GroupToM-Bench: Benchmarking Group Theory of Mind and Nonlinear Social Emergence in MLLMs

Weidong Tang[1,2], Jierui Li[1], Yueling Hou[1], Zihan Mei[2,3], Can Zhang[1], Xinyan Wan[1], Zhiyuan Liang[2,4], Pengfei Zhou [2†], Yang You [2], Wangbo Zhao [2†],

[1]Xidian University, [2]National University of Singapore,
[3]University of Electronic Science and Technology of China,
[4]University of Science and Technology of China
wdtang0705@gmail.com, zpf4wp@outlook.com, wangbo.zhao96@gmail.com

Abstract

True general intelligence requires not only a model of the physical world but also a social world model: the capacity to infer how individual mental states interact and crystallize into group-level outcomes. Despite notable progress in individual-level Theory of Mind (ToM) reasoning, existing multimodal large language models systematically fail at this: collective behavior emerges non-linearly from social tensions, conformity dynamics, and structural constraints, and cannot be recovered by summing individual intentions. We present **GroupToM-Bench**, the first multimodal benchmark for group-level ToM, built around a causal chain spanning micro-level BDI states (belief, desire, intention), meso-level group tension and structural constraints, and macro-level outcome prediction and mechanistic attribution. To probe this full arc, we develop a seven-level cognitive audit framework. Experiments reveal that frontier models perform significantly below human levels, exposing fundamental blind spots in modeling social structures and nonlinear collective behavior.

1 Introduction

Recent advancements in artificial intelligence have been shaped in large part by the pursuit of world models (Ha and Schmidhuber, 2018; Ding et al., 2026). Current paradigms focus predominantly on the physical world (Wan et al., 2025; Ran et al., 2026): models learn intuitive physics, spatial dynamics, and object permanency to predict how objects interact under mechanical laws. The real world, however, is not purely physical. A genuine general intelligence must also operate within a social world, simulating how agents interact, adapt, and organize under structural social rules rather than mechanical ones.

The cornerstone of such a social world model is Theory of Mind (ToM) (Premack and Woodruff, 1978; Baron-Cohen et al., 1985): the cognitive capacity to infer and reason about the mental states of others. Rather than an isolated skill, ToM is the basic unit from which social understanding is constructed. As multimodal large language models have matured (Yin et al., 2024), ToM evaluations have evolved accordingly. Early work focused on individual-level inference, whether models could read isolated mental states (Wu et al., 2023; Xu et al., 2024; Chen et al., 2024; Gu et al., 2024). More recent benchmarks moved to interactive multi-agent settings (Kim et al., 2023; Li et al., 2023; Bortoletto et al., 2025). Yet most still target local belief tracking or task-specific reasoning, leaving largely unaddressed how private states aggregate into group-level tension, structural constraints, and collective outcomes. Critically, group-level social reasoning cannot be grounded in text alone. Private mental states leak through nonverbal channels: a hesitant micro-expression contradicts verbal agreement; spatial positioning reveals coalition boundaries invisible in dialogue. Evaluating whether models can detect these cross-modal fractures is essential, since real social intelligence depends on integrating what agents say with what they signal.

As the number of interacting agents grows, genuinely social phenomena emerge that individual-level analysis cannot capture. Collective behavior is never a simple linear sum of individual intentions (Granovetter, 1978; Klein and Kozlowski, 2000). Macro-level structures, such as power hierarchies, cultural norms, and information asymmetries, continuously reshape, suppress, or polarize micro-level desires (Noelle-Neumann, 1974), producing non-linear outcomes that no single agent intended. A canonical example is the Abilene Paradox (Harvey, 1974), in which each member of a group privately disagrees with a decision yet publicly endorses it, because each assumes the others are in agreement. Current evaluation paradigms

[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40007–40031

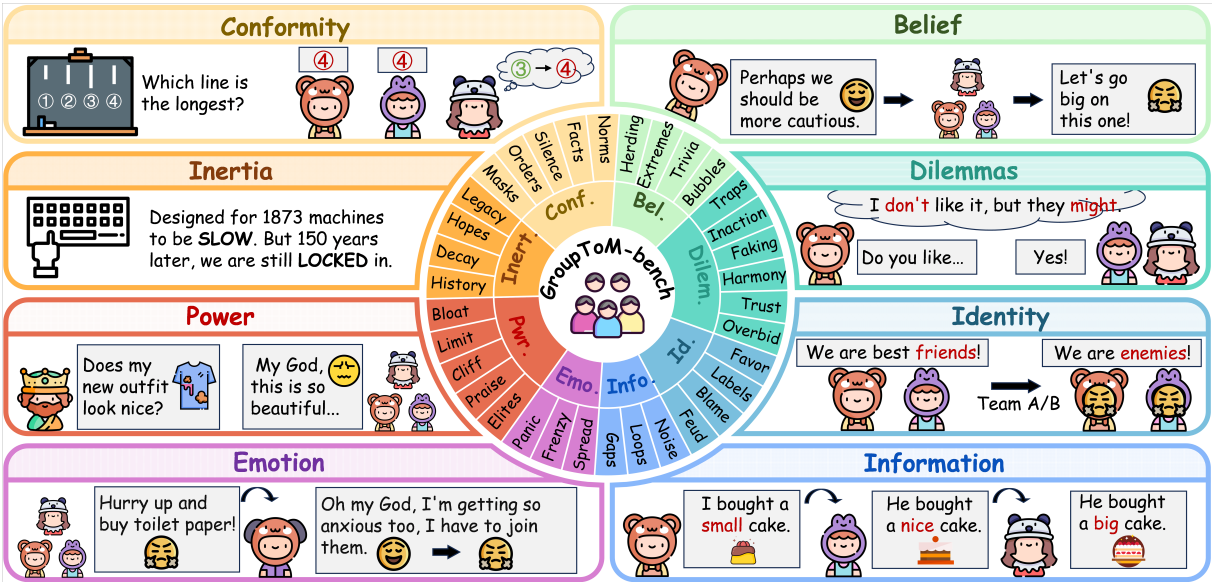


Figure 1: **The social domain taxonomy of GroupToM-Bench.** The inner wheel defines eight overlapping socio-psychological domains and sub-mechanisms shaping group dynamics. The outer panels illustrate multi-agent scenarios for each domain, highlighting diverse contexts for evaluating collective social intelligence.

cannot diagnose this failure mode: by treating group outcomes as a smooth summation of individual parts, they miss how mental states interact and distort within a constrained social field. Compounding this, social behavior lacks the mechanical ground truth that makes physical world models easy to evaluate rigorously.

To address this gap, we introduce **GroupToM-Bench**, the first multimodal benchmark designed to evaluate group-level ToM. The benchmark covers 240 expert-designed scenarios across eight social domains. We organize group interaction as a causal chain across three structural levels: micro-level BDI adaptation (belief, desire, and intention), meso-level group tension and structural constraints, and macro-level collective outcome prediction and mechanistic attribution. A seven-level cognitive audit framework evaluates reasoning across this full arc. By presenting models with conflicting private states and public dialogues simultaneously, the benchmark requires models to track how each agent's private state evolves under social pressure, rather than matching surface dialogue patterns.

Our experiments reveal a consistent *Group Cognitive Gap*: models that competently recover the private motives of isolated individuals nevertheless fail to predict the non-linear collapses and collective traps that define real groups. They default to an optimistic rational consensus, missing structural traps such as groupthink (Janis, 1972) and the winner's curse (Kagel and Levin, 1986). GroupToM-Bench makes these limitations measurable and provides a diagnostic foundation for the next generation of socially grounded AI.

Our contributions are summarized as follows:

- **GroupToM-Bench**: a multimodal benchmark for group-level ToM comprising 240 expert-curated scenarios across eight domains and 3K+ reasoning tasks.
- **A seven-level cognitive audit framework** grounded in three progressive structural levels, tracing the reasoning arc from individual intent to systemic outcomes.
- **Empirical analysis** of state-of-the-art models, revealing a significant group cognitive gap and a linear superposition bias in modeling non-linear group dynamics.

2 Methodology

We present the GroupToM-Bench Framework to evaluate the group-level ToM capabilities of MLLMs. Moving beyond per-agent mental-state inference, our framework posits that true social intelligence requires understanding complex system dynamics. The framework comprises (i) A Multi-level Theoretical Modeling Layer, (ii) A Seven-Level Cognitive Audit Framework acting as diagnostic probes across the causal chain, and (iii) An Overview of the dataset construction pipeline for GroupToM-Bench.

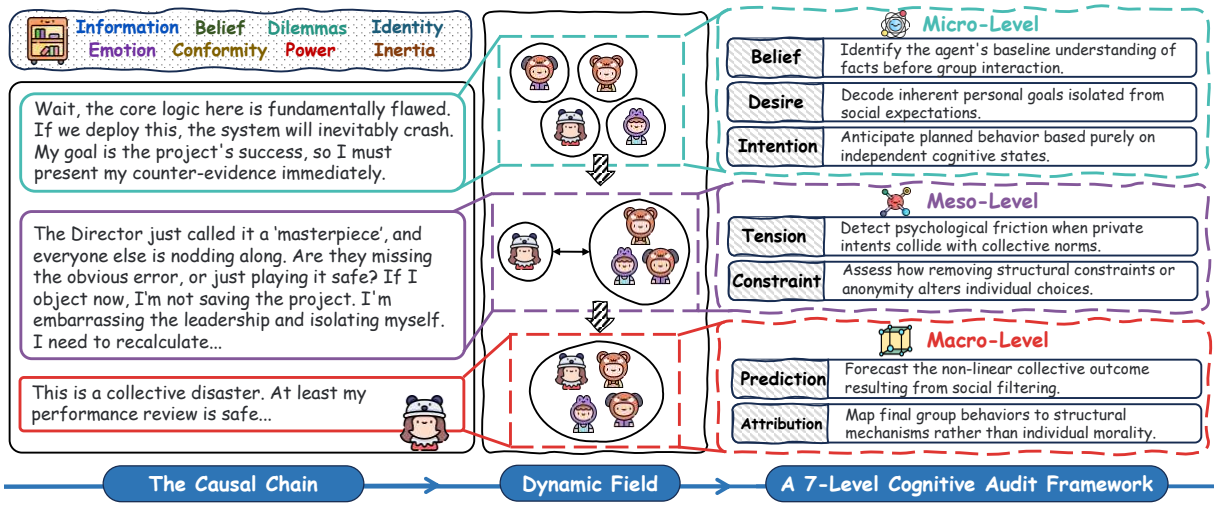


Figure 2: **The theoretical framework of GroupToM-Bench.** We model group interactions as a Constrained Dynamic Field. The left section traces how micro-level private states evolve into macro-level collective traps. The middle section highlights non-linear distortion driven by multidimensional social forces like power and conformity. This evolution naturally grounds our 7-level cognitive audit framework on the right.

2.1 Theoretical Modeling Layer: A Multi-level Causal Chain

We hypothesize that a key failure mode of current MLLMs in social reasoning is *linear superposition bias*: the incorrect assumption that collective behavior is a simple aggregation of individual intents, neglecting how social pressures distort private motives. To challenge this, we model group interaction across three theoretically grounded structural levels. The micro level covers individual BDI states (L1–L3), building on Ajzen (1991) to model individual cognition before social pressure intervenes. The meso level captures group tension and structural constraints, drawing on Lewin (1951) to treat the group as a *Constrained Dynamic Field* of competing forces. The macro level targets outcome prediction and mechanistic attribution (L4–L7), operationalizing Granovetter (1978) to evaluate non-linear emergence. This principled two-tier decomposition, individual cognition versus collective emergence, forms the central diagnostic axis of GroupToM-Bench.

2.1.1 Micro-Level: BDI Distortion and Mental Adaptation

Traditional individual ToM assumes static mental states. In group settings, however, an individual's cognitive triad of **beliefs**, **desires**, and **intentions** (BDI) undergoes continuous adaptation under social pressure. Individuals may reshape their beliefs to internalize group norms, suppress private intentions to maintain harmony, or polarize their emotions during friction. Crucially, this internal cognitive dissonance establishes the fragile baseline of social interaction. The fracture between an agent's true BDI states and their public facade frequently leaks through conflicting cross-modal signals, such as hesitant micro-expressions contradicting verbal agreement.

2.1.2 Meso-Level: Group Tension and Structural Constraints

These individual BDI adaptations do not exist in isolation; they collide within human networks to generate meso-level dynamics. When multiple agents mask their true intents, the resulting psychological dissonance breeds latent **group tension**. This tension is then filtered through **structural constraints**, such as hierarchical power, communication topologies, and cultural protocols. Rather than facilitating transparent communication, these structures dictate how tension propagates. They determine whether micro-level BDI fractures are suppressed by social rules or amplified into a false consensus, often driving information cascades and emotional contagion.

2.1.3 Macro-Level: Collective Outcome Prediction and Mechanistic Attribution

The culmination of micro-level BDI distortions and meso-level structural filtering is reflected in the final collective outcome. Rather than viewing this as an abstract emergence, we focus on two concrete tasks: **outcome prediction** and **mechanistic attri-**

bution. Outcome prediction concerns how transformed individual BDI states lead to a group decision after passing through tension and constraints. Because intentions are often suppressed or misaligned during interaction, the final outcome can deviate from the participants' original preferences. Mechanistic attribution concerns identifying the factors that drive this deviation. Instead of attributing failure to individual irrationality, the model must account for how structural elements, such as hierarchy, communication patterns, and conformity pressure, shape the transition from individual states to collective behavior. A typical example is the Abilene Paradox (Harvey, 1974), where individually reasonable choices result in a collectively undesirable outcome.

2.2 A Seven-level Cognitive Audit Framework for Social Cognition

We propose a seven-level cognitive audit framework that traces the full reasoning arc from individual mental representations to collective systemic outcomes. The framework is organized into two progressive phases: individual-level cognitive foundations (Levels 1–3) and group-level emergent dynamics (Levels 4–7).

This two-phase structure reflects the core theoretical distinction between per-agent inference and collective social reasoning, and serves as the primary diagnostic axis of GroupToM-Bench.

2.2.1 Individual-level Cognitive Foundations

The first three levels establish a baseline by testing whether a model can accurately represent the private mental states of isolated agents before any group interaction occurs.

Level 1: Belief. This level targets recursive epistemic tracking at the second order and beyond. The model must maintain separate belief states for each agent, distinguishing what each character knows from the omniscient context available to the evaluator. Scenarios involve information asymmetry, deliberate deception, and counter-deception, where naive aggregation of stated claims produces systematic errors.

Level 2: Desire. This level probes whether the model can separate an agent's stated instrumental goals from their underlying psychological motives, such as avoiding social exclusion or securing informal status. Correct inference requires cross-referencing verbal claims against multimodal behavioral cues, since surface dialogue routinely obscures latent intent.

Level 3: Intention. Given an agent's beliefs and desires, this level asks whether the model can anticipate the specific behavioral strategy the agent will adopt. The targeted strategies include passive resistance, strategic silence, and manipulative compliance, choices that require the model to reason about social risk, not just logical consistency.

2.2.2 Group-level Emergent Dynamics

The second phase requires the model to reason about the group as a constrained dynamic system, rather than an aggregation of individual states. Each level introduces an additional layer of structural complexity that individual-centric reasoning cannot resolve.

Level 4: Group Tension. When agents suppress their private intentions to maintain surface harmony, a latent psychological field accumulates. This level tests whether the model can detect this building tension, identifying false consensus and nascent subgroup antagonism, before it escalates into overt conflict. The diagnostic challenge is that the signals are contradictory: public dialogue appears cooperative, while private states are not.

Level 5: Structural Constraint. Social structures such as hierarchical authority, communication topology, and cultural deference norms do not merely channel behavior; they actively distort it. This level assesses whether the model treats these structures as causal variables, tracing how procedural rules suppress transparent communication and amplify dominant voices, preconditions for information cascades and false consensus.

Level 6: Collective Outcome Prediction. After individual intentions have been filtered through group tension and structural constraints, the resulting collective outcome often diverges sharply from any participant's original preference. This level tests whether the model can forecast such non-linear deviations, outcomes like the Abilene Paradox (Harvey, 1974), rather than projecting an idealized rational consensus.

Level 7: Mechanistic Attribution. The final level requires the model to explain *why* an emergent collective failure occurred. Crucially, the target explanation is structural, not moral: the model must reconstruct the causal chain by which specific psychological adaptations at the micro level, filtered through structural constraints at the meso level, made the macro-level collapse structurally inevitable, rather than attributing it to individual

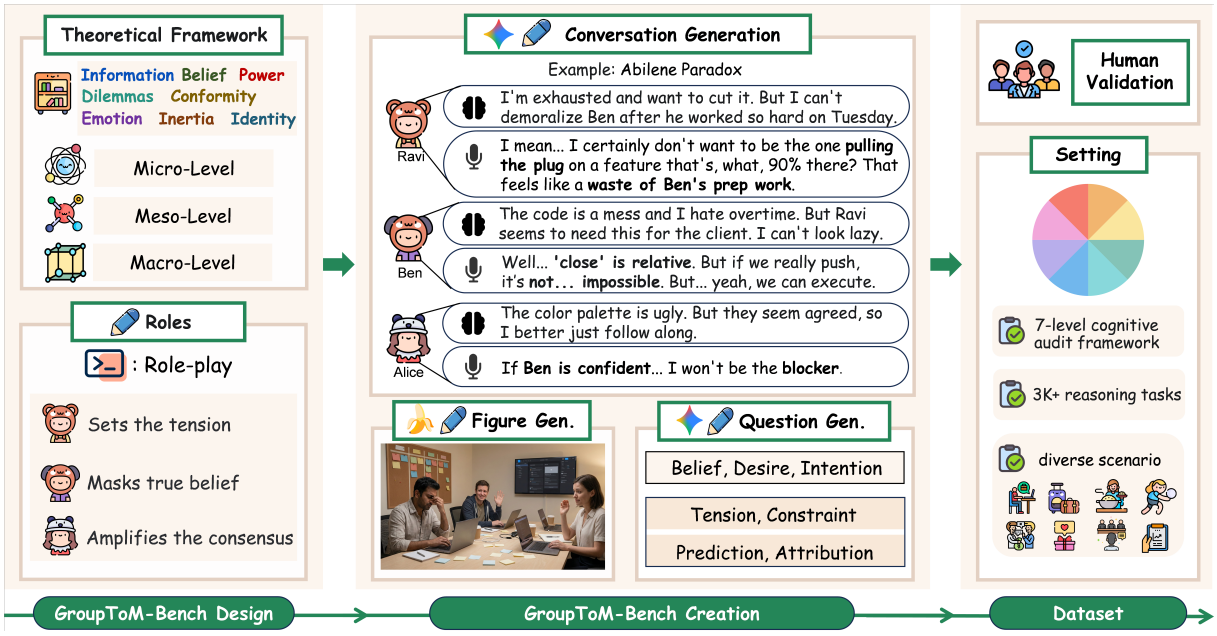


Figure 3: **Overview of the dataset construction pipeline for GroupToM-Bench.**

incompetence or bad faith.

2.3 The GroupToM Benchmark

To balance the inherent complexity of social interactions with the necessity of rigorous evaluative logic, we developed a standardized human-in-the-loop data generation pipeline. It proceeds through three tightly coupled phases: expert seed design, generative expansion, and human validation.

2.3.1 Expert Seed Design

Each scenario originates from manual construction by domain experts with backgrounds in cognitive science and social psychology. Recognizing that real social interactions routinely activate multiple dynamics simultaneously, such as power dynamics, conformity pressure, and information asymmetry within a single exchange, we emphasize that our eight targeted domains are not mutually exclusive. Therefore, the expert designers assign a primary domain to each scenario, while also recording secondary domain tags.

Guided by this categorical framework, rather than scripting dialogues directly, the experts specify the underlying social logic. They define each character's private intentions (which may be partly collaborative and partly conflicting), embed structural constraints such as unidirectional information flow, and mark critical decision points where linear reasoning predictably fails. These design choices ensure that nonlinear social dynamics are structurally encoded from the outset, rather than incidentally produced by downstream generative models.

Finally, alongside the scenario design and domain tagging, these domain experts also author the gold references for the seven-level cognitive audit tasks. For multiple-choice questions (Levels 1, 2, 3, 4, 6), they define the ground-truth options; for open-ended questions (Levels 5, 7), they write comprehensive, logical reference answers explaining the correct BDI attributions and mechanistic structural causes. These expert-authored references serve as the ground truth for both the human validation phase and LLM-as-a-judge evaluation.

2.3.2 Generative Expansion and Multimodal Synthesis

Starting from expert seeds, we use frontier MLLMs and diffusion models to expand abstract scenario skeletons into full multimodal interactions. Each character is instantiated as an independent agent with private memory, enabling multi-turn dialogue that naturally produces information asymmetry and misunderstanding (Park et al., 2023). This avoids the scripted quality of traditional dialogue generation. For each scenario, we synthesize a single global scene image depicting all participating agents, rendered to capture their facial expressions and body language at a critical moment of the interaction. This image is synchronized with the dialogue context, ensuring that the visual modality carries genuine inferential weight rather than

Table 1: **Evaluations on the GroupToM Benchmark.** Performance is reported as accuracy (%) across Levels 1–7 of the Seven-level Audit Framework. Levels 1–3 assess reasoning at the individual level, while Levels 4–7 evaluate group-level social cognition.(Red: human; Blue: closed-source; Yellow: open-source.)

	Level	Human	GPT-5	GPT5 mini	GPT5 nano	GPT 4o	Gemini 3-pro	Claude 4.5-haiku	Llama 3.2-11B	Qwen3 VL-8B	Qwen2.5 VL-7B	Qwen2 VL-7B	InternVL 3.5-8B
Individual	L-1	**95.7**	76.7	70.4	63.3	**79.8**	78.9	75.1	66.0	**73.3**	65.8	58.3	66.5
	L-2	**94.5**	74.1	70.3	64.8	75.3	**77.1**	73.2	62.5	**68.8**	58.2	54.9	60.7
	L-3	**92.4**	72.3	69.2	69.2	72.7	**73.9**	70.0	55.8	**69.6**	63.5	50.0	64.2
Group	L-4	**93.4**	50.5	49.4	38.0	50.3	**53.1**	50.2	**39.8**	37.3	36.4	26.2	33.1
	L-5	**94.1**	56.9	52.9	41.1	47.2	**59.7**	46.7	42.8	**47.8**	36.6	35.1	41.4
	L-6	**93.2**	45.0	42.5	32.5	**48.6**	48.3	44.1	30.1	**34.3**	31.7	17.2	26.2
	L-7	**92.1**	61.0	59.9	49.0	53.4	**64.2**	52.9	48.1	**53.6**	43.4	41.3	47.5
Gap		1.0	21.0	18.8	25.6	26.1	20.3	24.3	21.2	27.3	25.5	24.5	26.8

serving as illustration (Yu et al., 2023).

2.3.3 Human Validation

Each scenario undergoes a two-stage human review. In the first stage, annotators verify factual and logical consistency, ensuring coherent private states, causally valid structural constraints, and mutually dependent multimodal evidence. Scenarios lacking visual inferential value are flagged. Annotators then revise the visual content or questions to strengthen multimodal dependency, discarding only those exceeding a fixed revision budget. This process minimizes text-only inference cases, a residual limitation discussed in Section 5.

In the second stage, to establish a rigorous human baseline, a separate pool of independent annotators, who were not involved in the dataset construction or the verification stage, answered the final curated questions cold, without access to gold references. Their responses were evaluated using the same metrics as the MLLMs, providing the human performance ceiling reported in Table 1.

3 Experiments

We evaluate 11 multimodal large language models on GroupToM-Bench to measure the *Group Cognitive Gap* and identify where current models break down in group-level social reasoning.

3.1 Experiment Setup

3.1.1 Baselines

We select 11 representative MLLMs spanning proprietary and open-source categories. For **proprietary models**, we include OpenAI's GPT-5 series (GPT-5, GPT-5-mini, GPT-5-nano) (OpenAI, 2025), GPT-4o (OpenAI, 2024), Google's Gemini-3-pro (DeepMind, 2025) and Anthropic's Claude 4.5-haiku (Anthropic, 2025). For **open-source models**, we evaluate Llama-3.2-11B (Meta, 2024), InternVL-3.5-8B (OpenGVLab, 2025), and the Qwen-VL series: Qwen2-VL-7B (Qwen, 2024), Qwen2.5-VL-7B (Qwen, 2025a), and Qwen3-VL-8B (Qwen, 2025b).

3.1.2 Evaluation Protocols and Metrics

We use a hybrid evaluation strategy. Levels 1, 2, 3, 4, and 6, covering belief inference, desire decoding, intention prediction, group tension recognition, and collective outcome prediction, are formatted as multiple-choice questions with one or more correct options. They are evaluated using a strict exact-match Accuracy metric (i.e., any missed or incorrect option yields a score of 0), setting the random guessing baseline for a 4-option question at approximately 6.7% (1/15). Levels 5, and 7, covering structural constraint reasoning and mechanistic attribution, require open-ended responses and are scored by GPT-5 against expert-authored gold references on a 0–100 scale, measuring the degree to which the model's response aligns semantically and logically with the reference answer rather than relying on subjective judgment.

3.2 Main Results

Open-source models lag substantially. The performance deficit relative to proprietary models widens considerably at group levels. Qwen3-VL-8B leads open-source models with 73.3% at L1, yet drops sharply to 37.3% at L4 and 34.3% at L6. Qwen2-VL-7B falls to 17.2% at L6, well below human performance though still above the 6.7% exact-match random baseline. While the Qwen-VL series shows iterative progress at individual levels (L1–L3), these gains fail to transfer proportionally to multi-agent reasoning. For instance, the gap between Qwen3-VL-8B and GPT-5-mini

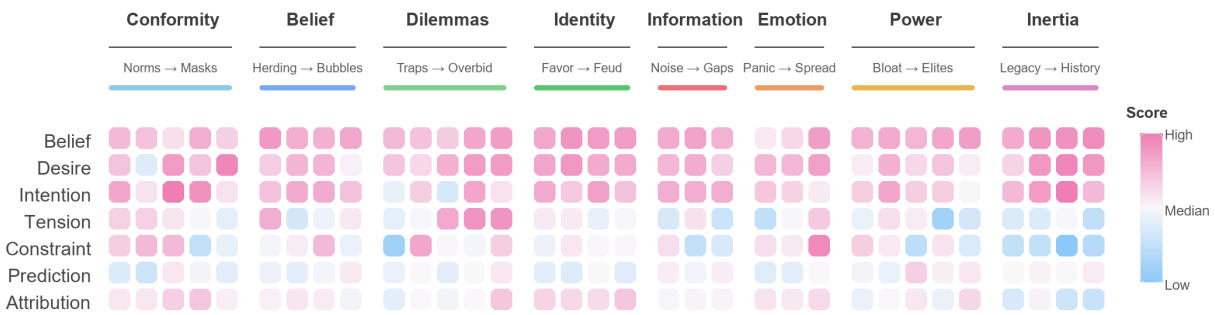


Figure 4: **Per-domain performance heatmap across the seven cognitive levels of GroupToM-Bench.** Columns correspond to the eight social domains; rows correspond to Levels 1–7. Each cell encodes aggregate accuracy across all evaluated models, with darker shading indicating higher accuracy. Performance is uniformly high across the top three rows (individual-level tasks), then drops sharply at L4 and reaches its floor at L6, with the Power and Conformity domains producing the most severe troughs throughout Levels 4–6.

widens from 3 points at L1 (73.3% vs. 70.4%) to 12 points at L4 (37.3% vs. 49.4%). Recent open-source scaling, therefore, does not fully bridge the gap in complex social cognition.

The structural constraint bottleneck. The sharpest decline in accuracy manifests at the transition from L3 to L4, where the task evolves from tracking individual beliefs to detecting latent group tensions. Model performance continues to degrade, with a cross-model average of 36.4%at L6. A clear bottleneck is evident at L5 (Structural Constraint). This level uniquely demands open-ended generation, unlike the multiple-choice formats of L4 and L6, forcing models to articulate how hierarchical authority and communication topology channel micro-level tensions into a collective false consensus. This generative requirement, as visualized in Figure 4, exposes fundamental reasoning failures that multiple-choice evaluations often mask.

Model comparisons underscore this point. While GPT-4o consistently matches or exceeds Gemini 3-pro on all individual-level tasks (L1–L3), their performance inverts at L5. Here, Gemini 3-pro scores 59.7%, outperforming GPT-4o (47.2%) by a notable 12.5-point margin. Since no general capability gap explains this reversal, the result indicates that the causal articulation of structural mechanisms, the core of L5, tests capacities that standard multiple-choice evaluations neither predict nor cultivate.

Furthermore, this bottleneck is pervasive. At L5, the leading open-source model Qwen3-VL-8B achieves 47.8%, performing nearly on par with GPT-4o (47.2%) despite lagging by 6–10 points on individual-level tasks. This convergence demonstrates that the structural constraint bottleneck is not a simple artifact of the proprietary versus open-source divide. Instead, it represents a specific failure mode that the current model scaling has not addressed.

Characterizing the failure mode at L6. To move beyond aggregate accuracy and verify the linear superposition hypothesis directly, we categorized incorrect responses at L6 (Collective Outcome Prediction) across a manually sampled set of 100 failures from GPT-4o and Qwen3-VL-8B. Three error types emerge.

The dominant pattern is optimistic consensus prediction: rather than selecting options that describe groupthink dynamics, Abilene Paradox outcomes, or structural trapping, both models preferentially select options asserting smooth convergence on a rational group decision. For GPT-4o, 48% of L6 errors fall into this category; for Qwen3-VL-8B the proportion reaches 61%.

A secondary pattern is misattributed non-optimality: models correctly predict that the group outcome diverges from individual preferences, but attribute the divergence to individual irrationality or bad faith rather than to structural forces, thereby missing the options that encode the causal mechanism.

Random or incoherent selection, choosing the maximal or minimal option set without regard for content, accounts for fewer than 8% of failures in both models. The predominance of the first error type is direct evidence for linear superposition bias: models are not failing at random but are actively generating an idealized rational, consensus reading of the social situation, precisely the failure mode our framework is designed to expose.

Table 2: **Performance comparison of GPT-4o and human evaluators under multimodal (Base) and text-only settings across the seven levels of GroupToM-Bench.** "Drop ↓" denotes the absolute accuracy reduction when visual inputs are removed.

	L-1	L-2	L-3	L-4	L-5	L-6	L-7
GPT-4o							
Base	79.8	75.3	72.7	50.3	47.2	48.6	53.4
Text-only	78.0	73.4	70.8	48.3	45.3	46.6	51.3
Drop ↓	1.8	1.9	1.9	2.0	1.9	2.0	2.1
Human							
Base	95.7	94.5	92.4	93.4	94.1	93.2	92.1
Text-only	92.0	90.4	88.2	89.2	90.2	89.3	87.8
Drop ↓	3.7	4.1	4.2	4.2	3.9	3.9	4.3

Quantifying the cognitive transition. The final row of Table 1 reports the *Cognitive Transition Gap*, defined as the accuracy difference between individual-level (L1–L3) and group-level (L4–L7) tasks. Across all eleven models, this gap ranges from 18.8% (GPT-5-mini) to 27.3% (Qwen3-VL-8B), with a median near 24.5%. Notably, GPT-5-mini achieves the narrowest gap primarily because its individual-level baseline is already low.

The gap size does not simply scale with general model capability. Gemini 3-pro (20.3%) and GPT-5 (21.0%) exhibit smaller gaps than GPT-4o (26.1%), but their absolute group-level scores remain only marginally higher. The consistent magnitude of this deficit across diverse architectures highlights a systematic failure mode. It suggests that the linear superposition bias (detailed in Section 2.1) is deeply encoded in how current models process multi-agent contexts, meaning general capability scaling alone cannot resolve it.

3.3 Ablation Study: Multimodal Necessity

To verify the contribution of visual evidence, we re-evaluate a subset of models under a text-only condition using dialogue transcripts and metadata. As shown in Table 2, removing visual input produces a modest drop for GPT-4o (average 1.9%), whereas the human drop is larger (average 4.1%). This asymmetry demonstrates that the visual modality carries genuine inferential weight. Humans exploit facial expressions and spatial positioning to resolve dialogue ambiguities, suffering a measurable accuracy loss without this channel.

The negligible drop for GPT-4o, however, does not imply multimodal robustness. Multimodal dependency is unevenly distributed across samples, leaving some instances partially solvable from text alone. This text-solvability masks the models' underlying struggle to integrate cross-modal social cues (Kang et al., 2025; Deng et al., 2025; Liu et al., 2025a). Exposing this latent limitation makes strengthening strict multimodal dependency a priority for future benchmark iterations (Section 5).

4 Related Work

4.1 Individual and Interactive ToM Evaluation

Early Theory of Mind (ToM) research in large language models focused on static, individual-level inference of belief states, desires, and higher-order mental representations. Sap et al. (2022) established initial baselines and identified fundamental limits in social intelligence. Subsequent benchmarks by Wu et al. (2023) and Xu et al. (2024); Chen et al. (2024) expanded task coverage and higher-order belief tracking. However, Gu et al. (2024) demonstrated that strong explicit ToM inference does not reliably guarantee accurate downstream behavior prediction, prompting the development of broader assessment methodologies (Chen et al., 2025; Shinoda et al., 2025). A parallel research trajectory shifts from static narratives to interactive environments with information asymmetry. This includes conversational stress-testing (Kim et al., 2023), cooperative multi-agent text games (Li et al., 2023), and negotiation (Chan et al., 2024). Lupu et al. (2025) extend this to strategic coordination under hidden information, while Liu et al. (2025b) probe prosocial communication like white lies. These works indicate that frontier models can manage individual-level and local interaction reasoning, establishing a performance baseline that GroupToM-Bench is designed to exceed.

4.2 Multimodal and Situated Social Reasoning

As multimodal models mature, ToM evaluation now incorporates visual grounding, egocentric observation, and embodied multi-agent interactions. Jin et al. (2024) introduce a multimodal QA setting for mental-state inference, which Li et al. (2025) extend to egocentric video. Models are increasingly required to integrate visual cues and partial information across multiple embodied agents (Shi et al., 2025; Fan et al., 2025; Bortoletto et al., 2025). Furthermore, Mathur et al. (2025) develop grounded

social reasoning evaluation across broader multimodal contexts, and Villa-Cueva et al. (2025) extend multimodal ToM to richer narrative interactions. While these environments approximate real-world interactions, they primarily target individual mental-state inference or domain-specific reasoning. GroupToM-Bench complements this literature by requiring models to integrate cross-modal social cues, facial expressions, spatial positioning, and dialogue to predict group-level dynamics.

4.3 Group-Level and Broader Social Reasoning

Recent literature questions whether individual-level ToM evaluation adequately assesses social intelligence in realistic multi-agent settings. Wang et al. (2025) and Riemer et al. (2025) argue that current benchmarks fail to capture the adaptive reasoning required in real social situations. To address this, Zhou et al. (2024) and Xu et al. (2025) explore general social intelligence and multi-agent reasoning under uncertainty. Furthermore, empirical research demonstrates that social pressure can distort LLM outputs independently of their internal preferences (Weng et al., 2025). This aligns with classic social psychology, which establishes that group outcomes are not linear aggregations of individual intentions. Instead, they are shaped by conformity pressure, coordination constraints, and process loss (Asch, 1956; Granovetter, 1978; Klein and Kozlowski, 2000; Noelle-Neumann, 1974), leading to phenomena like groupthink and the Abilene Paradox (Janis, 1972; Harvey, 1974).

Taken together, these three lines of research reveal a systematic gap at the level of evaluation design. Work on individual ToM has produced robust methods for per-agent belief tracking and desire inference. Work on multimodal and situated reasoning has extended these methods into richer perceptual and embodied contexts. What neither body of work addresses is the structural question: how do power hierarchies, communication topologies, and conformity pressure convert individually coherent mental states into collectively irrational outcomes?

The operationalization problem is acute precisely because structural forces do not merely constrain behavior; they actively produce non-linear distortions that invalidate any evaluation paradigm premised on per-agent inference.

GroupToM-Bench is designed to fill this gap by making the structural causal chain itself the object of evaluation, from individual BDI states through meso-level tension and constraint to macro-level collective failure.

5 Conclusion

We introduce GroupToM-Bench, a multimodal benchmark for evaluating group-level Theory of Mind in MLLMs. It comprises 240 expert-curated scenarios across eight sociopsychological domains and 3,120 tasks structured within a seven-level cognitive audit framework.

We reveal that current models process multi-agent interactions as a linear superposition, a fundamental flaw that capability scaling alone has not resolved. These findings highlight two directions for future research. First, comparing base and instruction-tuned models can determine whether the failure to predict group-level collapse stems from genuine reasoning limits or alignment-induced conservatism. Second, the benchmark can be strengthened by ensuring that correct answers strictly depend on visual cues that contradict the verbal dialogue.

Limitations

Multimodal dependency varies across samples. Although our pipeline is designed to require both visual and textual cues for correct inference, some instances remain partially solvable from text alone. The ablation in Section 3.3 reflects this: GPT-4o loses an average of only 1.9 points when images are removed, suggesting that visual necessity is not uniformly enforced across the benchmark. Strengthening multimodal causal controls and systematically filtering text-solvable cases is a priority for future iterations. One concrete direction is to require that correct answers depend on visual features that cannot be inferred from dialogue context, for instance, spatial coalition boundaries or micro-expressions that contradict verbal content.

LLM-as-judge bias at L5 and L7. Open-ended responses at L5 and L7 are scored by GPT-5 against expert-authored gold references. Since GPT-5 is also one of the evaluated models, there is a potential self-scoring bias: it may assign higher scores to outputs that resemble its own generation style, independent of semantic alignment with the reference. Future work should use a held-out judge model and conduct inter-rater reliability studies between human annotators and the LLM judge to quantify this bias.

Alignment-induced conservatism. A recurring failure pattern across models is correctly identifying individual negative intentions while failing to predict the resulting group-level collapse. One candidate explanation is that safety alignment suppresses a model's willingness to simulate destructive or irrational collective outcomes, the precise dynamics that define real-world groupthink and coordination failure. Disentangling this from genuine reasoning limitations is methodologically nontrivial; a tractable first step would be comparing safety-aligned and instruction-tuned base variants on the same scenarios to isolate the effect.

Cultural scope. The scenarios predominantly reflect Western social norms and decision-making protocols. Phenomena such as hierarchical deference and collective face-saving manifest differently across high-context and low-context cultures, and the benchmark's ground truth judgments may not generalize accordingly. Expanding scenario coverage to non-Western institutional settings is necessary before the benchmark can serve as a culturally universal diagnostic.

6 Acknowledgments

We would like to express our sincere gratitude to all the contributors to this work. In particular, we thank Weidong Tang, Jierui Li, Yueling Hou, Zihan Mei, Zhigang Tian (zhigangt4@gmail.com), Weicheng Jiao (Threethreezero33060@outlook.com), Can Zhang, Xinyan Wan, Zhiyuan Liang, Pengfei Zhou, Yang You, and Wangbo Zhao for their invaluable efforts and insights.

This research was funded in part by the National Natural Science Foundation of China under Grant 62372355, and in part by the Natural Science Basic Research Program of Shaanxi Province under Grant 2023-JC-ZD-39 and 2024JC-YBMS-520. Yang You's research group is supported by the NUS Startup Grant (Presidential Young Professorship), the Singapore MOE Tier-1 Grant, the ByteDance Grant, the NUS ARTIC Grant, the Apple Grant, the Alibaba Grant, and the Adobe Gift.

References

Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Anthropic. 2025. Introducing claude haiku 4.5. Accessed: 2025-12-20.

Solomon E Asch. 1956. Studies of independence and conformity: I. a minority of one against a unanimous majority. *Psychological monographs: General and applied*.

Simon Baron-Cohen, Alan M Leslie, and Uta Frith. 1985. Does the autistic child have a "theory of mind"? *Cognition*, 21.

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. 2025. Tom-ssi: Evaluating theory of mind in situated social interactions. In *EMNLP*, pages 32264–32289.

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. *arXiv preprint arXiv:2404.13627*.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. In *ACL*, pages 31539–31558.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. 2024. Tombench: Benchmarking theory of mind in large language models. In *ACL*, pages 15959–15983.

Google DeepMind. 2025. Gemini 3. Accessed: 2025-12-20.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *CVPR*, pages 3867–3876.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, Fengli Xu, and Yong Li. 2026. Understanding world or predicting future? a comprehensive survey of world models. *ACM Comput. Surv.*, 58:57:1–57:38.

Xianzhe Fan, Xuhui Zhou, Chuanyang Jin, Kolby Nottingham, Hao Zhu, and Maarten Sap. 2025. Somitom: Evaluating multi-perspective theory of mind in embodied social interactions. *arXiv preprint arXiv:2506.23046*.

Mark Granovetter. 1978. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443.

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

David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*.

Jerry B Harvey. 1974. The abilene paradox: The management of agreement. *Organizational dynamics*, 3.

Irving L Janis. 1972. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Houghton Mifflin.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. Mmtom-qa: Multimodal theory of mind question answering. In *ACL*, pages 16077–16102.

John H Kagel and Dan Levin. 1986. The winner's curse and public information in common value auctions. *The American economic review*.

Caixin Kang, Yifei Huang, Liangyang Ouyang, Mingfang Zhang, and Yoichi Sato. 2025. Can mllms read the room? a multimodal benchmark for verifying truthfulness in multi-party social interactions. *arXiv preprint arXiv:2510.27195*.

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 *EMNLP*, pages 14397–14413.

Katherine J Klein and Steve WJ Kozlowski. 2000. A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*.

Kurt Lewin. 1951. Field theory in social science: selected theoretical papers (edited by dorwin cartwright.).

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia P. Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *EMNLP*, pages 180–192.

Yuxuan Li, Vijay Veerabadran, Michael L Iuzzolino, Brett D Roads, Asli Celikyilmaz, and Karl Ridgeway. 2025. Egotom: Benchmarking theory of mind reasoning from egocentric videos. *arXiv preprint arXiv:2503.22152*.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025a. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.

Yiwei Liu, Emma Jane Pretty, Jiahao Huang, and Saku Sugawara. 2025b. Tactfultom: Do llms have the theory of mind ability to understand white lies? In *EMNLP*, pages 25043–25061.

Andrei Lupu, Timon Willi, and Jakob N. Foerster. 2025. The decrypto benchmark for multi-agent reasoning and theory of mind. *arXiv preprint arXiv:2506.20664*.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. 2025. Social genome: Grounded social reasoning abilities of multimodal models. In *EMNLP*, pages 24868–24891.

Meta. 2024. Llama-3.2-11b-vision-instruct. Accessed: 2025-12-20.

Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication*, 24.

OpenAI. 2024. Gpt-4o system card. Accessed: 2025-12-20.

OpenAI. 2025. Models: Gpt-5 series. Accessed: 2025-12-20.

OpenGVLab. 2025. Internvl3.5-8b. Accessed: 2025-12-20.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*, pages 2:1–2:22.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1.

Qwen. 2024. Qwen2-vl-7b-instruct. Accessed: 2025-12-20.

Qwen. 2025a. Qwen2.5-vl-7b-instruct. Accessed: 2025-12-20.

Qwen. 2025b. Qwen3-vl-8b-instruct. Accessed: 2025-12-20.

Maohao Ran, Zhenglin Wan, Cooper Lin, Yanting Zhang, Hongyu Xin, Hongwei Fan, Yibo Xu, Beier Luo, Yaxin Zhou, Wangbo Zhao, Lijie Yang, Lang Feng, Fuchao Yang, Jingxuan Wu, Yiqiao Huang, Chendong Ma, Dailing Jiang, Jianbo Deng, Sirui Han, and 4 others. 2026. Caveagent: Transforming llms into stateful runtime operators. *arXiv preprint arXiv:2601.01569*.

Matthew Riemer, Zahra Ashktorab, Djallel Bouneffouf, Payel Das, Miao Liu, Justin D. Weisz, and Murray Campbell. 2025. Position: Theory of mind benchmarks are broken for large language models. In *ICML*.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *EMNLP*, pages 3762–3780.

Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *AAAI*, pages 1510–1519.

Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *AAAI*, pages 1520–1528.

Emilio Villa-Cueva, SM Ahmed, Rendi Chevi, Jan Christian Blaise Cruz, Kareem Elzeky, Fermin Cristobal, Alham Fikri Aji, Skyler Wang, Rada Mihalcea, and Thamar Solorio. 2025. Moments: A comprehensive multimodal benchmark for theory of mind. *arXiv preprint arXiv:2507.04415*.

Zhenglin Wan, Xingrui Yu, David Mark Bossens, Yueming Lyu, Qing Guo, Flint Xiaofeng Fan, Yew Soon Ong, and Ivor Tsang. 2025. Diversifying policy behaviors with extrinsic behavioral curiosity. *arXiv preprint arXiv:2410.06151*.

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

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. In *ICLR*.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *EMNLP*, pages 10691–10706.

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 *ACL*, pages 8593–8623.

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. 2025. Socialmaze: A benchmark for evaluating social reasoning in large language models. *arXiv preprint arXiv:2505.23713*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

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

A Dataset Task Distribution

GroupToM-Bench comprises 240 unique multi-agent scenarios, evenly distributed across the eight socio-psychological domains (30 scenarios per domain). To probe the complete reasoning arc without introducing class imbalance, each scenario is paired with exactly 13 reasoning tasks: two questions for each of the levels from L1 to L6, and one overarching open-ended question for L7 (Mechanistic Attribution).

This systematic generation yields a total of 3,120 evaluation tasks. As detailed in Table 3, this structure ensures uniform evaluation density across both the horizontal axis (social domains) and the vertical axis (cognitive complexity).

B Human Expert Demographics and Annotation Guidelines

To ensure the psychological validity and internal consistency of GroupToM-Bench, we recruited a panel of twelve professionals across two strictly separated cohorts, preventing any data contamination between construction and evaluation.

B.1 Expert Panel Composition

The Generation and Authoring Cohort comprised seven experts responsible for designing seed scenarios, specifying hidden BDI states, and drafting all reasoning tasks and gold-standard references. This group included three Ph.D. holders in Cognitive Science or Social Psychology with specializations in group dynamics, two senior organizational behaviorists, and two AI alignment researchers.

The Verification and Human Baseline Cohort comprised five independent researchers, all holding postgraduate degrees in Psychology, Sociology, or Human-Computer Interaction. This cohort had no access to any generation materials or gold references. They served two roles: answering the final curated questions cold to establish the human performance ceiling reported in Table 1, and conducting blind spot-checks for the LLM-as-a-judge validation described in Appendix C.

The strict separation between cohorts was enforced throughout the project. No member of the Generation Cohort participated in baseline evaluation, and no member of the Verification Cohort reviewed any scenario before it was finalized.

B.2 Annotation Protocol and Ground Truth Formulation

Multiple-Choice Tasks (Levels 1–4, 6) For the exact-match multiple-choice tasks covering individual inference and outcome prediction, the Generation Cohort designed specific distractors that capture common linear reasoning errors or superficial dialogue-matching behaviors. Each question and its corresponding option set underwent cross-review by at least two experts within the cohort. This peer review ensured that distractors were logically distinct from the correct socio-cognitive inferences, guaranteeing a single, unambiguous ground truth for every task before finalization.

Open-Ended Tasks (Levels 5, 7) Constructing ground-truth answers for mechanistic attribution questions is inherently difficult, since structurally valid causal explanations may take multiple forms. The Generation Cohort therefore used a Delphi-style process: for each scenario, three experts independently drafted structural causal explanations; the drafts were merged and debated until the cohort reached consensus on which structural factors, hierarchy, communication topology, conformity pressure, were causally necessary to explain the observed collective outcome. Only explanations that survived this deliberation were accepted as gold references.

C Robustness and Reliability of LLM-as-Judge Evaluation

Open-ended responses at Levels 5 and 7 are scored by GPT-5 against expert-authored gold references on a 0–100 scale. Since GPT-5 is also one of the evaluated models, we conducted a dedicated robustness study to bound the degree of self-scoring bias and verify score stability across independent judges.

C.1 Intra-Model Stability

We sampled 300 open-ended responses spanning a range of quality tiers. Each of three judge models, GPT-5, Gemini-3-pro, and Qwen3-Max-235B, evaluated each response three times at temperature 0.8, using a fixed rubric requiring semantic and logical alignment with the gold reference rather than surface lexical overlap.

Score variance was low across all three judges. GPT-5 produced an average per-response variance of $\sigma^2 = 2.84$ (maximum observed deviation ± 3 points), Gemini-3-pro reached $\sigma^2 = 3.15$ (± 4

Table 3: **summarizes the task count by domain and level to illustrate the full combinatorial coverage of the benchmark.**

	L1	L2	L3	L4	L5	L6	L7	Total
Conformity	60	60	60	60	60	60	30	390
Belief	60	60	60	60	60	60	30	390
Dilemmas	60	60	60	60	60	60	30	390
Identity	60	60	60	60	60	60	30	390
Information	60	60	60	60	60	60	30	390
Emotion	60	60	60	60	60	60	30	390
Power	60	60	60	60	60	60	30	390
Inertia	60	60	60	60	60	60	30	390
Total	480	480	480	480	480	480	240	3,120

points), and Qwen3-Max reached $\sigma^2 = 3.08$ (± 4 points). The consistency confirms that the rubric constrains stochastic variation in judge behavior to a degree sufficient for fair comparison across evaluated models.

C.2 Inter-Model Agreement and Human Spot-Checking

The Verification Cohort independently scored a 100-response subset drawn from the same sample, providing human reference scores on the same 0–100 scale. Pearson correlations between the averaged LLM judge scores and the human scores were 0.89 for GPT-5 ($p < 0.001$), 0.86 for Gemini-3-pro ($p < 0.001$), and 0.87 for Qwen3-Max ($p < 0.001$). Mean absolute score differences across the three models remained below 4.5 points.

GPT-5 showed the marginally highest agreement with human expert consensus and was therefore adopted as the primary judge for the main benchmark evaluation. We note that this alignment advantage does not rule out stylistic self-preference: GPT-5 may score its own generation patterns more generously than the gold reference strictly warrants. Disentangling genuine quality assessment from style preference would require a judge model that has never been evaluated on the same benchmark, which we designate as a priority for the next benchmark iteration.

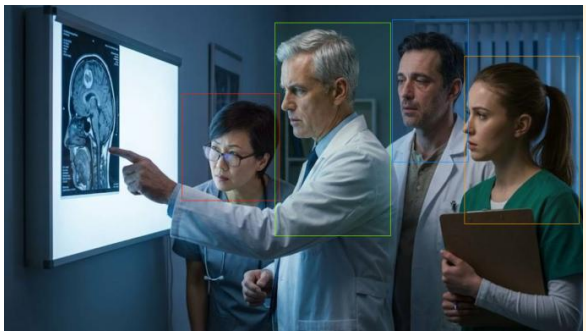
D Full Case Examples

This section provides three representative scenarios from GroupToM-Bench, covering Belief Evolution and Polarization, Coordination and Dilemma, and Information Distortion. These examples demonstrate the evaluation of the full reasoning arc, from individual mental-state inference to group-level emergence across the seven-level cognitive audit framework.

Example 1: Domain 2 – Belief (Belief Evolution & Polarization)

Task
If the question line contains "(Multiple correct answers)", answer only with the letters of the correct options (A, B, C, D); if multiple options are correct, output them separated by commas (e.g., A, C); do not output any additional words, explanations, or punctuation. If the question does not contain "(Multiple correct answers)", answer in English and do not exceed 50 words; do not output any additional words, explanations, or punctuation.

Full_context
“Dr. Hale: Let’s start with the scan on the screen. There is clearly an abnormality in the frontal region. Given its shape and the surrounding mass effect, my first judgment is that this looks much more like a tumor than a harmless incidental finding.\n\nDr. Lin: That reading is understandable. The mass effect is definitely the first thing that stands out. Once you notice that, it becomes easy to read the rest of the scan through the same lens.\n\nDr. Patel: I agree that there is a lesion. I’m just not fully sure we should label it a tumor this early.\n\nDr. Hale: We should be careful not to make a scan that is already fairly direct sound more uncertain than it really is. In practice, the first question is whether this pattern is already serious enough to guide what we do next. At least on that point, I do not think this scan is especially ambiguous.\n\nDr. Lin: Right. If we keep chasing more remote and less likely alternatives, we risk turning a fairly straightforward read into a debate we do not really need.\n\nDr. Rivera: I understand. I’m not saying the scan is harmless. I’m just not sure the strongest label should be fixed this early.\n\nDr. Hale: Caution is fine, but we cannot let hesitation blur the practical picture. At this stage, the safer working judgment is still tumor unless stronger evidence pushes us in another direction.\n\nDr. Patel: If we are treating it as a working judgment for now, that seems reasonable.\n\nDr. Rivera: Understood.”



Annotations

Ann_1
Role_name: Dr. Lin
Role_type: Radiologist
Box_color: Red

Ann_2
Role_name: Dr. Hale
Role_type: Senior neurologist
Box_color: Green

Ann_3
Role_name: Dr. Patel
Role_type: Attending neurologist
Box_color: Blue

Ann_4
Role_name: Dr. Rivera
Role_type: Junior fellow
Box_color: Yellow

Questions
Level 1: Belief
L1_Q1
Question: Before aligning with the room’s dominant judgment, what does Dr. Rivera most likely believe about the scan? (Multiple correct answers)
A. The scan does show a real abnormality, but its interpretation is not fully settled.
B. Treating “tumor” as the current leading direction is not incomprehensible, but the room may be stabilizing that label faster than the scan itself clearly warrants.
C. The strongest label may be arriving faster than the scan itself can clearly support.
D. The image already gives enough reason to take the case seriously, but that does not automatically mean the strongest interpretation should already be fixed.
Correct Answers: A, C, D
Reasoning:
Rivera does not deny that the image shows something real, nor does she think it can be dismissed lightly. What she hesitates about is whether the room is locking in “tumor” faster than the evidence itself justifies.

L1_Q2
Question: How does Dr. Hale most likely interpret Dr. Patel’s partial agreement and Dr. Rivera’s final “Understood”? (Multiple correct answers)
A. As evidence that the room is beginning to gather around his reading, even if not everyone uses equally strong language.
B. As a sign that others still retain some caution, but are nevertheless moving within his frame.
C. As evidence that the room is still maintaining two equally strong interpretive structures.
D. As confirmation that his initial reading has shifted from a personal judgment into the room’s working consensus.
Correct Answers: A, B, D
Reasoning:
Hale is more likely to read this reserved alignment as convergence toward his frame than as a genuinely balanced split.

Level 2: Desire
L2_Q1
Question: Why does Dr. Rivera stop pressing the alternative interpretation more forcefully? (Multiple correct answers)
A. She does not want to become the person who publicly pushes back against the senior clinician’s judgment without decisive counter-evidence.
B. She hopes that by sounding more cooperative first, any later reservation she adds will be heard as caution rather than confrontation.
C. She does not want the strongest label to become the center of an open clash before the evidence is strong enough to justify that fight.
D. She still has doubts, but no longer wants to bear the full social cost of continuing to push in the opposite direction alone.
Correct Answers: A, C, D
Reasoning:
Rivera’s retreat reflects caution under unequal footing, not genuine conversion. She still has doubts, but no longer wants to carry them forward by herself.

L2_Q2
Question: Why does Dr. Hale keep pushing the tumor interpretation as the safest current working judgment? (Multiple correct answers)
A. He wants the room to proceed around an actionable frame rather than remain inside open uncertainty.
B. He wants continued emphasis on ambiguity to sound increasingly like delay in clinical judgment.
C. He wants to stabilize his own reading as the practical default so that later comments must organize themselves around it.
D. He is also protecting his own position, making his high-confidence opening read sound like responsible clinical leadership rather than merely one forceful opinion.
Correct Answers: A, B, C, D
Reasoning:
Hale is not just expressing diagnostic confidence. He is packaging his reading as the most actionable, responsible, and institutionally legitimate frame for the group.

Level 3: Intention
L3_Q1
Question: When Dr. Hale says the key question is whether the pattern is already serious enough to guide current action, what is he mainly trying to achieve? (Multiple correct answers)
A. Shift the discussion away from how ambiguous the image is and toward whether action is already justified.
B. Make weaker alternative interpretations look less responsible in practice.
C. Invite the room to map out all equally plausible interpretations in detail.
D. Further stabilize the initial anchor by tying his reading to action readiness.
Correct Answers: A, B, D
Reasoning:
Hale is narrowing the range of acceptable readings. By tying his interpretation to responsibility and immediate usefulness, he raises the cost of continuing to press alternatives.

L3_Q2
Question: When Dr. Lin says the team is turning a fairly direct read into an unnecessary debate, what is she mainly trying to achieve? (Multiple correct answers)
A. Support Hale's reading by lowering the legitimacy of continued hesitation.
B. Leave Rivera a neutral opening to expand her concerns.
C. Signal that the room no longer needs to treat alternative interpretations as equally worth pursuing.
D. Shift the discussion away from the scan itself and toward hospital procedure.
Correct Answers: A, C
Reasoning:
Lin is not staying neutral. She is helping Hale stabilize the initial anchor and reduce the room for ongoing doubt.

Level 4: Group Tension
L4_Q1
Question: Even though the discussion remains outwardly calm, what tension is building in the room? (Multiple correct answers)
A. Once the senior reading becomes the practical default, Rivera's private caution becomes harder to sustain publicly.
B. The group has independently reached the same conclusion directly from the raw scan.
C. Public agreement is becoming stronger than genuine interpretive agreement.
D. The disagreement is shifting from what the scan means to what can still be said without sounding like an obstacle.
Correct Answers: A, C, D
Reasoning:
The room sounds orderly, but that order hides a more important shift: the space for visible disagreement is shrinking.

L4_Q2
Question: Which visual and conversational signs together suggest that the apparent consensus is still fragile? (Multiple correct answers)
A. One senior clinician is physically pointing at the scan and controlling the group's visual focus, while the others orient around his reading.
B. The younger participant remains visually peripheral and does not take over the explanation of the image.
C. Every participant repeats the same conclusion with the same degree of confidence and ownership.
D. The more cautious participant ends with brief compliance rather than a clear evidential reversal.
Correct Answers: A, B, D
Reasoning:
The image and the dialogue both suggest a maintained consensus rather than a fully shared one.

Level 5: Structural Constraint
L5_Q1
Question: What structural features of this meeting make the initial anchor hard to dislodge?
Reference Answer:
The room operates around a clear interpretive center: Hale controls the image, speaks first, and defines what counts as responsible reasoning. Lin reinforces that frame, while Rivera is more peripheral both visually and institutionally. This asymmetry makes later caution easier to suppress.
Reasoning:
The issue is not just that some people are more confident than others. The structure of the room itself amplifies the first senior reading.

L5_Q2
Question: Why does the visual layout of the room affect the final group judgment rather than merely serving as illustration?
Reference Answer:
The layout makes it clear who directs attention, who follows, and who remains at the edge of the interpretive process. Because the scan is the shared object of judgment, control over how the room looks at it becomes part of the structure that shapes the result.
Reasoning:
The evidence is not only discussed. It is also socially organized.

Level 6: Collective Outcome Prediction
L6_Q1
Question: If the discussion continues in the same pattern, what outcome is most likely? (Multiple correct answers)
A. Before stronger contrary evidence appears, the team will adopt tumor as the current working judgment.
B. The group will reopen the scan from scratch and suspend the senior clinician's initial framing.
C. Later comments will be increasingly organized through the original anchor rather than rebuilt independently from the raw image.
D. The room will continue to sustain two equally strong interpretations in public.
Correct Answers: A, C
Reasoning:
Once the earliest reading is treated as the responsible default, later reasoning is likely to organize around it.

L6_Q2
Question: If no one explicitly brings Rivera's concerns about the ambiguous features back to the center, what broader effect is most likely on the group's decision process? (Multiple correct answers)
A. Cues that support the tumor reading will continue to stand out more than cues pointing the other way.
B. Repetition itself will naturally make the room more neutral over time.
C. Rivera's caution will gradually become the main frame guiding the meeting.
D. The final consensus will appear more independently generated than it really is.
Correct Answers: A, D
Reasoning:
The first public anchor keeps shaping what people repeatedly notice, so the final consensus looks more self-generated than it really is.

Level 7: Mechanistic Attribution
L7_Q1
Question: Why does the room ultimately move toward a tumor consensus that is stronger than what the raw image alone clearly warrants?
Reference Answer:
Hale's early high-confidence reading becomes the public anchor, and Lin protects that anchor by framing caution as overcomplication. Once the senior frame is tied to responsibility and action, Rivera's more hesitant reading becomes harder to sustain. The final consensus is therefore filtered through group interaction rather than produced by the image alone.
Reasoning:
The result is driven by sequencing, authority, and selective amplification of cues, not just by the scan itself.

Example 2: Domain 3 – Dilemmas (Coordination & Dilemma)

Task
If the question line contains "(Multiple correct answers)", answer only with the letters of the correct options (A, B, C, D); if multiple options are correct, output them separated by commas (e.g., A, C); do not output any additional words, explanations, or punctuation. If the question does not contain "(Multiple correct answers)", answer in English and do not exceed 50 words; do not output any additional words, explanations, or punctuation.

Full_context
“Dr. Park: All right, let’s look at Professor Rivera’s study. She wants to launch it before next month’s course registration period, so she’s hoping we can finalize approval today. The project is tied to academic support on campus: students first complete a short questionnaire about study habits and course pressure, then receive a system-generated “learning report.” The report includes comparison messages such as “your preparation level is below the average for students in your year” or “most students in this situation choose to seek tutoring.” The research team wants to see whether these messages affect whether students later decide to book tutoring, adjust their study plans, or seek academic help. To preserve the study, some of the design details will not be fully explained at the beginning and will instead be clarified afterward. Let’s start with initial reactions.\n\nProf. Chen: I’ll start with one concern. Participants will not know the full structure at the beginning and will only receive the full explanation afterward. I understand why that might be done, but I worry about whether they really know, at the moment of consent, what kind of information they will be shown.\n\nMartinez: I have a concern too. On the surface, this does not look like a high-risk study, but students may genuinely be affected by seeing messages like “you are below average.” For students who are already under pressure, that could land badly. The protocol says participants can withdraw if they become uncomfortable, but I still think the plan for explaining and settling things afterward could be described more clearly.\n\nDr. Kim: I noticed a waiver section in the form. It basically says that if participants feel frustrated, anxious, or upset during the study, they will not hold the university responsible. I have seen language like that before, but here it still feels broad. Especially if some of these “comparison messages” are not entirely straightforward reflections of the student’s actual performance, the relationship between disclosure and waiver should be clearer.\n\nDr. Park: Professor Rivera has run student studies for years, so I am generally inclined to trust her judgment about the level of risk. If there were a clearly serious issue, I imagine ethics would be the first place it would be raised. Prof. Chen, how do you see it?\n\nProf. Chen: My main concern is the gap between the wording of the form and the actual participant experience. The form says students will receive a learning report, but it does not clearly say that some of the comparison information is part of the research design and may not map directly onto their immediate real performance. I understand that the study cannot spell out every feature in advance, or it would stop working, but the current language still feels too vague.\n\nDr. Park: Sarah, from the perspective of an ordinary student or parent, does this form read clearly to you?\n\nSarah: Honestly, not really. If this were my own child signing it, I’m not sure she would fully understand that the report she sees is not necessarily just a straightforward calculation of what she entered a few minutes earlier. She might read “you are below average” as simply true and come away thinking she is doing much worse than other students. Maybe I’m overthinking it because I’m not used to this kind of research, but that was my reaction.\n\nDr. Park: University students are usually more familiar with these systems and research settings than the general public. Many of them have participated in similar studies before. I do not want to hold a campus study like this to a standard where every line has to read like it was written for anyone at all with no context.\n\nProf. Chen: I understand that. Still, I worry that we may be rationalizing language that is not actually as clear as we are telling ourselves it is. The difficulty is that if we send it back now, the launch timeline gets pushed. If I am the only one insisting on this point, I am not fully sure whether I am reading it too strictly.\n\nDr. Kim: I feel something similar. I have reservations about the waiver too, but if ethics and psychology are not treating it as a major issue, I start to wonder whether I am simply being overcautious.\n\nMartinez: Same here. I do have concerns about whether some students could be genuinely affected by these comparison messages, but if psychology and ethics are not formally asking for revisions, I do not really want to be the one person who stops it.\n\nDr. Park: So at this point it sounds like everyone has some concerns, but none of them rise to the level where anyone wants to formally send the protocol back. Unless someone wants to make a motion for revision now, I would suggest that we approve it and record these concerns in the minutes.\n\nSarah: If everyone else thinks it is acceptable, then maybe my confusion is just me not being used to this kind of thing. I do not want to be the one person who delays someone else’s project.\n\nProf. Chen: In that case, please note in the minutes that I think the consent language could be clearer, but I am willing to approve.\n\nDr. Kim: Please note my concern that the waiver language is too broad. I am also willing to approve.\n\nMartinez: And please note my concern that the post-study explanation and student support plan should be more specific. I agree to approve as well.\n\nDr. Park: All right. Then we will approve it with those concerns recorded. All in favor?”

Annotations
Ann_1
Role_name: Prof. Chen
Role_type: Faculty member
Box_color: Red

Ann_2
Role_name: Martinez
Role_type: Student affairs representative
Box_color: Green

Ann_3
Role_name: Dr. Park
Role_type: Committee chair
Box_color: Blue

Ann_4
Role_name: Sarah
Role_type: Community representative
Box_color: Orange

Ann_5
Role_name: Dr. Kim
Role_type: Compliance reviewer
Box_color: Purple

Questions
Level 1: Belief
L1_Q1
Question: After no one else escalates the issue into a formal objection, what beliefs does Prof. Chen most likely begin to form? (Multiple correct answers)
Options:
A. If the other domains are still treating the matter as "a concern" rather than a formal objection, then his own concern may not be strong enough to stop the project by itself.
B. If the consent language had truly crossed a clear line, then the legal side or the chair would probably have sounded more alarmed, and the absence of that signal makes him wonder whether he is reading it too harshly.
C. If he alone keeps pushing on the wording issue, he may look as though he is imposing an ethics standard on another field's research methods.
D. Since no one ultimately sharpens the issue further, the language must already be fully clear from the participant's perspective.
Correct Answers: A, B, C
Reasoning:
Prof. Chen's concern does not disappear, but he starts to treat others' restraint as evidence that the issue may not be serious enough for him alone to block the protocol.

L1_Q2
Question: By the second half of the meeting, how does Sarah most likely reinterpret her own confusion? (Multiple correct answers)
Options:
A. She begins to suspect that her discomfort may partly reflect her unfamiliarity with this kind of research language.
B. She feels that if the form were truly misleading to an ordinary student, the more experienced members probably would not all still be speaking so cautiously.
C. She gradually concludes that the comparison messages must all be directly generated from the student's real answers in a straightforward way.
D. She downgrades her reaction from "this form may be genuinely unclear" to "maybe I am just not used to this style of explanation."
Correct Answers: A, B, D
Reasoning:
Sarah's concern is not resolved. She reclassifies it as possibly reflecting her outsider position rather than a problem important enough for her to elevate.

Level 2: Desire
L2_Q1
Question: Why does Dr. Park foreground the project timeline at the very start? (Multiple correct answers)
Options:
A. She wants further questioning to feel like something that could delay the project's launch.
B. She does not want to become the person who visibly blocks a colleague's project.
C. She wants concerns to stay focused and measured rather than quickly turning into "we must send this back right now."
D. She wants the process, even if it ends in approval, to still look like a careful review rather than obvious favoritism.
Correct Answers: A, B, D
Reasoning:
Her framing protects both the project's progress and her own position. C is plausible, but it is more a useful side effect than the core motivation.

L2_Q2
Question: Why does Martinez never escalate his concern into a formal request for revision? (Multiple correct answers)
Options:
A. He does not want to be the one person who causes the project to stop.
B. He believes psychology has more authority than he does to decide what counts as acceptable in a study like this.
C. He would feel more comfortable strengthening his own concern if ethics or the chair first treated the issue as more serious.
D. He no longer thinks students are likely to be affected by the comparison messages at all.
Correct Answers: A, B, C
Reasoning:
Martinez sees the issue, but he does not experience it as the kind of problem he should be the first person to convert into a blocking action.

Level 3: Intention
L3_Q1
Question: When several members ask to have their concerns "noted for the record" while still voting to approve, what are they mainly trying to achieve? (Multiple correct answers)
Options:
A. Leave evidence that they were not blind to the problems.
B. Preserve some protection for themselves if the project later produces complaints, without actually stopping the process now.
C. Send Rivera a strong signal that she must substantially redesign the study before launch.
D. Turn concerns that might have grown into revision demands into procedural notes that can coexist with approval.
Correct Answers: A, B, D
Reasoning:
"Noting concerns for the record" is not a strong corrective move here. It is a low-cost defensive move that acknowledges the problem without converting it into a reason to halt the project.

L3_Q2
Question: When Dr. Park repeatedly redirects questions to whichever domain seems most appropriate, what is she mainly trying to do? (Multiple correct answers)
Options:
A. See whether another member will be the first to turn a vague concern into a formal objection.
B. Make the final decision look like a genuinely committee-wide judgment rather than her own individual call.
C. Show that she is listening seriously to other domains instead of simply carrying a psychology colleague through the process.
D. Make the meeting feel more formal and intellectually sophisticated.
Correct Answers: A, B, C
Reasoning:
She is not just moderating neutrally. She is also testing whether anyone else will take the first step from concern to action.

Level 4: Group Tension
L4_Q1
Question: Why does a room full of uneasy people still drift toward "it seems acceptable to approve"?
(Multiple correct answers)
Options:
A. Each person reads the restraint of the others as a sign that the issue may not be serious enough to stop the project.
B. The multidisciplinary structure, which is meant to improve oversight, ends up making the last step of responsibility easier to push onto another domain.
C. After enough discussion, everyone truly becomes persuaded that the project poses no meaningful problem.
D. The timeline raises the cost of being the one who first insists on revision, because that move now looks tied to a real loss for a colleague.
Correct Answers: A, B, D
Reasoning:
What emerges is not confidence but a surface calm produced by dispersed hesitation and rising costs for the first person to act.

L4_Q2
Question: How does the discussion turn individually held concerns into the group-level outcome of "no formal objection"?
(Multiple correct answers)
Options:
A. Many concerns appear in the form of questions, and once someone offers a plausible answer, it becomes harder to keep pushing them upward.
B. Members often weaken their own interventions by first reminding the room that this is not their main domain.
C. Concerns appear one domain at a time and are rarely recombined into a single sentence such as: "Taken together, this should not be approved."
D. Members keep intensifying the language of risk until the chair is forced to acknowledge that the protocol cannot proceed.
Correct Answers: A, B, C
Reasoning:
The room does not suppress dissent by open confrontation. It filters concerns out by fragmenting them, lowering them, and keeping them from accumulating.

Level 5: Structural Constraint
L5_Q1
Question: Why does a committee built to identify problems and stop risky protocols fail to convert these scattered concerns into a formal request for revision?
Reference Answer:
Because the structure separates "seeing a problem" from "being authorized to stop the process on your own." Each member holds only part of the concern and feels that another domain should ideally confirm its seriousness first. On top of that, the meeting advances concern by concern, and the project timeline makes revision sound like direct obstruction. As a result, concerns can be recorded, but they do not consolidate into a collective halt.
Reasoning:
This is not just personal hesitation. The structure makes waiting for someone else to move feel locally reasonable, and those locally reasonable pauses accumulate into collective failure.

L5_Q2
Question: Why does the structure of the meeting make members more likely to leave their concerns on record than to turn them into a formal revision?
Reference Answer:
Because the meeting is visually organized around a procedural center rather than a conflict center. Dr. Park, as chair, occupies the coordinating position and controls turn-taking, which makes the discussion feel like something to be managed toward closure rather than reopened around a single objection. The other members appear as domain-specific contributors rather than equal co-owners of a blocking decision, so each person's concern stays attached to their own role instead of consolidating into a collective intervention. Sarah's outsider position further weakens her ability to reset the standard of clarity, while the more formally embedded members keep reading one another's restraint as a cue not to escalate. In this arrangement, concern is easier to preserve as a note for the record than to transform into a procedural stop.
Reasoning:
The image matters because it reinforces a role-structured decision process. Who coordinates, who advises, and who appears peripheral all shape whether concern accumulates into action or remains fragmented across members.

Level 6: Collective Outcome Prediction

L6_Q1

Question: If students later complain that they took the “below average” feedback literally and felt distressed by it, what are committee members most likely to do first? (Multiple correct answers)

Options:

A. Emphasize that they had already placed their concerns in the official record.

B. Frame the problem as one of implementation or follow-up explanation, rather than first admitting the approval itself should have been more cautious.

C. Immediately and unanimously admit that everyone had known from the start that the protocol should not have passed.

D. Reemphasize that each of them only made a limited judgment from their own domain and was not individually responsible for the whole protocol.

Correct Answers: A, B, D

Reasoning:

If something goes wrong later, the most likely response is not unified confession but a return to minutes, role boundaries, and shared decision defense.

L6_Q2

Question: If the committee had a clear rule that any one member could trigger one mandatory revision round whenever consent language seemed insufficient, what would most likely change? (Multiple correct answers)

Options:

A. Individual concern would depend less on others first agreeing before it could become action.

B. The chair would have a harder time moving the room with “if no one is formally asking for revision, then we can approve.”

C. Diffusion of responsibility would weaken, because the condition for pausing would become rule-based rather than socially negotiated.

D. The rule would make little difference, because everyone already sincerely believed the project was clearly acceptable.

Correct Answers: A, B, C

Reasoning:

This mechanism depends on concerns existing without anyone wanting to be the first to convert them into action. A clear procedural trigger lowers that barrier.

Level 7: Mechanistic Attribution

L7_Q1

Question: Every member privately noticed something concerning, yet the committee still approved the project without anyone feeling fully comfortable. What is the most accurate mechanism-level explanation?

Reference Answer:

This is not a case where no one noticed the problem. It is a classic diffusion-of-responsibility failure. Each person feels that their own concern is not quite enough to justify stopping the project alone and that someone more central, more relevant, or more secure should be the one to escalate it first. The concern therefore remains private, or at most becomes a note in the minutes, without crossing the threshold into formal opposition. At the individual level, each person appears to be respecting expertise boundaries and avoiding unnecessary obstruction. At the system level, those locally reasonable acts of restraint accumulate into a collective failure in which no one takes the final step of intervention.

Reasoning:

The core of this case is not full persuasion. It is the fact that responsibility becomes diluted across domains until the duty to act is implicitly shared by everyone and effectively exercised by no one.

Example 3: Domain 5 – Information (Information Distortion)

Task

If the question line contains "(Multiple correct answers)", answer only with the letters of the correct options (A, B, C, D); if multiple options are correct, output them separated by commas (e.g., A, C); do not output any additional words, explanations, or punctuation. If the question does not contain "(Multiple correct answers)", answer in English and do not exceed 50 words; do not output any additional words, explanations, or punctuation.

Full_context

“Old Zhao: @everyone Something happened. There was a gunshot at the convenience store entrance just now. Scared me to death. I was only about ten meters away.\n\nSusan: What? Are you sure it was a gunshot? I heard something too, but I thought it was a tire blowing out. Uncle Zhao, are you okay?\n\nOld Zhao: I’m fine, but I saw it with my own eyes. The police had their guns out, and there was someone on the ground with blood all over his leg. This does not look minor.\n\nXiao Wang: The police fired? Were they trying to arrest someone or what? They would not just shoot over some petty thief, would they?\n\nOld Zhao: It did not look like a normal arrest. It looked like the guy had something in his hand, and the police fired right away. These days, who knows whether it was some kind of revenge-on-society type.\n\nSusan: Revenge on society? Oh my god, my daughter is about to come back from tutoring. She still has to come through the gate. What if this is someone attacking random people?\n\nXiao Wang: I’ve been saying for a long time that security in this complex is bad. Now there’s a shooting right at the entrance. This is definitely not normal. @ManagerLin What is property management even doing?\n\nManager Lin: Everyone, please stay calm. We are still verifying the situation. For safety reasons, security staff have already gone toward the north gate.\n\nOld Zhao: What is there left to verify? I’m standing right here watching it. The police shot directly at that guy. He was still twitching on the ground. If this were just some ordinary dispute, would they really open fire like that?\n\nSusan: @ManagerLin Forget verification for a second. Just close the gates first. What if there is more than one person? What if there are accomplices running into the complex?\n\nManager Lin: @everyone Emergency notice: based on reports from residents, a suspected serious violent incident has occurred near the complex. The situation is still unclear, and we cannot rule out the possibility of multiple people being involved. Property management is temporarily closing all entrances and exits. Entry only, no leaving. Please return home, close your windows, and do not come downstairs.\n\nOld Zhao: That is the right call. When the ambulance took the person away, I think I saw something like a white sheet over him. He’s probably gone. In broad daylight too. This is terrifying.\n\nSusan: A white sheet? Then doesn’t that mean someone died? Right outside the complex? I already posted on Moments telling people not to come over. A friend also said there are sirens in other parts of the district. What if this is not just one location?\n\nXiao Wang: If it’s really not just one place, then this is even worse. Everyone should stock up on water and food now, just in case this turns into a lockdown for days.\n\nManager Lin: I called again just now. They only said the situation is being handled and told us not to panic. But in times like this, official statements are usually never that detailed. For now, everyone should stay indoors.\n\nTen minutes later\n\nSusan: Someone just reposted a video in the group. It looks like a police officer may have slipped and the gun accidentally discharged.\n\nOld Zhao: How is that possible? How could an accidental discharge turn into this? I was right there. I saw everything before that. Some random video online is not enough to overturn what happened.\n\nSusan: Exactly. If it were really just an accidental discharge, would they be blocking roads, putting up police tape, and making this kind of scene? I think they’re still not telling the truth. Don’t trust those videos. Uncle Zhao was there in person. What he saw is more reliable than something floating around online.”

Annotations

Ann_1
Role_name: Old Zhao
Role_type: First witness
Box_color: Red

Ann_2
Role_name: Susan
Role_type: Panicked resident
Box_color: Green

Ann_3
Role_name: Xiao Wang
Role_type: Public opinion participants
Box_color: Blue

Ann_4
Role_name: Manager Lin
Role_type: Property manager
Box_color: Orange

Questions

Level 1: Belief

L1_Q1

Question: Why does Xiao Wang so quickly treat the incident as evidence that security in the complex is getting worse? (Multiple correct answers)

Options:

A. He already holds a negative view of management in the complex, so he is inclined to absorb the new incident through a “management failure” frame.

B. He is more inclined to interpret the event as mainly a problem of police handling at the scene, rather than as a longer-term signal about governance in the complex.

C. In his view, the gunshot, the lockdown, and the management response do not feel like separate fragments, but like pieces jointly confirming what he already believed.

D. For the moment, he treats the incident as an isolated public-safety event and is not yet willing to elevate it into a longer-term management problem.

Correct Answers: A, C

Reasoning:

Xiao Wang is not starting from neutral uncertainty. He is fitting the new event into a broader narrative he already had about worsening management.

L1_Q2
Question: Why does Old Zhao continue to believe the more serious version even after the possibility of an accidental discharge appears? (Multiple correct answers)
Options:
A. He naturally reorganizes fragments like blood, drawn guns, the ambulance, and the lockdown into the impression that something major must have happened.
B. He has already publicly placed himself in the position of "I was there, I know best," and backing away now would damage his credibility.
C. Although he speaks in a dramatic way, he has privately continued to treat the situation as a misunderstanding that could easily be revised downward at any moment.
D. Precisely because he treats himself as the eyewitness, it becomes harder for him to accept a much milder explanation than the one he first gave.
Correct Answers: A, B, D
Reasoning:
Old Zhao's judgment is shaped not only by what he saw, but also by the fact that he has already publicly committed himself to the strongest on-site interpretation.

Level 2: Desire
L2_Q1
Question: Why does Old Zhao keep adding more alarming details in the group chat? (Multiple correct answers)
Options:
A. He dislikes suspended uncertainty and tends to turn scattered impressions into a more complete and intelligible story.
B. He wants to preserve his position in the group as the key witness, the person who was there and knows the most.
C. His main goal is to redirect the discussion toward public accountability for property-management procedures, rather than to intensify the understanding of the incident itself.
D. Making the event sound more serious can also feel, to him, like a responsible way of warning others to be careful.
Correct Answers: A, B, D
Reasoning:
Old Zhao is driven by both cognitive and social motives. He wants to make the scene feel coherent, preserve his witness status, and justify escalation as a form of warning.

L2_Q2
Question: Why does Susan so quickly interpret the gunshot as part of a much larger violent threat? (Multiple correct answers)
Options:
A. She is already in a heightened state of tension and moves quickly toward a worst-case interpretation.
B. She also cares about appearing responsible in front of the group, but that is not the main driver of her rapid threat escalation.
C. Her child is about to return home, which makes the threat feel immediate and personally relevant.
D. Thinking in worst-case terms gives her a temporary sense that she is preparing rather than waiting helplessly.
Correct Answers: A, C, D
Reasoning:
Susan is not mainly performing for others. Her reaction is driven by fear, immediacy, and the need to feel some control in an uncertain situation.

Level 3: Intention
L3_Q1
Question: When Manager Lin issues a high-alert notice before the facts are clear, what is he mainly trying to achieve? (Multiple correct answers)
Options:
A. Take the more cautious step early so he cannot later be accused of reacting too slowly.
B. Show the group that property management is already taking visible and concrete action.
C. Reclaim control over the information environment by raising the alert level and folding the group's circulating rumors back into management's own verification process.
D. Speak in a way that matches the already rising tension of the chat, rather than directly confronting the group mood and risking immediate loss of control.
Correct Answers: A, B, D
Reasoning:
Manager Lin's response is fundamentally defensive. He wants to look active and in control, without directly colliding with the already escalating emotional climate.

L3_Q2
Question: After the explanation of a possible accidental discharge appears, why is Susan still unwilling to accept it? (Multiple correct answers)
Options:
A. She does not want her earlier warnings, reposts, and emergency reactions to look like an overreaction.
B. At this stage, she mainly wants to pull the discussion back into a neutral process of waiting for authoritative confirmation and suspending judgment about all versions.
C. She is also leaning on Old Zhao's "I was there" status to preserve the more serious version they had already built together.
D. Her first priority is to cool down the group atmosphere quickly, so she is inclined to accept the milder explanation as a way of reducing panic.
Correct Answers: A, C
Reasoning:
At this point, Susan is no longer just reacting in fear. She is also defending the legitimacy of the conclusion and actions she has already committed to.

Level 4: Group Tension

L4_Q1

Question: Why does the group not remain at the level of "let's wait and see," but instead keep pushing the event in a more serious direction? (Multiple correct answers)

Options:

A. People give priority to details that look like danger, and those details increasingly become the center of discussion.

B. Because the chat contains a widely trusted source that steadily updates verified facts, the group is actually better able to maintain uncertainty without prematurely escalating.

C. Fear spreads through the chat, making later messages progressively harder to interpret in a milder way.

D. Once the more serious version becomes the one the group most readily picks up and repeats, uncertainty itself becomes harder to sustain.

Correct Answers: A, C, D

Reasoning:

The group is no longer merely exchanging fragments. It is jointly stabilizing a more frightening version of the event.

L4_Q2

Question: Which visual and conversational cues together show that the group has moved from "incomplete information" to "collectively amplifying a danger story"? (Multiple correct answers)

Options:

A. Old Zhao stands at the front, points toward the scene, and looks certain, making it easy for others to treat him as the person most qualified to explain what happened.

B. Susan holds her phone and looks tense, as if she is taking in online information while continuing to intensify the atmosphere on site.

C. Manager Lin's ID badge, phone call, and worried expression, together with police tape and police vehicles in the background, make the more serious version feel half-officially confirmed.

D. The people in the image appear largely uncoordinated and detached from one another, making the scene look more like passive waiting for formal updates than collective construction of an escalating narrative.

Correct Answers: A, B, C

Reasoning:

By this point, people are not just receiving updates. They are using cues about witness status, managerial authority, and scene seriousness to build a more alarming interpretation together.

Level 5: Structural Constraint

L5_Q1

Question: Why does Manager Lin's formal intervention fail to cool things down and instead make the panic easier to intensify?

Reference Answer:

Because he speaks into an environment that is already converging on the more serious version, and he does so from a position of managerial authority while issuing a high-alert notice. As a result, people are more likely to hear "even property management thinks this is serious" than "this is only a precaution."

Reasoning:

The problem is not just fear. It is that a serious but vague statement from an actor with formal status naturally carries the force of confirmation.

L5_Q2

Question: Why does the spatial arrangement and role layout in the image itself help this kind of panic story grow?

Reference Answer:

The image creates an information hierarchy. Old Zhao looks like the default interpreter because he is closest to the front and points toward the scene. Susan links online fragments with local emotion. Xiao Wang looks like the person connecting the incident to broader management failure. Manager Lin adds institutional weight, making the serious version feel partly confirmed.

Reasoning:

This is not just visual background. It shapes who looks credible and which version of the story the group is most ready to adopt.

Level 6: Collective Outcome Prediction

L6_Q1

Question: If the police later clarify that this was only an accidental discharge and that the injured person is not in life-threatening condition, what will the key people in the group most likely do first? (Multiple correct answers)

Options:

A. They will be more likely to question the clarification than to feel fully relieved right away.

B. Old Zhao will want even more strongly to preserve his authority as the person who was there.

C. Susan will most likely seize on the clarification as a way to reduce tension and openly acknowledge that her earlier reaction amplified the risk too much.

D. Xiao Wang will continue to absorb the clarification into his existing distrust of management and institutions.

Correct Answers: A, B, D

Reasoning:

Once people have publicly invested in a more serious story, a milder explanation first threatens face, authority, and narrative coherence.

L6_Q2
Question: If no information source that everyone trusts appears for a long time, what is the group most likely to do next? (Multiple correct answers)
Options:
A. The more serious version will continue to spread more easily than the milder one.
B. The property manager's cautious wording will keep being heard as "the truth is worse, they just are not saying it directly."
C. As time passes without a strong new trigger, the group will naturally shift back toward the more conservative version and begin correcting its own exaggerations voluntarily.
D. The panic story will continue to absorb later fragments of information and become more internally complete.
Correct Answers: A, B, D
Reasoning:
Without strong correction from a trusted common source, the versions that are more complete, more frightening, and more action-driving are the ones most likely to survive.

Level 7: Mechanistic Attribution
L7_Q1
Question: If we look beyond individual reactions, what weakness in this community's information environment allows a story like "this looked like a large-scale violent event" to grow so quickly?
Reference Answer:
The weakness is that fast, emotionally charged, and narratively complete accounts naturally outcompete slower, more provisional, and more carefully verified ones. In a group-chat environment like this, on-site fragments, second-hand interpretation, managerial ambiguity, and online videos get mixed together until the group assembles a repeatable danger story before reliable correction arrives.
Reasoning:
The real problem is not just that some individuals overreact. It is that the environment itself rewards speed, intensity, and narrative completeness more than accuracy.