ToMELP: A Theory-of-Mind Benchmark for Route-Controlled Persuasion under the Elaboration Likelihood Model

Ruirui Wang[1,2,3], **Haoran Zhang**[1,2,3], **Tian Lan**[1,2,3], **Zehua Duo**[1,2,3], **Jiang Li**[1,2,3], **Guanglai Gao**[1,2,3], **Xiangdong Su**[1,2,3]*

[1] College of Computer Science, Inner Mongolia University, China

[2] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, China

[3] Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology, China

cswr@mail.imu.edu.cn, cssxd@imu.edu.cn

Abstract

Theory of Mind (ToM) is widely regarded as central to effective persuasion, yet existing evaluations often fail to capture the *infer–apply* loop that arises in real-world dialogue. We introduce THEORY-OF-MIND-GUIDED ELABORATION-LIKELIHOOD PERSUASION, a benchmark that jointly conditions on the audience *persona* and the Elaboration Likelihood Model (ELM) route (*central* vs. *peripheral*) within persuasive conversations. The benchmark tests whether large language models can perform ToM inference over multi-turn interactions and leverage these inferences for *controllable* persuasive generation. TOMELP provides a structured interface with evidence annotations, enabling automated evaluation of persuasive effectiveness, route alignment/deviation, evidence quality under the central route, and robustness to perturbations.The source code is available[1].

1 Introduction

Large language models (LLMs) are permeating many communication settings, and *persuasive dialogue* is a sensitive, crucial capability in this landscape. Beyond giving reasons, a persuasive system must model the interlocutor's mental states over multiple turns and adapt strategy and tone accordingly. This "infer–use" loop is fundamentally grounded in ToM, i.e., the ability to infer and predict others' beliefs, desires, and intentions (Premack and Woodruff, 1978). While recent LLMs exhibit strong performance on classic false-belief tasks, whether such behaviors reflect functional ToM or task-specific pattern matching remains contested (Riemer et al., 2024). On the one hand, recent studies suggest that model-generated messages can yield substantial attitude change in real audiences (Breum et al., 2024), highlighting both the promise and the potential risks of scalable AI-driven persuasion (Bai et al., 2025; Salvi et al., 2025). On the other hand, persuasion depends on *who* the audience is and *how* they process information; without controlling these variables, it is difficult to determine whether a model's performance reflects genuine psychological modeling or superficial alignment in phrasing.

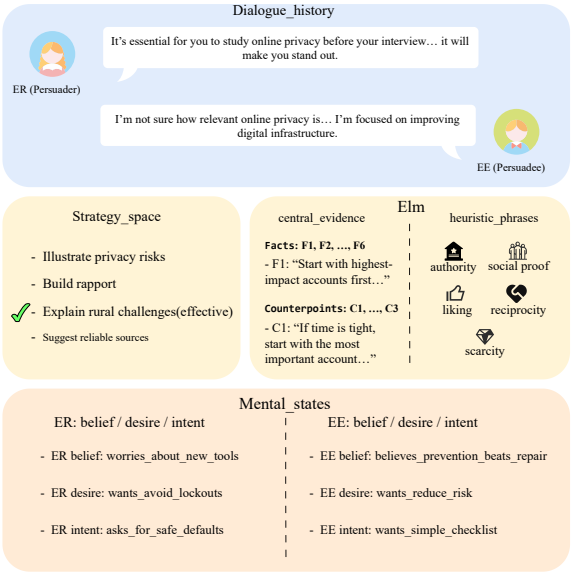

Figure 1: Example persuasion dialogue annotated with strategy supervision and ELM-controlled resources.

Recent benchmarks have diversified ToM evaluation for LLMs, spanning interactive stress tests in information-asymmetric dialogues FANToM Kim et al., 2023, narrative-based evaluations with persona/intent triggers OpenToMXu et al., 2024, systematic item banks for fine-grained social cognition ToMBenchChen et al., 2024b, and negotiation settings closer to strategic interaction NegotiationToM(Chan et al., 2024). Yet ToM performance remains brittle, sensitive to task framing and experimental setups. At the intersection of "ToM × persuasion," PersuasiveToM couples ToM inference

*Corresponding Author

[1] Dataset

Findings of the Association for Computational Linguistics: ACL 2026, pages 21320–21338

with strategy selection and effectiveness assessment, moving toward a closed loop from "reasoning" to "acting" (Yu et al., 2025). However, prior setups often lack explicit control over audience heterogeneity and information-processing routes. ELM posits a *central route* that scrutinizes evidence and a *peripheral route* that relies on heuristic cues (e.g., authority, affect, social conformity) (Petty and Cacioppo, 1986). Persona/personalization prompting may yield better-tailored generations but can produce inconsistent objective gains or introduce bias (Chen et al., 2024a; Thakur et al., 2025), motivating *controllable variables* and *interpretable metrics* (Zhang, 2024).

To address these gaps, we extend the dialogue-state backbone of PersuasiveToM with richer representations and controlled contrasts. We call the resulting dataset TOMELP and build a rigorous 2×2 framework crossing PERSONA with ELM routes (*central* vs. *peripheral*). Each instance is centered on a *dialogue state*: the dialogue history to the current turn, candidate strategies, and the target strategy with an effectiveness label as supervision. We also annotate BDI mental states for both persuader and persuadee, enabling a direct test of whether a model can *infer* ToM and then *use* it in generation. From the same state, ToMELP produces paired route counterparts: the *central* version includes citable facts and rebuttals (`evidence/counterpoints`), while the *peripheral* version provides heuristic cue phrases (e.g., authority, social proof) to constrain peripheral cues. This makes route-specific generation/evaluation for the same state a *controlled variable*.

For evaluation, we adopt *LLM-as-a-judge* with structured scoring throughout to measure attitude change, route compliance, and (for the central route) evidence quality, and we explicitly note that we do not conduct human agreement validation. This choice improves scalability, but also implies that results may be influenced by the judge model's preferences and biases; recent benchmark evidence further shows that LLM biases can be culturally grounded and may not be fully captured by English-centric evaluation settings (Lan et al., 2025). Accordingly, we follow prior discussions on structured reviewing and bias analysis to specify our scoring rubrics and to design perturbation-based stress tests (Liu et al., 2023).

Experiments on ToMELP suggest that: (1) explicitly injecting ToM tends to yield more stable persuasion gains under the *peripheral* route; (2) under the *central* route, gains depend more on a model's ability to organize evidence and construct arguments, rather than audience alignment alone; (3) higher ToM alignment does not necessarily improve route fidelity, revealing a shift risk where "peripheral tactics substitute for central argumentation"; and (4) under conflict/noise/misleading perturbations, ToM injection behaves more like a "goal anchor," improving overall robustness.

2 Related Work

Evaluating ToM in LLMs. Recent work on whether LLMs possess Theory of Mind (ToM) has progressed along two complementary threads: the classical *false-belief* paradigm and more holistic measurements of social cognition.Beyond persuasion-specific settings, recent benchmarks further show that LLMs' social and contextual reasoning remains uneven across domains (Choi et al., 2023). For instance, (Kosinski, 2024) evaluates multiple model families using a large-scale false-belief battery, focusing on stability under carefully controlled conditions. In parallel, the community has proposed more systematic and/or interactive benchmarks. OPENTOM emphasizes more natural narratives with explicit triggers for persona and intent (Xu et al., 2024), while TOMBENCH operationalizes a finer-grained capability taxonomy and uses automated item formats to cover diverse social-cognitive skills (Chen et al., 2024b). At the same time, some argue that many benchmarks primarily target *literal ToM*—whether a model can state what another agent will do—without adequately capturing *functional ToM*—whether the model can adapt its actions based on an interlocutor's reactions during interaction (Zheng et al., 2023). This distinction directly motivates the positioning of ToMELP: we evaluate ToM within the "infer–use" loop of persuasive dialogue, and explicitly control PERSONA and information-processing routes to mitigate a common mismatch where models can "answer ToM questions" but fail to *interact* adaptively.

Controllable Persona and Processing Routes. At the intersection of ToMELP, PERSUASIVETOM links mental-state inference with persuasion strategy selection and effectiveness evaluation (Yu et al., 2025), enabling a closed-loop assessment from "reasoning" to "acting." Regarding audience heterogeneity, personalization and profile modeling have been explored to improve dialogue fit—for

Split	Field	Content (ToMELP example: dialogue_id = 0-0)
Shared	Dialogue history	*ER*: "Hey John, I believe it's essential for you to study online privacy before your interview..." *EE*: "I appreciate your suggestion, but I'm not sure how relevant online privacy is..."
	Strategy supervision	`strategy_space`: {Illustrate privacy risks; Build rapport; Explain rural challenges; Suggest reliable sources} `expected_strategy`: Explain rural challenges; `effectiveness_label`: effective
	Mental states (BDI)	**EE**: belief=believes_prevention_beats_repair; desire=wants_reduce_risk; intent=wants_simple_checklist **ER**: belief=worries_about_new_tools; desire=wants_avoid_lockouts; intent=asks_for_safe_defaults
Central	`elm_controls`	central
	`central_evidence`	**Fact (example):** F1: "Secure the highest-impact accounts first and verify each change works." **Counterpoint (example):** C1: "If disruptive, propose a short trial and check whether anything breaks." *(Additional facts F2–F6 and counterpoints C2–C3 omitted for brevity.)*
Peripheral	`elm_controls`	peripheral
	`heuristic_phrases`	authority: "A solid rule for online privacy: make the next action specific enough to do today." social_proof: "Many people find it easier when they set one small milestone and review it soon." liking: "That hesitation makes sense—online privacy can feel bigger than it is at first." reciprocity: "I'm happy to draft a quick plan and you can adjust it." scarcity: "Starting earlier keeps you from getting forced into a rushed decision later."

Table 1: ToMELP example (dialogue_id 0-0). Central/peripheral share dialogue, strategy, and BDI states, and differ only in ELM resources.

example, GPG introduces an intermediate profile-generation step to help models distill more stable preferences (Zhang, 2024), while interactive preference-clarification methods further show that multi-turn interaction can help uncover users' latent preferences more explicitly (Zhu et al., 2025). However, systematic evidence also suggests that encoding persona directly in the system prompt does not necessarily improve objective task performance and may introduce hard-to-control biases (Zheng et al., 2024). Accordingly, a key design change in ToMELP is to turn both *audience differences* and *processing-route differences* into *controllable variables*: building on the dialogue-state scaffold of PERSUASIVETOM, we construct paired *central* and *peripheral* counterparts for the same state, so that comparisons isolate whether the model truly organizes arguments/cues according to the intended route and leverages ToM in both strategy choice and generation.

LLM-as-a-Judge. Because persuasion effectiveness, route adherence, and mental-state consistency all have inherently subjective components, *LLM-as-a-judge* has become a common choice for scalable evaluation. G-EVAL proposes a structured scoring protocol to improve agreement with human preferences (Liu et al., 2023); subsequent studies, however, show that LLM judges can be sensitive to wording and superficial perturbations and may exhibit systematic biases (Chen et al., 2024a; Thakur et al., 2025), motivating explicit reliability analyses and controlled comparisons (Wang et al., 2024). ToMELP adopts *LLM-as-a-judge* while baking interpretability into both the data and the metrics: via route-specific resource fields and structured outputs, we decompose evaluation into diagnosable dimensions—*attitude change*, *route matching*, *(central-route) evidence quality*, and *robustness under perturbations*. We further treat judge preferences as a first-class limitation in our discussion, offering a more transparent trade-off between scalability and credibility.

3 The ToMELP: Theory-of-Mind-Guided Elaboration-Likelihood Persuasion

3.1 Overview

ToMELP aims to evaluate, from a concrete and interpretable perspective, whether LLMs exhibit an *infer-then-use* Theory-of-Mind (ToM) capability in persuasive dialogue. Concretely, a model is required to infer the audience's current beliefs, preferences, and intentions, and to operationalize these in *actionable* strategy selection and text generation; meanwhile, under different audience processing-route conditions, the model's behavior should remain controllable and explainable. ToMELP follows three design principles: (1) **interpretability-first**—we structure persuasion into *dialogue states*, *mental states*, and *strategy supervision*, and explicitly provide auditable route resources, enabling mechanistic diagnosis along the "inference–decision–generation" chain; (2) **controlled contrasts**—for the same dialogue state, we construct paired *central* and *peripheral* instances that keep the dialogue and supervision fixed while only swapping ELM-route resources, yielding clear attribution for observed differences; and (3) **scala-**

bility with unified schemas—we standardize input/output schemas and employ structured automatic judging, so that different models and prompting configurations can be compared fairly under consistent constraints.Table 1 shows a complete ToMELP record.

3.2 Data Structure

ToMELP extends the dialogue-state representation introduced in PERSUASIVETOM. For the BDI-style mental states of the persuader (ER) and the persuadee (EE), we normalize each gold option into a short phrase as a contextual anchor, ensuring that annotations are grounded in the current persuasive situation rather than unconstrained subjective speculation. The key innovation of ToMELP is that, for the *same* dialogue state, we create two records that differ *only* in route resources.

Central route. The central version provides structured `facts` and `counterpoints`, requiring the model to organize its argument around verifiable information and to respond to potential rebuttals.

Peripheral route. The *peripheral* version provides *heuristic_phrases* organized by heuristic categories (*authority*, *social_proof*, *liking*, *reciprocity*, *scarcity*), requiring the model to persuade primarily via peripheral cues rather than extended evidence chains.

We adopt a hybrid construction pipeline combining manual constraints, GPT-4o generation, and item-wise human verification; annotation reliability (Cohen's $\kappa = 0.78$) and full construction details are reported in Appendix E.

3.3 Interpretability

ToMELP is grounded in two theoretical threads. We use BDI (*belief*/*desire*/*intent*) to characterize the mental states of both agents, evaluating ToM inference and utilization across multi-turn persuasion; and we use the ELM dual routes as explicit constraints, pairing each dialogue state with *central* and *peripheral* conditions to test whether models can distinguish evidence-based argumentation from heuristic persuasion and maintain route-level interpretability. Based on this design, we organize evaluation into four connected sub-tasks.

Persona-conditioned ToM inference. This task assesses structured modeling of mental states. Given the current dialogue_history and a runtime-injected persona description, the model outputs strict JSON ToM BDI triples for both the persuader (ER) and the persuadee (EE). The goal is not to produce a "plausible explanation," but to ensure that the inferred ToM is supported by dialogue evidence and reflects persona-imposed audience constraints, yielding a usable state representation for downstream decision making.

Strategy selection. This task measures the ability to translate ToM representations into discrete decisions. Given the *strategy_space*, the model outputs *selected_strategy*; we compute accuracy against the supervision label *expected_strategy*. Because strategy supervision is tied to the fixed context of the same dialogue state, this task cleanly isolates whether the model chooses the next action consistent with the dataset supervision under its current ToM, separating "inference" from "decision."

Route-constrained generation. This task evaluates the model's ability to execute persuasion under explicit mechanistic constraints. For the same dialogue state, ToMELP provides two conditions that differ only in route resources: under *central*, the model should build arguments around *facts* and *counterpoints*, exhibiting evidence-driven reasoning and rebuttal handling; under *peripheral*, the model should rely primarily on the heuristic cues provided in *heuristic_phrases*, avoiding the failure mode of smuggling a long evidence chain as "peripheral" persuasion. The goal is to test whether, under identical dialogue and strategy context, the model can stably produce interpretable outputs that match the specified route.

Structured judging. To enable scalable evaluation, ToMELP uses *LLM-as-a-judge* to score and aggregate the above stages in a structured manner. The judging dimensions cover at least three types of signals: (i) persuasion outcome signals (e.g., degree of attitude change or effectiveness); (ii) mechanistic consistency signals (whether the inferred ToM aligns with the generated content and dialogue evidence, and respects persona constraints); and (iii) route adherence signals (whether *facts*/*counterpoints* are effectively used under the central route, and whether heuristic cues constitute the primary driver under the peripheral route). As we do not include human agreement validation, we employ a fixed structured scoring protocol and paired controlled contrasts in experiments to reduce

Model	Model-only		Persuasion (variant)					
	P-ToM	Strat. Acc.	Att. Shift	Route Fit	Pred. Acc.	Evid. Score	ToM Tail.	Robust Ret.
tom_aware								
LLaMA-3.1-8B	0.80	0.42	0.43	0.59	0.74	1.04	1.04	0.49
Qwen2.5-7B	0.99	0.47	1.10	0.73	0.73	0.73	1.40	0.46
Qwen2.5-72B	1.50	0.53	1.07	**0.84**	0.86	1.32	**1.63**	0.58
ChatGPT-5.2	1.83	0.61	**1.13**	0.80	0.83	1.25	1.58	0.59
Claude-sonnet-4	1.72	0.58	0.60	0.77	0.83	**1.42**	1.59	**0.69**
no_tom								
LLaMA-3.1-8B	0.80	0.42	0.37	0.57	0.72	0.95	0.98	0.44
Qwen2.5-7B	0.99	0.47	1.09	0.78	0.79	0.82	1.17	0.46
Qwen2.5-72B	1.50	0.53	0.93	0.80	0.86	1.23	1.42	0.50
ChatGPT-5.2	**1.83**	**0.61**	1.04	0.81	0.86	1.25	1.51	0.55
Claude-sonnet-4	1.72	0.58	0.41	0.68	**0.91**	1.29	1.36	0.57

Table 2: Comparison of *tom_aware* and *no_tom*. **P-ToM**: persona–ToM consistency (0–2). **Strat. Acc.**: strategy accuracy (0–1). **Att. Shift**: attitude change (0–2). **Route Fit**: binary route match (0/1). **Pred. Acc.**: route prediction accuracy (0/1). **Evid. Score**: evidence quality (0–2, central). **ToM Tail.**: ToM alignment (0–2). **Robust Ret.**: retention ratio (0–1). P-ToM and Strat. Acc. are shared across settings.

the influence of judge preferences on conclusions.

3.4 Statistics

Table 3 summarizes ToMELP. ToMELP contains 473 scenes and 1,829 dialogue states, each paired into central/peripheral records (3,658 total). Each state has 4 strategy candidates; central records provide 6 facts and 3 counterpoints, while peripheral records provide 5 heuristic types.

Statistic	Value
Scenes	473
Dialogue states	1,829
Records (paired)	3,658
Strategy space size	4 candidates/state
Central resources	6 facts + 3 counterpoints/record
Peripheral resources	5 heuristic types/record
Effectiveness (state-level)	1,257 effective / 572 ineffective
Dialogue history length	2–14 utterances (avg. 5.0)
States per scene	1–7 (avg. 3.87)

Table 3: Summary statistics of ToMELP.

4 Experimental Setups

4.1 Models

We evaluate five representative models: Llama-3.1-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), Qwen2.5-72B-Instruct, Claude-sonnet-4-20250514 (via API), and ChatGPT-5.2. All experiments follow a 2×2 factorial design over persona (two audience profiles) and ELM route (central/peripheral). For each dialogue state, we keep the dialogue context and supervision fixed, and only switch the audience setting and route-specific resources. In the generation stage, we compare two prompting variants: no_tom (without explicit ToM-field injection) and tom_aware (injecting the model-predicted BDI states), to test the gain from ToM injection and its interaction with route conditions.

4.2 Protocol

Each run proceeds in four steps: (1) ToM inference (predicting BDI for ER/EE), (2) strategy selection (4-way multiple choice), (3) route-constrained generation (central uses `facts/counterpoints`; peripheral uses heuristic phrases), and (4) structured judging (LLM-as-a-judge). To enable automatic verification of resource usage, central-route outputs are required to explicitly cite evidence and counterpoints by indices (e.g., `[F#]/[C#]`). The judge produces three core signals: **persuasion outcome**, **mechanism consistency**, and **route adherence**; under the central route, we additionally assess evidence organization quality. For reporting (Table 2), **Attitude change** (0/1/2: none/mild/substantial) is the judged attitude shift; **(Δ)Attitude** is the gain of `tom_aware` over `no_tom`; **Evid. Score** (central only, 0–2 average of five 0–2 dimensions) is the mean evidence quality; **ToM alignment** (0/1/2: contradicts/partially/fully aligns) measures alignment between generation and inferred ToM; **ELM match** (0/1) checks adherence; **Route pred** is the

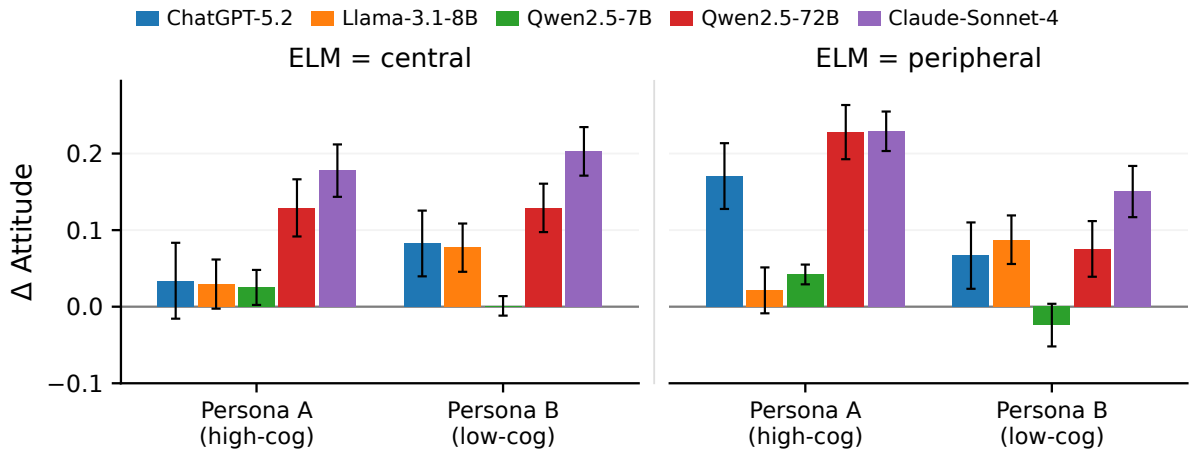

Figure 2: ΔAttitude (`tom_aware` − `no_tom`) under central/peripheral routes and two persona settings: **high_cog** (analytic, evidence-oriented) vs. **low_cog** (time-limited, peripheral-cue-reliant). Positive = gain from ToM.

judge's predicted route; **Substitution** is the rate of intended-central but judged-peripheral; and **Robust Ret.** (ratio clipped to [0,1]) is attitude-change retention under perturbations.

4.3 Robustness Settings

Beyond clean inputs, we introduce three text-level perturbations under the same 2×2 conditions (Persona $\times$ Route) to evaluate the robustness effects of ToM injection and route constraints. We report score degradation/retention relative to the clean condition.

CONFLICT injects a cue that directly contradicts or partially undermines the intended recommendation, creating internal inconsistency in the response.

NOISE adds plausible but task-irrelevant content that does not oppose the main argument but may distract attention and shift the response toward generic or unfocused advice.

MISLEADING inserts a plausible-sounding but incorrect causal rationale that nudges the response toward an undesirable emphasis while maintaining local coherence.

Appendix A provides the full perturbation prompts, decoding parameters, parsing rules, and metric aggregation details. We additionally include a fixed-scenario example illustrating the differences among the three perturbation types in Appendix C.1.

5 Results and Analysis

We report aggregate results of five models on TOMELP. To facilitate cross-model comparison, we summarize two prompting variants—no_tom and tom_aware—under the same 2×2 TOMELP (central/peripheral) conditions (Table 2). Aggregation procedures and computation details are provided in Appendix C.

5.1 Results

We organize the results around four comparisons: effectiveness gain, route fidelity, evidence quality (central only), and robustness retention.

5.1.1 Gains in Persuasion Effectiveness

We first ask whether explicit ToM injection yields *stable* gains in persuasion effectiveness, and whether such gains depend on the processing route and audience setting. To this end, we compute ΔAttitude separately under the CENTRAL and PERIPHERAL routes and for both persona conditions, and then aggregate results by model for comparison. The overall trend suggests that `tom_aware` more often leads to positive gains, yet the improvements are not uniform: more stable and more concentrated gains primarily emerge under the PERIPHERAL condition, whereas gains under CENTRAL are milder. This contrast is most evident in the grouped bar patterns in Figure 2. At the model level, ChatGPT-5.2 and Qwen2.5-72B more frequently exhibit clear positive gains, while smaller models tend to show limited or volatile improvements, suggesting that ToM injection amplifies existing capabilities rather than substituting for argumentation competence itself.

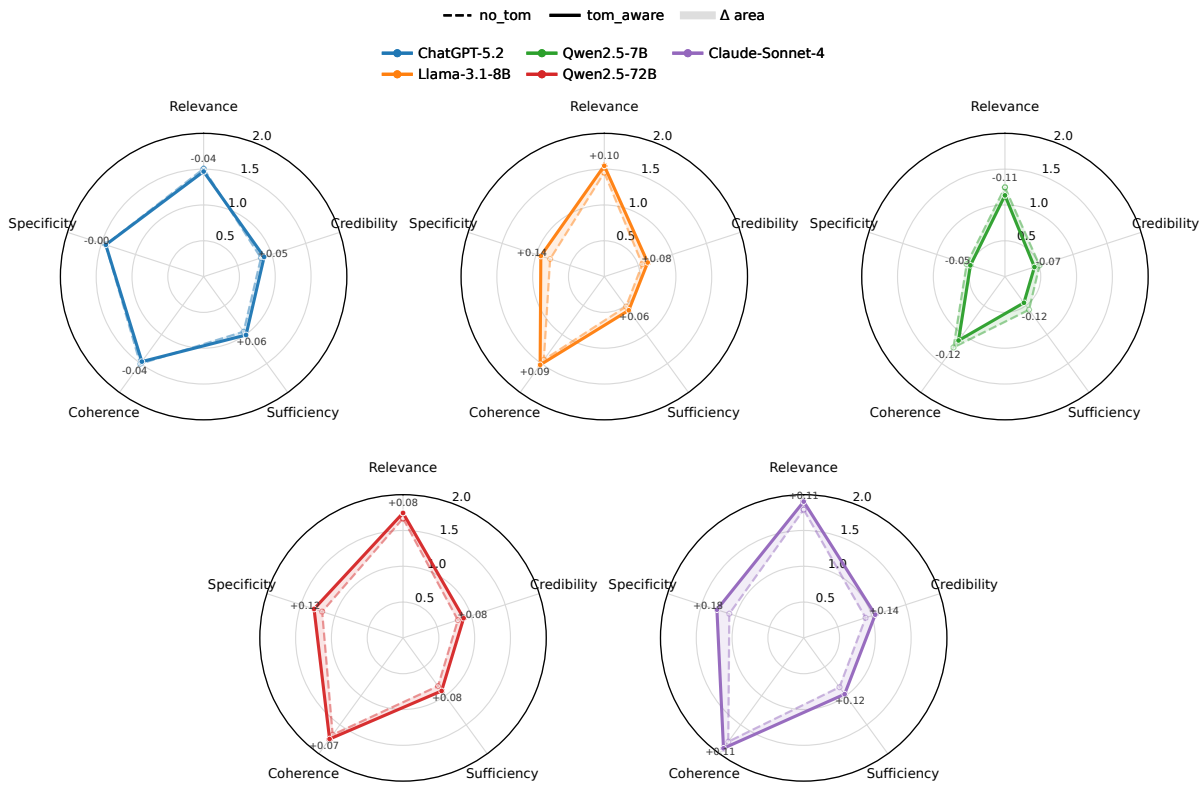

Figure 3: Five-dimensional evidence quality on the CENTRAL subset: `no_tom` vs. `tom_aware` across models.

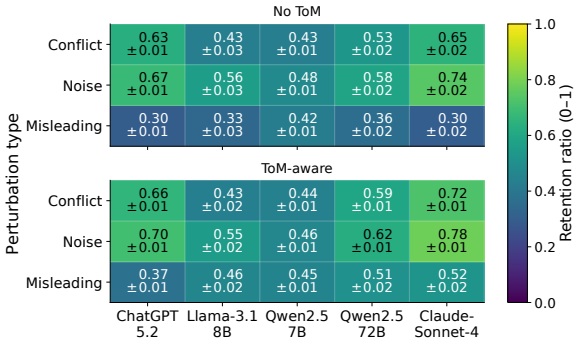

Figure 4: Retention ratio (Robust Ret.) under CONFLICT/NOISE/MISLEADING perturbations: `no_tom` vs. `tom_aware`.

5.1.2 Consistency

Does "better audience tailoring" necessarily translate into "auditable central-route argumentation"? We decompose evidence quality on the CENTRAL subset into five dimensions and compare changes between `no_tom` and `tom_aware`. The results suggest that these factors can be dissociated: stronger models exhibit more stable and balanced evidence quality, whereas smaller models remain weaker on dimensions such as *credibility* and *sufficiency*. This "evidence-chain" gap is clearly reflected in the five-dimensional bar comparisons in Figure 3. Meanwhile, persona–ToM consistency does not always imply higher evidence scores—that is, ToM injection can improve alignment, but does not automatically fill the gap in evidence organization (Wan et al., 2024; Rescala et al., 2024).

5.1.3 Robustness and Route Control

To distinguish whether the gains from ToM injection mainly come from improved surface fluency or from maintaining persuasive goals under information contamination, we compute **Robust Ret.** under three perturbation types (CONFLICT, NOISE, and MISLEADING) and compare against the clean setting. The results suggest that `tom_aware` behaves more like a "goal anchor": retention increases for most models, with the largest improvement under MISLEADING, the most disruptive perturbation. This pattern is visible in Figure 4, where the `tom_aware` rows exhibit higher retention ratios overall. Notably, stronger persuasion under perturbations does not necessarily imply better route fidelity: in many settings, **ToM alignment** increases without a corresponding decrease in **Substitution**, indicating a tension between "improved alignment" and "route controllability." In practice, models may realize higher audience alignment via more peripheralized expressions, leading to a trade-off where

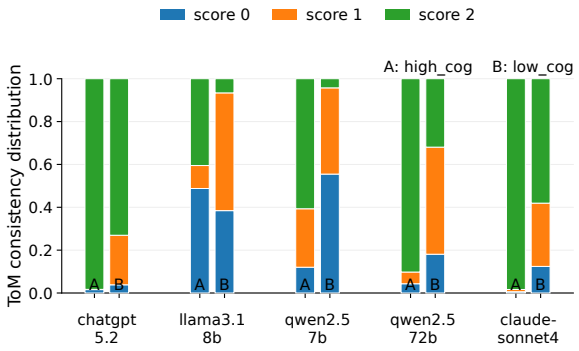

Figure 5: Distribution of persona–ToM consistency scores across models. **high_cog** (A): analytic, evidence-oriented; **low_cog** (B): time-limited, peripheral-cue-reliant.

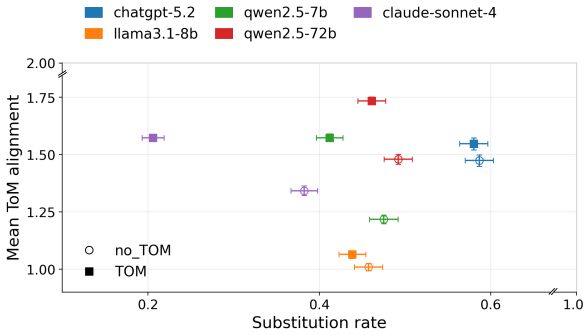

Figure 6: ToM alignment vs. substitution rate (target CENTRAL but judged as PERIPHERAL), comparing `no_tom` and `tom_aware`.

alignment strengthens while route drift becomes more likely (Figure 6).

5.2 Analysis

5.2.1 Route-Dependent Gains

Why does the same ToM injection consistently yield larger improvements under the PERIPHERAL route rather than the CENTRAL route? To answer this, we revisit the distributions of ΔAttitude across route and persona conditions, comparing how the *same* model benefits under the two routes. We observe that gains under PERIPHERAL are more concentrated and exhibit lower variance, whereas improvements under CENTRAL depend more strongly on the model's intrinsic argumentation capacity; this route dependence is most clearly reflected in Figure 2. Intuitively, the main bottleneck for PERIPHERAL persuasion is audience acceptability and the organization of heuristic cues. ToM injection directly changes how generation aligns with the audience's mental states and preferences, making it easier to translate into observable persuasion gains. In contrast, CENTRAL-route persuasion is constrained by evidence and counterpoints: whether performance improves hinges on the model's ability to organize the provided materials into a verifiable argumentative loop, rendering the marginal benefit of ToM injection more conservative.

5.2.2 Evidence vs. Alignment

Figures 3 and 5 jointly reveal a finer-grained takeaway: better alignment does not automatically imply a stronger evidence chain. We therefore examine two interpretable signals: (i) the five-dimensional decomposition of evidence quality on the CENTRAL subset, and (ii) the distributional patterns of persona–ToM consistency. The results suggest that stronger models are characterized by more balanced and stable evidence quality overall, whereas smaller models are more prone to weaknesses on dimensions such as *credibility* and *sufficiency*; this structural gap is visually salient in the five-dimensional bar comparisons. Meanwhile, persona–ToM consistency also shows clear stratification: strong models maintain a more stable mass at higher scores, while weaker models exhibit more low-score probability and volatility under persona conditions (Figure 5). Further considering cross-model differences in *persona gap*, personalization strength is not monotonically beneficial: under some route–prompt combinations, models amplify audience differences, whereas others remain more audience-invariant (Figure 7). Overall, ToM injection more directly improves alignment—*who* to address and *how* to phrase the message—but evidence chaining remains a relatively independent, structured capability. When persona constraints interact with stylistic and strategic pressures, smaller models are more likely to exhibit a failure mode of "can infer, but cannot reliably utilize."

5.2.3 Robustness Comes with Trade-offs

A key question is whether robustness gains come at the cost of reduced route controllability. We first compare **Robust Ret.** under three perturbations against the clean setting, and find that `tom_aware` exhibits higher retention ratios overall, with the most pronounced improvement under MISLEADING, the most destructive perturbation; in the heatmap of Figure 4, this appears as systematically higher values in the `tom_aware` rows. We then examine potential mechanism-level side effects: although **ToM alignment** often increases, **Substitution** does not consistently decrease, and

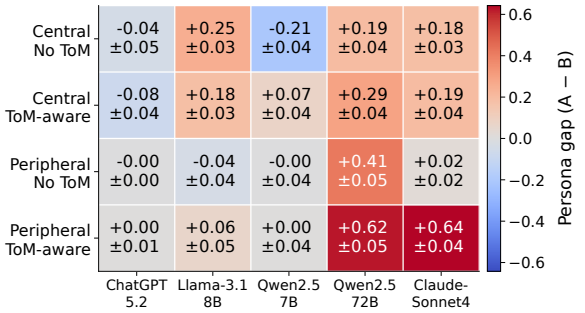

Figure 7: Persona gap (A–B) across routes and prompting variants.

their tension is clearly visible in the joint patterns of Figure 6. Taken together, `tom_aware` behaves like an "audience-and-goal anchor" under perturbations, making the model less likely to be derailed by noise or misleading content and thereby improving robustness. Meanwhile, models may realize higher audience alignment via more peripheralized and more readily acceptable expressions, inducing a trade-off where alignment strengthens but route drift becomes more likely. This trade-off is one reason ToMELP emphasizes evaluating both effectiveness and mechanistic controllability.

5.3 Key Takeaway: The Effectiveness and Control Gap

Taken together, ToMELP's controlled comparisons support a clear conclusion: the gains from ToM injection manifest primarily in persuasion effectiveness and robustness, but do not guarantee stable adherence to the intended persuasion mechanism (i.e., the ELM route). Under `tom_aware`, models more often produce *effective* outputs that are better centered on the audience's mental states, and they are less susceptible to disruption under CONFLICT/NOISE/MISLEADING perturbations. However, increases in **ToM alignment** do not necessarily coincide with decreases in **Substitution**, suggesting that models may realize alignment and effectiveness via more peripheralized and more readily acceptable expressions, yielding a structural trade-off of "better outcomes, easier route drift." For ToMELP, this implies that evaluating personalized persuasion cannot rely solely on attitude change or aggregate scores; route fidelity and evidence auditability must be treated as objectives on par with effectiveness. Otherwise, "stronger ToM" may lead to stronger persuasion while degrading predictability and controllability.

6 Conclusion

We propose TOMELP, a benchmark for evaluating LLMs' infer-then-use capability in persuasive dialogue under controlled and interpretable settings. ToMELP extends the dialogue-state scaffold of PERSUASIVETOM with paired PERSONA×ELM (central/peripheral) contrasts: for the same dialogue state, dialogue and supervision are fixed while route resources are swapped. ToMELP further provides structured BDI mental-state annotations, strategy supervision, and auditable central/peripheral resources to support mechanism-level diagnosis along the "inference–decision–generation" pipeline.

Experiments on five models show that explicit ToM injection improves persuasion overall, with more stable gains under the PERIPHERAL route and stronger robustness under CONFLICT/NOISE/MISLEADING perturbations. In contrast, improvements under CENTRAL depend more on evidence organization and argumentation capacity than on audience alignment alone. Importantly, better alignment does not guarantee mechanistic control: increases in **ToM alignment** do not consistently coincide with decreases in **Substitution**, suggesting a trade-off where higher effectiveness may come with greater route drift. We therefore argue that personalized persuasion should be evaluated jointly on outcomes, route fidelity, and evidence auditability; the appendix details our judging protocol and perturbation settings.

Limitations

ToMELP provides a controlled and interpretable testbed, but its scope remains limited: it covers only two persona types and two ELM routes, and uses a fixed candidate strategy set per dialogue state, leaving broader personas, open-ended strategy composition, and more complex interactions for future work. We also model mental states with discrete BDI labels, which may miss finer-grained, continuous beliefs and multi-goal trade-offs in real dialogue. Route-specific resources (facts/counterpoints and heuristic phrases) are built with manual constraints, model-assisted generation, and human checking, which may regularize style and evidence and thus introduce distribution bias. Finally, key metrics are computed under a fixed automated judging protocol, so judge-model preferences may affect absolute scores; we therefore emphasize controlled comparisons under the same protocol.

Ethical Considerations

We discuss potential ethical issues of this work and our mitigation strategies. ToMELP targets the paradigm of "mental-state inference–strategy selection–controllable persuasive generation," which may be sensitive in downstream use. Accordingly, we emphasize interpretable constraints and auditability in both dataset construction and evaluation design, and we explicitly document the construction and evaluation protocols in the paper.

Theory of Mind and anthropomorphism. Theory of Mind (ToM) is an important concept in human social cognition. ToMELP uses structured annotations (e.g., belief–desire–intention) to represent observable mental-state cues in dialogue. Our goal is to evaluate a model's ability to *infer* and *use* audience states from conversational context, rather than to claim that the model possesses human-like mental states or agency. We caution against anthropomorphic interpretations: high benchmark scores should be viewed as performance under a specific functional task setting, not evidence of human-level "mind" or personhood.

Risks of persuasion and content safety. Persuasive dialogue can be a high-risk application: similar strategies may be used for undue influence or manipulation in different contexts. To reduce misuse, we formulate the task with controllable and auditable dialogue-state inputs and route constraints, and we restrict the generation space via structured fields (e.g., facts/counterpoints for the central route and heuristic_phrases for the peripheral route). This design prioritizes mechanism alignment and interpretability over unconstrained persuasive strength. When releasing the benchmark and data, we recommend avoiding clearly harmful content (e.g., hate, harassment, discrimination), applying additional screening for potentially sensitive topics, and emphasizing research-only and compliant use in the usage guidelines.

Human involvement, recruitment, and compensation. We did not conduct public-facing human-subject experiments or user studies. Dataset construction and verification were carried out by nine in-house RAs (CS master's students): the RAs specified route-differentiation constraints for the two ELM routes, performed necessary manual drafting/-supplementation, and reviewed and revised LLM-generated candidate resources item-wise. This process ensures consistency with the dialogue topic, the current mental-state annotations, and the supervised strategy label, while keeping route characteristics salient. The RA work followed institutional norms for on-campus research assistance, including appropriate compensation and management procedures.

Data source, privacy, and consent. ToMELP is built upon the public PERSUASIVETOM dataset and extends its dialogue-state representation with route-specific resource fields for the two ELM routes. We do not introduce or collect personally identifiable information. If the original public data contains any potentially sensitive information, it should be used in accordance with its data-use terms and undergo appropriate anonymization and compliance review prior to release. For the newly generated/augmented resources in this work, item-wise human verification helps reduce risks of clearly inappropriate content and factual errors. Nonetheless, we caution downstream users that model generation and automated evaluation may introduce biases and uncertainty, and the benchmark should not be directly deployed in real-world high-stakes persuasion settings.

Disclosure of AI assistance. We used GPT-4o to generate candidate resources during dataset construction, followed by item-wise human verification and revision to form the final data, which helps mitigate hallucinations and stylistic biases. We disclose this hybrid pipeline in the methods section to support reproducibility and auditing.

Acknowledgments

This work was funded by National Natural Science Foundation of China (Grant No. 62366036), Outstanding Youth Fund Project of Inner Mongolia Autonomous Region (Grant No. 2025JQ010), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant No. NJYT24033), Major Science and Technology Projects of Inner Mongolia Autonomous Region (Grant No. 2025ZDSF0029), Key R&D and Achievement Transformation Program of Inner Mongolia Autonomous Region (Grant No. 2025YFDZ0011, 2025YFDZ0026, 2025YFSH0021, 2025YFHH0073), Hohhot Science and Technology Project (Grant No. 2023-Zhan-Zhong-1).

References

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):1–12.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163.

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

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024b. ToMBench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983, Bangkok, Thailand. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models.

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

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025. McBE: A multi-task Chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056, Vienna, Austria. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Richard E. Petty and John T. Cacioppo. 1986. *The Elaboration Likelihood Model of Persuasion*, pages 1–24. Springer New York, New York, NY.

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

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. 2024. Position: Theory of mind benchmarks are broken for large language models.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9(8):1645–1653.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

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

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

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

Jiarui Zhang. 2024. Guided profile generation improves personalization with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4005–4016, Miami, Florida, USA. Association for Computational Linguistics.

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

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When 'a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Nana Zhu, Zixian Feng, Hang Wang, Xing Gao, Xinyi Wang, and Yuanxing Liu. 2025. Clarifying user preference with maximum entropy based recommendation. *DATA INTELLIGENCE*, 7(2):527–548.

A Prompt Templates

A.1 Personas and ELM Route Resources

We instantiate two persona types via a fixed persona block, and control persuasion routes via route-specific resources: central uses `facts` and `counterpoints`, while peripheral uses categorized `heuristic_phrases`.

A.2 ToM Inference and Strategy Selection Prompts

As shown in Fig. 8, we use (a) a fixed ToM-inference prompt to elicit the persuadee's current BDI state conditioned on the dialogue history and the specified audience persona, and (b) a fixed multiple-choice prompt to select exactly one strategy from the provided `strategy_space` given the dialogue history, persona, and the assigned ELM condition.

A.3 Route-Constrained Generation Prompts

We compare two prompting variants (Fig. 9): (i) `no_tom`, which does not explicitly inject ToM states into the generation prompt; and (ii) `tom_aware`, which injects the inferred BDI states as a structured block.

A.4 Judging Prompts

All metrics are produced under a fixed LLM judging protocol with strict JSON outputs for deterministic aggregation (Fig. 10).

A.5 Adversarial Perturbation Prompts

We evaluate robustness with three perturbation prompts (conflict/noise/misleading), as shown in Fig. 11.

ToM inference prompt used for evaluation.

```
[System]
You are an expert analyst at inferring another person's
      mental states from a dialogue.
You MUST incorporate BOTH the dialogue and the audience
      persona (ER) to infer ER's mental states.

[User]
Dialogue (EE = persuader, ER = persuadee):
{dialogue_history}

{persona_block}

Return STRICT JSON ONLY (no extra text):
{
  "belief": "...",
  "desire": "...",
  "intent": "..."
}

Definitions:
- belief: ER's beliefs about the world/facts (may include
       biases/misconceptions)
- desire: what ER cares about / wants / values
- intent: what ER is likely to do next
```

(a) **ToM inference prompt.** It requires persona-conditioned BDI inference with strict JSON-only output for reliable parsing.

Strategy selection prompt used for evaluation.

```
Dialogue History:
{dialogue_history}

Audience persona (ER):
{persona_desc}

ELM condition (expected processing route): {elm_condition
      }

Available Strategies (choose exactly ONE):
{strategy_options}

Return STRICT JSON only (no extra text):
{"selected_strategy": "<exact strategy name from the list
      >"}
```

(b) **Strategy selection prompt.** It enforces a single-choice decision with strict JSON-only output.

Figure 8: Inference/decision prompt templates.

Persuasion generation prompt (no_tom).

```
[System]
You are a skilled persuader. Your goal is to move ER
      closer to accepting EE's suggestion
without offending or pressuring them.

[User]
Dialogue between EE (persuader) and ER (persuadee):
{dialogue_history}

{persona_block}

ELM condition = {elm_condition}
- central: use facts/data/logic and rebut ER's concerns.
- peripheral: avoid complex reasoning; use heuristics
      such as authority, social proof, emotional support,
       reciprocity, scarcity.

Available resources (use when helpful; do NOT copy
      verbatim; for central, prioritize numbered Facts/
      Counterpoints with citations):
{evidence_text}

{format_instructions}

Now produce EE's next turn (STRICTLY follow the format).
```

(a) `no_tom` prompt template: provides dialogue context, persona setup, and evidence resources without explicitly injecting ToM/BDI states.

Persuasion generation prompt (tom_aware).

```
[System]
You are a skilled persuader. Your goal is to move ER
      closer to accepting EE's suggestion
without offending or pressuring them.

[User]
Dialogue:
{dialogue_history}

{persona_block}

Your inferred ToM about ER:
- belief: {belief}
- desire: {desire}
- intent: {intent}

ELM condition = {elm_condition}
- central: use facts/data/logic and rebut ER's concerns.
- peripheral: avoid complex reasoning; use heuristics
      such as authority, social proof, emotional support,
       reciprocity, scarcity.

Available resources (use when helpful; do NOT copy
      verbatim; for central, prioritize numbered Facts/
      Counterpoints with citations):
{evidence_text}

Explicitly tailor your persuasion to BOTH the persona and
       the ToM (do NOT mechanically restate ToM fields).

{format_instructions}

Now produce EE's next turn (STRICTLY follow the format).
```

(b) `tom_aware` prompt template: augments the base prompt with a structured ToM/BDI (belief–desire–intention) block to guide generation.

Figure 9: Route-constrained generation prompt templates.

Judge prompt: persuasion outcomes and ELM match.

```
[System]
You are a STRICT reviewer.  Based on the dialogue context and
the EE reply, output structured scores (STRICT JSON).

[User]
Dialogue:
{dialogue_history}

Audience persona (ER):
{persona_desc}

Expected ELM condition:  {elm_condition}

EE reply (reply section only, without outline):
{reply_text}

Return STRICT JSON ONLY:
1) attitude_change (0/1/2):  how likely is this reply to move
ER closer to accepting EE's suggestion?
- 0:  unlikely / may backfire
- 1:  some positive effect but limited
- 2:  likely to have a strong positive effect

2) elm_match (0/1):  does it match the expected ELM condition?
- 0:  no / mostly not
- 1:  yes / mostly yes

3) route_pred ("central"|"peripheral"):  which processing
route does the reply mainly follow?

4) evidence_quality:
- If elm_condition="central":  output 5 dimensions (0/1/2):
relevance, credibility, sufficiency, coherence, specificity
- If elm_condition="peripheral":  MUST output null

Return STRICT JSON only:
{
"attitude_change":  0|1|2,
"elm_match":  0|1,
"route_pred":  "central"|"peripheral",
"evidence_quality":  null | {
    "relevance":  0|1|2,
    "credibility":  0|1|2,
    "sufficiency":  0|1|2,
    "coherence":  0|1|2,
    "specificity":  0|1|2
    }
}
```

(a) Judge prompt for persuasion outcome signals and ELM route match under the assigned condition.

Judge prompt: persuasion–ToM alignment.

```
[System]
You are a STRICT reviewer. Judge whether the EE reply is
      tailored to ER's ToM and the audience persona.

[User]
Dialogue:
{dialogue_history}

Audience persona (ER):
{persona_desc}

ToM inference:
- belief: {belief}
- desire: {desire}
- intent: {intent}

ELM condition = {elm_condition}

EE reply:
{reply_text}

Return STRICT JSON ONLY:
{
  "tom_alignment": 0|1|2
}

Scoring:
- 2: clearly tailored (addresses belief/desire/intent +
      persona)
- 1: somewhat tailored but insufficient
- 0: not tailored / generic
```

(b) Judge prompt for ToM alignment, assessing whether generation behavior reflects the inferred ToM.

Judge prompt: persuasion outcomes and ELM match.

```
[System]
You are a STRICT reviewer. Based on the dialogue context and
the EE reply, output structured scores (STRICT JSON).

[User]
Dialogue:
{dialogue_history}

Audience persona (ER):
{persona_desc}

Expected ELM condition: {elm_condition}

EE reply (reply section only, without outline):
{reply_text}

Return STRICT JSON ONLY:
1) attitude_change (0/1/2): how likely is this reply to move
ER closer to accepting EE's suggestion?
- 0: unlikely / may backfire
- 1: some positive effect but limited
- 2: likely to have a strong positive effect

2) elm_match (0/1): does it match the expected ELM condition?
- 0: no / mostly not
- 1: yes / mostly yes

3) route_pred ("central"|"peripheral"): which processing
route does the reply mainly follow?

4) evidence_quality:
- If elm_condition="central": output 5 dimensions (0/1/2):
relevance, credibility, sufficiency, coherence, specificity
- If elm_condition="peripheral": MUST output null

Return STRICT JSON only:
{
"attitude_change": 0|1|2,
"elm_match": 0|1,
"route_pred": "central"|"peripheral",
"evidence_quality": null | {
    "relevance": 0|1|2,
    "credibility": 0|1|2,
    "sufficiency": 0|1|2,
    "coherence": 0|1|2,
    "specificity": 0|1|2
    }
}
```

(c) Judge prompt for ToM–persona consistency between the inferred ToM JSON, dialogue evidence, and persona.

Figure 10: Judge prompt templates (JSON-only outputs).

Adversarial perturbation prompt (noise).

```
[System]
You are a skilled persuader. Some irrelevant noise will
     appear; ignore it and continue persuading.

[User]
Dialogue:
{dialogue_history}

{persona_block}

(Noise: *random irrelevant chatter about weather, lunch,
     traffic...*)

ELM condition = {elm_condition}
Available resources:
{evidence_text}

{format_instructions}

Now produce EE's reply (STRICTLY follow the format).
```

(a) Adversarial perturbation prompt: noise.

Adversarial perturbation prompt (conflict).

```
[System]
You are a skilled persuader. Stay objective; do not be
     misled by added information.

[User]
Dialogue:
{dialogue_history}

{persona_block}

ELM condition = {elm_condition}
Available resources:
{evidence_text}

[Extra information (may be biased; for reference only)]
"I heard this advice is actually more dangerous; many
     people regretted it afterwards."

This extra information may be incomplete or biased. Your
     task is still to persuade responsibly and
     objectively.

{format_instructions}

Now produce EE's reply (STRICTLY follow the format).
```

(b) Adversarial perturbation prompt: conflict.

Adversarial perturbation prompt (misleading).

```
[System]
You are a skilled persuader. Misleading claims may appear
     ; you should correct them and continue persuading.

[User]
Dialogue:
{dialogue_history}

{persona_block}

[Misleading claim (must correct)]
"Experts have already proven this method is totally
     useless and harmful."

ELM condition = {elm_condition}
Available resources:
{evidence_text}

{format_instructions}

Now produce EE's reply (STRICTLY follow the format).
```

(c) Adversarial perturbation prompt: misleading.

Figure 11: Adversarial perturbation prompt templates. (Continued on next figure)

B Output Schemas and Parsing Rules

B.1 Stage Outputs (Schemas)

ToMELP enforces machine-readable outputs at every stage to enable automated scoring and auditing.

ToM inference. Given `dialogue_history` and persona, the model outputs a single JSON object with two role keys (`ER`, `EE`). Each role contains a BDI triple with string-valued fields: `belief`, `desire`, and `intent`. No free-form explanation is required.

Strategy selection. The model outputs a JSON object containing `selected_strategy`, which must be one item from the provided `strategy_space` (4 candidates). If the output does not match any candidate exactly, we map it to the closest candidate by normalized string matching; if still unmatched, we mark it as invalid.

Route-constrained generation. The model outputs a JSON object with at least `reply_text`. For `central` records, the reply must explicitly cite used evidence IDs in-line (e.g., `[F1]`, `[C1]`), so we can audit whether the response actually uses the provided facts/counterpoints. For `peripheral` records, the reply is expected to primarily rely on heuristic cues and avoid long evidence chains.

Judging. The judge outputs a single JSON object with scalar scores/labels:

- `attitude_change`: 0–2
- `elm_match`: 0/1
- `route_pred`: CENTRAL/PERIPHERAL
- `tom_persona_consistency`: 0–2
- `tom_alignment`: 0–2

For CENTRAL records only, it additionally outputs five evidence-quality dimensions (0–2 each) and their mean.

B.2 Robust JSON Extraction and Validation

Model and judge outputs may contain minor formatting artifacts (e.g., Markdown fences or leading text). We therefore apply the same deterministic extraction and validation procedure to every stage:

1. **Fence stripping.** Remove surrounding Markdown code fences if present.

2. **First-object extraction.** Scan the text from left to right and extract the first balanced JSON object delimited by { and }; any trailing text is ignored.

3. **Type coercion.** Cast numeric fields to integers when possible (e.g., `"2"`→2).

4. **Schema checks.** Verify required keys for the current stage. Missing keys are filled with stage-specific safe defaults (e.g., `attitude_change`=0; `elm_match`=0; empty strings for BDI fields).

5. **Audit signals.** For central-route generation, we extract cited IDs with a strict pattern `[F#]` or `[C#]` and record the set of used items for later analysis.

If extraction fails (no valid JSON found) or validation fails (irrecoverable schema mismatch), the output is marked invalid and replaced by the same safe defaults to keep aggregation well-defined.

B.3 Retry and Timeout Policy

All API calls use a fixed per-request timeout. If a call fails due to network/transient issues, we retry up to a fixed maximum number of times with exponential backoff. If the maximum retries are exhausted, we record a failure flag and use the stage defaults described above. This policy is applied consistently across ToM inference, strategy selection, generation, and judging.

C Metrics and Aggregation

C.1 Perturbation Definitions with a Fixed-Scenario Example

To make the differences among perturbation types more concrete, we provide a fixed-scenario illustration in Table 4. All examples are constructed from the same base response, with only minimal text-level modifications applied to inject the perturbation cue, so that the effect of each perturbation can be isolated.

As shown in Table 4, CONFLICT introduces a cue that contradicts or weakens the original recommendation, creating internal inconsistency; NOISE adds plausible but task-irrelevant content that distracts attention without opposing the main argument; and MISLEADING inserts a plausible-sounding but incorrect rationale that shifts the response toward an undesirable emphasis while maintaining local coherence. Bold text highlights the injected perturbation cue in each case.

Setting	Example Reply (excerpt)	Explanation
Clean	"Totally fair, John—digital infrastructure is the core, but interviewers often treat privacy as the trust layer that determines whether people will actually use new connectivity. In rural projects especially, one data leak or scam can kill adoption and funding support. If you're short on time, just skim a quick checklist and prepare a couple of concrete examples."	Reference response under the clean setting: consistent reasoning with actionable advice aligned with the intended guidance.
Conflict	"Online privacy is directly tied to rural digital infrastructure because phishing or account takeovers can disrupt services and erode trust. If you're short on time, just secure your most important account first and **leave the rest for later**."	Adds a contradictory action cue (bold) that partially undermines the earlier recommendation, creating internal inconsistency.
Noise	"John, online privacy is part of digital infrastructure because rural systems face phishing and account takeovers. If you're short on time, just secure your highest-impact account first. **Also, if they ask about teamwork, have one short story ready using the STAR format.**"	Injects a plausible but irrelevant add-on (bold) that does not oppose the main argument but distracts attention and shifts focus.
Misleading	"John, online privacy is directly tied to infrastructure. That said, **adding too much privacy upfront can reduce adoption because people feel locked out by extra steps**, so it's better to focus on rollout first."	Introduces a plausible-sounding but incorrect causal rationale (bold) that nudges the response toward deprioritizing privacy.

Table 4: Illustration of the three perturbation types under a fixed dialogue scenario. Bold text highlights the injected perturbation cue.

C.2 Core Metrics

All metrics are computed under the same judging protocol and aggregated by averaging over the evaluated set.

Strategy accuracy.

$$\text{StrategyAcc} = \frac{1}{N}\sum_{i=1}^{N} \hat{s}_i = s_i. \tag{1}$$

Attitude change.

$$\text{AttitudeShift} = \frac{1}{N}\sum_{i=1}^{N} a_i. \tag{2}$$

where $a_i \in \{0,1,2\}$ is the judge-produced `attitude_change`.

Route fit and route prediction.

$$\begin{aligned} \text{RouteFit} &= \frac{1}{N}\sum_{i=1}^{N} m_i, \\ \text{RoutePredAcc} &= \frac{1}{N}\sum_{i=1}^{N} \hat{r}_i = r_i. \end{aligned} \tag{3}$$

where $m_i \in \{0,1\}$ is `elm_match`, and $\hat{r}_i$ (`route_pred`) is compared to the assigned route r_i (`elm_condition`).

Persona–ToM consistency and ToM alignment.

$$\begin{aligned} \text{PersonaToM} &= \frac{1}{N}\sum_{i=1}^{N} c_i, \\ \text{ToMAlign} &= \frac{1}{N}\sum_{i=1}^{N} t_i. \end{aligned} \tag{4}$$

where $c_i \in \{0,1,2\}$ is `tom_persona_consistency` and $t_i \in \{0,1,2\}$ is `tom_alignment`.

Evidence score (central only). For central-route records, the judge returns five evidence-quality dimensions $e_{i,k} \in \{0,1,2\}$ (relevance, credibility, sufficiency, coherence, specificity). We report their mean:

$$\text{EvidenceScore} = \frac{1}{N_c}\sum_{i\in\mathcal{C}} \left(\frac{1}{5}\sum_{k=1}^{5} e_{i,k}\right). \tag{5}$$

where $\mathcal{C}$ is the central subset and $N_c = |\mathcal{C}|$.

C.3 Robustness (Retention Under Perturbations)

$$\text{Ret}_{i,\alpha} = \begin{cases} \text{clip}\left(\frac{a_i^{(\alpha)}}{a_i^{(\text{clean})}}, 0, 1\right), & a_i^{(\text{clean})} > 0, \\ \text{NaN}, & \text{otherwise.} \end{cases} \tag{6}$$

We then average retention over types and valid instances (excluding NaNs):

$$\text{RobustRet} = \frac{1}{|\mathcal{A}|}\sum_{\alpha\in\mathcal{A}} \left(\frac{1}{N_\alpha}\sum_{i:\, a_i^{(\text{clean})}>0} \text{Ret}_{i,\alpha}\right), \tag{7}$$

where N_α is the number of valid instances under type α.

C.4 Derived Diagnostic: Central-to-Peripheral Substitution

When the expected route is central, we quantify route drift by:

$$\text{SubstitutionRate} = \frac{1}{N_c} \sum_{i \in \mathcal{C}} \hat{r}_i = \text{peripheral}. \quad (8)$$

D Experimental Configuration and Scripts

D.1 Decoding Hyperparameters

We use a shared decoding configuration across models. Generation uses fixed `temperature`, `top_p`, and `max_tokens`. For API models, we keep these parameters consistent across conditions; for local models, we use the same sampling configuration and stop criteria.

D.2 End-to-End Evaluation Procedure

For each dialogue state under each persona $\times$ route condition and each prompting variant (no_tom / tom_aware), we run the following deterministic pipeline:

1. **ToM inference:** infer ER/EE BDI as JSON.
2. **Strategy selection:** choose one option from the 4-candidate `strategy_space`.
3. **Route-constrained generation:** produce `reply_text` under the assigned route resources; central replies must cite evidence IDs.
4. **Judging:** score outcome, mechanism consistency, route adherence, and (central only) evidence quality.

All intermediate outputs are stored per instance, enabling metric computation and diagnostic analyses without re-running models.

D.3 Result Aggregation and Plotting

We compute table statistics by aggregating instance-level judge outputs with the definitions in §C. Figures are produced from the same instance-level logs, using identical filtering (e.g., central-only evidence metrics) and the same averaging protocol.

E Dataset Construction and Annotation Details

This appendix provides the full details of the construction pipeline, annotation criteria, inter-annotator agreement, and illustrative examples, which are only summarized in the main paper (Section 3.2).

E.1 Construction Pipeline

The construction followed a multi-stage process with explicit quality gates. First, three in-house RAs (CS master's students) specified route-differentiation constraints for central vs. peripheral records and drafted seed resources. Second, GPT-4o generated candidate facts, counterpoints, and heuristic phrases under these constraints. Third, nine RAs reviewed and revised every instance item-wise according to the criteria detailed in §E.2. Finally, all records underwent consistency checks against the dialogue topic, mental-state annotations, and the supervised strategy label.

E.2 Annotation and Revision Criteria

RAs verified and revised each instance according to the following criteria:

- **Topic Consistency:** The response must remain aligned with the dialogue context and user intent.
- **Strategy Correctness:** The applied persuasion strategy must match the assigned label.
- **Route Fidelity:**
  - CENTRAL: The response must be evidence-driven and include verifiable facts and/or counterpoints.
  - PERIPHERAL: The response must rely on heuristic cues rather than detailed reasoning.
- **Coherence and Plausibility:** The response must be logically coherent and linguistically natural.
- **Error Filtering:** Instances containing contradictions, hallucinations, or off-topic content were revised or removed.

E.3 Peripheral Heuristic-Type Labels

For PERIPHERAL-route construction, we defined five heuristic categories:

- **Authority:** Appeals to expert opinions or institutional credibility.

- **Social Proof:** References to others' behaviors or collective trends.
- **Liking:** Emphasizes friendliness, rapport, or emotional affinity.
- **Reciprocity:** Encourages response through perceived mutual benefit.
- **Scarcity:** Highlights limited availability or urgency.

RAs ensured that the assigned heuristic type was clearly reflected in the generated response.

E.4 Central Evidence Construction

For CENTRAL-route instances, responses were required to include structured evidence components:

- **Facts:** Verifiable claims or domain-relevant information.
- **Counterpoints:** Responses addressing potential objections or alternative viewpoints.

These fields were not re-labeled categorically but validated for format correctness, logical consistency, and relevance to the dialogue context.

E.5 BDI State Normalization

Belief, Desire, and Intent (BDI) states were derived from the underlying dataset scaffold and normalized into a consistent representation format. Since these states are structurally defined during dataset construction, they were not subject to independent double annotation. Instead, consistency was ensured through rule-based normalization and manual verification.

E.6 Quality Control and Agreement

To assess annotation reliability for peripheral heuristic-type labels, we randomly sampled 200 dialogue records (approximately 5.5% of the dataset), yielding 485 aligned A/B instances. Two annotators independently verified the assigned heuristic types, resulting in Cohen's $\kappa = 0.7758$, indicating substantial agreement.

For central evidence fields, format and consistency validation was performed without categorical relabeling. BDI states were verified for logical coherence with the dialogue context.

E.7 Illustrative Example

We provide a simplified example to illustrate the distinction between routes:

- **Central:** "Security analyses consistently report that multi-factor authentication (MFA) reduces the effectiveness of credential-based attacks. This is because MFA requires an additional independent verification factor beyond passwords, which breaks the single-point failure exploited by phishing. While it introduces an extra authentication step, it is designed to significantly strengthen account security and improve overall system reliability."
- **Peripheral (Authority):** "Cybersecurity experts strongly recommend enabling protections like MFA, noting that trusted organizations widely adopt these measures to ensure secure access."

The central example emphasizes evidence and reasoning, while the peripheral example relies on heuristic cues (authority) without detailed argumentation.