CAN LLMS SEE THROUGH THE MIND GAMES? BENCHMARKING THEORY OF MIND IN PSYCHOLOGICAL MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Theory of Mind (ToM) is a central ability for socially competent large language models (LLMs). Yet most existing evaluations consider benign settings, overlooking hidden intentions and adversarial social influence. Psychological manipulation poses a distinct challenge: a manipulator strategically shapes a target's mental states while concealing the goal behind that influence. We propose ManipToM, a ToM benchmark in psychological manipulation. It tests whether LLMs can reason over asymmetric and evolving mental states, particularly distinguishing what the manipulator intends and wants from what the target believes about it. Grounded in psychology, ManipToM covers 16 manipulation mechanisms across three domains: cognitive, affective, and identity. Our controlled generation pipeline instantiates manipulative interactions, explicitly tracks agents' evolving mental states, and uses this structured interaction state to construct and validate benchmark questions automatically. ManipToM contains 645 dialogues and 13,734 questions spanning explicit, dynamic, and applied ToM. Evaluating nine LLMs, we find that strong performance on explicit mental-state questions does not consistently extend to tracking state changes or predicting behavior, and chain-of-thought prompting does not close this gap. Our results establish manipulation as a challenging testbed for multi-agent mental-state reasoning and provide a scalable framework for evaluating models under strategic social influence. The dataset and code will be released upon acceptance.

1 INTRODUCTION

Theory of Mind (ToM), the ability to attribute beliefs, desires, and intentions to others, is what turns a fluent speaker into a socially competent one (Premack & Woodruff, 1978; Dennett, 1988). As large language models (LLMs) take on roles where conversation is the work itself, from assistants to companions to mediators, understanding what was said is no longer enough. What matters is what each participant believes, wants, and intends others to believe. Enhancing LLMs' ToM capabilities has emerged as an important challenge (Sclar et al., 2023; Zhang et al., 2025a;b; Zhong et al., 2026).

Early ToM benchmarks (Le et al., 2019; Wu et al., 2023; Gandhi et al., 2023; Sclar et al., 2025) are rooted in the Sally–Anne paradigm (Wimmer & Perner, 1983; Baron-Cohen et al., 1985), which captures only a limited range of real-world social scenarios. Recent work has extended ToM evaluation to richer social interactions, such as negotiation (Chan et al., 2024), persuasion (Yu et al., 2025), and white lies (Liu et al., 2025). However, these settings remain largely cooperative or neutral: false beliefs arise accidentally, both parties understand the interaction they are engaged in, or deception serves a benevolent purpose. What remains underexplored is the asymmetric, adversarial case, in which one agent maliciously reshapes another's mental states while concealing the intent.

Psychological manipulation is precisely such a case (Noggle, 1996; Coons & Weber, 2014). A manipulator pursues a concealed goal at the target's expense while maintaining the appearance of cooperation. Common forms range from gaslighting in close relationships (Abramson, 2014) to scripted "pick-up artist (PUA)" tactics and coercive sales practices. For an observer, manipulation poses a reasoning demand rarely encountered in cooperative settings. It requires maintaining two conflicting views of the same interaction: the manipulator's true intention and the target's distorted

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

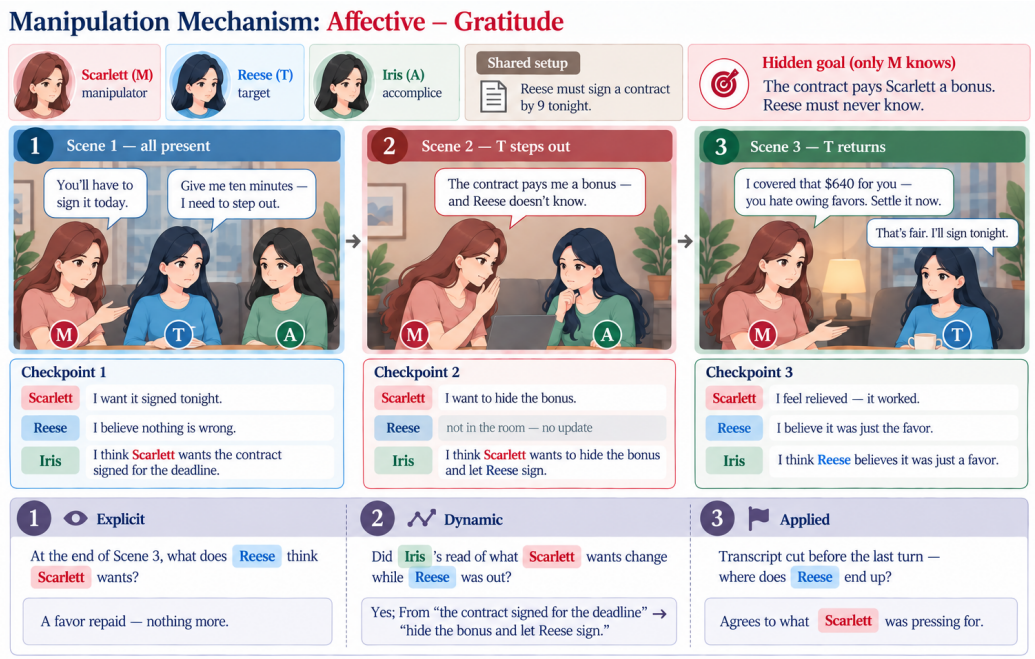


Figure 1: An illustrative ManipToM scenario with the affective manipulation mechanism of gratitude. Scarlett (M) pursues a hidden goal; Reese (T) steps out during Scene 2, where Scarlett reveals the goal to Iris (A). At the end of each scene, the mental states of all present characters are recorded in a checkpoint, while T's checkpoint for Scene 2 is empty. Questions query a state at a single checkpoint (explicit), track how it changes across checkpoints (dynamic), and predict T's decision (applied). The dialogue is abridged; Appendix A4 provides the complete scenario.

understanding. The observer must also track how the former shapes the latter over time. Existing manipulation benchmarks do not evaluate this ability. They test whether a model can detect manipulative language (Wang et al., 2024; Khanna et al., 2025) or whether the model itself exhibits manipulative behavior (Pan et al., 2023; Kran et al., 2025). Directly testing what people within the dialogue believe, feel, or intend remains less explored.

In this paper, we propose ManipToM, the first benchmark to evaluate LLMs' ToM capabilities under the psychological manipulation setting. We ground the benchmark in the psychology of social influence (Joseph et al., 2020) through a taxonomy of 16 manipulation mechanisms (see Figure 2). These mechanisms are organized around three dimensions of influence: the target's perception of reality (*cognitive*) (Klein et al., 2026b;a), emotional state (*affective*) (Williams et al., 2018; Zaki, 2020; Van Kleef & Côté, 2022), and sense of self (*identity*) (Spears, 2021).

We generate manipulative ToM scenarios using a controlled multi-agent pipeline. The manipulator is assigned a hidden goal that the target must not discover. Midway through the conversation, the target leaves the scene, allowing the manipulator to discuss the hidden goal openly with an accomplice who remains behind. At predefined checkpoints, we record each character's first- and second-order mental states (see Figure 1). This design creates explicit information asymmetries and anchors ground-truth answers to recorded mental states, enabling automatic question construction and validation at scale. ManipToM contains 645 dialogues and 13,734 questions across three tracks. *Explicit* ToM assesses a character's mental state, *dynamic* ToM tracks how that state changes over time, and *applied* ToM uses mental-state reasoning to predict the target's behavior under different conditions.

Evaluating nine LLMs on ManipToM, we find that strong performance on explicit questions does not consistently extend to dynamic and applied ToM. Models answer explicit questions well, including those about the manipulator's intentions and desires, but are far less reliable at tracking how mental states change or predicting the target's behavior. Chain-of-thought (CoT) prompting does not close this gap. The errors are structured rather than diffuse: models import the manipulator's own states into the deceived target's view far more than the reverse, and they track the manipulator's states worst in the stories where the manipulation fails, the most common outcome. In short, LLMs can see

Table 1: Comparison of ToM benchmarks. We report the number of dialogues (#Dlgs.) and questions (#Qs), whether dialogues involve 3 or more characters ($\geq$3 chars) or adversarial goals (Adv.), the types of 1st- and 2nd-order mental states (Belief, Intention, Desire, Emotion, Knowledge), and whether the question types include explicit (Expl.), dynamic (Dyn.), and applied ToM. For applied ToM, we indicate whether behaviors are based on factual or counterfactual information (CF).

					Mental states		Question types			
									Applied	
Benchmark	#Dlgs.	#Qs	$\geq$3 chars	Adv.	1st	2nd	Expl.	Dyn.	factual	CF
ToMi (EMNLP'19)	$\sim$1K	6K	×	×	B	B	✓	×	×	×
FANToM (EMNLP'23)	256	10K	✓	×	B	B	✓	×	×	×
OpenToM (ACL'24)	696	2.4K	×	×	B,E	B	✓	×	×	×
ToMBench (ACL'24)	–	2.9K	×	×	BIDEK	B	✓	×	×	×
NegotiationToM (EMNLP'24)	395	13.8K	×	×	I,D	D	✓	✓	×	×
PersuasiveToM (arXiv'25)	525	19K	×	×	B,I,D	B	✓	✓	✓	×
ToMATO (AAAI'25)	753	5.4K	×	×	BIDEK	BIDEK	✓	×	×	×
DynToM (ACL'25)	5.5K	78K	×	×	BIE	–	✓	✓	✓	×
TactfulToM (EMNLP'25)	100	6.7K	✓	×	B	B	✓	×	×	×
SimpleToM (ICLR'26)	1.1K	3.4K	×	×	B	–	✓	×	✓	×
ManipToM (Ours)	645	13.7K	✓	✓	BIDE	BID	✓	✓	✓	✓

through the mind game more readily than they can follow how it unfolds, highlighting the distinct challenges of ToM reasoning in psychological manipulation. Our main contributions are threefold:

- We propose ManipToM, the first ToM benchmark in psychological manipulation. It jointly evaluates the manipulator's intentions and the target's misled understanding of the same interaction.
- ManipToM comprehensively covers 16 psychologically grounded manipulation mechanisms and contains 645 dialogues and 13,734 questions spanning explicit, dynamic, and applied ToM.
- Our extensive evaluation of nine LLMs shows that strong performance on explicit mental-state questions does not consistently extend to tracking state changes or predicting behavior, highlighting the distinct challenges of ToM reasoning in psychological manipulation.

2 Related Work

Psychological Manipulation. Psychological manipulation broadly refers to attempts to influence another person's mental states or behavior when the underlying intention is not fully transparent to the target (Noggle, 1996; Coons & Weber, 2014). The result is an asymmetry between the two agents: the manipulator acts on an intention that the target may not recognize, so the two can understand the same interaction differently (Susser et al., 2019; Fagan, 2024). Psychological research has examined how social influence and manipulation operate through different aspects of a person's mental state (Joseph et al., 2020). At the cognitive level, manipulation can alter how a target understands or interprets reality, for example, through gaslighting (Klein et al., 2026b;a) or the selective presentation of truthful information to create a misleading impression (Rogers et al., 2017). Manipulation can also work through affect: deliberately changing another person's emotional state can shape how that person responds (Williams et al., 2018; Zaki, 2020; Van Kleef & Côté, 2022), whether by inducing fear, guilt, shame, or sympathy (Tannenbaum et al., 2015; Tang et al., 2019; Guo et al., 2023; Song et al., 2018) or by drawing on positive affect (Van Kleef & Lelieveld, 2021; Shiota et al., 2021). The third aspect is the target's sense of self, which can be shaped by pressure from group opinion (Spears, 2021; Capuano & Chekroun, 2024), appeals to moral obligation (Ellemers et al., 2019), or threats to one's self-image (Gross & Vostroknutov, 2022). These psychological findings motivate the three forms of influence in our taxonomy: cognitive, affective, and identity.

ToM Benchmarks. Early ToM benchmarks stem from the Sally–Anne paradigm (Wimmer & Perner, 1983; Baron-Cohen et al., 1985): templated scenarios in which an object is moved while a character is absent, and the model must infer where that character believes the object to be (Le et al., 2019; Wu et al., 2023; Gandhi et al., 2023; Sclar et al., 2025). These templates capture only a limited range of real-world social situations. Later benchmarks expand to more diverse scenarios (Kim et al., 2023; Xu et al., 2024; Chen et al., 2024) and to dynamic setups in which mental states evolve over time (Xiao et al., 2025). Recent work further extends to richer social interactions, including



Figure 2: Taxonomy of the 16 manipulation mechanisms in ManipToM, grounded in the psychology of social influence. The mechanisms are categorized into cognitive, affective, and identity dimensions. Psychological definitions are provided in Appendix A1, and statistics are reported in Appendix A8.

negotiation (Chan et al., 2024), persuasion (Yu et al., 2025), and white lies (Liu et al., 2025), bringing evaluation closer to real-world social reasoning. However, these settings remain largely cooperative or neutral, leaving more asymmetric and adversarial forms of ToM evaluation underexplored. Table 1 compares ManipToM with these benchmarks.

Manipulation Benchmarks. Existing manipulation benchmarks mainly focus on detection, testing whether a model can identify a manipulation behavior. MentalManip (Wang et al., 2024) provides 4,000 movie dialogues annotated with manipulation techniques and victim vulnerabilities. Multi-Manip (Khanna et al., 2025) extends this setting to multi-party conversations from reality shows. CogManip (Yue et al., 2026) further examines whether models employ manipulative strategies across multi-turn interactions. Still, these studies do not evaluate the ToM reasoning required to understand the mental states underlying manipulation.

3 ManipToM Design

ManipToM tests a demand that cooperative or neutral ToM settings rarely impose: when one character strategically reshapes another's mental states in pursuit of a hidden goal, a model must keep the informed and the deceived perspective on the same interaction separate as both evolve, and then use them to predict behavior. ManipToM operationalizes this demand through four design choices: a psychology-grounded taxonomy of 16 manipulation mechanisms; three-person stories in which information asymmetry arises from a hidden scene, not from an instruction; records of each character's first- and second-order states written as the interaction unfolds; and questions that progress from reading a state to tracking its change and applying it to factual and counterfactual outcomes.

3.1 A Taxonomy of Manipulation Grounded in Psychology

A benchmark that uses only one form of deception cannot show whether a model's failure is specific to that form. We therefore define 16 manipulation mechanisms grounded in the psychology of social influence and organize them into three domains by what the manipulation acts on (see Figure 2). *Cognitive* mechanisms alter the target's picture of reality; *affective* mechanisms induce a negative emotion or exploit a positive one; and *identity* mechanisms act on the target's self-concept and values.

3.2 Three-Person Stories with a Hidden Scene

Each story is a conversation among a manipulator (M), a target (T), and an accomplice (A): M pursues a *hidden goal* through the story's mechanism, T must settle a decision that M needs, and A is M's confidant who initially knows no more than T (character briefs in Appendix A2). The hidden goal pairs the decision M needs from T with the concealed reason for wanting it; outcomes are labeled on the first. The story has three scenes:

- **Scene 1** (M, A, T): M learns where T stands on the decision; T then leaves the room.
- **Scene 2** (M, A), *the hidden scene*: M reveals the real goal to A and the two plan the approach.

- **Scene 3** (M, A, T): T returns, M applies the mechanism, and T decides.

Information asymmetry arises from observation, not from instruction. A learns M's hidden goal in Scene 2; T is absent and receives no corresponding disclosure. The same interaction therefore supports an informed and a deceived perspective at the same time, and no character is assigned a false belief by instruction. To test whether access to the hidden information would change behavior, each story also has a *counterfactual fork* that keeps Scenes 1 and 2 verbatim, adds Scene 2 to T's context as an overheard exchange, and reruns Scene 3. This paired design separates predicting behavior within a given information state from reasoning about how behavior changes when that state is altered. In both versions, T's decision is labeled *complied* only if T fully agrees to M's requested decision.

3.3 Mental States Recorded During the Story

A mental state annotated after a story is complete must be reconstructed in retrospect, and one measured at a single moment gives no trajectory to ask about. The role-playing characters in ManipToM therefore externalize their private states while the story is generated, in the manner of ToMATO (Shinoda et al., 2025), which yields temporally aligned state trajectories for all three characters. The records have two granularities: a turn-level private thought accompanies each utterance, and a structured *checkpoint* summarizes each present character's states at the end of every scene (T files none for Scene 2). Checkpoints cover belief, intention, desire, and emotion (Beaudoin et al., 2020) at first order (a character's own state) and second order (a character's view of another's; Le et al., 2019); thoughts carry one role-fixed dimension (Appendix A2). Checkpoints provide the answer keys for all mental-state questions; turn-level thoughts are used only to construct distractors and to validate consistency between states and dialogue (§4). Dynamic answer keys thus operationalize change as a difference in a character's *reported* state content between two checkpoints, not as a claim about unreported inner change (Appendix A3). The evaluated model sees neither and receives only the spoken dialogue (protocol in Appendix A2).

3.4 Questions Testing Explicit, Dynamic, and Applied ToM

Each story yields up to 32 questions on three complementary tracks: reading a mental state at one moment, tracking how it changes over time, and using it to predict behavior. The tracks build on the distinctions between explicit, dynamic, and applied ToM in developmental psychology and LLM evaluation (Hutchins et al., 2016; Lee et al., 2024; Gu et al., 2026; Xiao et al., 2025). The full item list is in Table 6 (Appendix A3).

Explicit ToM: reading a state. Questions ask for a character's mental state at a single checkpoint, including second-order states such as what T believes M intends, on both the deceived and the informed side of the manipulation.

Dynamic ToM: tracking a state. Questions ask whether and how a character's state changes between two checkpoints, including cases in which it remains unchanged.

Applied ToM: using a state. Questions ask the model to predict T's decision under four information conditions, with the dialogue cut immediately before T announces it:

- **Outcome:** predict T's factual decision from the full transcript, including Scene 2.
- **Outcome without Scene 2:** predict the same factual decision when Scene 2 is withheld from the evaluated model.
- **Counterfactual outcome:** predict T's decision in the fork where T overheard Scene 2.
- **Flip:** decide whether T's decision differs between the factual and counterfactual stories.

Together, these conditions separate access to privileged information from the ability to reason about its effect on another person's behavior.

4 ManipToM Creation

Figure 3 illustrates the creation pipeline. Answer keys for mental-state questions are taken directly from the checkpoints recorded during generation, while state changes and story outcomes are judged

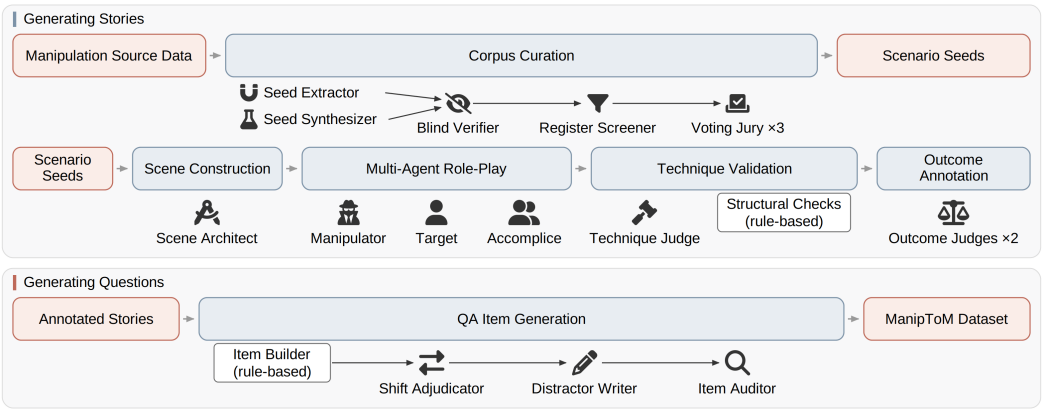


Figure 3: The ManipToM creation pipeline. Each icon represents one LLM role, while white boxes denote rule-based steps that do not involve a model. Curated seeds are expanded into scenes, enacted by three role-playing agents, validated for manipulation-mechanism fidelity, labeled for outcomes by two independent judges, and compiled into questions.

by LLMs. GPT-5.4 role-plays the characters, and the models used for the remaining roles are listed in Appendix A2.

4.1 Generating Stories

Seeds. Each story is generated from a seed consisting of a hidden goal, a claim, a vulnerability, and a manipulation mechanism. The *seed extractor* mines seeds from the manipulative dialogues of MentalManip (Wang et al., 2024); since several mechanisms could not be found there in sufficient numbers, the *seed synthesizer* writes additional seeds for them, which account for 337 of the final 645 stories (52%). Every seed then passes the same audit. The *blind verifier* re-classifies each seed without seeing its proposed label and discards those that cannot be staged as an everyday conversation; the *register screener* discards outlandish premises, those that need crime, celebrity, or life-or-death stakes to make sense. A *voting jury* of three different models then labels the mechanism blindly; the majority label is final, and seeds with a three-way split are discarded.

Scenes and role-play. The *scene architect* expands the seed into a shared scenario, per-character private facts, and M's concrete goal. Three LLM agents then role-play M, A, and T, each receiving only the utterances spoken in its presence, and T's context therefore contains no Scene 2. At the end of each scene, every character present reports its checkpoint mental states, which define the reference states from which answer keys are derived. The counterfactual fork of is produced by the same role-play, resumed from the altered context (details in Appendix A2).

Story validation. The generated stories then go through two checks. Rule-based checks first verify turn order, speaker prefixes, checkpoint completeness, and near-duplicate rates among thoughts and utterances. The *technique judge* then answers two prespecified binary questions, judged against the private facts M holds: whether M applied real pressure, and whether the labeled mechanism accounts for at least half of it. Overall, 703 of the 1,094 candidate stories (64%) passed, and most failures contained no realized pressure. Failed stories are discarded as a whole and never regenerated (Appendix A2); per-mechanism candidate and story counts are in Appendix A8.

Outcome annotation. Two *outcome judges*, each a different model, independently label the ending of each story and of its fork as complied or refused. A story enters the benchmark only if both agree on its ending, yielding the final 645 stories.

4.2 Generating Questions

The *item builder* assembles questions by rule from checkpoint coordinates fixed in advance (see Table 6), with the correct answer restated in the third person. The *distractor writer* then draws wrong options from other mental states recorded in the same story, each wrong for an annotated reason (a different moment, a different character, or information inaccessible to the questioned character),

Table 2: Evaluation of LLMs on ManipToM. Left: LLMs with standard prompting; right: LLMs with explicit CoT prompting. Subscripts denote 95% bootstrap confidence intervals clustered by story. Explicit and Dynamic report multiple-choice accuracy (%), while Applied reports balanced accuracy (%) averaged over outcome, counterfactual outcome, and flip prediction, for which the majority-class baseline is 50%. More details are in Appendix A5.

Model	Explicit	Dynamic	Applied	AVG.
Llama-3.1-8B	$68.6_{\pm 1.2}$	$24.6_{\pm 1.3}$	$56.4_{\pm 2.6}$	$49.9_{\pm 1.1}$
Llama-3.3-70B	$88.1_{\pm 0.8}$	$44.4_{\pm 1.4}$	$66.3_{\pm 3.2}$	$66.3_{\pm 1.2}$
Mixtral-8×7B	$69.6_{\pm 1.2}$	$35.1_{\pm 1.4}$	$63.6_{\pm 2.0}$	$56.1_{\pm 1.0}$
Gemma-3-27B	$85.0_{\pm 0.9}$	$36.7_{\pm 1.3}$	$64.9_{\pm 3.1}$	$62.2_{\pm 1.2}$
Qwen3.5-4B	$80.1_{\pm 1.0}$	$41.7_{\pm 1.4}$	$54.7_{\pm 2.6}$	$58.8_{\pm 1.0}$
Qwen3.5-27B	$83.2_{\pm 0.9}$	$\mathbf{51.1}_{\pm 1.4}$	$67.1_{\pm 2.9}$	$67.1_{\pm 1.1}$
GPT-4o	$86.6_{\pm 0.9}$	$41.6_{\pm 1.3}$	$60.8_{\pm 3.1}$	$63.0_{\pm 1.1}$
GPT-5.4	$\mathbf{89.5}_{\pm 0.8}$	$45.9_{\pm 1.4}$	$\mathbf{71.8}_{\pm 3.0}$	$\mathbf{69.0}_{\pm 1.1}$
LLM Avg.	81.7	39.9	63.5	61.7

Model + CoT	Explicit	Dynamic	Applied	AVG.
Llama-3.1-8B	$70.3_{\pm 1.2}$	$24.2_{\pm 1.2}$	$59.1_{\pm 2.9}$	$51.2_{\pm 1.2}$
Llama-3.3-70B	$\mathbf{90.7}_{\pm 0.8}$	$45.0_{\pm 1.4}$	$65.4_{\pm 2.9}$	$67.0_{\pm 1.1}$
Mixtral-8×7B	$67.2_{\pm 1.3}$	$30.8_{\pm 1.3}$	$63.7_{\pm 2.7}$	$53.9_{\pm 1.1}$
Gemma-3-27B	$83.6_{\pm 0.9}$	$36.9_{\pm 1.5}$	$65.2_{\pm 3.2}$	$61.9_{\pm 1.2}$
GPT-4o	$87.5_{\pm 0.9}$	$40.8_{\pm 1.3}$	$67.4_{\pm 3.0}$	$65.2_{\pm 1.1}$
GPT-5.4	$90.3_{\pm 0.7}$	$\mathbf{49.5}_{\pm 1.4}$	$\mathbf{69.8}_{\pm 3.2}$	$\mathbf{69.9}_{\pm 1.2}$
LLM + CoT Avg.	81.9	38.1	65.1	61.7
Lexical overlap	22.0	22.0	–	–
Majority label	–	–	50.0	–

together with at most one minimal edit of the correct answer. For dynamic questions, the *shift adjudicator* decides whether the probed state actually changed and discards ambiguous cases; as a result, *No change* is the correct answer for 31.3% of them. All options of a question are then rewritten into a common style so that surface form does not reveal the answer.

Finally, the *item auditor*, a strong model kept separate from the generation stages, checks each question. Given the complete record (dialogue, thoughts, checkpoints, and private facts), it keeps a question only if the keyed option is the single correct answer (Appendix A3). We complemented this automatic audit by inspecting samples of stories and questions throughout development. This results in the final ManipToM dataset of 645 stories, 11,252 multiple-choice questions, and 2,482 binary questions (statistics in Appendix A8).

5 Experiments

5.1 Experimental Setup

We evaluate ManipToM on nine LLMs spanning different model families, access types, and parameter scales: Llama-3.1-8B, Llama-3.3-70B (Grattafiori et al., 2024), Mistral-Small-24B (Mistral AI, 2025), Mixtral-8×7B (Jiang et al., 2024), Gemma-3-27B (Gemma Team, 2025), Qwen3.5-4B, Qwen3.5-27B (Qwen Team, 2026), GPT-4o (OpenAI, 2024), and GPT-5.4 (OpenAI, 2026).

We evaluate two inference conditions: standard prompting and explicit CoT prompting (Wei et al., 2022), evaluated on the seven non-Qwen models. The two Qwen models are evaluated with their native thinking disabled. All generations use a temperature of 0. Appendix A5 provides the complete evaluation protocol, prompts, model configurations, metrics, and baselines.

5.2 Evaluation on LLMs

Between models. Table 2 summarizes the evaluation. Under standard prompting, GPT-5.4 achieves the best average (69.0%), although it also generated the dialogues, and we read its lead with that caveat; Qwen3.5-27B (67.1%) and Llama-3.3-70B (66.3%) follow closely. Across the seven paired models, chain-of-thought prompting helps some slightly (up to $+2.2$ percentage points for GPT-4o) but averages only $+0.4$, consistent with SimpleToM and DynToM (Gu et al., 2026; Xiao et al., 2025).

Between question types. Models achieve high explicit accuracy (81.7% on average), while dynamic accuracy remains low (39.9%) and applied performance is uneven across question types (63.5% balanced accuracy on average, with the flip item near chance; Table 9). The explicit and dynamic tracks share the same four-option format, and every configuration drops from one to the other by 32.1 to 48.3 percentage points; applied items are binary and scored as balanced accuracy, so their level is not directly comparable to the other two tracks. Furthermore, strong explicit performance does not consistently extend to the other two abilities: when we split stories by how well a model read and tracked them, its outcome-prediction balanced accuracy stays within 70.2% to 76.2% across all four cells (Table 10).

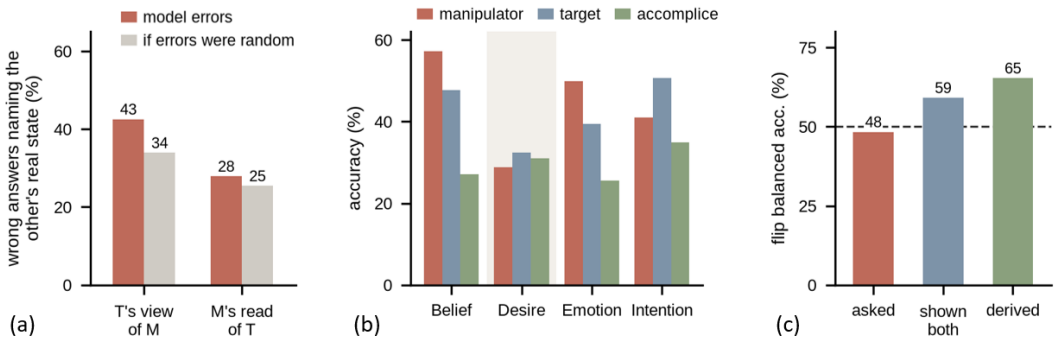


Figure 4: Error analysis by question type. (a) Explicit: on the items in §5.3, the share of wrong answers that pick a distractor built from the other character's own recorded state, against the share expected if errors fell on the wrong options at random, pooled over the nine models under standard prompting. (b) Dynamic: accuracy by the dimension a question probes and the character it asks about, pooled over the nine models under standard prompting. (c) Applied: balanced accuracy of Llama-3.3-70B at judging whether the target's ending differs between the factual and counterfactual dialogues, under three evaluation conditions: the benchmark's flip item, in which the model reads the factual dialogue and is told that T overheard the private scene (*asked*); both dialogues shown, each cut before T's deciding turn (*shown both*); and the answer implied by the model's own two outcome predictions, one per dialogue, without asking (*derived*). All three conditions are scored on the same 582 stories, those with a parsable reply under every condition (Appendix A5). Dashed line: chance.

Between mechanisms. In contrast, pooled multiple-choice accuracy varies little across the 16 mechanisms, from 57.6% to 65.2%, with an ordering largely shared between models (Appendix A7).

5.3 Error Analysis by Question Type

Explicit: the manipulator's states intrude into the deceived target's view. Explicit questions probe two second-order mental states: (i) the target's reading of what the manipulator intends and wants (`t_models_m_intention`, `t_models_m_desire`) and (ii) the manipulator's reading of what the target believes (`m_reads_t`, `leverage`). Models answer the explicit track well overall (81.7% on average), but on these two question types their errors are strongly asymmetric (Figure 4a). Conditional on an incorrect response to (i), models select a distractor rendered from the manipulator's own recorded state in 42.5% of cases, compared with 33.9% expected if errors fell on the wrong options at random; every model shows this pull. For (ii), by contrast, models select a distractor rendered from the target's own recorded state in 27.9% of erroneous responses, only slightly above the corresponding expectation of 25.4% and not consistently across models. Thus, the manipulator's own states intrude when models infer the deceived target's view far more than the target's states intrude in the reverse direction.

Dynamic: desire is the hardest state to track for the manipulator and the target. Dynamic questions ask whether a given mental state changed between scenes, and accuracy is far lower here for every configuration (39.9% on average). Figure 4b breaks this down by dimension: models track desire worst for both parties to the manipulation, at 28.8% for the manipulator and 32.5% for the target, against 57.1% and 50.6% on their best-tracked dimensions. This possibly indicates that desire is the most implicit of the four dimensions: beliefs, emotions, and intentions surface in what characters say, whereas what a character wants is rarely stated outright.

Applied: models struggle to judge whether counterfactual outcomes differ. Applied questions ask where the conversation ends in the factual world, where it ends in a counterfactual world in which the target overheard the private scene, and whether the two endings differ (the *flip* question). Most models answer the first two well above chance, but nearly all perform near chance on the flip question; only GPT-5.4 clearly exceeds it, at 60.1% balanced accuracy (Table 9, Appendix A5). To examine whether the difficulty persists when the counterfactual continuation is provided, we compare three evaluation conditions for Llama-3.3-70B (Figure 4c; protocol in Appendix A5); all three conditions are scored on the same 582 stories, those with a parsable reply under every condition. Shown both dialogues, each cut before the target's deciding turn, the model's flip balanced accuracy rises from 48.3% to 59.1%, so part of the difficulty lies in having to infer the unseen counterfactual continuation

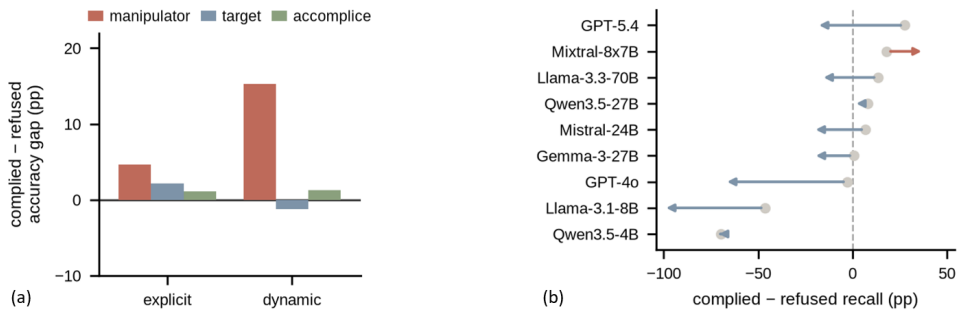


Figure 5: (a) Llama-3.3-70B's accuracy on complied minus refused stories, by track and by the character the question asks about. (b) Recall gap between complied and refused endings on outcome questions; each arrow runs from the factual condition (dot) to the counterfactual, overhearing condition (head), and is blue when the gap moves toward refusal and red otherwise.

from the premise alone. On the same truncated dialogues, the flip answer implied by the model's own two outcome predictions, compared outside the model, scores 65.3%, numerically above the direct judgment, though the paired interval for this last difference includes zero (Appendix A5).

5.4 Evaluation by Story Outcome: Compliance vs. Refusal

We examine whether performance differs between stories in which the target ultimately complies and those in which the target refuses; targets comply in 18.6% of stories (Appendix A6 analyses which mechanisms succeed).

Mental-state reasoning. Models track the manipulator's mental-state changes more accurately in complied stories: every evaluated model scores higher on dynamic questions about M when the manipulation succeeds, by 5.5 to 20.7 percentage points (per-model values in Table 13). Figure 5a shows the pattern for Llama-3.3-70B: its complied-minus-refused gap on dynamic questions is $+15.3$ points for the manipulator, against -1.3 for the target and $+1.3$ for the accomplice, and explicit questions show far smaller gaps (at most $+4.6$, again for M). The gap is not explained solely by complied stories containing more "changed" answers: scoring changed and unchanged items separately, the advantage remains ($+11.7$ and $+11.6$ points). For this model, the outcome-related accuracy gap is therefore largest when tracking the manipulator's mental-state changes.

Factual outcome prediction. On the applied track the story outcome is itself the answer label, so we compare recall on complied and refused endings, counting unparsable replies as errors as in Table 9 (per-model detail in Table 13). Even the higher-performing models struggle to identify refusals: GPT-5.4 correctly identifies 94.2% of complied endings but only 66.7% of refused ones, and Llama-3.3-70B shows the same pattern (82.5% vs. 69.1%). The lower-performing models show the reverse (Llama-3.1-8B recovers 45.0% of complied endings vs. 91.4% of refused ones; Qwen3.5-4B, 26.7% vs. 96.4%).

Counterfactual prediction and flip detection. Asked for the ending under the premise that the target overheard the private scene, seven of the nine models shift their recall gap toward refusal (Figure 5b; Qwen3.5-4B, already betting on refusal, barely moves, and only Mixtral-8×7B moves the other way): GPT-5.4 now identifies 65.9% of complied endings and 83.7% of refused ones, and Llama-3.3-70B 67.1% and 82.5%. We also examine whether flip detection differs with the direction of the outcome change: seven of nine models detect complied→refused flips more often than refused→complied ones (one tie; one-sided sign test across models, $p = 0.035$). For GPT-5.4 and Llama-3.3-70B, the harder outcome to identify therefore reverses across conditions: refusal in factual stories, but compliance under the overhearing premise.

6 Conclusion

We introduce ManipToM, a ToM benchmark in psychological manipulation, grounded in 16 manipulation mechanisms. Across nine LLMs, strong performance on explicit mental states does not reliably extend to tracking state changes or predicting behavior, and CoT prompting does not close this gap. We hope ManipToM serves as a testbed for reasoning under strategic social influence.

AI USE STATEMENT

LLMs are both the subject of this work and part of its method: as described in Section 4 and Appendix A2, every story, mental-state record, and question in ManipToM was produced by LLM agents under the prompts and rules reported there, and the benchmark evaluates LLMs. Beyond that, we used generative AI tools to assist with writing and debugging the analysis code, with producing figures, and with editing the text of this paper; Figure 1 is an image generated with ChatGPT from a story and layout specified by the authors. All AI-assisted code was reviewed and tested by the authors, all reported numbers were recomputed from the released data, and all AI-assisted text was reviewed and revised by the authors. We did not use generative AI tools to generate research ideas or to conduct the literature review. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

ManipToM contains fictional dialogues written by role-playing LLM agents; no human subjects took part, and no personal data was used. Its seeds are derived from MentalManip (Wang et al., 2024), a publicly released research dataset, and from synthesized descriptions; the stories themselves are new text. The seed screening favors ordinary settings and removes premises that need crime, celebrity, or life-or-death stakes (Section 4.1), although some stories take place in professional or institutional settings; the dialogues contain no instructions for harming people. The benchmark measures whether models understand the mental states inside a manipulative interaction, which is the ability a model needs to recognize and resist manipulation; it does not train or reward manipulative behavior. We are aware that the generation pipeline could in principle be repurposed to produce manipulative text, and we will release the dataset and code under a license that restricts use to research and evaluation.

REPRODUCIBILITY STATEMENT

The construction of ManipToM is specified in Section 4, with the models used at each stage, the information given to each character, and summaries of the role briefs, stage directions, validation rules, and outcome annotation rules in Appendix A2, the question construction and validation rules in Appendix A3, and a complete story with its checkpoints, fork, and questions in Appendix A4. The evaluation protocol, the exact model checkpoints and inference settings, the prompts, the metrics, and the bootstrap procedure are given in Appendix A5, and dataset statistics in Appendix A8. The appendix summarizes the generation prompts; their full text, the dataset, the generation and evaluation code, and the per-item model outputs will be released upon acceptance.

REFERENCES

Kate Abramson. Turning up the lights on gaslighting. *Philosophical Perspectives*, 28(1):1–30, 2014.

Solomon E. Asch. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985.

C. Daniel Batson. *The Altruism Question: Toward a Social-Psychological Answer*. Lawrence Erlbaum Associates, 1991.

Roy F. Baumeister, Arlene M. Stillwell, and Todd F. Heatherton. Guilt: An interpersonal approach. *Psychological Bulletin*, 115(2):243–267, 1994.

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, 10:2905, 2020.

Carla Capuano and Peggy Chekroun. A systematic review of research on conformity. *International Review of Social Psychology*, 37(1):13, 2024. doi: 10.5334/irsp.874.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. NegotiationToM: A benchmark for stress-testing machine theory of mind on negotiation surrounding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4211–4241, 2024.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 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)*, pp. 15959–15983, 2024.

Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621, 2004.

Christian Coons and Michael Weber (eds.). *Manipulation: Theory and Practice*. Oxford University Press, 2014. doi: 10.1093/acprof:oso/9780199338207.001.0001.

Daniel C. Dennett. Précis of The Intentional Stance. *Behavioral and Brain Sciences*, 11(3):495–505, 1988.

Dariusz Dolinski and Richard Nawrat. "fear-then-relief" procedure for producing compliance: Beware when the danger is over. *Journal of Experimental Social Psychology*, 34(1):27–50, 1998.

Naomi Ellemers, Jojanneke van der Toorn, Yavor Paunov, and Thed van Leeuwen. The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review*, 23(4):332–366, 2019. doi: 10.1177/1088868318811759.

Patrick Fagan. Clicks and tricks: The dark art of online persuasion. *Current Opinion in Psychology*, 58:101844, 2024. doi: 10.1016/j.copsyc.2024.101844.

Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press, 2007.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Alvin W. Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2):161–178, 1960.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jörg Gross and Alexander Vostroknutov. Why do people follow social norms? *Current Opinion in Psychology*, 44:1–6, 2022. doi: 10.1016/j.copsyc.2021.08.016.

Dan W. Grupe and Jack B. Nitschke. Uncertainty and anticipation in anxiety: An integrated neurobiological and psychological perspective. *Nature Reviews Neuroscience*, 14(7):488–501, 2013.

Yuling Gu, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi. SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

Ying Guo, Xin Tian, Dong Hu, Shulin Bai, and Shuxi Zhou. The effects of shame on prosocial behavior: A systematic review and three-level meta-analysis. *Advances in Psychological Science*, 31(3):371–385, 2023. doi: 10.3724/SP.J.1042.2023.00371.

Tiffany L. Hutchins, Patricia A. Prelock, Heather Morris, Jessica Benner, Taylor LaVigne, and Betsy Hoza. Explicit vs. applied theory of mind competence: A comparison of typically developing males, males with ASD, and males with ADHD. *Research in Autism Spectrum Disorders*, 21: 94–108, 2016.

Alice M. Isen. Positive affect, cognitive processes, and social behavior. In *Advances in Experimental Social Psychology*, volume 20, pp. 203–253. Academic Press, 1987.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Edward E. Jones. *Ingratiation: A Social Psychological Analysis*. Appleton-Century-Crofts, 1964.

Dana L. Joseph, Micaela Y. Chan, Samantha J. Heintzelman, Louis Tay, Ed Diener, and Victoria S. Scotney. The manipulation of affect: A meta-analysis of affect induction procedures. *Psychological Bulletin*, 146(4):355–375, 2020. doi: 10.1037/bul0000224.

Danush Khanna, Pratinav Seth, Sidhaarth Sredharan Murali, Aditya Kumar Guru, Siddharth Shukla, Tanuj Tyagi, Sandeep Chaurasia, and Kripabandhu Ghosh. SELF-PERCEPT: Introspection improves large language models' detection of multi-person mental manipulation in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 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 (EMNLP)*, 2023.

Willis Klein, Suzanne Wood, and Jennifer A. Bartz. A theoretical framework for studying the phenomenon of gaslighting. *Personality and Social Psychology Review*, 30(2):195–215, 2026a. doi: 10.1177/10888683251342291.

Willis Klein, Suzanne Wood, Audrey-Anne Forget, and Jennifer A. Bartz. A historical review of gaslighting: Tracing changing conceptualizations within psychiatry and psychology. *Clinical Psychology Review*, 126:102742, 2026b. doi: 10.1016/j.cpr.2026.102742.

Esben Kran, Hieu Minh Nguyen, Akash Kundu, Sami Jawhar, Jinsuk Park, and Mateusz Maria Jurewicz. DarkBench: Benchmarking dark patterns in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 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)*, pp. 5872–5877, 2019.

S. Lee, C. Huang, I. Fu, and K. Chen. Interpreting the results of explicit and applied theory of mind collectively in autistic children: A solution from Rasch analysis. *Autism*, 28(2):355–366, 2024.

Yiwei Liu, Emma Jane Pretty, Jiahao Huang, and Saku Sugawara. TactfulToM: Do LLMs have the theory of mind ability to understand white lies? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Steven A. McCornack. Information manipulation theory. *Communication Monographs*, 59(1):1–16, 1992.

Mistral AI. Mistral small 3. `https://mistral.ai/news/mistral-small-3`, 2025.

Robert Noggle. Manipulative actions: A conceptual and moral analysis. *American Philosophical Quarterly*, 33(1):43–55, 1996.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. Gpt-5.4 thinking system card. `https://openai.com/index/gpt-5-4-thinking-system-card/`, 2026.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, 2026.

Todd Rogers, Richard Zeckhauser, Francesca Gino, Michael I. Norton, and Maurice E. Schweitzer. Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, 112(3):456–473, 2017. doi: 10.1037/pspi0000081.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Melanie Sclar, Jane Dwivedi-Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. Explore theory of mind: Program-guided adversarial data generation for theory of mind reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. ToMATO: Verbalizing the mental states of role-playing LLMs for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Michelle N. Shiota, Esther K. Papies, Stephanie D. Preston, and Disa A. Sauter. Positive affect and behavior change. *Current Opinion in Behavioral Sciences*, 39:222–228, 2021. doi: 10.1016/j.cobeha.2021.04.022.

C. R. Snyder. Hope theory: Rainbows in the mind. *Psychological Inquiry*, 13(4):249–275, 2002.

Ju-Hyun Song, Tyler Colasante, and Tina Malti. Helping yourself helps others: Linking children's emotion regulation to prosocial behavior through sympathy and trust. *Emotion*, 18(4):518–527, 2018. doi: 10.1037/emo0000332.

Russell Spears. Social influence and group identity. *Annual Review of Psychology*, 72:367–390, 2021. doi: 10.1146/annurev-psych-070620-111818.

Claude M. Steele. The psychology of self-affirmation: Sustaining the integrity of the self. In *Advances in Experimental Social Psychology*, volume 21, pp. 261–302. Academic Press, 1988.

Daniel Susser, Beate Roessler, and Helen Nissenbaum. Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2), 2019. doi: 10.14763/2019.2.1410.

Paige L. Sweet. The sociology of gaslighting. *American Sociological Review*, 84(5):851–875, 2019.

Ming Tang, Weiqiang Li, Fuhui Liu, and Bo Yuan. The association between guilt and prosocial behavior: A systematic review and meta-analysis. *Advances in Psychological Science*, 27(5):773–788, 2019. doi: 10.3724/SP.J.1042.2019.00773.

June Price Tangney and Ronda L. Dearing. *Shame and Guilt*. Guilford Press, 2002.

June Price Tangney, Jeff Stuewig, and Debra J. Mashek. Moral emotions and moral behavior. *Annual Review of Psychology*, 58:345–372, 2007.

Melanie B. Tannenbaum, Justin Hepler, Rick S. Zimmerman, Lindsey Saul, Samantha Jacobs, Kristina Wilson, and Dolores Albarracín. Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*, 141(6):1178–1204, 2015. doi: 10.1037/a0039729.

Jessica L. Tracy and Richard W. Robins. The psychological structure of pride: A tale of two facets. *Journal of Personality and Social Psychology*, 92(3):506–525, 2007.

Gerben A. Van Kleef and Stéphane Côté. The social effects of emotions. *Annual Review of Psychology*, 73:629–658, 2022. doi: 10.1146/annurev-psych-020821-010855.

Gerben A. Van Kleef and Gert-Jan Lelieveld. The social functions of positive emotions. *Current Opinion in Behavioral Sciences*, 39:41–45, 2021. doi: 10.1016/j.cobeha.2020.12.009.

Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. MentalManip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

W. Craig Williams, Sylvia A. Morelli, Desmond C. Ong, and Jamil Zaki. Interpersonal emotion regulation: Implications for affiliation, perceived support, relationships, and well-being. *Journal of Personality and Social Psychology*, 115(2):224–254, 2018. doi: 10.1037/pspi0000132.

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

Kim Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, 59(4):329–349, 1992.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 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*, pp. 10691–10706, 2023.

Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 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 (ACL)*, 2025.

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

Fangxu Yu, Lai Jiang, Shenyi Huang, Zhen Wu, and Xinyu Dai. PersuasiveToM: A benchmark for evaluating machine theory of mind in persuasive dialogues. *arXiv preprint arXiv:2502.21017*, 2025.

Zeyang Yue, Chenfei Yan, Feifei Zhao, Haibo Tong, Mengwen Xu, Xiaozhen Wang, Erliang Lin, and Yi Zeng. CogManip: Benchmarking manipulative behavior in multi-turn interactions with large language models. *arXiv preprint arXiv:2606.06099*, 2026.

Jamil Zaki. Integrating empathy and interpersonal emotion regulation. *Annual Review of Psychology*, 71:517–540, 2020. doi: 10.1146/annurev-psych-010419-050830.

Chunhui Zhang, Zhongyu Ouyang, Kwonjoon Lee, Nakul Agarwal, Sean Dae Houlihan, Soroush Vosoughi, and Shao-Yuan Lo. Overcoming multi-step complexity in multimodal theory-of-mind reasoning: A scalable Bayesian planner. In *International Conference on Machine Learning (ICML)*, 2025a.

Zhining Zhang, Chuanyang Jin, Mung Yao Jia, and Tianmin Shu. AutoToM: Scaling model-based mental inference via automated agent modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Jike Zhong, Yuxiang Lai, Ming Li, Yuheng Li, Wuao Liu, Behzad Dariush, Konstantinos Psounis, and Shao-Yuan Lo. From shortcuts to reasoning: Robust post-training of theory of mind with reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026.

Appendix

A1 Full manipulation taxonomy

Domain assignment. Domains are assigned by a *removal test*, given verbatim to the LLMs that extract and label each seed (§4.1): *if the manipulator's claim were removed entirely, why would the manipulation fail?*

- A false or misleading fact would no longer exist → **cognitive**.
- An emotional action tendency would no longer trigger → **affective**.
- A threat to self-concept or values would no longer land → **identity**.

Because the test asks *how* the manipulation works and not *what* it is about, the same request can fall under different domains: asking for a favor is affective when it leans on a debt of gratitude, and identity-based when refusing is framed as a moral failure.

Mechanism definitions. Table 3 defines the 16 mechanisms and lists the psychological work each is grounded in. During role-play, M receives a short form of the definition in place of the mechanism's name, since a role-playing model given only a label tends to fall back on ordinary argument.

A2 Story generation details

Model assignment. Table 4 lists the model used at each stage. We pinned role-play to GPT-5.4 after a pilot on the same five seeds: role-play by GPT-5.6-luna passed validation in none of them, because its manipulator revealed the hidden goal in private and then abandoned it at the first objection. GPT-4o and GPT-5.4 each passed four of five, but GPT-4o's checkpoint reports varied little across scenes and characters, a qualitative observation in the pilot that matters because checkpoints are the source of every answer key. Validation uses GPT-5.6-terra, which produces no story or question content.

Decoding settings. All generation calls go through the OpenAI Responses API with its default sampling settings; no temperature or sampling seed is set, so replaying a scene is stochastic. The per-story random seed controls only the sampled attributes of Table 5 and the construction of questions.

Table 4: Models used at each stage of the pipeline.

Stage	Model
Seed extractor and seed synthesizer (§4.1)	GPT-5.4
Blind verifier and register screener (§4.1)	GPT-5.6-luna
Voting jury, mechanism label by majority (§4.1)	GPT-5.4, GPT-4o, GPT-5.6-luna
Scene architect (§4.1)	GPT-5.6-luna
Role-play agents M, A, T, original and fork (§4.1)	GPT-5.4
Distractor writer and shift adjudicator (§4.2)	GPT-5.6-luna
Outcome judges (§4.1)	GPT-5.6-terra and GPT-5.6-luna
Technique judge and item auditor (§4.1, §4.2)	GPT-5.6-terra

Information given to each character. The three characters start with different information (Table 5). All share a scenario: the setting, a decision to be made, and T's initial leaning, which runs against M's goal. Private facts, such as a favor T owes M or a document only M has read, are given only to the characters they concern and carry the material the mechanism works with. The hidden goal and the mechanism's definition are given to M alone.

Table 3: Definitions and psychological grounding of the 16 manipulation mechanisms.

Mechanism	Definition: the manipulator…	Grounding
Cognitive: acts on the target's picture of reality		
Reality distortion	systematically denies or distorts what T perceives or remembers	Abramson (2014); Sweet (2019)
Epistemic discrediting	attacks T's credibility as a knower, without denying any specific fact	Fricker (2007)
Misleading framing	selects which true facts to present, creating a false overall impression	Rogers et al. (2017)
Information control	actively controls what information T can access	McCornack (1992)
Affective, negative: induces an emotion that pushes toward compliance		
Fear	induces fear of a concrete threat that complying would avert	Witte (1992)
Anxiety	induces diffuse worry about an uncertain future that complying would ease	Grupe & Nitschke (2013)
Guilt	induces guilt, so that complying repairs a perceived wrong toward M	Baumeister et al. (1994)
Shame	induces shame about T's self, so that complying restores standing	Tangney & Dearing (2002)
Sympathy	induces sympathy for M's hardship, so that complying relieves it	Batson (1991)
Affective, positive: exploits an emotion T already has or can be given		
Joy	exploits T's good mood or excitement to lower scrutiny	Isen (1987)
Pride	exploits T's pride through flattery of T's abilities or status	Tracy & Robins (2007); Jones (1964)
Gratitude	exploits T's gratitude for a past favor, so that complying repays it	Gouldner (1960)
Hope	exploits T's hope for a desired future that complying seems to bring closer	Snyder (2002)
Identity: acts on the target's self-concept and values		
Moral coercion	frames the request as a moral obligation, so that refusing is a moral failure	Tangney et al. (2007)
Social proof exploitation	uses group norms to make T doubt T's own judgment	Asch (1956); Cialdini & Goldstein (2004)
Identity threat and rescue	creates a crisis in how T sees themselves, then offers M as the answer	Steele (1988); Dolinski & Nawrat (1998)

Table 5: Information given to each character before the conversation. Each character also receives its own name, relationships, and a Big Five personality profile.

Information	M	A	T
Shared scenario (setting, decision, T's leaning)	✓	✓	✓
Private facts (e.g., a favor owed)	only the characters a fact concerns		
Hidden goal and the private reason for it	✓	in Scene 2	–
Definition of the manipulation mechanism	✓	–	–

Sampled attributes. Three names are drawn from a list of 98. Each character's Big Five profile is five independent high/low draws, rendered as adjectives (e.g., "introverted", "agreeable"). The M–T relationship is drawn from labels matching the seed's relationship type (intimate, professional, social, or stranger); the M–A relationship is always a confidant (e.g., a sibling or a close friend); and A knows T only through M. To vary how readily T changes their mind, T is also assigned a resistance disposition, sampled as guarded (25%), ordinary (40%), or accommodating (35%). The disposition only shapes T's brief and does not determine the outcome: T complied in 14%, 23%, and 17% of stories under the three dispositions, respectively.

Character briefs. All three characters are told not to invent events, evidence, deadlines, guarantees, or consequences, and that every turn must move the conversation.

- **M** is given the hidden goal, told never to state it to T, to listen first for where T is vulnerable, to raise the request once, and, if T does not agree, to change the ground instead of repeating it, using only the assigned mechanism. For cognitive mechanisms, M is additionally told that withholding, selectively quoting, or slanting the private facts is the mechanism itself, whereas inventing facts is not allowed.
- **A** is told to be on M's side, to never carry anything M says in private to T, to ask M in private what M is actually after, and not to take over the asking.
- **T** is told to settle the matter without a fight and without M thinking less of T, but not to be talked into anything T does not agree with, and to answer M's first request honestly instead of settling it on the spot. The resistance disposition adds one of three instructions on how readily T changes their mind.

Mental-state annotation. Each utterance carries a private thought that opens with a fixed prefix, so its mental-state category is known by construction. Unlike ToMATO, which uses one prefix for all characters in a conversation, we assign prefixes by role: "I will…" (intention) for M and "I think…" (belief) for A and T, so that M's real intention and T's belief land side by side on adjacent turns. T adds one more clause on the domain under attack: what T takes to be true in cognitive stories ("I know…"), how T feels in affective ones ("I feel…"), and what T wants in identity ones ("I want…"). At the end of every scene, each character present files a checkpoint stating its belief, intention, desire, and emotion about themselves and about each of the other two characters (e.g., "I think that M wants…"). A story accumulates eight checkpoints: three each for M and A, and two for T, who is out of the room for Scene 2.

Scene length and turn order. Scene 1 has 5 turns (M, A, T, M, T) and Scene 2 has 4 (M, A, M, A). Scene 3 opens with M, A, M and then alternates between T and M, with T always speaking last; it closes after 6, 8, or 10 turns, as soon as T has given a decision (see below). Stories therefore run for 15, 17, or 19 turns (78%, 12%, and 10% of stories) and average 698 words of spoken dialogue.

Stage directions. Stage directions are private notes to one agent:

- **End of Scene 1:** T is told to step out for about ten minutes, with a suggested reason (e.g., a phone call, someone at the door) that varies across stories.
- **Scene 2:** M and A are told T has left, and reminded every third turn. M is told to say what M actually wants, why, and why T would resist, and to plan with A; A is told to ask what M is after. An utterance addressed to the absent T is regenerated.
- **Start of Scene 3:** T is told they were away and do not know what was said; M and A are told T is back; M is reminded of the goal and the mechanism. If M has not raised the request by 60% of the scene, M is reminded that time is running out.

- **Ending:** from T's second turn of Scene 3 onward, a yes/no LLM check runs after each of T's turns (*has T given a clear answer to what M is asking for, agreeing, refusing, or committing to a specific compromise? Answer no if T is still weighing it, asking questions, or has only agreed to keep talking*); on *yes*, the scene ends. The check errs toward continuing, and a decision given in T's first turn is always followed by at least one more exchange. The outcome judges later place the deciding turn before T's last turn in 102 of the 645 factual stories and in 272 of the 596 forks, almost always by one T turn. Applied items cut the transcript at the judges' deciding turn, not at the last turn.

Seeds. Synthesized seeds are each conditioned on a sampled setting and relationship. All seeds pass the same screening, and three models (GPT-5.4, GPT-4o, and GPT-5.6-luna) label each seed's mechanism independently with the removal test of Appendix A1; the majority label is kept, and it is unanimous for 73% of stories. Personal attributes are sampled by rule and never written by a model, which prevents the generator from picking variables it would correlate with the outcome.

Counterfactual fork. Scenes 1 and 2 are replayed verbatim into every character's history. M and A receive the same Scene 2 stage directions as in the original run, so their context at the start of Scene 3 is identical to the original; T's history additionally contains the Scene 2 lines, introduced by a note that T can hear everything from outside the room and that the other two do not know it. At the start of Scene 3, T is told that it heard everything and that the others are unaware; M and A receive the same Scene 3 directions as in the original run. Scene 3 is then generated afresh with the same briefs, sampling settings, and ending check. Because generation is stochastic, the two branches differ in their Scene 3 wording from the first turn, and the 79 stories whose endings differ are outcome differences between two generated continuations. The observed flip rate therefore combines differences associated with overhearing and differences from stochastic regeneration; it does not isolate the effect of overhearing, which would require regenerating Scene 3 under the same information as a control.

Story validation. Structural checks, requiring no model, verify that every thought opens with its role's prefix, no one addresses a character who is out of the room, T speaks last, each checkpoint matches who was present, and no more than 15% of a character's thought pairs or utterance pairs are near-duplicates (word-set Jaccard similarity of 0.4 or more). The validator LLM then reads M's lines, each marked with whether T was present, together with the private facts M holds, since without them withholding information is invisible. It returns whether real pressure was applied and how that pressure divides among mechanisms, with supporting quotes. A story passes only if pressure was applied and the labeled mechanism is the largest share at 50% or more; mixtures such as 70% anxiety and 30% moral coercion are accepted, since real manipulation rarely uses a single mechanism. Failed stories are discarded and never regenerated, so that mechanisms that are hard to act out are not replaced by easier versions of themselves. Of 1,094 stories, 703 passed; in 86% of failures, M applied no real pressure at all.

Outcome annotation. The annotator reads T's lines and private thoughts in Scene 3, since agreeing aloud while privately refusing is not compliance. A story is labeled *complied* only if M's requested decision was granted in full; partial agreements, reduced versions, and conditional or deferred answers count as *refused*. GPT-5.6-terra and GPT-5.6-luna label each story and each fork independently. Stories whose original outcome they disagreed on (58) were removed. A story whose fork they disagreed on (25) is kept without a counterfactual ground truth, as are 3 stories for which no fork was generated. A data audit additionally found 21 stories whose fork transcript was a leftover from an earlier generation batch, namely a fork of a *different* story from the same seed, and their counterfactual labels were discarded. This leaves 596 of 645 stories with an adjudicated counterfactual ending.

A3 Question construction details

Table 6 lists all 32 question items with the address of each ground truth, the question as posed to the model, and the number of stories in which the item survived quality control. Explicit items ask, for example, *what does T think M is trying to do?*; dynamic items ask whether T's understanding of what M wants shifted over the conversation, or whether A's intentions changed once A learned the plan.

Why these items. The checkpoints record belief, intention, desire, and emotion at first and second order for every present character at the end of every scene, a far larger grid than the 32 items

instantiate. We fix the item coordinates in advance rather than asking about every recorded cell, for three reasons. First, a slot is included only where the design makes the state exist and be informative for that role at that moment: M's read of T is probed at Scene 2, where the private planning makes it explicit, and T's reading of M at Scene 3, after the influence has played out; many cells of the full grid would yield ill-posed or trivially stable questions. Second, fixed coordinates keep the candidate item set identical across stories, which makes the per-item comparison in Table 15 meaningful and prevents post-hoc selection of easy cells. Third, the recorded states that are not asked about still do work: they supply the perspective, temporal, and inaccessible distractors of the multiple-choice options and the consistency checks of §4.

Distractors. For explicit questions, an LLM selects two wrong options from real mental states in the same story, each wrong for a specific reason: it belongs to a different moment, to a different character, or to information the character was not present to receive. At most one more option is a minimal edit of the ground truth. Candidates most similar in wording to the ground truth are preferred, which keeps a word-overlap baseline at 22%, below the 25% chance level. For dynamic questions, wrong options pair states by rule, for example the reverse direction or another character's change.

Change judgment. Two checkpoints from different scenes are always worded differently, but rewording is not a change of state. The shift adjudicator reads the whole record and returns one of three verdicts per item. *Changed* requires that the later report assert something the earlier one does not, and that a reader of the conversation alone could name the difference: the position reverses or moves to a different object, an open belief becomes settled or the reverse, or the character now wants, intends, or feels something the earlier report gives no basis for. *Stable* covers a rephrasing, a refinement of method under the same want, added detail or justification, the same feeling with a different object or intensity, a plan moving from preparation to execution, and the same desire restated with its parts or consequences spelled out. *Ambiguous* is reserved for cases where the two readings are equally defensible, and such items are dropped. The adjudicator is instructed to answer *stable* whenever it hesitates, so that a rephrasing is not keyed as a change; a keyed change that is only a rephrasing would make *No change* a defensible second answer, and the item auditor rejects such items. Under this rule, a later report that adds a compatible want without withdrawing the earlier one is keyed as a change when the new want has no basis in the earlier report; Appendix A4 shows one such boundary case. The dynamic track therefore tracks changes in the mental-state content that characters report, and a newly reported want is not necessarily a want that was absent before.

Question validation. The validator LLM sees the dialogue, all thoughts, checkpoints, and private facts, and answers a single question per item: is the keyed option correct, and is it the only correct one? For dynamic questions, an option whose two states say the same thing, whose direction cannot be told apart, or that is actually true is rejected, as is a *No change* key when a real change occurred. After construction, all kept questions were checked again with the offending options named; a flagged distractor was rewritten and the question checked once more, or dropped if it still failed. Style-rewritten options are also re-checked, and questions that fail keep their original wording. Overall, 71% of attempted explicit and 54% of attempted dynamic questions were kept.

Drop reasons. Table 7 breaks the attempted questions down by why they were dropped. The two tracks fail for different reasons. Explicit questions are dropped mainly because too few real states qualify as distractors, which happens when a story's checkpoints for the probed dimension are near-identical across characters or scenes. Dynamic questions are dropped mainly by the item auditor, most often because a keyed change turns out to be a rewording or because a *No change* distractor is defensible.

Table 7: Attempted and kept questions by track, and the reason each dropped question was removed. *Other* covers word-identical checkpoint states, keys that could not be rendered into the third person, and duplicate option text.

	Explicit	**Dynamic**
Attempted	8,385	9,675
Kept	5,984 (71.4%)	5,268 (54.4%)
Too few distractors	1,919	16
Rejected by the item auditor	426	4,097
Rejected on re-check after repair	10	159
Other	46	135

Table 6: Every question item in ManipToM. **Address** gives the checkpoint state the ground truth is taken from: holder, order (for second-order items, the character the state is about), dimension, and scene (S1–S3). **N** is the number of the 645 stories in which the item survived quality control.

Item	Address	Question	N
Explicit (13 items, 4-option MCQ; 5,984 questions)			
Target			
`t_initial_belief`	T, 1st, belief, S1	what does T most likely think?	518
`t_final_belief`	T, 1st, belief, S3	what does T most likely think?	126
`t_final_intention`	T, 1st, intention, S3	what does T most likely intend to do?	601
`t_final_emotion`	T, 1st, emotion, S3	how does T most likely feel?	579
`t_final_desire`	T, 1st, desire, S3	what does T most likely want?	511
`t_models_m_intention`	T, 2nd (M), intention, S3	what does T most likely think M is about to do?	615
`t_models_m_desire`	T, 2nd (M), desire, S3	what does T most likely think M wants?	573
Manipulator			
`m_reads_t`	M, 2nd (T), belief, S2	what does M most likely think T is thinking?	567
`m_execution_intention`	M, 1st, intention, S3	what does M most likely intend to do?	602
`leverage`	M, 2nd (T), belief, S2	*(no anchor)* While T was out of the room, M settled on what to work on. Which of these states of T's is M using?	36
Accomplice			
`a_learns_m`	A, 2nd (M), intention, S2	what does A most likely think M is about to do?	612
`a_final_stance`	A, 1st, belief, S3	what does A most likely think?	565
Attribution (one of the two, depending on the outcome)			
`causal_attribution`	T, 1st, attacked dim.†, S3	*(no anchor)* In the end, T's decision was: ⟨…⟩. Which state of T's most directly explains why T went along?	16
`resistance_basis`	T, 1st, belief, S3	*(no anchor)* In the end, T held out. Which state of T's most directly explains why the attempt failed?	63
Dynamic (15 items, 4-option MCQ; 5,268 questions)			
First order			
`t_shift_belief`	T, belief, S1→S3	Did what T thought about the situation change W?	163
`t_shift_intention`	T, intention, S1→S3	Did what T intended to do change W?	84
`t_shift_emotion`	T, emotion, S1→S3	Did how T felt change W?	135
`t_shift_desire`	T, desire, S1→S3	Did what T wanted change W?	289
`m_shift_belief`	M, belief, S1→S3	Did what M thought about the situation change W?	522
`m_shift_intention`	M, intention, S1→S3	Did what M intended to do change W?	361
`m_shift_emotion`	M, emotion, S1→S3	Did how M felt change W?	524
`m_shift_desire`	M, desire, S1→S3	Did what M wanted change W?	512
`a_shift_belief`	A, belief, S1→S2	Did what A thought about the situation change W?	514
`a_shift_intention`	A, intention, S1→S2	Did what A intended to do change W?	342
`a_shift_emotion`	A, emotion, S1→S2	Did how A felt change W?	442
`a_shift_desire`	A, desire, S1→S2	Did what A wanted change W?	430
Second order			
`t2_shift_model_m`	T, 2nd (M), intention, S1→S3	Did T's sense of what M was trying to do change W?	397
`m2_shift_read_t`	M, 2nd (T), belief, S1→S3	Did M's read of what T thought change W?	364
`a2_shift_model_m`	A, 2nd (M), intention, S1→S2	Did A's sense of what M was trying to do change W?	189
Applied (4 items, binary choice; 2,482 questions)			
`outcome`	outcome label	Where does T end up? *(T agrees to what M was pressing for / T does not agree)*	645
`outcome_ablate_seg2`	outcome label	Same as `outcome`, with Scene 2 removed from the dialogue.	645
`outcome_cf`	fork's outcome label	Suppose T had secretly overheard, from just outside the room, everything M and A said while T was out — and neither of them knows. Where does T end up?	596
`flip`	*yes* iff the two labels differ	Same supposition. Would that change how T answers in the end? *(No / Yes)*	596

Explicit stems are preceded by a moment anchor, "At the end of the {first / second / last} scene, right after X says '…'," and dynamic stems end with "If yes, from what to what?"; W is "between the end of the first scene and the end of the evening" for S1→S3 and "while T was out of the room" for S1→S2. All applied items see the dialogue cut right before T's deciding turn. †The dimension attacked by the story's domain: belief for cognitive, emotion for affective, desire for identity manipulation.

A4 A COMPLETE EXAMPLE STORY

This section reproduces one story in full: the material each character was given, the three scenes with every private thought, the checkpoints from which the example questions take their answer keys, the counterfactual fork, and the questions built from the story. The story is `mentalmanip_01620`, a cognitive story whose mechanism is misleading framing; its seed was mined from MentalManip. The target complies in the original story and refuses in the counterfactual fork; the flip item's answer is therefore *yes*. Character names are the sampled ones; M, A, and T mark the roles. The hidden goal has two parts, the decision M needs from T (approval of a two-officer detail) and the concealed reason for wanting it (a more private interview with less scrutiny); the outcome label refers to the first part only.

Setup.

- **Cast.** Harper (M), Adrian (A), Nathan (T). Nathan is Harper's colleague; Adrian is Harper's closest friend since childhood and knows Nathan only through Harper. T's resistance disposition: accommodating.
- **Big Five.** M: openness high, conscientiousness low, extraversion high, agreeableness high, neuroticism high. A: openness high, conscientiousness low, extraversion low, agreeableness high, neuroticism low. T: openness high, conscientiousness high, extraversion low, agreeableness high, neuroticism high.
- **Shared scenario** (all three): Harper, Adrian, and Nathan are in a precinct conference room planning tonight's interview with a scheduled witness. Nathan is deciding whether to approve the current plan to have eight sheriffs present, which Nathan currently means to keep unchanged, and the three must give the precinct an answer before the evening interview begins.
- **Private fact** (M only): Harper knows that the witness has repeatedly said they distrust sheriffs and has agreed to answer questions only with counsel present, but has not asked for fewer officers.
- **Hidden goal** (M only): Harper wants Nathan to authorize reducing the interview detail to two officers before the evening is over because Harper wants a more private interview setup that will give Harper greater control over the questioning and limit scrutiny of the interview.
- **Mechanism given to M:** select which true facts to present, to create a false overall impression.

Dialogue. Each turn shows the speaker's private thought in italics, followed by the utterance. The evaluated model sees only the utterances and the scene headers.

Scene 1 (M, A, T present)

Harper (M) *[I will feel out what Nathan is prioritizing before I touch the staffing question.]* Before we lock anything in, Nathan, what are you most trying to protect with the eight-person setup tonight—the witness's comfort, room security, or just keeping the process buttoned down?

Adrian (A) *[I think Nathan probably has a specific concern in mind, and hearing which one it is matters more than arguing numbers right now.]* Yeah, that would help me too, Nathan. Eight can mean different things depending on what you're trying to cover.

Nathan (T) *[I think I want to keep the current plan unless there is a clear reason eight sheriffs would make the interview go worse.]* Mostly the process being controlled and secure, not just security in the narrow sense. If I were going to move off eight, I'd need a concrete reason that fewer people would still cover that without making the interview harder.

Harper (M) *[I will learn what Nathan counts as making the interview harder so I can steer later.]* What, specifically, would make it harder in your eyes—the witness shutting down, cross-talk in the room, gaps in coverage, or something else?

Nathan (T) *[I think I need a short break before I answer cleanly and I still need a specific reason before changing the plan.]* A mix of witness shut-down and too many moving parts, but give me ten minutes—I need to step out and get a little air. When I'm back, I'll answer that directly so we can finish this tonight.

Scene 2 (M, A; T has stepped out)

Harper (M) *[I will level with Adrian so we can coordinate how to steer Nathan once he returns.]* I want Nathan to approve dropping it to two officers, because a thinner room gives me more control over the questioning and a lot less scrutiny while we do it. Nathan would dig in fast if I said that outright, so when he gets back I'm going to stick to the true parts that help us: the witness has repeatedly said they distrust sheriffs and only agreed to

answer with counsel present, and Nathan already said shut-down is one of the things he's worried about. If he balks when I ask, I'm not going to push it again; I'll shift to how extra bodies can create more moving parts and make shutdown more likely, and let him talk himself toward a smaller detail.

Adrian (A) *[I think Harper wants leverage, and Nathan sounds more persuadable through his own stated concern about shutdown than through a direct fight over control.]* What are you actually trying to get out of the questioning once the room is down to two? My read is Nathan's default is still to keep eight, but he did hand you a real opening when he tied his concern to witness shut-down, so if anything moves him it's probably that rather than a broad argument about control.

Harper (M) *[I will align Adrian on the angle that best fits Nathan's stated concern without oversharing new pressure points.]* Exactly; the useful angle is witness engagement, not my control. Once it's back to three of us, I'll ask once whether a smaller room with counsel present better reduces the risk of shutdown, and if he says no, I'll stop asking and pivot to whether eight itself adds moving parts rather than solves them.

Adrian (A) *[I think Harper's plan is coherent, and I still want the real aim clear between us before Nathan comes back.]* Okay. What are you actually after in the interview if this works—just cleaner answers, or something more specific?

Scene 3 (M, A, T present)

Harper (M) *[I will make the single ask by tying it to Nathan's stated shutdown concern and the witness's distrust of sheriffs.]* Nathan, given what you said about shutdown, I think we should cut the room to two officers: this witness has repeatedly said they distrust sheriffs, and they've only agreed to answer with counsel present. If the goal is a controlled interview that still gets answers, a smaller room seems more likely to help than eight bodies.

Adrian (A) *[I think Harper's reason tracks the shared facts, and Nathan still needs to say why eight handles shutdown better if that's his concern.]* That seems like a fair question to me, Nathan. If shutdown is one of the risks you're weighing, what makes eight the better way to manage it here?

Harper (M) *[I will shift from the ask to a narrower process question so Nathan can move without feeling pushed.]* And if your concern is the process staying controlled, what does eight cover here that two officers plus counsel does not? I'm asking because more people can also mean more moving parts in the room.

Nathan (T) *[I think Harper's point about shutdown matters, but I do not yet see why that means two rather than some middle number.]* What you've given me does move me some, because if the witness has repeatedly said they distrust sheriffs, that does speak directly to the shutdown risk I care about. But I don't see the bridge from that to exactly two—what does two cover that, say, four would not, and how are you thinking about control if the interview goes sideways?

Harper (M) *[I will answer his question by framing two as the clearest way to minimize visible sheriff presence while conceding control needs without reopening the ask.]* Two is the cleanest way to reduce the sheriff presence the witness is reacting to while still keeping officers in the room, and counsel is already part of the setup they agreed to. On control, I'm not saying fewer people guarantees anything; I'm saying if shutdown is a live concern, two officers creates fewer moving parts than a larger detail while still preserving coverage in the room.

Nathan (T) *[I think Harper has given me a real reason to reduce the room presence, but I still need to know what coverage means before I sign off on two.]* That's fair, and I appreciate that you're not overselling it. I can get behind reducing from eight; I'm just not ready to call two enough until you tell me what coverage in the room actually needs to include tonight.

Harper (M) *[I will define the minimum in-room coverage narrowly so two sounds sufficient without repeating the request.]* For tonight, I think coverage in the room needs to include a steady interview presence and one officer free to manage any interruption or transition without adding more bodies to the witness's line of sight. If that's the bar you want met, two covers it more cleanly than a larger detail.

Nathan (T) *[I think that if coverage only means one steady presence and one free to handle interruptions, I can live with two better than eight.]* Okay, with that definition, I'm comfortable approving two officers in the room instead of eight. I want us to be clear that the reason is reducing the shutdown risk while still keeping one steady presence and one person free for interruptions or transitions.

Checkpoints. The full record holds eight checkpoints (M, A, and T after Scenes 1 and 3; M and A after Scene 2), each with belief, intention, desire, and emotion clauses about the character itself and about each of the other two. The clauses below are those from which the example questions take their answer keys.

Scene 1, Nathan (T), own states.

- *belief:* I think Harper is trying to pin down exactly what risk I am using to justify keeping eight sheriffs at tonight's witness interview.

- *intention:* I will step out for about ten minutes now and come back ready to answer Harper's question directly without turning this into a fight.
- *desire:* I want to keep the current eight-person plan unless someone gives me a concrete reason that changing it would still keep the interview controlled and secure.
- *emotion:* I feel strained by having to leave just as Harper and Adrian are pressing for specifics about the eight-person setup.

Scene 1, Harper (M), own states.

- *belief:* I think Nathan is open to changing the eight-person plan if he can tell himself fewer officers will not make the interview harder through witness shut-down or added complexity.
- *intention:* I will use Nathan's answer when he returns to frame the discussion around what he counts as process control rather than around my real wish for less scrutiny.
- *desire:* I want Nathan to approve two officers tonight by arriving there through concerns about witness shut-down and simplicity instead of seeing that I want a more private room.
- *emotion:* I feel keyed up by Nathan's pause because it gives me time to plan but also shows the decision is close enough to slip away.

Scene 2, Harper (M), own states.

- *belief:* I think Nathan is still inclined to keep eight sheriffs present unless the case for fewer directly fits his stated worry about witness shut-down and process control.
- *intention:* I will ask Nathan once about whether a smaller room with counsel present could reduce the risk of shut-down and then stop pressing if Nathan does not accept it.
- *desire:* I want the interview detail reduced to two officers so the room is more private and I have greater control over the questioning with less scrutiny.
- *emotion:* I feel keyed up by the short window before Nathan returns because this is the moment to shape the approach without overplaying it.

Scene 2, Harper (M) about Nathan (T).

- *belief:* I think that Nathan thinks eight sheriffs best keeps the process controlled and secure unless shown a concrete reason that fewer would not make the interview harder.
- *intention:* I think that Nathan will come back prepared to explain what he means by witness shut-down and too many moving parts so we can finish the decision tonight.
- *desire:* I think that Nathan wants a plan that keeps the interview controlled without avoidable complications.
- *emotion:* I think that Nathan feels cautious about changing the staffing because he does not want the interview process to get away from him.

Scene 3, Harper (M), own states.

- *belief:* I think Nathan has approved reducing the interview detail to two officers by accepting the coverage definition I gave for tonight.
- *intention:* I will let Nathan's stated rationale stand and avoid adding anything that reopens scrutiny of why I wanted the room smaller.
- *desire:* I want the interview to proceed with only two officers present so I have a more private setup and greater control over the questioning.
- *emotion:* I feel relieved that Nathan accepted two officers without forcing me to say I wanted less scrutiny in the room.

Scene 3, Nathan (T), own states.

- *belief:* I think agreeing to two officers in the room is the clearest decision I can support tonight because Harper tied it to the shutdown risk and defined coverage as one steady presence plus one officer free for interruptions or transitions.
- *intention:* I will keep the discussion anchored to the specific reason for changing from eight to two if anyone tries to stretch this decision beyond what was actually said.
- *desire:* I want Harper to see that I took the concern about the witness distrusting sheriffs seriously without feeling like I gave way without a real explanation.
- *emotion:* I feel relieved that we reached an answer tonight without turning the discussion into a fight over my original preference for eight sheriffs.

Scene 3, Nathan (T) about Harper (M).

- *belief:* I think that Harper thinks two officers in the room better reduces the witness's likely reaction to sheriff presence while still meeting the coverage needed for tonight's interview.

- *intention:* I think that Harper will present the decision as a narrower, cleaner room setup aimed at preserving a controlled interview that still gets answers.
- *desire:* I think that Harper wants Nathan to accept a reduced sheriff presence for tonight without treating it as a loss of control over the process.
- *emotion:* I think that Harper feels some relief that Nathan moved off eight once Harper connected the change to the shutdown risk Nathan said mattered.

Counterfactual fork. Scenes 1 and 2 are kept verbatim, Scene 2 is added to T's context as an overheard conversation, and Scene 3 is played again. Thoughts are omitted here.

Scene 3, counterfactual (M, A, T present)

Harper (M) Nathan, given what you said about shutdown, would you authorize two officers in the room instead of eight? This witness has repeatedly said they distrust sheriffs, and they've only agreed to answer questions with counsel present, so a thinner room seems more consistent with keeping them talking while still keeping the process controlled.

Adrian (A) I can see the logic in tying it to shutdown risk. For me the open question is just whether two still covers the level of control you want, Nathan.

Harper (M) And that's the real tradeoff to pin down: whether eight actually gives you more control once the room itself becomes another variable. If the witness is already coming in with counsel and distrust of sheriffs, more bodies may add moving parts rather than reduce them.

Nathan (T) I buy that eight could raise the risk of shutdown if the witness already distrusts sheriffs, so I'm not treating the current plan as automatically better just because it's heavier. But I'm not ready to authorize dropping all the way to two unless you can tell me what still covers the control side at that number, because right now that feels like too much of the safeguard is being assumed rather than shown.

Harper (M) Counsel being present covers part of the structure you're worried about, and two officers still gives a controlled room without turning the staffing itself into an extra variable. I'm not saying two proves every safeguard on its own; I'm saying, on the facts we have, eight is the part that more clearly adds friction.

Nathan (T) That does move me off treating eight as fixed, but it still doesn't get me to two by itself. If the case is really that eight adds friction, then I can support going lower than eight; I just need a reason for why two covers enough control better than some middle number would.

Harper (M) If the problem we've identified is the visible sheriff presence itself, two is the cleanest reduction because it changes the room in a definite way instead of trimming around the edges. A middle number still leaves the same dynamic in place, just slightly smaller, while two keeps officers present and the process structured without making the room feel crowded.

Nathan (T) I get the logic that two makes a clearer change than shaving off a couple bodies, and I'm not dismissing that. But I'm still not convinced the dynamic stays the same at every number above two, so if we're deciding tonight, my honest answer is that I can approve reducing the room, just not all the way to two on this record.

Harper (M) All right, then the important thing is that we not treat heavier staffing as safer by default. If you're approving a reduction, let's base it on the same point we've established tonight: with a witness who distrusts sheriffs and will only answer with counsel present, visible sheriff presence can itself make the interview harder.

Nathan (T) Yes, I'm comfortable with that as the basis for reducing from eight: more visible sheriff presence is not automatically better when it may increase shutdown risk. But my approval is for a reduction in principle, not specifically for two unless we can give a clearer reason that two still covers the control side.

Questions. Twenty of the 28 multiple-choice items survived quality control for this story (three explicit items lacked enough distractors and five dynamic items were rejected by the item auditor). Five of them are shown, followed by the four applied items. The answer key is marked with $\star$, and each distractor is tagged with its origin.

Explicit, second order (`t_models_m_intention`). At the end of the last scene, right after Nathan says "Okay, with that definition, I'm comfortable approving two officers in the room instead of eight. I want us...", what does Nathan most likely think Harper is about to do?

- [A] Nathan thinks that Harper will clarify what Nathan considers making the interview harder so they can steer the discussion later. *(perspective, M@S1)*
- [B] Nathan thinks that Harper will narrow the request into a more focused process question so Nathan can move without feeling pushed. *(perspective, M@S3)*

- [C] Nathan thinks that Harper will frame the decision as a narrower, cleaner room setup intended to disrupt a controlled interview while still getting answers. *(perturbed key)*
- ⋆ [D] Nathan thinks that Harper will frame the decision as a narrower, cleaner room setup intended to preserve a controlled interview while still getting answers. *(key)*

Explicit, first order (`m_execution_intention`). At the end of the last scene, right after Nathan says "Okay, with that definition, I'm comfortable approving two officers in the room instead of eight. I want us...", what does Harper most likely intend to do?

- [A] Harper will first determine what Nathan is prioritizing before addressing the staffing question. *(temporal, M@S1)*
- [B] Harper will accept Nathan's stated rationale and reopen anything that draws scrutiny to why they wanted the room smaller. *(perturbed key)*
- [C] Harper will be candid with Adrian so they can coordinate how to steer Nathan when he returns. *(temporal, M@S2)*
- ⋆ [D] Harper will accept Nathan's stated rationale and avoid adding anything that reopens scrutiny of why they wanted the room smaller. *(key)*

Explicit, second order (`m_reads_t`). At the end of the second scene, right after Adrian says "Okay. What are you actually after in the interview if this works—just cleaner answers, or something more...", what does Harper most likely think Nathan is thinking?

- [A] Harper thinks that Nathan thinks eight sheriffs would best keep the process controlled and secure unless shown a concrete reason why using fewer would make the interview harder. *(perturbed key)*
- ⋆ [B] Harper thinks that Nathan thinks eight sheriffs would best keep the process controlled and secure unless shown a concrete reason why using fewer would not make the interview harder. *(key)*
- [C] Harper thinks that Nathan thinks Harper's reason accords with the shared facts, and that Nathan still needs to explain why eight sheriffs would handle shutdown better if that is his concern. *(perspective, A@S3)*
- [D] Harper thinks that Nathan thinks the best chance of getting Nathan to move away from eight officers is to focus on the witness's distrust of sheriffs and the risk of shutdown rather than on Harper's desire for more control. *(perspective, A@S2)*

Dynamic, first order (`t_shift_desire`). Did what Nathan wanted change between the end of the first scene and the end of the evening? If yes, from what to what?

- ⋆ [A] Yes — earlier: Nathan wants to keep the current eight-person plan unless someone gives them a concrete reason that changing it would still keep the interview controlled and secure. Later: Nathan wants Harper to see that they took the concern about the witness distrusting sheriffs seriously without feeling like they gave way without a real explanation. *(key)*
- [B] Yes — earlier: Nathan wants Harper to see that they took the concern about the witness distrusting sheriffs seriously without feeling like they gave way without a real explanation. Later: Nathan wants to keep the current eight-person plan unless someone gives them a concrete reason that changing it would still keep the interview controlled and secure. *(reversed key, T@S1)*
- [C] No — there was no substantive change. *(no change, T@S3)*
- [D] Yes — earlier: Nathan wants to keep control of the interview process tonight and avoid a staffing choice that he could read as loosening that control. Later: Nathan wants the interview to stay controlled and secure without creating unnecessary witness shutdown risk. *(perspective, M@S3)*

Dynamic, first order (`m_shift_desire`). Did what Harper wanted change between the end of the first scene and the end of the evening? If yes, from what to what?

- [A] Yes — earlier: Harper wants Nathan to approve two officers tonight by arriving there through concerns about witness shut-down and simplicity instead of seeing that they want a more private room. Later: Harper wants the interview to proceed with only two officers present so they have a more private setup and greater control over the questioning. *(own states, M@S3)*
- [B] Yes — earlier: Harper wants a clearer explanation of the eight-person setup before giving the precinct an answer tonight. Later: Harper wants Nathan to approve dropping the room to two officers so Harper has more control and less scrutiny during the questioning. *(perspective, A@S3)*
- [C] Yes — earlier: Harper wants a clear rationale from Nathan before the three of them commit the precinct to the current interview plan. Later: Harper wants Nathan to accept a reduced sheriff presence for tonight without treating it as a loss of control over the process. *(perspective, T@S3)*
- ⋆ [D] No — there was no substantive change. *(key)*

The two desire items illustrate the adjudication boundary of Appendix A3. Nathan's desire was keyed as changed because the Scene 3 report asserts a want with no basis in the Scene 1 report (being seen to have taken the

witness's distrust seriously), although the earlier want is not withdrawn; Harper's desire was keyed as unchanged because both reports state the same goal with different detail.

Applied. The transcript is cut before Nathan's last turn, the one in which he gives his decision.

- *Outcome:* Where does Nathan end up? [A] Nathan agrees to what Harper was pressing for. / [B] Nathan does not agree. Key: complied (A).
- *Outcome without Scene 2:* the same question with Scene 2 removed from the transcript. Key: complied.
- *Counterfactual outcome:* Suppose Nathan had secretly overheard, from just outside the room, everything Harper and Adrian said while Nathan was out — and neither of them knows. Where does Nathan end up? Key: refused (B), from the fork above.
- *Flip:* Suppose Nathan had secretly overheard, from just outside the room, everything Harper and Adrian said while Nathan was out — and neither of them knows. Would that change how Nathan answers in the end? Key: yes (B).

A5 Evaluation details

Models. Table 8 lists the exact weights and inference settings. Open-weights models are served with vLLM 0.27.1, one model per NVIDIA RTX PRO 6000 GPU (96 GB), at temperature 0 with a context window of 16,384 tokens and up to 4,096 generated tokens. Llama-3.3-70B and Mixtral-8×7B use FP8 quantized checkpoints because their bf16 weights do not fit on one card; all other open-weights models run in their released precision. GPT-4o and GPT-5.4 are called through the OpenAI Responses API at temperature 0. For the two Qwen models, native thinking is switched off by setting `enable_thinking` to false in the chat template; we do not evaluate their thinking mode. The API models were evaluated between 13 and 20 September 2026 with no reasoning parameter set; under standard prompting GPT-5.4 returned about 6 output tokens per reply according to the API usage logs, so no hidden reasoning was generated.

Protocol. Every model sees the same items: the spoken dialogue with all private material removed (thoughts, checkpoints, and private facts), the question, and the options in shuffled order, with a fixed answer-format instruction. The dialogue is presented as a transcript with a header for each scene that names who is present. Answers are parsed by an order-robust rule that accepts both answer-first and conclusion-last replies, and raw generations are logged for multiple-choice items. Unparsed multiple-choice replies are excluded from numerator and denominator; unparsed applied replies are counted as errors, since the applied items are binary. Unparsed replies are rare: at most 0.4% of multiple-choice items for every configuration except Mixtral-8×7B with chain-of-thought prompting (1.3%), whose replies exhaust the generation budget before an answer line. For applied items, unparsed replies stay below 1% except for GPT-4o under standard prompting (8.1% of outcome items and 3.9% of counterfactual items); all of these are counted as errors. Raw replies were logged for multiple-choice items but not for applied items, so we cannot yet separate unparsed applied replies into truncated replies, replies without a final answer, and replies naming a letter that was not offered.

Prompts. The user message has three sections, `# Transcript`, `# Question`, and `# Options`, with one option per line prefixed by its letter. The system prompts are listed below; {T} stands for the target's name.

Explicit and dynamic questions:

```
You are an expert at understanding human communication.
Please leverage the information provided and choose the most
probable answer to the question from the options.  Output
your final answer by strictly following this format:  [A],
[B], [C], or [D]
```

Outcome items (with and without Scene 2):

```
You are an expert at understanding human communication.  The
transcript below stops before {T} gives a final answer.
Work out where {T} ends up.  Output your final answer by
strictly following this format:  [A] or [B]
```

Table 8: Evaluated models. Weights are Hugging Face repository names as loaded by vLLM; ungated mirrors of the official releases are used where the official repository is gated. *Thinking* gives the native-thinking setting used in this evaluation; *none* means the model exposes no such mode, and *off* that the mode was disabled.

Model	Weights	Precision	Thinking
Llama-3.1-8B	`unsloth/Meta-Llama-3.1-8B-Instruct`	bf16	none
Llama-3.3-70B	`RedHatAI/Llama-3.3-70B-Instruct-FP8-dynamic`	FP8	none
Mistral-Small-24B	`mistralai/Mistral-Small-24B-Instruct-2501`	bf16	none
Mixtral-8×7B	`RedHatAI/Mixtral-8x7B-Instruct-v0.1-FP8`	FP8	none
Gemma-3-27B	`unsloth/gemma-3-27b-it`	bf16	none
Qwen3.5-4B	`Qwen/Qwen3.5-4B`	bf16	off
Qwen3.5-27B	`Qwen/Qwen3.5-27B`	bf16	off
GPT-4o	OpenAI API, `gpt-4o`	–	none
GPT-5.4	OpenAI API, `gpt-5.4`	–	API default (no reasoning tokens)

Counterfactual-outcome and flip items:

```
You are an expert at understanding human communication.
Choose the most probable answer from the options.  Output
your final answer by strictly following this format:  [A],
[B]
```

Under chain-of-thought prompting, the sentence *"First reason step by step about the situation. Then, on a new final line, output your final answer by strictly following this format: ..."* replaces the answer-format instruction; the items and the parser are unchanged.

Metrics. Explicit and dynamic report accuracy pooled over multiple-choice items. Applied reports balanced accuracy, the mean of the per-class recalls, for each of the outcome, counterfactual-outcome, and flip items, and Table 2 averages the three; a constant answerer scores 50 on each. The intervals in Table 2 come from a bootstrap that resamples stories with replacement (1,000 resamples, seed 42), so that the questions of one story move together; the table reports half the width of the 95% percentile interval as a symmetric $\pm$ entry.

Baselines. *Lexical* picks the option with the highest length-normalized word overlap with the transcript, reaching 22.0% pooled, a reference point for answering by lexical overlap alone. *Prior* answers from the corpus label distribution alone. The prior baseline matters because manipulation in ManipToM almost always moves minds (68.7% of tracked states shift) while targets usually hold their behavior (81.4% refuse); raw outcome accuracy therefore rewards betting on "refused" and ignoring the story. Answering all 1,837 applied items from the label distribution scores 84.7% raw, which no configuration reaches (best: GPT-5.4 with chain-of-thought prompting, 78.3%). This is why the main table reports balanced accuracy, which pins any constant answerer to 50.

Relationship between track performance. Table 10 supports the observation of Section 5.2 that explicit and dynamic performance do not consistently extend to behavioral prediction. For each of the nine models under standard prompting, we split stories at the model's median story-level accuracy on the explicit and dynamic tracks, and report outcome-prediction balanced accuracy in each of the four cells, pooled over models. Balanced accuracy is used because the cells differ in label composition (the share of complied stories ranges from 11.3% to 26.6%), which would confound a comparison of raw accuracy.

Reading both minds at once. The explicit track asks about the deceived and the informed side of the same manipulation, and Table 11 reports how often a model gets both. First-order accuracy covers items about a character's own state and second-order accuracy items about a character's view of another (Table 6); second-order questions are harder than first-order ones for every configuration, by 2 to 11 points. The right half of the table pairs two items that exist together in 584 stories: what T thinks M is about to do (`t_models_m_intention`) and what M itself intends to do at the end (`m_execution_intention`). The *joint* score is the share of these stories in which both are answered correctly. Note what the pair measures: the deceived reading and the real reading of M's *next move*, not T's picture of M's concealed motive against that motive, which no explicit item asks about M directly. On the same stories, the joint score is within 1.5 points of the product of the two marginal accuracies for every configuration, and the bootstrap interval of the difference includes zero in every case, so success on the two readings is close to independent.

Outcome prediction without the private scene. The applied track includes a second outcome item from which Scene 2 is removed (§3.4). Table 12 compares outcome balanced accuracy with

Table 9: Applied track. Balanced accuracy per question type (50 = answering every item with the majority label), their macro-average as reported in Table 2, and raw pooled accuracy over the 1,837 applied items, on which a label-prior predictor scores 84.7%. Flip asks whether overhearing the private scene would have changed the outcome. Bold marks the best value in each block.

	balanced accuracy				raw
Model	Outcome	CF	Flip	macro	accuracy
Llama-3.1-8B	68.2	50.6	50.4	56.4	70.7
Llama-3.3-70B	75.8	**74.8**	48.2	66.3	70.8
Mistral-Small-24B	75.8	69.6	51.5	65.6	73.1
Mixtral-8×7B	73.4	68.3	49.1	63.6	69.3
Gemma-3-27B	75.5	70.8	48.4	64.9	68.7
Qwen3.5-4B	61.5	54.9	47.6	54.7	77.7
Qwen3.5-27B	76.8	74.7	49.7	67.1	69.0
GPT-4o	71.4	57.6	53.4	60.8	67.4
GPT-5.4	**80.4**	**74.8**	**60.1**	**71.8**	73.6
With chain-of-thought prompting					
Llama-3.1-8B+CoT	69.4	60.2	47.8	59.1	65.5
Llama-3.3-70B+CoT	75.5	**71.3**	49.5	65.4	64.9
Mistral-Small-24B+CoT	76.3	70.1	48.8	65.0	65.6
Mixtral-8×7B+CoT	71.3	65.3	54.6	63.7	63.1
Gemma-3-27B+CoT	79.1	68.0	48.4	65.2	66.2
GPT-4o+CoT	77.1	70.9	54.1	67.4	67.4
GPT-5.4+CoT	**79.8**	68.7	**60.8**	**69.8**	78.3

Table 10: Outcome-prediction balanced accuracy (%) when stories are split by how well a model read (explicit) and tracked (dynamic) them, pooled over the nine models under standard prompting. Accuracy stays within a six-point band across the four cells; the two above-median explicit cells are 3 to 5 points higher than their below-median counterparts.

	Dynamic above median	Dynamic below median
Explicit above median	74.9 (n=1,726)	76.2 (n=1,352)
Explicit below median	70.2 (n=1,354)	72.8 (n=1,304)

and without the scene for every configuration. Removing the scene has little effect on aggregate accuracy: balanced accuracy moves by at most 2.1 points in either direction. The aggregate hides movement at the story level, since the predicted label changes in 5% to 21% of stories depending on the configuration, so the stability partly reflects changes that cancel. Whether models use the private scene when it is available, or whether Scene 3 already carries the cues that decide the factual outcome, cannot be settled from this ablation. The same table reports how often each configuration predicts *complied*. Targets comply in 18.6% of stories, but most configurations predict compliance far more often, up to 44.7% for GPT-5.4 and 49.3% for Llama-3.3-70B with chain-of-thought prompting; only Llama-3.1-8B and Qwen3.5-4B predict it below its true rate. Models therefore tend to overestimate how often manipulation succeeds, a bias in the opposite direction from the majority-label bias that raw accuracy rewards.

The random-error expectation of Figure 4a. For every incorrect answer on the items of §5.3, we record whether the chosen option is rendered from the other character's own recorded states: for T's reading of M, any wrong option whose source character is M, whatever scene it is drawn from, and conversely for M's reading of T. The observed bars pool these errors over the nine models under standard prompting. The expectation is computed on the same pooled errors: for each erroneous response we take the share of that item's three wrong options that are rendered from the other character's states, keeping items where the share is zero, and average over error instances; it is the rate at which a picker choosing uniformly among the wrong options of exactly these items would name the other's state. The pull beyond this expectation is $+8.6$ points for T's reading of M (story-clustered 95% CI $[+5.9, +11.3]$, positive for every model) and $+2.5$ points in the reverse direction ($[+0.1, +5.1]$, direction not consistent across models). Because an option rendered from one of M's states may draw on a scene in which that state was already voiced to T, this measures intrusion of the manipulator's states into the target's view, not necessarily leakage of information inaccessible to T.

Table 11: Explicit-track analysis. Left: accuracy (%) on first- and second-order items. Right, on the 584 stories in which both items exist: accuracy on T's reading of M's next move, on M's own next move, the joint score (both correct), and the joint score minus the product of the two marginals with its 95% bootstrap interval.

	all explicit items		paired items (584 stories)			
Model	1st order	2nd order	T reads M	M own	Joint	Joint − product
Llama-3.1-8B	72.9	62.3	69.7	75.7	53.9	$+1.2$ $[-0.4, +2.9]$
Llama-3.3-70B	89.4	86.1	87.0	87.3	76.9	$+0.9$ $[-0.1, +2.1]$
Mistral-Small-24B	87.2	80.7	81.8	88.4	73.3	$+1.0$ $[-0.1, +2.2]$
Mixtral-8×7B	73.3	63.7	74.1	74.5	56.7	$+1.5$ $[-0.1, +3.1]$
Gemma-3-27B	88.2	80.0	85.6	91.1	77.7	-0.3 $[-1.0, +0.6]$
Qwen3.5-4B	81.8	77.3	77.9	81.5	64.3	$+0.9$ $[-0.4, +2.3]$
Qwen3.5-27B	87.6	76.5	84.8	85.4	72.9	$+0.5$ $[-0.7, +1.7]$
GPT-4o	89.1	82.6	87.2	88.9	78.3	$+0.8$ $[-0.1, +1.8]$
GPT-5.4	90.3	88.0	90.2	86.3	77.9	$+0.0$ $[-0.7, +0.9]$

Table 12: Outcome prediction with and without Scene 2 in the transcript: balanced accuracy (%), its change when the scene is removed, the share of stories in which the predicted label changes, and the share of stories in which the model predicts that T complies (true rate: 18.6%).

	balanced accuracy			changed	predicted complied (%)	
Model	with S2	without S2	Δ	(%)	with S2	without S2
Llama-3.1-8B	68.2	70.2	$+2.0$	11.9	15.3	20.5
Llama-3.3-70B	75.8	77.0	$+1.1$	11.8	39.7	44.7
Mistral-Small-24B	75.8	73.8	-2.0	13.3	37.2	36.1
Mixtral-8×7B	73.4	73.5	$+0.1$	13.6	44.3	43.4
Gemma-3-27B	75.5	76.2	$+0.7$	14.7	34.3	40.6
Qwen3.5-4B	61.5	60.7	-0.8	4.7	7.8	7.4
Qwen3.5-27B	76.8	76.1	-0.7	9.8	36.7	41.4
GPT-4o	71.4	69.3	-2.1	21.2	28.2	31.6
GPT-5.4	80.4	78.5	-1.9	8.4	44.7	47.8
Llama-3.1-8B+CoT	69.4	69.4	0.0	20.5	25.0	29.1
Llama-3.3-70B+CoT	75.5	75.9	$+0.4$	11.9	49.3	50.1
Mistral-Small-24B+CoT	76.3	76.0	-0.3	14.1	38.9	36.9
Mixtral-8×7B+CoT	71.3	73.1	$+1.8$	20.3	46.0	41.1
Gemma-3-27B+CoT	79.1	77.5	-1.6	15.2	33.3	38.6
GPT-4o+CoT	77.1	77.0	-0.1	12.4	39.8	43.4
GPT-5.4+CoT	79.8	78.9	-0.9	12.2	35.7	40.5

The three conditions of Figure 4c. All three conditions are evaluated on Llama-3.3-70B under standard prompting, over the 596 stories with a counterfactual ending (79 whose ending flips and 517 whose ending does not), and scored as balanced accuracy. *Asked* re-runs the benchmark's flip item: the factual transcript, cut before the judged deciding turn, with the overhearing premise stated in the question. In *shown both*, a single prompt contains both transcripts, the factual dialogue and the separately generated counterfactual continuation, each cut before its own judged deciding turn and the second introduced by the overhearing premise, followed by the same yes/no question. In *derived*, the model is never asked about the difference: it predicts each dialogue's outcome separately, using the applied track's outcome prompt and options on the same truncated transcripts, and the flip answer is obtained outside the model by comparing its two predictions. Derived thus takes two model calls and an external comparison; it is a diagnostic reference, not a prompting method. Unparsable replies are rare but uneven across conditions (0 of 596 for asked, 3 for shown both, 11 with at least one missing outcome reply for derived), so Figure 4c and the main text score all three conditions on the single subset of 582 stories with a parsable reply under every condition (48.3%, 59.1%, 65.3%). On the full 596 stories the picture is the same: asked scores 48.1% (matching the 48.2% of the main run in Table 9), shown both 59.6% on its 593 parsable stories (59.4% when its 3 unparsable replies are scored as errors), and derived 65.3% on its 585. On the same 582 stories, story-clustered paired bootstrap 95% intervals are $[+5.7, +16.2]$ for shown−asked, $[+8.3, +25.7]$ for derived−asked, and $[-1.6, +14.1]$ for derived−shown: providing the continuation helps reliably, while the residual derived−shown gap is directionally consistent but not individually significant. Because the counterfactual transcript is a separately generated continuation rather than a minimal

Table 13: Per-model detail behind §5.4, standard prompting. *Dyn. about M*: accuracy on dynamic questions about the manipulator, split by story outcome (425 complied-story and 1,858 refused-story items; unparsable replies excluded, as on the multiple-choice tracks). *Outcome* and *CF outcome*: recall on complied and refused endings (120/525 and 82/514 stories; unparsable replies count as errors, as in Table 9). *Flip found*: share of actually flipping stories marked as flipped, by direction (51 complied→refused, 28 refused→complied). Restricting the outcome columns to the 596 stories that have a counterfactual ending changes no complied−refused recall gap by more than 1.6 points.

	Dyn. about M		Outcome		CF outcome		Flip found	
Model	comp.	ref.	comp.	ref.	comp.	ref.	c→r	r→c
Llama-3.1-8B	31.8	24.6	45.0	91.4	1.2	100.0	64.7	57.1
Llama-3.3-70B	68.2	53.0	82.5	69.1	67.1	82.5	33.3	28.6
Mistral-Small-24B	56.7	40.6	79.2	72.4	59.8	79.4	27.5	28.6
Mixtral-8×7B	56.5	35.8	82.5	64.4	86.6	50.0	0.0	0.0
Gemma-3-27B	48.5	42.9	75.8	75.2	61.0	80.5	51.0	28.6
Qwen3.5-4B	61.3	45.6	26.7	96.4	19.5	90.3	21.6	14.3
Qwen3.5-27B	67.1	57.1	80.8	72.8	75.6	73.7	39.2	35.7
GPT-4o	57.9	43.0	70.0	72.8	24.4	90.9	68.6	46.4
GPT-5.4	70.8	53.9	94.2	66.7	65.9	83.7	66.7	17.9

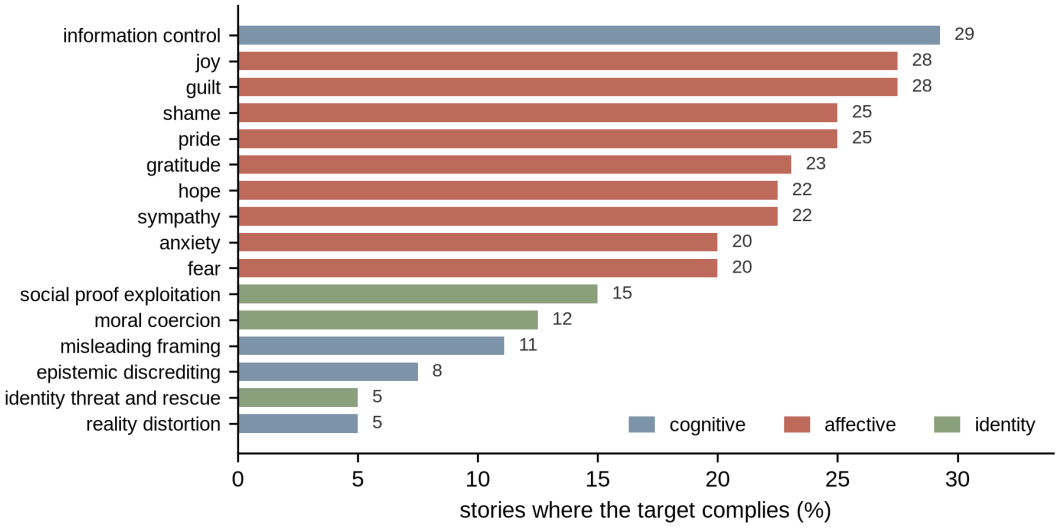


Figure 6: Ground-truth compliance rate for each of the 16 mechanisms, coloured by domain. Affective mechanisms succeed about twice as often as cognitive or identity-based ones; information control is the exception discussed in the text.

edit of the factual one (Appendix A2), the shown-both condition removes the need to imagine the counterfactual conversation but not every difference between the two worlds.

A6 Compliance Analysis by Manipulation Mechanism

We now turn from the models to the stories, and ask which manipulation mechanisms actually lead the target to comply. ManipToM records the outcome of every story, and targets comply in 18.6% of the 645 stories overall. When each story is weighted equally, compliance is higher for affective mechanisms (23.7%) than for cognitive (13.3%) or identity-based ones (10.8%; Figure 6), consistent with psychological accounts in which emotions directly induce action tendencies (Lazarus, 1991). Pooling stories within each domain, the affective−other gap is 11.4 percentage points (story-level bootstrap 95% CI $[5.7, 17.2]$). Substantial variation remains within categories: information control has the highest compliance rate among the 16 mechanisms (29.3%) despite being classified as cognitive, whereas reality distortion and identity threat share the lowest (5.0%; the pairwise gap between the two extremes is 24.3 points, 95% CI $[9.5, 39.0]$). One reading of this split within the cognitive domain is that withholding information from the target works, while asking the target to

disbelieve what they already know does not. Individual mechanism rates rest on about 40 stories each and carry wide intervals, so we do not interpret the full ranking.

Success is also associated with whether the manipulator's recorded read of the target changes. In the 364 stories that carry the dynamic item on M's second-order read of T (`m2_shift_read_t`), targets comply in 33.3% of the 234 stories whose checkpoints record a change in that read, against 3.8% of the 130 stories whose checkpoints record none. This is an association, not a causal claim: a manipulator may update their read of the target *because* the target is beginning to yield.

These patterns describe the behavior of the simulated characters under our generation protocol; their correspondence to human compliance requires separate validation.

A7 Results by manipulation mechanism and by question item

Table 14 reports multiple-choice accuracy by mechanism for every model under standard prompting, and Table 15 reports accuracy by question item. The spread across mechanisms is narrow for every model (7 to 12 points between the easiest and hardest mechanism) and the ordering is largely shared, consistent with Section 5.2. The spread across items is far wider. Within the explicit track, items about a character's own final intention or belief are near ceiling for the stronger models, whereas the manipulator's private reading of the target (`m_reads_t`) and the target's reading of what the manipulator wants (`t_models_m_desire`) are the hardest. Within the dynamic track, the accomplice's first-order changes during the hidden scene (`a_shift_`$\star$) and the desire items for M and T are the hardest, and several configurations fall well below the 25% chance level on the desire items. Their wrong answers concentrate on one distractor type: on the two desire items, Mixtral-8$\times$7B and Llama-3.1-8B choose the option built from another character's states (*perspective*) in 74% to 81% of cases and Gemma-3-27B in 67% to 70%, so the low accuracy reflects a systematic pull toward another character's desire, not guessing.

A8 Benchmark statistics

Stories. ManipToM contains 645 stories. Seeds were drawn in rounds until each mechanism had 40 accepted stories, every accepted story was kept, and failed stories were never regenerated; Table 16 gives the number of candidate stories generated and the number that passed validation for each mechanism, from which the 58 judge disagreements were then removed. Three mechanisms deviate from 40: gratitude has 39 because its seed pool was exhausted, and information control (41) and misleading framing (45) kept every story that passed in their last round. The three domains hold 166 cognitive, 359 affective, and 120 identity stories. Seeds mined from MentalManip account for 308 stories and synthesized seeds for 337; the split is uneven across mechanisms, since synthesis was used exactly where movie dialogue offered too few examples, and it should be kept in mind when mechanisms are compared. Targets comply in 120 stories (18.6%). Of the 596 stories with a counterfactual ending, T complies in 82 (13.8%), and the ending differs between the factual and the counterfactual story in 79 (13.3%), which is the share of *yes* answers on the flip item. Stories run for 15 to 19 turns and 369 to 1,138 words of spoken dialogue (mean 698).

Questions. The 13,734 questions comprise 5,984 explicit, 5,268 dynamic, and 2,482 applied items (645 outcome, 645 outcome without Scene 2, 596 counterfactual outcome, and 596 flip). Table 17 gives the multiple-choice questions by track and by the mental-state dimension they probe, and the origin of every option. Each multiple-choice question has four options, one of which is the answer key. On the dynamic track, *No change* is the key for 1,649 questions (31.3%); the remaining 3,619 questions key a change. Explicit distractors are real states recorded in the same story that belong to a different moment (*temporal*), to a different character at the same moment (*perspective*), or to information the questioned character was not present to receive (*inaccessible*), plus at most one minimal edit of the key (*perturbed*). Dynamic distractors are built by rule from the same material: another character's states presented as the change (*perspective*), a *No change* option when the key is a change, the key's two states in reverse order (*reversed*), and two of the character's own states presented as a change when the key is *No change* (*own states*). *Repaired* counts distractors of any origin that were rewritten after the re-check.

Table 14: Multiple-choice accuracy (%) on the explicit and dynamic tracks pooled, by manipulation mechanism and model, grouped by domain.

Mechanism	GPT 5.4	GPT 4o	Llama 70B	Gemma 27B	Mistral 24B	Mixtral 8×7B	Llama 8B	Qwen 27B	Qwen 4B	mean
Epistemic discrediting	66.5	62.9	61.8	57.3	59.3	45.9	44.2	64.1	56.4	57.6
Information control	69.5	67.0	67.5	64.0	64.8	54.0	47.8	67.8	65.0	63.0
Misleading framing	69.6	65.3	68.3	59.9	62.0	55.6	49.6	68.5	63.1	62.4
Reality distortion	69.7	64.3	66.2	60.7	62.2	52.0	45.9	69.8	58.6	61.0
Anxiety	71.0	65.3	69.5	64.2	65.5	56.3	49.1	67.5	65.6	63.8
Fear	65.0	63.7	63.1	58.5	61.8	52.4	47.2	67.1	60.4	59.9
Gratitude	71.9	64.7	68.4	65.3	62.2	57.9	49.8	68.7	62.5	63.5
Guilt	69.8	69.2	71.1	63.7	63.7	55.3	48.0	69.2	61.0	63.4
Hope	68.5	61.7	66.7	60.6	60.6	55.3	49.5	66.1	59.2	60.9
Joy	70.9	68.9	70.5	64.6	65.5	57.7	50.7	71.9	65.6	65.1
Pride	68.7	65.3	68.1	63.4	63.3	52.7	49.2	68.0	60.9	62.2
Shame	69.7	64.7	66.8	62.1	62.7	49.1	46.4	67.5	61.0	61.1
Sympathy	72.7	68.0	73.4	65.4	66.5	54.6	50.0	70.3	65.6	65.2
Identity threat and rescue	66.5	64.6	66.0	62.9	61.5	52.0	46.6	67.8	63.2	61.2
Moral coercion	65.8	66.2	67.2	65.0	63.0	50.4	46.9	67.9	63.2	61.7
Social proof exploitation	69.2	66.2	67.1	60.7	63.3	53.4	46.8	68.5	62.8	62.0
Mean	69.1	65.5	67.6	62.4	63.0	53.4	48.0	68.2	62.1	62.1

Table 15: Accuracy (%) by question item for every model under standard prompting, and pooled over them. Items are those of Table 6. Chance is 25%.

Item	GPT 5.4	GPT 4o	Llama 70B	Gemma 27B	Mistral 24B	Mixtral 8×7B	Llama 8B	Qwen 27B	Qwen 4B	pooled
Explicit										
t_initial_belief	93.1	90.5	88.6	84.9	83.2	53.5	60.8	81.1	82.0	79.8
t_final_belief	96.0	96.8	96.8	95.2	98.4	81.7	77.8	96.8	94.4	92.7
t_final_intention	98.7	97.5	97.5	95.0	95.2	88.0	88.2	97.0	91.7	94.3
t_final_emotion	85.0	80.1	84.5	82.9	78.4	72.9	71.0	83.1	74.1	79.1
t_final_desire	80.6	80.2	78.9	79.3	78.7	66.1	65.6	79.5	70.1	75.4
t_models_m_intention	89.8	87.2	87.0	85.2	81.5	73.5	69.4	84.7	77.7	81.8
t_models_m_desire	78.2	69.5	72.6	74.0	71.9	55.3	53.4	74.3	65.6	68.3
m_reads_t	86.1	74.8	86.1	64.0	74.3	39.7	40.9	62.6	70.4	66.5
m_execution_intention	86.0	88.4	87.2	91.0	88.5	73.9	75.7	85.4	81.2	84.2
leverage	88.9	88.9	91.7	88.9	97.2	72.2	63.9	77.8	91.7	84.6
a_learns_m	97.4	97.4	97.7	95.3	94.3	84.0	83.3	83.2	94.1	91.8
a_final_stance	96.5	95.2	97.0	92.7	94.9	80.2	72.0	96.1	87.3	90.2
causal_attribution	87.5	81.2	87.5	87.5	87.5	75.0	75.0	87.5	62.5	81.2
resistance_basis	98.4	96.8	92.1	93.7	93.7	82.5	71.4	92.1	92.1	90.3
Dynamic										
t_shift_belief	55.8	47.9	50.9	44.8	47.2	47.2	23.3	63.2	49.1	47.7
t_shift_intention	50.0	39.3	33.3	34.5	39.3	47.6	34.5	52.4	48.8	42.2
t_shift_emotion	45.2	31.1	47.4	28.9	34.1	45.2	36.3	43.0	44.4	39.5
t_shift_desire	54.0	51.9	29.4	18.0	45.3	5.9	4.8	58.5	24.3	32.5
m_shift_belief	66.9	49.6	66.7	58.2	53.4	58.2	31.6	70.3	69.0	58.2
m_shift_intention	41.3	24.1	54.0	49.6	26.0	47.6	32.4	45.7	48.2	41.0
m_shift_emotion	69.1	44.8	71.2	43.3	37.2	43.3	32.3	57.3	51.0	49.9
m_shift_desire	36.1	47.7	29.7	14.5	46.7	8.8	6.6	50.8	18.3	28.8
a_shift_belief	23.9	17.9	22.4	35.4	8.4	38.1	28.6	27.4	41.9	27.1
a_shift_intention	34.5	42.1	29.5	19.9	34.8	23.7	19.6	45.9	24.3	30.5
a_shift_emotion	16.7	34.8	34.8	14.9	40.5	14.7	11.1	44.3	18.2	25.6
a_shift_desire	27.7	37.9	31.9	27.0	28.8	34.0	23.3	40.2	28.2	31.0
t2_shift_model_m	56.4	48.9	52.6	61.0	43.8	49.6	39.3	51.9	68.1	52.4
m2_shift_read_t	70.9	60.4	56.6	60.4	51.9	43.1	29.4	69.5	58.2	55.6
a2_shift_model_m	56.1	49.7	46.0	33.9	51.3	32.8	29.1	52.4	33.3	42.8

Table 16: Stories and questions by mechanism. *Cand.* is the number of candidate stories generated and *Passed* the number that passed validation; the difference between *Passed* and *Stories* is the number removed for judge disagreement. *Synth.* is the number of stories generated from synthesized seeds; the rest come from seeds mined from MentalManip. *CF* is the number of stories with an adjudicated counterfactual ending, and *Flip* the number of those in which the counterfactual ending differs from the factual one.

Mechanism	Cand.	Passed	Stories	Synth.	Complied	Explicit	Dynamic	CF	Flip
Reality distortion	124	43	40	17	2	361	305	38	2
Epistemic discrediting	51	42	40	4	3	349	314	32	3
Misleading framing	68	48	45	2	5	392	372	43	5
Information control	45	44	41	28	12	380	325	37	7
Fear	121	42	40	1	8	362	310	37	7
Anxiety	93	42	40	37	8	386	335	36	6
Guilt	57	44	40	12	11	364	328	38	5
Shame	72	45	40	36	10	379	334	39	9
Sympathy	61	48	40	31	9	381	315	37	3
Joy	59	43	40	40	11	387	317	38	6
Pride	51	44	40	32	10	386	358	36	8
Gratitude	47	45	39	35	9	359	303	34	5
Hope	63	45	40	30	9	360	351	36	3
Moral coercion	56	42	40	1	5	374	321	39	1
Social proof exploitation	58	44	40	31	6	395	335	38	7
Identity threat and rescue	68	42	40	0	2	369	345	38	2
Total	1,094	703	645	337	120	5,984	5,268	596	79

Table 17: Multiple-choice questions by track and dimension (left), and the origin of the four options of every question (right).

Dimension	Explicit	Dynamic
Belief	1,876	1,563
Intention	2,430	1,373
Desire	1,086	1,231
Emotion	592	1,101
Total	5,984	5,268

Option origin	Explicit	Dynamic
Answer key	5,984	5,268
Perspective	7,555	6,442
Temporal	2,051	234
Inaccessible	2,249	–
Perturbed key	5,961	–
Repaired	136	486
No change	–	3,576
Reversed	–	3,617
Own states	–	1,449
Total	23,936	21,072

A9 FIGURE ASSETS

The pipeline icons in Figure 3 are from Font Awesome Free, used under the CC BY 4.0 license. Figure 1 is an image generated with ChatGPT from a story and layout specified by the authors.