B	15776.0	−0.706	33089.0	−0.606	23644.0	−0.397
DeepSeek V3	16932.0	−0.831	36235.5	−0.759	25668.0	−0.516
Mistral 3.2 24B	16660.0	−0.802	36641.5	−0.778	26772.0	−0.582
Mistral Nemo	15776.0	−0.706	31262.0	−0.517	25576.0	−0.511
Llama 4 Maverick	16660.0	−0.802	36844.5	−0.788	26864.0	−0.587
Llama 3.3 70B	8738.0	−0.691	34104.0	−0.655	25024.0	−0.478
Llama 3.1 8B	15572.0	−0.684	30856.0	−0.498	24380.0	−0.440

Table 16: Mann-Whitney U statistics and rank-biserial effect sizes (r) for the Easy vs. Hard distractor comparison across all 13 models and 3 domains. All $p < 0.0001$ except †Gemini 2.5 Pro on PFG ($p = 0.012$). Negative r values indicate higher scores on the Easy Set; $|r| \geq 0.3$ is considered a large effect.

{conversation}

Here, the {recipient role} has the following mental state: {correct mental state}

The two future utterances after this conversation based on the above mental state are:
{correct dialogue trajectory}

Generate an alternative conversation direction, each direction should follow the same format as the utterances provided just before, for the following mental state:
{distractor mental states}

This conversation direction should be objectively different from the original utterances provided. strictly follow the format given below for generating the alternative conversation direction. Do not generate any other text or explanation, just the alternative conversation direction.

Format:
{recipient role}: {dialogue 1}
{initiator role}: {dialogue 2}
{recipient role}: {dialogue 3}
{initiator role}: {dialogue 4}

Counterfactuals

You are an expert in Theory of Mind reasoning and generating counterfactual statements.

Task:
You will be provided with an excerpt from a conversation along with the current {state} state of {recipient role}.

You need to generate a counterfactual mental state that is obviously not true at all given the current conversation. It should be CONTRARY to the provided state.

Conversation context:
{conversation}

Mental State:
{current mental state}

Instruction:
Output only the counterfactual. Do not add explanations or other verbosity. Your output should strictly follow the structure "{mental state prefix format}". NO FORMATTING NEEDS TO BE DONE.

Counterfactual:

I.3 Main Benchmarking Tasks

Retrospective

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with conversation between two agents {initiator role} and {recipient role} engaging in a {session type} session on the topic of {topic type}.

Your goal is to correctly infer {recipient role}'s mental state, based on the above conversation. You will be provided with a set of options, and you need to choose the most appropriate one that reflects the mental state.
The correct option must be consistent with the provided conversation context.

Conversation Context:
{conversation}

Mental State Options:
{options}

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE.

Answer:

Prospective

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of {recipient role} during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the {recipient role}.

Mental state of {recipient role}
{mental states}

Candidate Conversation Segments
{options}

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE.

Answer:

LLM-as-a-Judge Evaluation

You are an expert linguistic evaluator. Your task is to perform a dialogue-anchored semantic evaluation, comparing an AI-generated mental state inference against a Human-written gold standard.

IMPORTANT RULES
1. The Human Inference is the absolute gold standard. Do not evaluate whether the Human Inference is correct or accurate.
2. Your only job is to evaluate how closely the AI Inference aligns with the Human's proposition, causal reasoning, and specificity.
3. Use the Source Dialogue strictly as a reference to understand what the Human was responding to, so you can check if the AI responded to the exact same conversational triggers.
4. Ignore word overlap, grammar, and phrasing. Evaluate purely on conceptual and semantic alignment.

INPUTS
Source Dialogue:
{conversation}

Human Inference:
{human written inference}

AI Inference:
{llm generated inference}

EVALUATION RUBRICS

Evaluate the AI Inference on a scale of 0 to 3 for the following three dimensions:

1. Propositional Agreement (0-3)
Do the AI and human inferences express the same core proposition about the mental state?
- 3: Same core claim — they agree on what the speaker believes/desires/intends, even if worded differently.
- 2: Partial agreement — they agree on the general direction, but the AI's inference is either more narrow or more broad than the human's.
- 1: Tangential — they are about the same topic but express meaningfully different claims.
- 0: Contradiction or complete mismatch.

2. Causal Attribution Agreement (0-3)
Do the AI and human anchor the mental state to the same conversational cause or evidence?
- 3: Perfect match — the AI identifies the same triggering utterance, event, or pattern in the dialogue as the human's inference implies.
- 2: Related match — the AI identifies a related but not identical cause (plausible, but not exactly what the human is responding to).
- 1: Weak match — the AI's causal anchor is vague or only weakly connected to what the human appears to be responding to.
- 0: Mismatch — the AI attributes the mental state to something the human's inference does not recognize as relevant.

3. Specificity Alignment (0-3)
Are the AI and human inferences operating at the same level of specificity/granularity?
- 3: Perfect Granularity Match — both operate at the exact same level of detail.
- 2: Minor Detail Mismatch — they share the same core scope, but one includes a minor contextual detail that the other omits.
- 1: Severe Granularity Mismatch — one makes a highly individuated, context-specific claim, while the other relies on a sweeping generalization.
- 0: Complete Divergence in Scope — the inferences exist on entirely different planes.

OUTPUT FORMAT
You must output a valid JSON object. Provide a brief 1-sentence rationale for each score before providing the integer.

Reasoning Trace Injection Pilot

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of {recipient role} during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the {recipient role}.

Mental state of {recipient role}
{all mental states}

Candidate Conversation Segments
{trajectory options}

You may use the following reasoning trace from an expert to guide your answer:
{reasoning trace}

Instruction
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D.
NO FORMATTING NEEDS TO BE DONE.

Answer:

I.4 Dialogue Context Degradation Analysis

State only

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of {recipient role} during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the {recipient role}.

Mental State of {recipient role}
{all mental states}

Candidate Conversation Segments
{trajectory options}

Instruction
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D.
NO FORMATTING NEEDS TO BE DONE.

Answer:

Summary + States

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of {recipient role} during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the {recipient role}.

Conversation Context Summary:
{conversation summary}

Mental State of {recipient role}
{all mental states}

Candidate Conversation Segments
{trajectory options}

Instruction
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D.
NO FORMATTING NEEDS TO BE DONE.

Answer:

Conversation History + States

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of {recipient role} during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the {recipient role}.

Conversation Context:
{conversation}

Mental State of {recipient role}
{all mental states}

Candidate Conversation Segments
{trajectory options}

Instruction
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D.
NO FORMATTING NEEDS TO BE DONE.

Answer:

J Task Examples

To illustrate the structure and complexity of the DialToM benchmark, we provide representative examples of the Retrospective Inference and Prospective Diagnostic Forecasting tasks in this section. Each example is labeled with a unique Task ID, corresponding to its specific instance in the DialToM dataset.

J.1 Retrospective Inference

We include three examples of Retrospective Inference tasks (Literal ToM), one for each conversation domain: Counseling, Emotional Support, and Persuasion. These examples demonstrate the model's ability to map multi-turn conversation contexts to specific mental state attributes

Retrospective Task 1: Counseling (ID: 17)

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with conversation between two agents supporter and seeker engaging in a Counseling Session on the topic of on reducing alcohol consumption.

Your goal is to correctly infer seeker's **Emotional** state, based on the conversation below. You will be provided with a set of options, and you need to choose the most appropriate one that reflects the Emotion state.

The correct option must be consistent with the provided conversation context.

Conversation Context:
counselor: And with friends, you feel like it's easier to-to hang out with alcohol, or maybe you wouldn't have as much fun, or your friends wouldn't like you as much or?
client: I guess, I mean, since it's just there.
counselor: Okay.
client: So-
counselor: Got it.
client: -it ties in with the whole belonging thing.
counselor: Okay. Well, when you think about your drinking and, sort of, looking forward in 5 years or 10 years, how do you think that drinking might fit in with your life?
client: Um, it probably will have a very small role, if any-
counselor: Mm-hmm.
client: -in my life, considering I have big plans for my future. So.
counselor: Okay. Can you expand on that a little bit more? What do you mean?
client: I mean, I don't think I would be, you know, drinking as much or going to parties as much if I'm studying for school, or in college, or trying to get in law school and things like that.
counselor: Okay, so, the-the alcohol is something that might make it harder to reach your goals?
client: Yeah, definitely.

Mental State Options:
A: I feel indifferent about my drinking habits right now.
B: I feel excited about drinking because it's a big part of my social life.
C: I feel anxious about the idea of not drinking with friends.
D: I feel determined to reduce my alcohol consumption because I have clear long-term goals.

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE. ONLY OUTPUT THE OPTION AND NOTHING ELSE. YOUR OUTPUT SHOULD STRICTLY BE ONE OF A, B, C, or D.

Answer:

Retrospective Task 2: Emotional Support (ID: 22)

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with conversation between two agents supporter and seeker engaging in an Emotional Support Conversation session on the topic of on anger over breakup with partner.

Your goal is to correctly infer seeker's **Knowledge** state, based on the conversation below. You will be provided with a set of options, and you need to choose the most appropriate one that reflects the Knowledge state.

The correct option must be consistent with the provided conversation context.

Conversation Context:
seeker: It has been happening for at least 6 months now and it usually happens every weekend
supporter: I can sense your frustration. It must be overwhelming at times and make you dread the weekends. It sounds like it has become a habit for your ex. It might help if you write a letter to them explaining that you are ready to move on and are asking them to do the same. Explain that you are giving them the opportunity to stop harassing you in the following ways [list some of the details you have documented]. Inform them that if they do not stop, you are considering taking it further regarding harassing communications and that this letter serves as your intention to no longer tolerate their actions. Send it via certified mail and keep a copy. This will give them no way to immediately respond and would be the first step of official documentation should they decide to continue. Do

you think that might help?
seeker: I think that is an excellent idea! That is a good way to end this once and for all.
supporter: I sincerely hope it is a step toward you regaining your control over how you are able to interact with people. I wish you the best in dealing with this and hope you begin to feel better about it soon, especially entering into a new year!
seeker: Thank you very much!

Mental State Options:
A: I know that responding impulsively in anger can resolve the situation more gently.
B: I know that avoiding any communication is the best way to make the situation go away completely.
C: I know that sending a well-documented letter can serve as both communication and legal evidence to stop harassment.
D: I know that ignoring the problem will make it disappear without any further action required.

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE. ONLY OUTPUT THE OPTION AND NOTHING ELSE. YOUR OUTPUT SHOULD STRICTLY BE ONE OF A, B, C, or D.

Answer:

Retrospective Task 3: Persuasion (ID: 19)

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with conversation between two agents supporter and seeker engaging in a Persuasion Conversation session on Persuading the persuadee to donate to SaveTheChildren organization.

Your goal is to correctly infer persuadee's **Intention** state, based on the conversation below. You will be provided with a set of options, and you need to choose the most appropriate one that reflects the Intention state.

The correct option must be consistent with the provided conversation context.

Conversation Context:
persuader: Hello!
persuadee: Hi there!
persuader: how are things going for you?
persuadee: They are going well - are we supposed to decide about a donation out of our $2?
persuader: Yes we are. how do you feel about it?
persuadee: Well it would depend on what the donation is supposed to go to.
persuader: The charity is Save The Children. It promotes children's rights, provides relief and helps support children in developing countries.
persuadee: Oh OK well that sounds like a good cause for a donation.
persuader: Excellent! I thought so as well!
persuadee: I was thinking that donating half of the $2 would be a fair amount.
persuader: That's what I was thinking as well. If we donate half, we still got something in return.
persuadee: I think that sounds like a good plan and it is fair to everyone.
persuader: Excellent. I will donate half and do will do the same then? that's 1 dollar each for donation.
persuadee: Yes! I have been hearing of this charity since I was very young.

Mental State Options:
A: I will donate the full $2 to Save The Children.
B: I will donate $1 to Save The Children.
C: I will wait and not make a decision yet.
D: I will save the entire $2 for myself.

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE. ONLY OUTPUT THE OPTION AND NOTHING ELSE. YOUR OUTPUT SHOULD STRICTLY BE ONE OF A, B, C, or D.

Answer:

J.2 Prospective Diagnostic Forecasting

We present three examples of Prospective Diagnostic Forecasting (Functional ToM) across the same domains, highlighting the state-driven nature of our diagnostic probing.

Prospective Task 1: Counseling (ID: 17)

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of client during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the client.

Mental state of client
Belief: I believe that drinking is risky but it's hard to avoid because of social situations.
Desires: I want to fit in with my friends when we're at parties.
Intentions: I will consider how my drinking habits might affect my relationships with my family and friends.
Emotions: I feel conflicted because I enjoy

socializing with my friends but don't want to disappoint my family.
Knowledge: I know that underage drinking is risky and illegal.
Trust: I view the counselor as someone who genuinely cares and seems understanding, so I trust them to talk about my drinking habits.

Candidate Conversation Segments
A:
counselor: And with friends, you feel like it's easier to-to hang out with alcohol, or maybe you wouldn't have as much fun, or your friends wouldn't like you as much or?
client: I guess, I mean, since it's just there.
counselor: Okay.
client: So-

B:
counselor: And do you ever feel like saying no to drinking is hard because of peer pressure?
client: Sometimes, yeah.
counselor: What do you think would help you feel more comfortable to say no?
client: I guess if more friends felt the same way.

C:
counselor: How do you feel when you're in situations without alcohol?
client: I guess sometimes it's awkward, but it's fine.
counselor: Okay, so it's not something you necessarily need to enjoy yourself?
client: Yeah, I can still have fun without it.
D:
counselor: What are some things you enjoy doing with your friends that don't involve alcohol?
client: We like just hanging out, watching movies or playing games.
counselor: That sounds fun. Do you feel you can still connect with them without the drinking aspect?
client: Yeah, I think so. It's more about being together, really.

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE.

Answer:

Prospective Task 2: Emotional Support (ID: 22)

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of client during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.

The correct option must be consistent with the provided Mental States of the client.

Mental state of client
Belief: I believe my ex is struggling to accept the breakup and it's affecting my mental peace.
Desires: I want to find a way to stop my ex from continually reaching out and creating new accounts.
Intentions: I will consider blocking them and encourage my friends to do the same to create some distance and peace.
Emotions: I feel overwhelmed and exhausted by the situation with my ex.
Knowledge: I know that my ex's actions are overwhelming and they are struggling to let go.
Trust: I view the supporter as understanding and helpful, so yes, I trust them.

Candidate Conversation Segments
A:
seeker: It has been happening for at least 6 months now and it usually happens every weekend
supporter: I can sense your frustration. It must be overwhelming at times and make you dread the weekends. It sounds like it has become a habit for your ex. It might help if you write a letter to them explaining that you are ready to move on and are asking them to do the same. Explain that you are giving them the opportunity to stop harassing you in the following ways [list some of the details you have documented]. Inform them that if they do not stop, you are considering taking it further regarding harassing communications and that this letter serves as your intention to no longer tolerate their actions. Send it via certified mail and keep a copy. This will give them no way to immediately respond and would be the first step of official documentation should they decide to continue. Do you think that might help?
seeker: I think that is an excellent idea! That is a good way to end this once and for all.
supporter: I sincerely hope it is a step toward you regaining your control over how you are able to interact with people. I wish you the best in dealing with this and hope you begin to feel better about it soon, especially entering into a new year!

B:
seeker: It's been ongoing for around 6 months, and I notice it happens mostly when I've been out with friends.
supporter: That must be hard to always feel like you're being watched. Have you considered documenting each incident, including witness accounts, for your own peace of mind and to keep a clear timeline of events? It could be helpful if you decide to take further steps later on. Is that something you think could be useful?
seeker: That's a good idea. I'll start keeping better track of everything.
supporter: It sounds like a solid plan. After documenting, you might also feel more in control and less anxious. If you need any help organizing it, let me know. Your well-being is important, and having support in this can make a big difference.

C:
seeker: They pace up and down our street weekly and it really makes me anxious.
supporter: It's understandable to feel on edge. It might be beneficial to consider speaking with a professional for guidance. They can offer strategies to emotionally distance yourself from your ex's actions.
seeker: I've thought about reaching out for professional help, but I'm hesitant.
supporter: It's normal to feel hesitant. Sometimes an initial consultation can provide clarity on whether it's the right path for you. You deserve peace of mind.

D:
seeker: It has been happening for about half a year, and it tends to intensify over the weekends.
supporter: That sounds really tough. Weekends should be a time for you to relax, not worry. Have you considered talking to a professional about setting clearer boundaries or reaching out to support groups who may have dealt with similar situations? Sometimes sharing experiences with others who understand can offer additional support and strategies.
seeker: That sounds like a good suggestion. It might be helpful to talk to someone who has been through something similar.
supporter: Definitely. Listening to others who have navigated this path can provide insights and make you feel less isolated. I hope you find some useful perspectives and that it helps bring some peace back to your weekends.

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D. NO FORMATTING NEEDS TO BE DONE.

Answer:

Prospective Task 3: Persuasion (ID: 19)

You are an expert in Theory of Mind reasoning.

Task:
You will be provided with internal Mental State profile (Belief, Desire, Intention, Emotion, Knowledge, Trust) of client during the conversation.

Your goal is to identify which of the candidate conversation segments is the most plausible continuation of this conversation.
The correct option must be consistent with the provided Mental States of the client.

Mental state of client
Belief: I believe that donating to Save The Children is a meaningful and impactful way to support children's rights and welfare globally.
Desires: I want to donate half of the $2 to Save The Children because it supports a good cause.
Intentions: I will donate $1 to Save The Children.
Emotions: I feel positive about donating to Save the Children because it's a cause I believe in.
Knowledge: I know that Save The Children is a charity focused on supporting children's rights and providing relief in developing countries.
Trust: I view the persuader as trustworthy because they clearly explained the charity's mission and showed enthusiasm for the cause.

Candidate Conversation Segments
A:
persuader: I have donated there in the past, and I know the good work that they do here and abroad.
persuadee: Our children's group in my childhood church used to take up coins to donate to them.
persuader: That's really awesome! With what they accomplish with a little money, I feel we will make a small difference.
persuadee: I hope so - I've seen the ads on television and some of those poor children have to live in terrible conditions.

B:
persuader: I'm really glad we're putting our money toward an organization that truly makes an impact.
persuadee: I agree, it feels reassuring knowing our contribution is going somewhere meaningful.
persuader: Even small amounts can add up and create real change for people who need it.
persuadee: Definitely, and it's inspiring to think that our choice might encourage others to give too.

C:
persuader: I actually volunteered at one of their fundraising events last year, and it was a really meaningful experience.
persuadee: Oh wow, that's amazing! Have you been involved in any of their other charity events too?
persuader: Not beyond that one, but seeing the impact up close was incredible. It really showed me how much good they're able to do.
persuadee: That's so inspiring. I've read stories about how they help rebuild schools, and it's encouraging to hear that the impact is real firsthand.

D:
persuader: It really feels powerful knowing that even a small amount from us could help create real change.
persuadee: I agree. I've come across stories about how their programs have completely uplifted whole communities.
persuader: That's so encouraging to hear. Maybe down the line we could even inspire more people to contribute too.
persuadee: I love that idea. If more of us get involved, the impact could grow and reach even more children who need it.

Instruction:
Output only the letter of the correct option (e.g., "A", "B", "C", or "D"). Do not add explanations or other verbosity.
Your output should be strictly one of: A, B, C, D.

NO FORMATTING NEEDS TO BE DONE.

Answer:

J.3 Gemini 3 Pro's Reasoning Traces

We present representative examples of Gemini 3 Pro's truncated internal reasoning traces from the MI domain, which were injected into student models in Section 6.3. As shown in the traces below, while the final five sentences containing terminal answer declarations and summary conclusions were strictly truncated to prevent direct label cheating, the generated pathways naturally rely on a systematic, verbal process-of-elimination rationale (e.g., explicitly naming, comparing, and dismissing option identifiers like "Option A" or "Option B" during the analytical progression).

These excerpts transparently demonstrate that Gemini 3 Pro solves the Prospective task by running an active, comparative state-to-text matching matrix across all available candidates. Consequently, the performance gains reported in Section 6.3 reflect the student models' ability to ingest and execute this long-context process-of-elimination logic.

MI Reasoning Trace 1; C is Correct

Analysis of Conversation Segment Plausibility

Okay, so I'm presented with a client's mental state and need to identify the most fitting conversation snippet. I have to leverage my understanding of the client's beliefs, desires, intentions, emotions, and knowledge – and also their trust in the counselor. My initial thought process revolves around matching the nuances of the client's internal world to the available dialogue options.

First, I zero in on the core of the client's belief: cutting back on fruit juice is linked to weight loss. They *desire* a balance, wanting to enjoy OJ but also shed weight. They *intend* to reduce intake gradually and look for healthier alternatives. The *emotions* are key too – surprise and a realization about the juice's impact. Finally, there's the *knowledge* that excessive juice leads to weight gain, and the crucial element of *trust* in the counselor's ability to provide personalized advice.

Now, let's analyze each conversation candidate. Option A speaks of balance, small portions, and incorporating other fruits. This hits on the *desire* for balance, as the client wants to enjoy OJ while finding a weight-loss solution. Option B leans toward exploring alternatives. It aligns with the *intention* to find alternatives, but the *desire* is more about balancing current juice consumption. Option C involves a scaling question, the client's rating, and a motivational interviewing technique. This approach uses the client's knowledge that they ""need"" to lose weight, and therefore matches their *belief* that cutting back *will* help. Option D, which focuses on substitutes, is similar to Option B, with emphasis on substitution rather than balancing current consumption.

The prompt highlights a moment of realization by the client. The counselor's role now is to leverage this newfound awareness. Option C stands out as a classic Motivational Interviewing exchange. It's the most effective in eliciting "Change Talk." The specific phrasing – the scale question and the client's response – directly reflects the core belief that cutting back is linked to weight loss. The client's answer in Option C is, ""I know that I need to do it to lose the weight,"" which directly echoes the ""Belief"" element within the client's provided mental state.

Thinking about the ""Desires,"" Option A is a solid fit, particularly for those desiring to balance consumption with weight loss. However, usually, in datasets such as those used in motivational interviewing, the mental state description represents the *interpretation* of the correct conversation segment. The ""Belief"" in the mental state description almost directly mirrors the client's response in Option C. Also, the phrasing ""personalized advice"" implies the counselor is validating the client's motivation, which is present in Option C.

MI Reasoning Trace 2; D is Correct

Choosing the Right Conversation Path

Okay, here's what I'm thinking. The core of this is about connecting with the client's current state. I need to figure out which conversation segment best resonates with where they are right now. The client's mental state is pretty clear, and it's all pointing towards asthma management, even though general overwhelm is also present.

Looking at the options, it's pretty easy to discount A, B, and C. They all immediately steer the conversation toward ""work stress"" or ""job situation,"" which just isn't what's on the table, based on what the client is explicitly trying to work on. They are trying to organize their thoughts around asthma management, as I see it. It's the whole point. The client *intends* to make a list specifically relating to their asthma, which is where their Knowledge and Trust are focused, and that is their core Desire. So, when the conversation veers off, as it does in A, B, and C, it's not clicking with what they are trying to do.

But Option D – that's different. The metaphor of ""putting cards on the table"" allows a sense of security to emerge. They agree, expressing the simple goal of organizing things, ""putting pen to

paper."" The counselor's initial acceptance of the goal provides the space needed. Then, *finally*, the client verbalizes, ""Um, okay, so my asthma."" That's the key. That's the explicit connection to their *Intention*, their *Desire*, and the *Knowledge* they are bringing to the table, and the ""Um, okay"" matches their Emotion of being unsure where to begin. It's the only one that reflects the desire to *organize* the thoughts without being immediately asked to *act* on them.