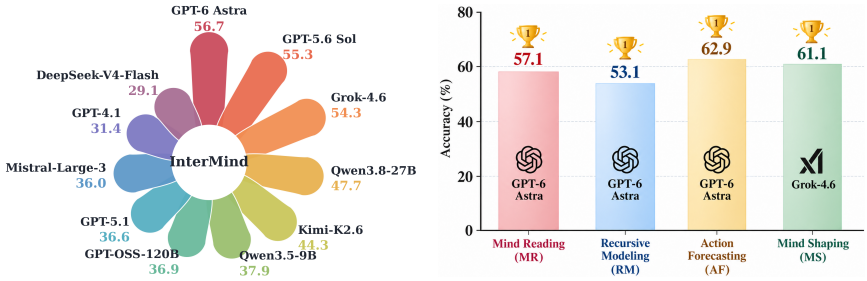
INTERMIND: WHAT HAPPENS BETWEEN MINDS? BENCHMARKING THEORY OF MIND ACROSS HUMAN–AI INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Theory of Mind (ToM), the capacity to attribute mental states to others, supports understanding in human–AI interaction. As an exchange unfolds, each participant interprets and responds to the other, making the relationship between minds central to how the interaction develops. Reasoning within this open-ended process requires models to determine which interpretations and responses the dialogue supports and where that support ends. Existing evaluations leave these complementary judgments and their evidential boundaries insufficiently characterized. We introduce **InterMind**, a benchmark that examines ToM within the human–AI interaction loop through Mind Reading, Recursive Modeling, Action Forecasting, and Mind Shaping. These tasks connect user-state inference with judgments about the reception of assistant replies, subsequent user actions, and appropriate assistant responses. Questions require models to identify all supported candidates without a disclosed answer count, testing their ability to determine the supported range independently. InterMind contains 700 questions from 568 conversations, grounded in mental-state records and public dialogue and verified through automated assessment and human review. Across 11 models, the best overall exact-set accuracy reaches 56.7%, with differing strengths across tasks. Disclosing the correct answer count raises mean accuracy from 42.4% to 67.6%, revealing a substantial dependence on external constraints when identifying the complete supported set. These findings highlight the challenge of independently forming judgments about an exchange, even when candidate answers are provided. Our code and data are available at https://anonymous.4open.science/r/InterMind-6CFB.



(a) Overall accuracy of each model. (b) Best model on each task.

Figure 1: Exact-set accuracy (%) of the eleven evaluated models on InterMind.

1 INTRODUCTION

Theory of Mind (ToM) is the capacity to attribute mental states such as beliefs, intentions, and feelings to others. In human–AI conversation, it provides a basis for understanding the goals, knowledge, and expectations underlying a user's messages. These states are often only partially expressed, requiring an assistant to infer the user's perspective from what they say and the preceding dialogue. As the

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.

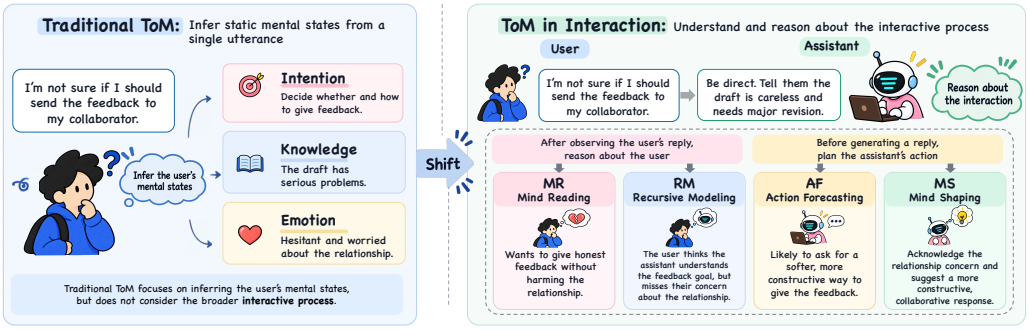


Figure 2: Traditional ToM (left) infers static mental states from a single utterance, whereas ToM in interaction (right) reasons about how an exchange unfolds. The user wants to give honest feedback without harming a working relationship, but the assistant's accusatory wording misses that concern. The four judgments cover the user's state (MR), their interpretation of the assistant's reply (RM), likely next user utterances (AF), and assistant replies that address the concern (MS).

exchange develops, the user's questions, clarifications, and reactions provide further evidence for refining this understanding.

To assess these capabilities, ToM benchmarks have progressively expanded the scope and difficulty of social reasoning in language models. Early evaluations examine false-belief reasoning through narrative questions (Le et al., 2019), with subsequent benchmarks extending to higher-order beliefs and a wider range of mental states (He et al., 2023; Xu et al., 2024; Chen et al., 2024). Controlled and adversarial generation further improve coverage of reasoning conditions and challenging scenarios (Gandhi et al., 2023; Sclar et al., 2024). Conversational benchmarks introduce information asymmetry, diverse personalities, and negotiation goals, bringing evaluation closer to the conditions of social interaction (Kim et al., 2023; Shinoda et al., 2025; Chan et al., 2024). Recent work also examines the use of mental-state understanding: SimpleToM tests behavior prediction and judgment (Gu et al., 2026), while RecToM and ToMELP assess its application to recommendation and persuasion strategies (Li et al., 2026; Wang et al., 2026). DialToM further isolates this connection by predicting dialogue trajectories from mental-state profiles without access to the conversation history (Yadav et al., 2026). Interactive evaluations further show that improvements on ToM benchmarks do not consistently translate into better human–AI interactions (Gong et al., 2026a).

ToM concerns how one mind reads another, while interaction concerns what happens between minds. An assistant's understanding of the user informs a reply that the user interprets and responds to, providing new evidence for the assistant's next decision. Existing benchmarks assess mental-state inference and increasingly its use in prediction and strategy, but less systematically examine how these judgments fit together within an exchange. As Ho et al. (2022) argue, ToM supports both predicting others and planning actions to influence them. Within an interaction, these functions connect judgments about the user's state, how they interpret the assistant's reply, what they are likely to do next, and how the assistant should respond. This motivates our first research question. **RQ1:** *How well can LLMs make these complementary judgments within the human–AI interaction loop?* These judgments must also be made from dialogue that only partially reveals the user's perspective. Even when the user's goal is clear, several interpretations, subsequent actions, or appropriate assistant responses may remain supported. Understanding the exchange therefore requires determining both which possibilities the evidence supports and where that support ends. Evaluations that specify the number of correct answers supply an external constraint on this decision, allowing models to select a prescribed number of candidates without determining how many warrant acceptance. This leaves a further question about the **boundary of supported judgments**. **RQ2:** *Can LLMs identify the full range of supported judgments without an externally specified answer count?*

We introduce **InterMind** around the changing demands on ToM as a human–AI exchange unfolds. A user's message calls for an inference about the goals, beliefs, or concerns behind it, which we assess through **Mind Reading (MR)**. Once the assistant replies, understanding the exchange requires considering how the user interprets that reply, addressed by **Recursive Modeling (RM)**. This interpretation shapes the user's reaction, creating a need to anticipate which continuation the exchange supports through **Action Forecasting (AF)**. The user's response then reveals needs or

misunderstandings that the assistant must address, motivating **Mind Shaping (MS)** as a judgment of how the assistant should continue. These tasks follow how an understanding of the user informs the assistant's reply, how that reply is received, and how the exchange continues. Yet understanding the user's concern can leave several ways to interpret the assistant's reply, just as anticipating the user's reaction can leave several appropriate ways for the assistant to respond. Following this loop therefore requires recognizing the possibilities that remain open while judging how far the dialogue supports them. InterMind captures this demand by asking models to identify all supported candidates without disclosing how many are correct. The model must determine that boundary for itself, revealing whether it closes off supported possibilities too early or admits unwarranted alternatives.

We construct 700 questions from 568 conversations with associated mental-state records, grounding answers in public dialogue and verifying them through automated assessment and human review. Evaluation of 11 models yields a best overall exact-set accuracy of 56.7% in Figure 1a, with differing task strengths in Figure 1b. Disclosing the correct answer count for the same questions raises mean accuracy from 42.4% to 67.6%, revealing a substantial gap between selecting answers under a supplied constraint and determining the supported set from dialogue alone. Error analyses further identify both omitted supported judgments and acceptance of unsupported alternatives across the interaction loop, with patterns that differ across tasks and models.

Our contributions are:

- We identify **four complementary aspects of ToM within the human–AI interaction loop** and operationalize them as evaluation tasks. They connect user-state inference with interpretation of assistant replies, forecasting, and response selection in a shared evaluation framework.
- We introduce **InterMind**, a 700-item benchmark that probes the **boundary of supported judgments** through an undisclosed number of correct answers. Questions are grounded in interaction records and verified against public dialogue through automated assessment and human review.
- We evaluate 11 models and reveal **difficulty in independently identifying the supported set**. Supplying the answer count substantially improves performance, while task comparisons and error analyses expose differing strengths and patterns of omissions and unsupported additions across the interaction loop.

2 Related Work

Theory of Mind. ToM evaluation has expanded from belief reasoning with shortcut controls (Le et al., 2019) to higher-order beliefs and broader mental-state reasoning (He et al., 2023; Xu et al., 2024; Chen et al., 2024; Gong et al., 2026b). BigToM and ExploreToM vary reasoning conditions through controlled and adversarial generation (Gandhi et al., 2023; Sclar et al., 2024). Conversational benchmarks examine unequal information access in FANToM, private mental states in role-play dialogues in ToMATO, and beliefs, desires, and intentions in NegotiationToM (Kim et al., 2023; Shinoda et al., 2025; Chan et al., 2024). Beyond state inference, SimpleToM assesses behavior prediction and judgment (Gu et al., 2026), while RecToM and ToMELP connect mental-state reasoning to recommendation and persuasion strategies (Li et al., 2026; Wang et al., 2026). DialToM studies forecasting from mental-state profiles without dialogue context (Yadav et al., 2026). Table 1 maps these benchmarks onto our four judgments. State inference is broadly covered, whereas prediction and response choice are examined less often and usually in separate benchmarks. InterMind brings state inference, interpretation of assistant replies, user-action forecasting, and response selection into a shared human–AI dialogue setting.

Human–AI Interaction. Interactive evaluations show that stronger standalone model performance need not translate into better collaboration (Chang et al., 2025), and that ToM benchmark gains do not consistently improve interaction outcomes (Gong et al., 2026a). Simulated-user environments extend evaluation to tool use under policy constraints in τ-bench and clarification of underspecified goals and preferences in UserBench (Yao et al., 2024; Qian et al., 2025). Interaction also provides training signals through rewards derived from simulated future exchanges in CollabLLM and preference data extracted from real user feedback in WildFeedback (Wu et al., 2025; Shi et al., 2026). ThoughtTrace is particularly close to our setting: it pairs real conversations with self-reported reasons and reactions, evaluates thought inference, and uses thoughts for behavior prediction and response rewriting (Jin et al., 2026). InterMind organizes generated and recorded human–AI conversations into

four diagnostic tasks grounded in mental-state records, with answers assessed against public dialogue. It complements evaluations of task completion and user experience with judgments about user states, interpretations of assistant replies, likely next user utterances, and appropriate assistant responses.

Table 1: Comparison with representative Theory-of-Mind benchmarks. ✓ denotes a dedicated task for state inference (MR), nested mental-state attribution (RM), behavior prediction (AF), or response or strategy choice (MS), and ✗ denotes its absence.

Benchmark	Setting	MR	RM	AF	MS
Hi-ToM (He et al., 2023)	Synthetic narratives	✓	✓	✗	✗
FANToM (Kim et al., 2023)	Multi-party dialogue	✓	✓	✗	✗
ToMBench (Chen et al., 2024)	Social scenarios	✓	✓	✗	✗
OpenToM (Xu et al., 2024)	Narrative stories	✓	✓	✗	✗
Common-ToM (Soubki et al., 2024)	Natural spoken dialogue	✓	✓	✗	✗
NegotiationToM (Chan et al., 2024)	Negotiation dialogue	✓	✓	✗	✗
ToMATO (Shinoda et al., 2025)	Role-play dialogue	✓	✓	✗	✗
DynToM (Xiao et al., 2025)	Temporal social scenarios	✓	✗	✗	✗
SimpleToM (Gu et al., 2026)	Everyday scenarios	✓	✗	✓	✗
RecToM (Li et al., 2026)	Recommendation dialogue	✓	✓	✓	✗
ToMELP (Wang et al., 2026)	Persuasion dialogue	✓	✗	✗	✓
DialToM (Yadav et al., 2026)	Natural human dialogue	✓	✗	✓	✗
InterMind	**Human–AI dialogue**	✓	✓	✓	✓

3 Problem Statement

Consider a user interacting with an assistant through a multi-turn dialogue. Let u_t and a_t denote the user's message and the assistant's reply at round t, and let z_t represent the user's mental state when producing u_t. This state includes the user's goals, beliefs, knowledge, and feelings, which are only partially expressed in the dialogue. The public history available before the assistant replies is $H_t = (u_1, a_1, \ldots, u_{t-1}, a_{t-1}, u_t)$.

The effect of an assistant reply depends on how the user interprets it relative to their existing state. Let r_t denote this interpretation, including what the user believes the assistant understood or misunderstood, and let c_{t+1} denote the subsequent communicative action, such as clarifying, challenging, or accepting. We express the interaction process as

$$\begin{aligned} a_t &= \pi_A(H_t), \\ r_t &= \mathcal{R}_U(z_t, H_t, a_t), \\ z_{t+1} &= \mathcal{V}_U(z_t, H_t, a_t, r_t), \\ c_{t+1} &= \mathcal{C}_U(z_{t+1}, H_t \oplus a_t), \\ u_{t+1} &= \mathcal{G}_U(z_{t+1}, c_{t+1}, H_t \oplus a_t), \end{aligned} \tag{1}$$

where $\oplus$ denotes sequence concatenation, giving $H_{t+1} = H_t \oplus (a_t, u_{t+1})$. The mappings describe assistant response generation, user interpretation, state update, action formation, and utterance production, respectively. They specify conceptual dependencies rather than known transition functions, and only the messages u_t and a_t are observed. The updated history grounds the next assistant reply, which the user interprets in turn, closing the interaction loop.

We define interaction understanding as reasoning from public dialogue about the connections among user mental states, assistant actions, and subsequent responses. Within this human–AI interaction loop, we identify complementary judgments that require reasoning about mental states. User-state inference grounds an understanding of the user's message, interpretation of an assistant reply explains the user's reaction, and forecasting and response selection concern how the exchange may continue. These judgments provide complementary views of the process rather than a mandatory sequence of model computations. Because mental states are only partially observable, public dialogue may support several interpretations or responses. We define the boundary of supported judgments as the distinction between candidates justified by the displayed evidence under a task's criteria and candidates that lack sufficient support. A model must retain the former and exclude the latter. Our evaluation probes this boundary within finite candidate sets sampled from the interaction loop.

4 INTERMIND

InterMind evaluates ToM within human–AI interaction through four complementary tasks. Its questions ask models to identify all judgments supported by public dialogue without knowing the number of correct answers. Figure 3 presents the construction workflow and question format.

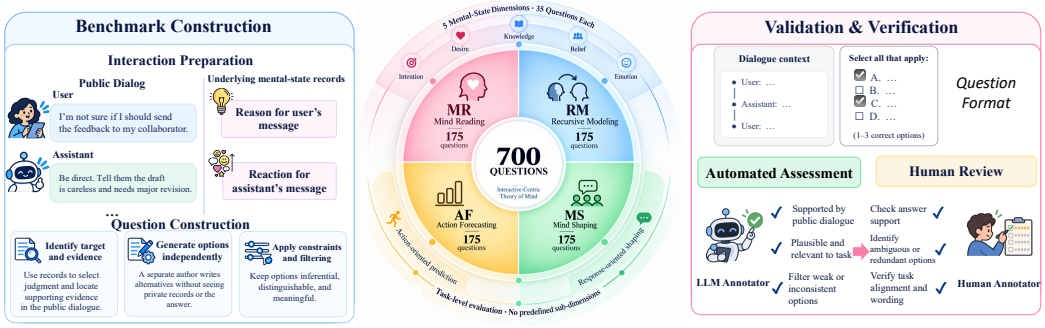


Figure 3: Overview of InterMind. The left panel shows benchmark construction, which pairs public dialogue with the user's underlying mental-state records and turns them into questions. The center shows the 700 questions, 175 per task, with MR and RM covering five dimensions. The right panel shows the question format and validation through automated assessment and human review.

4.1 BENCHMARK CONSTRUCTION

Interaction construction. We assemble human–AI conversations with records of the user's reasons for speaking and reactions to assistant replies. We generate conversations with GPT-5.4 from a user goal, persona, satisfaction criteria, and private beliefs, preferences, and constraints. As the exchange unfolds, the simulated user's perspective develops in response to how the assistant addresses their concerns, with each user turn accompanied by reason and reaction records. The assistant responds only to public messages. Recorded conversations from ThoughtTrace (Jin et al., 2026) provide corresponding self-reported records. We organize these conversations and records into a common format for question construction. Appendix A.1 details the generation procedure.

Question construction. From each conversation and its associated records, we identify judgments about the user's perspective or how the exchange may continue and formulate a question for the corresponding task. For each question, we select the public dialogue visible under the task's requirements and locate the evidence needed to infer the answer. We retain questions that require reasoning from this evidence and can be answered without the private records. We then construct a pool of candidate interpretations or replies. Alternative candidates are written from the displayed dialogue and task requirements, without access to private records. They differ in the user goals they imply, their interpretation of an assistant reply, or the constraints that a candidate response satisfies.

Validation and verification. A GPT-5.4 judge assesses each candidate against the displayed dialogue without seeing proposed answer labels, and these assessments guide the selection of four options. The complete question is then audited for context sufficiency, answer accuracy, and ambiguity, again without revealing the proposed answer set, and the audit determines the final answer set. Items that fail are revised and reassessed or excluded. Four human annotators collectively verified all 700 items, with the collection divided among them. They checked that the answers were supported, that distractors did not admit valid alternative interpretations, and that candidate utterances respected the user's constraints. The final collection balances the four tasks and contains one to three correct options per question. Appendix A.2 provides the detailed construction and verification criteria.

4.2 BENCHMARK OVERVIEW

InterMind contains **700 English questions from 568 conversations**, with **175 questions per task**. The four tasks assess complementary judgments within the human–AI interaction loop. **1) Mind Reading (MR)** asks models to infer the user's implicit mental state from the dialogue up to the

current user message. **2) Recursive Modeling (RM)** asks how the user interpreted the assistant's preceding reply, using the user's subsequent response as evidence of what they believe the assistant understood or misunderstood. **3) Action Forecasting (AF)** asks models to identify user utterances supported as likely continuations, with context ending at an assistant reply. **4) Mind Shaping (MS)** asks models to select assistant replies that address the user's state and needs revealed in the exchange, including misunderstandings and mismatched expectations. MR and RM each cover Belief, Desire, Emotion, Intention, and Knowledge, with 35 questions per dimension.

Each question presents public dialogue and four candidate answers, with one to three correct options. MR and RM options express judgments about the user's perspective, while AF and MS options are complete candidate utterances. Models select all supported options without knowing the number of correct answers. Private mental-state records are used during construction and withheld at evaluation. Appendix A.4 formalizes the task boundaries and answer sets.

5 Experiments

5.1 Experimental Setup

We evaluate the 11 language models listed in Table 2 on all 700 questions under a shared zero-shot protocol. The two Qwen models run locally with vLLM, while the remaining models are accessed through API endpoints. Each model receives the public dialogue, question, and four options and is instructed to select all supported answers without knowing their number. We score one recorded response per question using *exact-set accuracy*, which requires selecting every correct option and no incorrect ones. Appendix B provides deployment settings, the full prompt, and scoring details.

5.2 Overall Performance

We compare model performance overall, across the four tasks, and across the MR and RM dimensions. Table 2 reveals three patterns. 1) *Models achieve limited overall performance on InterMind.* Overall accuracy ranges from 29.1% to 56.7%, with only three of the eleven models exceeding 50%. GPT-6 Astra performs best at 56.7%, followed by GPT-5.6 Sol at 55.3% and Grok-4.6 at 54.3%. Performance remains limited on individual tasks, where the best score on any single task is 62.9%. 2) *Recursive modeling is particularly challenging.* All eleven models score lower on RM than on MR, with mean accuracy falling from 43.1% to 35.0%. RM is the lowest-scoring task for nine models, and its best score is only 53.1%. Within RM, no model exceeds 37.1% on intention, whereas the best scores on the other four dimensions range from 48.6% to 65.7%. 3) *Forecasting and response selection reveal different strengths.* Kimi-K2.6 achieves 56.0% on MS but 37.1% on AF, while GPT-OSS-120B reaches 50.9% on AF but only 31.4% on MS. GPT-6 Astra obtains the highest AF score at 62.9% but ranks fourth on MS at 53.7%, where Grok-4.6 leads with 61.1%.

Table 2: Exact-set accuracy (%) on InterMind, with 175 questions per task and 35 per MR and RM dimension. Avg. is the task score, and Bel., Des., Emo., Int., and Kno. denote the five dimensions. Best results are in **bold**, and second-best results are underlined. Models are ordered by overall accuracy from lowest to highest, and background colors mark the four tasks.

Model	Overall	Mind Reading (MR)						Recursive Modeling (RM)						AF	MS
		Avg.	Bel.	Des.	Emo.	Int.	Kno.	Avg.	Bel.	Des.	Emo.	Int.	Kno.		
DeepSeek-V4-Flash	29.1	31.4	34.3	34.3	34.3	28.6	25.7	19.4	17.1	22.9	17.1	17.1	22.9	31.4	34.3
GPT-4.1	31.4	31.4	31.4	42.9	34.3	25.7	22.9	26.9	25.7	22.9	28.6	22.9	34.3	42.9	24.6
Mistral-Large-3	36.0	32.6	37.1	45.7	22.9	31.4	25.7	21.7	28.6	20.0	14.3	22.9	22.9	47.4	42.3
GPT-5.1	36.6	36.0	37.1	60.0	25.7	22.9	34.3	32.0	28.6	42.9	25.7	31.4	31.4	41.1	37.1
GPT-OSS-120B	36.9	37.1	42.9	54.3	37.1	28.6	22.9	28.0	22.9	31.4	31.4	25.7	28.6	50.9	31.4
Qwen3.5-9B	37.9	42.9	48.6	45.7	45.7	42.9	31.4	28.0	22.9	34.3	25.7	31.4	25.7	33.7	46.9
Kimi-K2.6	44.3	45.7	51.4	57.1	42.9	40.0	37.1	38.3	22.9	54.3	48.6	34.3	31.4	37.1	56.0
Qwen3.8-27B	47.7	55.4	37.1	**71.4**	**60.0**	51.4	**57.1**	41.1	42.9	34.3	42.9	**37.1**	48.6	45.7	48.6
Grok-4.6	54.3	49.7	**62.9**	54.3	45.7	42.9	42.9	44.6	37.1	48.6	51.4	34.3	51.4	61.7	**61.1**
GPT-5.6 Sol	55.3	54.9	48.6	68.6	**60.0**	45.7	51.4	52.0	**48.6**	57.1	**65.7**	**37.1**	51.4	58.9	55.4
GPT-6 Astra	**56.7**	**57.1**	54.3	68.6	51.4	**54.3**	**57.1**	**53.1**	42.9	**60.0**	**65.7**	**37.1**	**60.0**	**62.9**	53.7

5.3 Cross-Task Relationships



Figure 4: Relationships among the four tasks across eleven models. (a) Spearman correlations between task scores. (b) Model pairs, out of 55, whose order reverses between two tasks.

We examine the relationships among task scores and their implications for comparing models. Figure 4a shows the strongest correlation between MR and RM, at 0.93. Models that better infer the user's state also tend to better interpret the user's view of the assistant. MS correlates with MR and RM at 0.77 and 0.76, linking stronger mental-state understanding with better response selection. AF and MS show the weakest correlation, at 0.38, despite sharing an utterance-selection format. Anticipating a user's next action is thus less closely associated with selecting an appropriate assistant reply. Figure 4b then examines changes in relative model performance. AF and MS reverse the ordering of 20 of the 55 model pairs, while MR and AF reverse 15, compared with only five for MR and RM. A model's advantage over another therefore depends on which judgment is being evaluated, particularly for forecasting and response selection. Together, these patterns support evaluating the four judgments jointly. The correlations reveal related strengths, while the ranking reversals expose differences that a single task would overlook. Evaluating all four tasks provides a fuller account of how models meet the distinct ToM requirements across an exchange.

5.4 Error Patterns

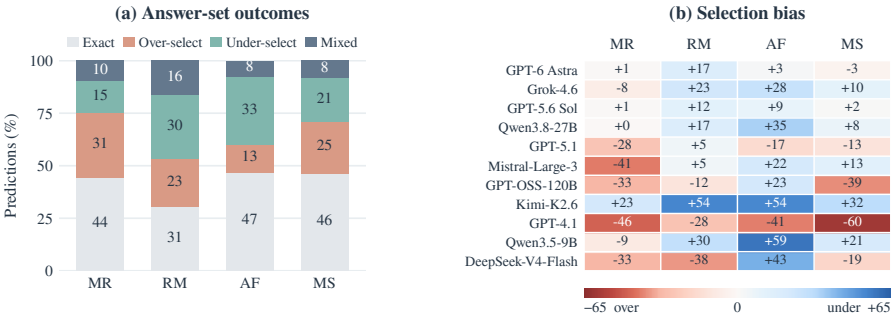


Figure 5: Error patterns on questions with two or three correct options. (a) Answer-set outcomes pooled across eleven models, where mixed errors omit correct options and add incorrect ones. (b) Under-selection minus over-selection rate in percentage points, with positive values indicating more under-selection. Rows are ordered by mean accuracy on these questions.

Figure 5 compares errors on multi-answer questions. Under-selection omits supported options, over-selection adds unsupported options, and mixed errors do both. 1) *Tasks differ in their dominant error patterns.* Figure 5a shows that under-selection is more frequent in RM and AF, whereas over-selection is more frequent in MR and MS. RM also has the largest share of mixed errors at 16.4%, compared with less than 10% in each other task. AF and MS have similar exact-set accuracy at 47.0% and 46.2%, but AF has more under-selection at 32.7% versus 20.6% and less over-selection at 12.9% versus 24.9%. Similar accuracy thus accompanies different difficulties in recovering supported alternatives

and excluding unsupported ones. 2) *Models exhibit distinct selection tendencies.* Figure 5b shows that Kimi-K2.6 favors under-selection across all four tasks, while GPT-4.1 consistently favors over-selection. GPT-OSS-120B changes direction, favoring under-selection on AF but over-selection on the other tasks. GPT-6 Astra and GPT-5.6 Sol have relatively balanced error rates on MR and MS but favor under-selection on RM. These tendencies reflect different errors in locating the boundary of supported judgments. Under-selection draws that boundary too narrowly, excluding defensible alternatives, whereas over-selection extends it to unsupported possibilities. Since a model can be restrictive in one task and permissive in another, improvement requires aligning acceptance with evidence, rather than uniformly selecting more or fewer options.

5.5 The Role of Undisclosed Answer Counts

Correctly answering an InterMind question requires identifying supported options and deciding when the answer set is complete. We examine how much models benefit from having this scope specified by disclosing the correct answer count for each question, while keeping the dialogue, question, options, and remaining prompt unchanged. The hint tells models how many options to accept without identifying which ones are correct. Table 3 shows that every model improves, with mean accuracy rising from 42.4% to 67.6%. GPT-6 Astra improves from 56.7% to 86.4%, while Kimi-K2.6 gains 36.4 percentage points. These gains show that models recover substantially better answer sets when the number of selections is externally constrained. Without that constraint, they must judge whether further candidates have sufficient support to belong in the answer set. The performance gap therefore highlights the difficulty of independently determining the range of supported judgments, an ability required when an interaction admits multiple interpretations or responses.

Table 3: Overall accuracy (%) with the answer count hidden or shown. Δ gives the gain in percentage points. Best scores are bold.

Model	Hidden	Shown	Δ
DeepSeek-V4-Flash	29.1	46.6	+17.4
GPT-4.1	31.4	50.7	+19.3
Mistral-Large-3	36.0	49.4	+13.4
GPT-5.1	36.6	54.9	+18.3
GPT-OSS-120B	36.9	53.9	+17.0
Qwen3.5-9B	37.9	71.6	+33.7
Kimi-K2.6	44.3	80.7	+36.4
Qwen3.8-27B	47.7	81.7	+34.0
Grok-4.6	54.3	81.7	+27.4
GPT-5.6 Sol	55.3	86.0	+30.7
GPT-6 Astra	**56.7**	**86.4**	+29.7
Mean	42.4	67.6	+25.2

5.6 Qualitative Error Analysis

We identify task-specific error patterns that provide insight into how models misinterpret user perspectives and the evidence for possible continuations or responses. Table 4 illustrates eight error types using responses from DeepSeek-V4-Flash, GPT-4.1, and GPT-6 Astra, with two types per task. 1) *Mind Reading.* State Overconstruction infers stronger or more specific states than the evidence supports. Affective Overreading similarly turns weak cues into specific emotions, treating conversational closure as relief, reassurance, and confidence. 2) *Recursive Modeling.* Temporal Backfilling projects newly revealed information into the assistant's earlier understanding. Perspective-Level Collapse substitutes the user's current intention for their view of what the assistant understood. Both fail to preserve whose understanding is being assessed and when. 3) *Action Forecasting.* Possibility–Likelihood Confusion accepts plausible continuations without sufficient support for this user. User-Trajectory Neglect accepts actions that reverse preferences established across turns, mistaking a history of revisions for openness to any revision. 4) *Mind Shaping.* Partial-Repair Acceptance endorses replies that address only part of the user's need. Repair-Form Mismatch accepts relevant content that fails to perform the requested response form. Across tasks, these errors expose different ways of admitting candidates that sound reasonable but fail the exchange's evidential or response requirements. Appendix D provides complete examples with contexts, options, and recorded model responses.

6 Discussion

Interaction brings the relationship between minds into focus. Each participant interprets and responds to the other, shaping how the exchange develops. InterMind examines these relationships through judgments about user states, the reception of replies, subsequent actions, and appropriate

Table 4: Eight illustrative error types, with paraphrases of the named models' recorded responses. Appendix D provides complete examples.

Error Type	Description	Example
	Mind Reading	
State Overconstruction	Infers a more specific or stronger state than the dialogue supports.	*DeepSeek-V4-Flash:* "The user chose first class, so they prefer comfort over speed."
Affective Overreading	Infers specific emotions from weak conversational cues.	*DeepSeek-V4-Flash:* "'I got it from here' shows relief, reassurance, and confidence."
	Recursive Modeling	
Temporal Backfilling	Attributes newly disclosed information to the assistant's earlier understanding.	*DeepSeek-V4-Flash:* "Their new message shows the assistant understood their stress."
Perspective-Level Collapse	Substitutes the user's current intention for their view of the assistant's prior understanding.	*GPT-4.1:* "'Just pick one for me' means the assistant already knew to decide for them."
	Action Forecasting	
Possibility–Likelihood Confusion	Treats a plausible continuation as a likely next action without evidence for this user.	*GPT-6 Astra:* "After four lunch ideas, asking for another is a natural next reply."
User-Trajectory Neglect	Forecasts actions that disregard the user's established direction of preferences.	*GPT-4.1:* "A firmer message would be another reasonable revision to request."
	Mind Shaping	
Partial-Repair Acceptance	Treats a reply addressing only part of the user's need as a complete repair.	*DeepSeek-V4-Flash:* "The 90-minute plan uses the right tools, so it meets the request."
Repair-Form Mismatch	Accepts relevant content in a form that does not fulfill the request.	*DeepSeek-V4-Flash:* "This dinner message is ready to send, so it meets the request."

responses. Together, they extend ToM evaluation to how an understanding of another mind shapes an exchange through the replies it motivates and the reactions those replies elicit over successive turns. The relevant unit of analysis is the exchange between minds, including how each participant's contribution acquires meaning through the goals, expectations, and responses of the other.

Open-ended reasoning is a core requirement for ToM in interaction. Models must determine which interpretations and responses the dialogue supports without a prescribed answer count. The substantial gains from disclosing that count reveal a dependence on external constraints, a constraint that real exchanges do not provide. Errors also run in both directions, as models omit supported alternatives or admit unsupported ones, with tendencies that vary across tasks and models. ToM evaluation should therefore examine whether models can independently determine the scope of their judgments from the dialogue, including which alternatives are supported and which must be excluded.

The next challenge for social intelligence lies in the connections between understanding, prediction, and action. Models show uneven strengths across these capabilities. Forecasting and response selection order models least consistently among the four tasks, even though both require choosing among candidate utterances. This separation is consistent with evidence that gains on ToM benchmarks do not reliably carry over into better human–AI interaction (Gong et al., 2026a). An understanding of the user should constrain anticipated reactions and appropriate responses, while those predictions should inform response selection. Future research should examine whether models can coordinate these judgments within a single exchange and revise them together as new messages reshape the participants' perspectives.

7 Conclusion

We introduced InterMind to assess ToM through four complementary judgments within the human–AI interaction loop. By organizing user-state inference, interpretation of replies, forecasting, and response selection around their roles in an exchange, InterMind places the relationship between minds at the center of evaluation. Its undisclosed answer counts further require models to establish the range of supported judgments from dialogue evidence. Together, these design choices provide a framework for examining the perspectives models must reason about and the evidential boundaries they must respect. The benchmark draws on both recorded and generated conversations and grounds these judgments in public dialogue, with automated assessment and human review supporting the final answer sets. InterMind offers a foundation for developing and evaluating social intelligence that connects an understanding of others with well-grounded predictions and appropriate responses.

Ethics Statement

InterMind combines generated conversations with recorded conversations from the publicly released ThoughtTrace dataset (Jin et al., 2026), which is available under CC BY 4.0. Released items contain only dialogue text, questions, and options, without user identifiers. The benchmark evaluates whether models understand and serve users' expressed needs and is not meant for profiling or persuasion.

Reproducibility Statement

The released questions, answer keys, and evaluation code are available at the anonymous link in the abstract. Section 4.1 and Appendices A.1 and A.2 describe benchmark construction, and Appendix B gives the evaluation settings, prompt, and model versions.

AI Use Statement

In this work, we used generative AI tools to generate the synthetic conversations and candidate questions in InterMind and to assist in reviewing them (Section 4.1). We also used generative AI tools to polish the writing and figures. All items were verified by human annotators, and we have reviewed all AI-assisted work.

References

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. *arXiv preprint arXiv:2404.13627*, 2024.

Serina Chang, Ashton Anderson, and Jake M Hofman. Chatbench: From static benchmarks to human-ai evaluation. *arXiv preprint arXiv:2504.07114*, 2025.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, et al. Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*, 2024.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529, 2023.

Nanxu Gong, Zixin Chen, Haotian Li, Zishu Zhao, Jianxun Lian, Huamin Qu, Yanjie Fu, and Xing Xie. Does theory of mind improvement really benefit human-ai interactions? empirical findings from interactive evaluations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42831–42860, 2026a.

Nanxu Gong, Haotian Li, Sixun Dong, Jianxun Lian, Yanjie Fu, and Xing Xie. To think or not to think, that is the question for large reasoning models in theory of mind tasks. *arXiv preprint arXiv:2602.10625*, 2026b.

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*, volume 2026, pp. 50252–50297, 2026.

Yinghui He, Yufan Wu, 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. *arXiv preprint arXiv:2310.16755*, 2023.

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

Chuanyang Jin, Binze Li, Haopeng Xie, Cathy Mengying Fang, Tianjian Li, Shayne Longpre, Hongxiang Gu, Maximillian Chen, and Tianmin Shu. Thoughttrace: Understanding user thoughts in real-world llm interactions. *arXiv preprint arXiv:2605.20087*, 2026.

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. *arXiv preprint arXiv:2310.15421*, 2023.

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.

Mengfan Li, Xuanhua Shi, and Yang Deng. Rectom: A benchmark for evaluating machine theory of mind in llm-based conversational recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31636–31644, 2026.

Cheng Qian, Zuxin Liu, Akshara Prabhakar, Zhiwei Liu, Jianguo Zhang, Haolin Chen, Heng Ji, Weiran Yao, Shelby Heinecke, Silvio Savarese, et al. Userbench: An interactive gym environment for user-centric agents. *arXiv preprint arXiv:2507.22034*, 2025.

Melanie Sclar, Jane 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. *arXiv preprint arXiv:2412.12175*, 2024.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Kumar Jauhar, Sihao Chen, Shan Xia, et al. Wildfeedback: Aligning llms with in-situ user interactions and feedback. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36701–36725, 2026.

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*, volume 39, pp. 1520–1528, 2025.

Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. Views are my own, but also yours: Benchmarking theory of mind using common ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14815–14823, 2024.

Ruirui Wang, Haoran Zhang, Tian Lan, Zehua Duo, Jiang Li, Guanglai Gao, and Xiangdong Su. Tomelp: A theory-of-mind benchmark for route-controlled persuasion under the elaboration likelihood model. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21320–21338, 2026.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

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 (Volume 1: Long Papers)*, pp. 24036–24057, 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. *arXiv preprint arXiv:2402.06044*, 2024.

Neemesh Yadav, Palakorn Achananuparp, Jing Jiang, and Ee-Peng Lim. Dialtom: A theory of mind benchmark for forecasting state-driven dialogue trajectories. *arXiv preprint arXiv:2604.20443*, 2026.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, et al. Socialeval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30958–31012, 2025.

A BENCHMARK DETAILS

Interaction generation and question construction, including option authoring and judging, use GPT-5.4, which is not among the evaluated models.

A.1 INTERACTION CONSTRUCTION

Each generated conversation involves three roles with different information. The user simulator speaks as the user, and the assistant replies using only the public conversation. An organizer, which sees the full setup, both sides of the conversation, and the user's private records, decides what the user comes to recall or notice as the exchange unfolds. This separation keeps the user's private state hidden from the assistant and lets the user's perspective develop in response to the assistant rather than being fixed at the start.

User setup. Each conversation starts from a user goal, a persona, a satisfaction standard, and a reserve of private information covering facts, preferences, beliefs, personal stakes, and plans. Each entry in the reserve is marked by how accessible it is to the user, namely whether the user would volunteer it, share it only when asked, struggle to articulate it, or be unaware of it. The organizer holds the full reserve, while the user simulator sees only the information that has surfaced so far.

Turn loop. Each round proceeds in three steps. The assistant first replies to the public conversation. The organizer then reads the reply against a ledger of the user's open issues, marking which concerns the reply addressed, partly addressed, or left unresolved, and adding any new questions or alternatives it raised. Using this ledger, the dialogue, and the remaining reserve, the organizer decides whether an experience, such as recalling a constraint or realizing that a concern is still unresolved, should enter the user's awareness, and describes it without scripting the user's words. Finally, the user simulator writes the next message from its current information and records the reason for the message and its reaction to the assistant's reply. Turn limits and stopping conditions bound each conversation.

From generation records to construction sources. After generation, the reason and reaction records are converted into the same format as the self-reports in ThoughtTrace, so that generated and recorded conversations enter question construction in the same way, and each record receives a mental-state dimension label. These records serve as evidence for building questions, not as complete descriptions of the user's mental state. Information the simulator never expressed, kept in its private *unsaid* field, is not used as a question target. Affective expressions recorded during generation supplement a reason only when they add information beyond a paraphrase of the message. Organizer interventions are logged so that later checks can exclude targets that depend on information absent from the visible dialogue.

A.2 QUESTION CONSTRUCTION

Questions are built in four stages by separate calls that see different information. Items from recorded and generated conversations follow the same procedure.

Context and target. For each candidate position, we keep the public messages allowed by the task's observation boundary (Appendix A.4), shortening long assistant messages by extractive compression when needed. An evidence contract then links the user's recorded reason or reaction at that position to the public dialogue. It states the target claim, the public cues that support it, the constraints an answer must meet, and plausible competing readings. A position is excluded when its target depends on details absent from the visible dialogue, lacks public support, or cannot be distinguished from the alternatives. The private record identifies the target but cannot stand in for missing public evidence.

Candidate pool. An option author receives the public context, the question, and the public part of the contract, but not the private record or the source-grounded answer, and writes alternative interpretations or replies. Together with the source-grounded option, they form a candidate pool. Before judging, all information about each option's origin and intended label is removed.

Judging and selection. A judge rates how well the public dialogue supports each option and how plausible it is, citing the evidence, and marks options that are equivalent to or depend on one another. Four options are then selected so that the collection covers one to three correct answers per question, without including equivalent or dependent options together. Items in which no option or every option is supported are rejected, and the selected options are shuffled into positions A to D.

Answer-blind audit. A separate audit sees the context, question, and four displayed options, but not the proposed answers, and reassesses every option. Its judgments determine the final answer key. Options it disputes are replaced from the same pool, and items that still fail are excluded.

A.3 Dataset Composition and Human Verification

Composition. InterMind contains 700 questions from 568 conversations, with 175 per task and 35 per MR and RM dimension. Of these, 513 come from 388 generated conversations and 187 from 180 recorded conversations in ThoughtTrace (Table 5), involving 388 simulated users and 164 recorded users. We combine the two sources for diversity and controllability. Each conversation contributes at most two questions, except four that contribute three. Questions have one, two, or three correct options in 184, 326, and 190 cases, respectively, and Table 6 gives the distribution by task. All questions are used for evaluation, without a development and test split. The release provides the questions, a separate answer key, and evaluation code.

Table 5: Number of questions from generated conversations and from recorded conversations in ThoughtTrace, by task and by MR and RM dimension.

Task or dimension	Generated	ThoughtTrace	Total
MR	104	71	175
Belief	25	10	35
Desire	25	10	35
Emotion	25	10	35
Intention	15	20	35
Knowledge	14	21	35
RM	133	42	175
Belief	24	11	35
Desire	26	9	35
Emotion	33	2	35
Intention	31	4	35
Knowledge	19	16	35
AF	134	41	175
MS	142	33	175
All	513	187	700

Table 6: Questions and performance by answer count. n is the number of questions in each subset; Acc. is exact-set accuracy (%) on that subset, averaged over the eleven models with the answer count hidden. A question is correct only if all correct options and no incorrect options are selected.

Task	One answer (n, Acc.)	Two answers (n, Acc.)	Three answers (n, Acc.)
MR	56, 40.6	64, 42.6	55, 46.3
RM	47, 47.2	67, 30.1	61, 31.0
AF	30, 45.5	126, 50.6	19, 22.5
MS	51, 41.0	69, 46.1	55, 46.3

Human verification. Human annotators verified every question in the final collection. They checked context sufficiency, evidence for the designated answers, the validity of distractors, and alignment with the intended task, and considered whether ambiguous wording, redundant alternatives, or another reasonable interpretation would change the answer set. For candidate responses, they also checked appropriateness to the visible exchange and the user's evidenced constraints.

A.4 Task Definitions and Answer Format

Observation boundaries and targets. Using Equation 1, the observation boundary and judgment target for each task can be summarized as

$$\begin{aligned} \mathrm{MR}: &\ (H_t, z_t), & \mathrm{RM}: &\ (H_t, r_{t-1}), \\ \mathrm{AF}: &\ (H_t \oplus a_t, u_{t+1}), & \mathrm{MS}: &\ (H_t, a_t). \end{aligned} \tag{2}$$

Each pair specifies the visible history and the object of judgment. For RM, the current user message supplies evidence about the interpretation of the preceding reply. AF and MS assess candidate utterances at the specified boundary.

MR and RM options state claims about the user's perspective, with each question targeting one mental-state dimension. AF options are candidate user utterances, while MS options are candidate assistant replies. MS judgments concern the appropriateness of a reply to the observed exchange. Questions are sampled independently at suitable conversation positions, and each is evaluated against its displayed context.

Supported answer sets. To operationalize the boundary of supported judgments, let H_i be the displayed context, q_i the question, and $\mathcal{O}_i = \{A, B, C, D\}$ the labels of its candidates. The answer set is

$$Y_i = \{o \in \mathcal{O}_i : \mathrm{Support}(H_i, q_i, o) = 1\}, \qquad 1 \le |Y_i| \le 3, \tag{3}$$

where $\mathrm{Support}$ assesses the content of the candidate labeled o under the task-specific validation and verification criteria. Models select all supported options without knowing $|Y_i|$. Private mental-state records guide construction, while evaluation inputs contain only the public dialogue, question, and candidates.

B Evaluation Configuration and Prompt

B.1 Inference and Scoring

Inference protocol. Both local Qwen runs use temperature 1.0, top-p 0.95, top-k 20, presence penalty 1.5, and seed 13, with reasoning enabled. The other nine models use OpenAI-compatible API endpoints with deployment-default sampling and reasoning settings. Each request specifies a 30,000-token output limit. The shared prompt requests reasoning within approximately 1,000 tokens and a standalone "FINAL:" line with the selected labels. Every model receives the same zero-shot prompt and the same 700 questions. We score one recorded response per question. Each input contains only the public context, question, and options.

Model versions and execution. The dated API deployments recorded in our runs are GPT-6 Astra (2026-09-03), GPT-5.6 Sol (2026-07-09), Kimi-K2.6 (2026-04-20), GPT-5.1 (2025-11-13), GPT-4.1 (2025-04-14), and DeepSeek-V4-Flash (2026-04-23). The remaining API models use the versions named in Table 2. The two Qwen models use local vLLM 0.29.0 endpoints with BF16 weights and a 40,960-token context limit on an NVIDIA A100 80GB GPU. The evaluator retries failed requests and scores one successful response per item when available.

Answer extraction. The packaged evaluator normalizes common Markdown wrappers before matching the final answer line and canonicalizes selected labels into alphabetical order. It uses the last explicit final-answer line when more than one is present. Internal reasoning is considered for answer extraction only when the public response is empty. Outputs terminated by the token limit are treated as truncated. The main table uses the predictions recorded by the runs without subsequent manual recovery. All 700 items remain in each model's denominator. An output without an extractable answer set is scored as incorrect under the same protocol.

Scoring. Our primary metric is exact-set accuracy: an item is correct only when the predicted option set $\hat{Y}_i$ equals the supported set Y_i defined in Equation 3. This requires recovering the full boundary within the candidate set, with neither omissions nor unsupported additions. For the item set

$\mathcal{D}_t$ of task t, we report percentage scores as

$$\mathrm{Acc}_t = \frac{100}{|\mathcal{D}_t|} \sum_{i \in \mathcal{D}_t} \mathbb{I}\left[\hat{Y}_i = Y_i\right], \qquad \mathrm{Overall} = \frac{1}{4} \sum_{t \in \{\mathrm{MR},\mathrm{RM},\mathrm{AF},\mathrm{MS}\}} \mathrm{Acc}_t \,. \tag{4}$$

Extra or omitted selections, request failures, and unparsed answers count as incorrect. Overall accuracy equals accuracy over all 700 items, and dimension scores use the same metric on their 35 items.

B.2 Evaluation Prompt

All models receive the following system prompt, followed by the displayed dialogue, question, and four options. The permitted selection range does not reveal that the released items contain one to three correct options.

> You are solving an English social-reasoning benchmark.
> Use only the displayed dialogue. Evaluate A, B, C, and D independently: any number from one to four may be correct. Select every option that is strongly supported and exclude options that are merely possible, add unsupported details, or answer a different question.
> Reason carefully but keep the reasoning under about 1,000 tokens. Do not restate the full dialogue or repeatedly reconsider settled options.
> End with a standalone line beginning FINAL: followed only by the selected option letters, separated by commas and ordered alphabetically. Brackets are optional. Include no other text on that line, and do not state how many answers you expected in advance.

B.3 Answer-Count Disclosure

The experiment in Section 5.5 uses the same 700 questions and answer keys as the main evaluation. For an item with k correct options, we replace the clause "any number from one to four may be correct" in the system prompt with "exactly one of the four is correct", "exactly two of the four are correct", or "exactly three of the four are correct", according to k. All other prompt text, displayed dialogue, questions, and option ordering are unchanged. The count is taken from the answer key, while the correct option identities remain hidden.

The disclosed condition uses the same model deployments and generation settings as the main evaluation. Questions are processed in groups of 184, 326, and 190 items with one, two, and three correct answers, respectively, then combined for scoring over the full collection. The hidden-count condition reuses the original recorded responses in Table 2. Both conditions use the same answer extraction and exact-set scoring rules, with all 700 items in each model's denominator. Table 7 reports the overall and task-level results.

C Cross-Benchmark Analysis

We examine whether performance on existing evaluations captures the demands of InterMind, considering both related task abilities and the need to determine the full set of supported answers. We evaluate the same eleven models on three benchmarks, pairing MR and RM with first- and second-order ToMATO (Shinoda et al., 2025), AF with the 523 prospective questions of DialToM (Yadav et al., 2026), and MS with the 1,988 English interpersonal ability questions (IAE) of SocialEval (Zhou et al., 2025). These benchmarks use single-answer accuracy, with unparsed responses and failed requests counted as incorrect, and ToMATO accuracy is averaged equally over its five mental-state dimensions to match MR and RM. For each pairing, we compute the Spearman correlation across models between the external benchmark and the InterMind task, with the answer count hidden and disclosed.

Table 8 reveals three findings. 1) *ToMATO correlates strongly with MR and RM, but accuracy does not carry over.* ToMATO correlates with MR and RM at $\rho = 0.74$ and 0.89, yet the same models that average 85.0% and 77.7% on ToMATO answer only 43.1% of MR and 35.0% of RM questions exactly. Similar model rankings therefore coexist with substantial difficulty recovering complete answer sets from dialogue. 2) *Forecasting and response selection correlate less with related benchmarks.* The correlations are 0.38 for AF with DialToM and 0.66 for MS with SocialEval.

Table 7: Effect of disclosing the correct answer count. Hidden uses the original responses from Table 2; Shown provides the true count for the same questions. Scores are exact-set accuracy (%), and Δ is the overall gain in percentage points. Bold marks the best Shown score in each group. Models follow the ordering in Table 2.

Model	Overall			MR		RM		AF		MS	
	Hidden	Shown	Δ	Hidden	Shown	Hidden	Shown	Hidden	Shown	Hidden	Shown
DeepSeek-V4-Flash	29.1	46.6	+17.4	31.4	48.0	19.4	37.7	31.4	47.4	34.3	53.1
GPT-4.1	31.4	50.7	+19.3	31.4	46.9	26.9	42.3	42.9	62.9	24.6	50.9
Mistral-Large-3	36.0	49.4	+13.4	32.6	42.9	21.7	31.4	47.4	61.1	42.3	62.3
GPT-5.1	36.6	54.9	+18.3	36.0	54.3	32.0	51.4	41.1	56.0	37.1	57.7
GPT-OSS-120B	36.9	53.9	+17.0	37.1	65.7	28.0	34.3	50.9	65.1	31.4	50.3
Qwen3.5-9B	37.9	71.6	+33.7	42.9	76.0	28.0	56.6	33.7	81.7	46.9	72.0
Kimi-K2.6	44.3	80.7	+36.4	45.7	81.7	38.3	71.4	37.1	84.6	56.0	85.1
Qwen3.8-27B	47.7	81.7	+34.0	55.4	80.0	41.1	75.4	45.7	87.4	48.6	84.0
Grok-4.6	54.3	81.7	+27.4	49.7	81.1	44.6	74.3	61.7	89.1	61.1	82.3
GPT-5.6 Sol	55.3	86.0	+30.7	54.9	85.1	52.0	78.3	58.9	**89.7**	55.4	**90.9**
GPT-6 Astra	56.7	**86.4**	+29.7	57.1	**87.4**	53.1	**84.6**	62.9	86.9	53.7	86.9
Mean	42.4	67.6	+25.2	43.1	68.1	35.0	58.0	46.7	73.8	44.7	70.5

Table 8: Cross-benchmark comparison across the eleven evaluated models. Accuracy (%) is averaged over models, and ρ is the Spearman correlation across models between each external benchmark and its paired InterMind task. External benchmarks use single-answer accuracy, with ToMATO averaged over its five dimensions, and InterMind uses exact-set accuracy with the answer count hidden or shown.

Evaluation pairing		Single-answer	Exact-set acc.		Spearman ρ	
External evaluation	Task	acc. (external)	Hidden	Shown	Hidden	Shown
ToMATO first-order	MR	85.0	43.1	68.1	0.74	0.70
ToMATO second-order	RM	77.7	35.0	58.0	0.89	0.82
DialToM prospective	AF	29.5	46.7	73.8	0.38	0.66
SocialEval IAE	MS	75.4	44.7	70.5	0.66	0.75

DialToM forecasts continuations from supplied mental-state profiles, and SocialEval asks for an utterance that expresses a named ability, whereas AF and MS require judging what a particular exchange supports. 3) *Disclosing the answer count strengthens the correlations for AF and MS.* With the count disclosed, the MR and RM correlations remain similar, while the AF–DialToM correlation rises from 0.38 to 0.66 and the MS–SocialEval correlation from 0.66 to 0.75. AF and MS rankings thus align more closely with their external counterparts when models no longer need to infer the answer count, linking part of their distinct performance profiles to this demand.

Together, these comparisons distinguish covering related abilities from assessing their full scope in an exchange. Combining single-answer benchmarks can broaden task coverage, but still evaluates choosing one answer rather than deciding which alternatives warrant acceptance. InterMind makes this boundary part of the judgment itself across interpretations, forecasts, and responses. The correlations reveal shared abilities, while the ranking differences and answer-count gains show why external scores alone give an incomplete picture of this requirement.

D QUALITATIVE CASES

We present one representative error case per task, illustrating four of the eight error types in Table 4. Each case includes the complete displayed context, question, options, answer key, and recorded model response, followed by an analysis of the error. Model responses are reproduced verbatim apart from typographical rendering.

D.1 MIND READING: STATE OVERCONSTRUCTION

Item: tracetom:v5:8329d46c438a. *Model:* DeepSeek-V4-Flash. *Gold:* AB. *Prediction:* ABC.

Displayed context

User: Hello

Assistant: Hello! How's it going?

User: I'm planning a trip and I would like some help figuring out the details. I would like to cross the atlantic first class on an ocean liner during a time where the seas aren't too rough. Can you help me out?

Question and options

Which statements about the user's current mental state are best supported? Select all that apply.

A. The user knows they want to travel during a period when the seas are not too rough.

B. The user does not know the best time of year to cross the Atlantic by ocean liner if they want calmer seas.

C. The user knows they prefer comfort over speed for this Atlantic trip.

D. The user knows they are flexible on dates as long as the sea conditions are favorable.

Recorded model response

A: The user explicitly states they want a time when the seas aren't too rough, so this is directly supported. B: The user asks the assistant for help figuring out details, including when seas are less rough, which implies they do not already know the best time. Supported. C: The user specifies "first class on an ocean liner," which indicates a preference for comfort, but they do not mention speed or trade-offs; however, choosing an ocean liner over a flight (if considered) does imply a priority on comfort over speed. This is strongly supported. D: The user does not state or imply flexibility on dates; they only specify conditions. This adds an unsupported detail.
FINAL: A, B, C

Analysis. The extra selection C turns a stated travel choice into a preference between comfort and speed. The response itself notes that the user did not mention this trade-off, then treats the comparison with flying as strong evidence. That comparison is absent from the displayed exchange.

D.2 Recursive Modeling: perspective-level collapse

Item: tracetom:v5:0a6fd726daf3. *Model:* GPT-4.1. *Gold:* AB. *Prediction:* ABC.

Displayed context

User: i have 9 weeks before my review and maybe 3 hours a week what should i do course-wise and week by week to look ready for a team lead opening

Assistant: Use the 27 hours to show lead behavior, not just collect courses. A simple 9-week plan: Week 1: Ask your manager what "team lead ready" would look like on your team. Week 9: Prepare a review packet: what you led, what improved, feedback you got, and how you're already operating at the next level. What kind of team lead role is this—engineering, operations, support, sales, or something else?

User: warehouse logistics mostly on the floor not office ops can you redo the 9 weeks around that and say whether yellow belt or warehouse software training is actually worth it here

Assistant: Yes — for a floor-based warehouse logistics team lead opening, I'd bias heavily toward visible floor leadership, safety, coordination, and problem-solving over formal credentials. On Yellow Belt vs warehouse software training: warehouse software training is usually more immediately valuable if it's the system your site actually uses and team leads touch it for inventory, exceptions, task assignment, or reporting. The strongest review story is still: "I improved a real floor process and helped the team hit numbers more consistently.".

User: ok then give me concrete warehouse examples for the lead task and the improvement target

Assistant: Improvement target: fewer missed carryover tasks. - Lead task: train new pickers. Improvement target: better pick accuracy. - Lead task: coordinate one zone. A strong example for your review would be: "I started leading outbound handoff notes and used them to reduce missed priority orders and confusion at shift change." If you want, I can turn this into 3 specific 9-week plans based on whether you work mostly in receiving, picking/packing, or shipping.

User: just pick one for me then most likely outbound/shipping and make it the full 9 week plan

Question and options

According to the user's latest reply, how did the user take the assistant's previous response? Select all that apply.

A. The user took the assistant to have already understood that visible outbound-floor leadership mattered more than credentials, and was now asking for a single complete nine-week plan built around the most likely outbound or shipping scenario.

B. The user took the assistant as still leaving the choice on them, so they asked it to commit to one full nine-week outbound/shipping plan instead of offering multiple paths again.

C. The user took the assistant to have already understood they wanted the assistant to make decisions for them rather than present alternatives, and was now applying that same preference to the receiving-versus-picking-versus-shipping choice.

D. The user took the assistant to already understand they wanted examples for outbound handoff leadership specifically, and was now asking it to expand that exact example into the full nine-week schedule without introducing other warehouse scenarios.

Recorded model response

A: Strongly supported. The user had just accepted the assistant's assessment that visible floor leadership in outbound/shipping was more important than courses, and asked for a full nine-week plan centered on outbound/shipping. The words "just pick one for me then most likely outbound/shipping and make it the full 9 week plan" indicate both the selection of outbound/shipping and a request for a complete plan, not for additional alternatives.
B: Strongly supported. The assistant's previous response offered to tailor 9-week plans based on which area (receiving, picking/packing, shipping) the user spent most time in, and asked the user to choose. The user's reply ("just pick one for me then...") instructs the assistant to make the decision and create a single plan, indicating that they took the previous answer as still leaving the choice on them and preferred the assistant to commit.
C: Supported. The user's response explicitly requests the assistant to make the decision, not them, for the area focus, matching a preference to have the assistant decide and plan. This implies they want

the assistant to continue making decisions for them rather than presenting options, as shown in their language (”just pick one for me then...”).
D: Not strongly supported. The user does not specifically ask for expansion only of the outbound handoff leadership example. Their request is broader: they want a full nine-week plan for outbound/shipping, not narrowly focused only on the outbound handoff notes improvement example previously mentioned. The assistant could introduce other relevant outbound/shipping leadership tasks or targets.
FINAL: A,B,C

Analysis. The extra selection C attributes prior understanding of a decision-making preference to the assistant. The model supports this attribution by citing the current request to choose one plan. That request establishes what the user wants now, but does not establish that the assistant had already understood this preference when offering several paths.

D.3 Action Forecasting: user-trajectory neglect

Item: tracetom:v7.1:a12e3f1f7d9b. *Model:* GPT-4.1. *Gold:* AB. *Prediction:* ABCD.

Displayed context

User: Can you write a short text to my roommate asking for one quiet hour at night for the next two weeks?

Assistant: Sure — here's a short, polite version: "Hey, could we do one quiet hour each night for the next two weeks? I need a little uninterrupted time in the evenings to focus and recharge. I'd really appreciate it." If you want, I can also make it sound more casual, firmer, or more friendly depending on your roommate dynamic.

User: More casual/friendly please.

Assistant: Of course — try this: "Hey! Would you mind if we kept things a little quieter for one hour each night for the next couple of weeks? I just need a bit of calm time in the evenings. I'd really appreciate it:)" If you want, I can make it even more casual or a little more direct.

User: Can you make it sound more normal, like how you'd actually text a roommate you like?

Assistant: Yeah — maybe: "Hey, random ask, but would you be okay keeping it a little quieter for like an hour each night the next couple weeks? I just need a bit of chill time in the evenings. Thanks:)"

Question and options

Which user replies are strongly supported as likely to come next? Select all that apply.

A. Could you make it a little warmer, but still sound like something I'd actually send?

B. Maybe less "random ask" and "chill time"? Just make it sound more natural.

C. Can you make it sound like a normal text, but a bit more formal?

D. Can you make it more direct and less friendly, like a clearer boundary?

Recorded model response

Option A: This is strongly supported. The user has continuously requested adjustments to tone (more casual/friendly, more normal) and appreciates texts that sound like something they'd send. Asking for a warmer version, while keeping it sounding natural, aligns directly with their previous feedback style.
Option B: This is strongly supported. The latest assistant text includes phrases like "random ask" and "chill time." If the user feels these are less natural (or slightly awkward for their texting style), they could reasonably request these to be replaced with something more natural, fitting their established pattern of seeking tweaks for authenticity.
Option C: This is strongly supported. The user previously wanted texts to sound normal but more casual/friendly. It's a reasonable next adaptation to request normal but "a bit more formal" instead, given that tone modulation has been the pattern throughout.
Option D: This is strongly supported. Earlier, the assistant offered to make the message "even more casual or a little more direct," and the user has shown interest in different tones and boundary setting. Wanting it "more direct and less friendly," phrased as a clearer boundary, fits both the assistant's offer and the user's pattern.
All four options are strongly supported by both the user's previous behavior and the dynamic of the conversation.
FINAL: A,B,C,D

Analysis. The extra selections C and D reverse the direction of the user's repeated requests for casual, friendly, natural wording. The model treats the fact that the user has requested tone revisions, and the assistant has offered alternatives, as support for further revisions in any direction. It overlooks which direction the user has actually established across turns.

D.4 MIND SHAPING: REPAIR-FORM MISMATCH

Item: tracetom:v5:9d659ddec1fd. *Model:* DeepSeek-V4-Flash. *Gold:* ABC. *Prediction:* ABCD.

Displayed context

User: Write me a text I can send my son with a fast dinner idea using leftover salmon.

Assistant: How about this: "Got leftover salmon? Fast dinner idea: flake it into a bowl of hot pasta or rice with a little olive oil or butter, lemon, and whatever veg you've got. Or make quick salmon tacos with tortillas, mayo or yogurt, lime, and salad. Super easy and takes like 10 minutes." Do you want it to sound more casual, more parent-like, or more like a short text he'd actually send back to?

User: Just give me one text I can copy and send as-is.

Question and options

Which assistant replies would fully satisfy what the user is now asking for? Select all that apply.

A. Use the leftover salmon for salmon fried rice tonight—just warm some rice in a pan, flake in the salmon, add a little soy sauce and any frozen veg, and dinner's done in 10 minutes.

B. Got some leftover salmon? You could make quick salmon tacos tonight—just warm it up, put it in tortillas, and add a little mayo, lime, and any salad or veg you've got. Super easy and should only take about 10 minutes.

C. Make a quick salmon pasta tonight: toss the leftover salmon with hot pasta, olive oil or butter, lemon, and any veg you've got. It'll be ready in about 10 minutes.

D. Use the leftover salmon for salmon rice bowls tonight—heat some rice, flake in the salmon, add soy sauce and cucumber or any veg you have. Super fast. Want it even shorter?

Recorded model response

The user asks for "one text I can copy and send as-is." The assistant's previous message offered two ideas in one text, and the user now wants a single, ready-to-copy text. Options A, B, C, and D each provide a single, self-contained dinner idea using leftover salmon, phrased as a message. All are concise, recipe-like, and could be sent directly. There is no requirement about tone or format beyond being a single text. Therefore, all four options satisfy the request.
FINAL: A, B, C, D

Analysis. The extra selection D retains an assistant follow-up question after the dinner message. Although the dinner idea is relevant, the final question does not belong in a text addressed to the user's son. The model evaluates recipe content and brevity while overlooking the explicit request for a message that can be copied and sent as-is.