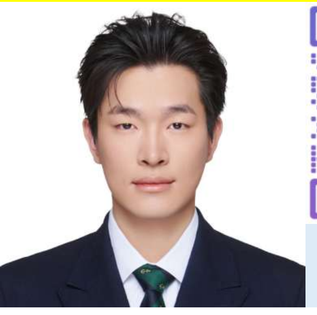
Looking for PhD Position



WeChat Rednote

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

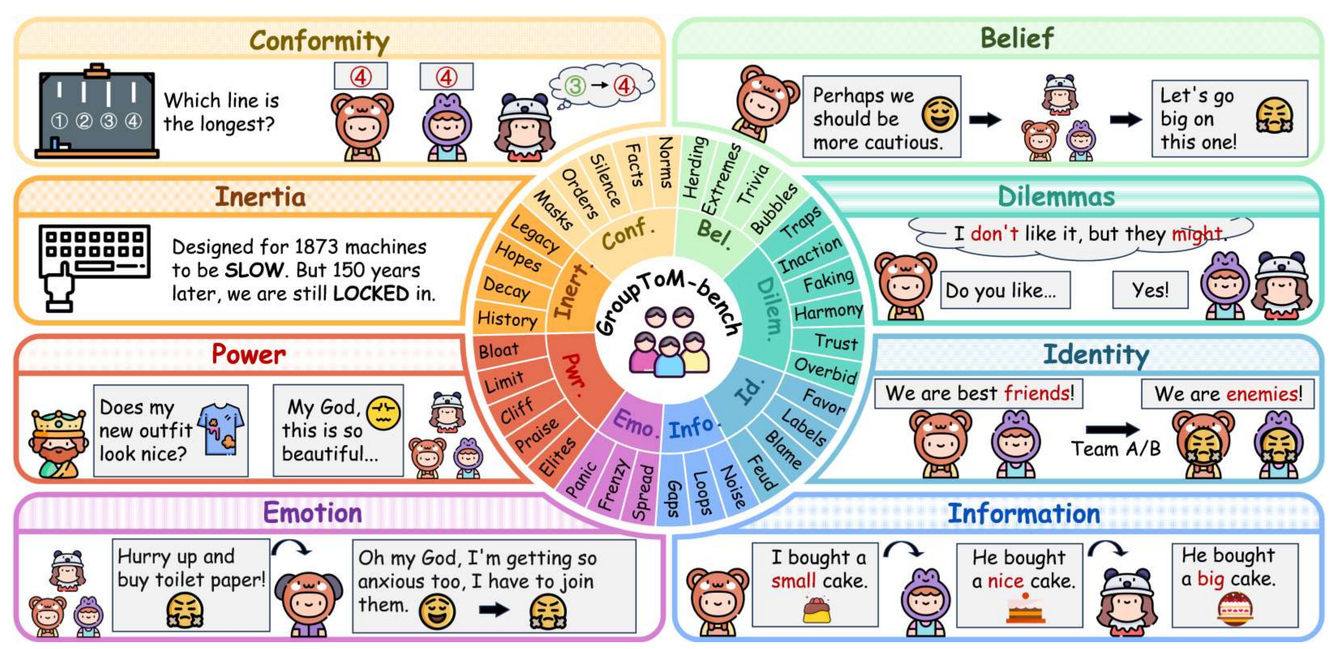
Weidong Tang, Jierui Li, Yueling Hou, Zihan Mei, Can Zhang,Xinyan Wan,
Zhiyuan Liang, Pengfei Zhou†, Yang You, Wangbo Zhao†

Contact: wdtang0705@gmail.com Homepage: wd7ang.github.io

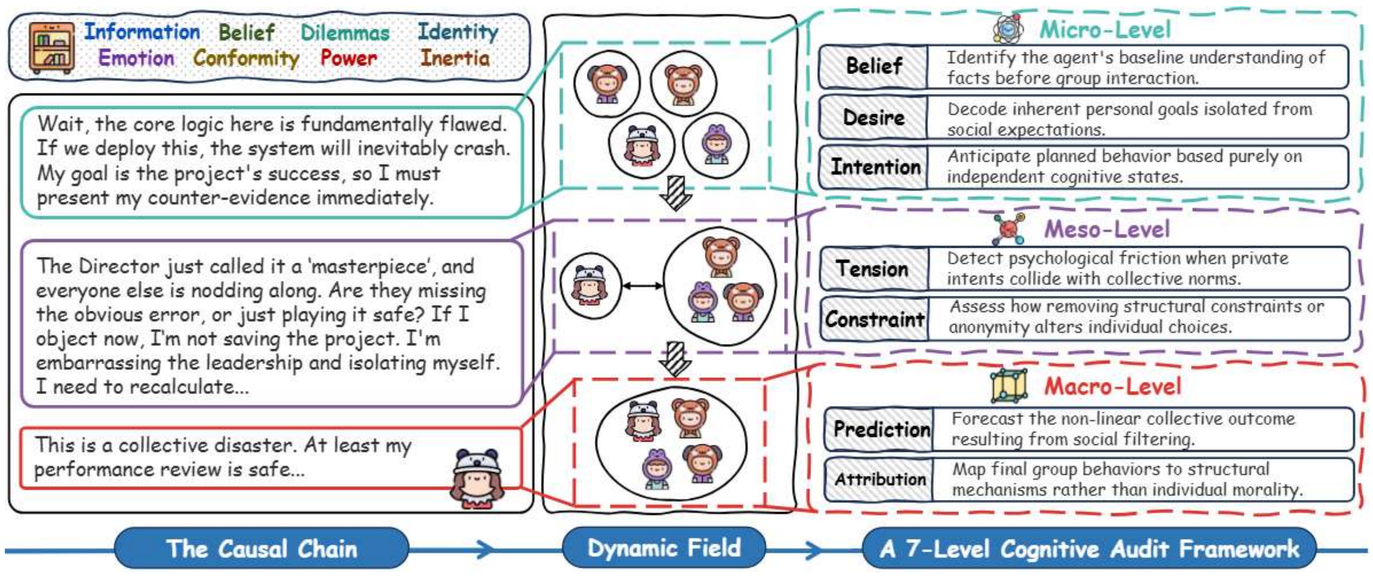
Introduction

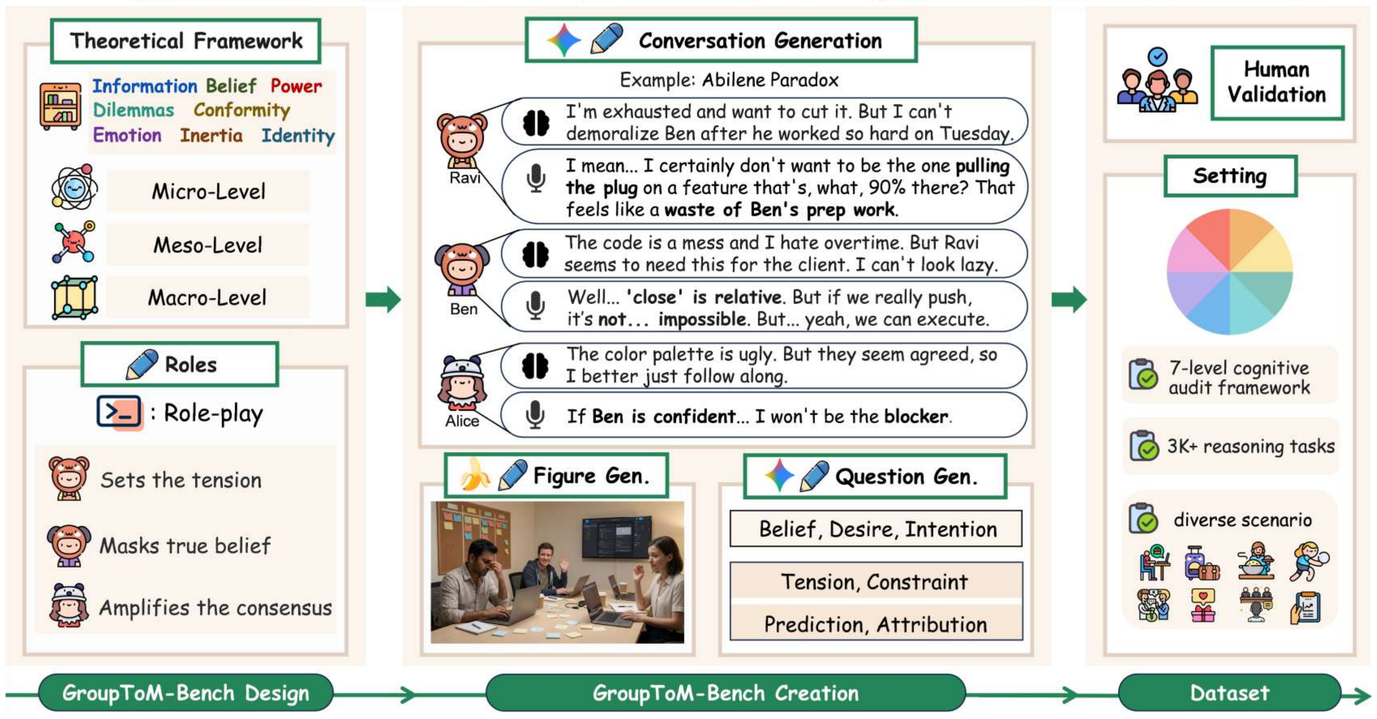
- True general intelligence requires not only modeling the physical world, but also understanding the social world: the ability to infer how individual mental states interact and give rise to group-level outcomes. While multimodal large language models have made progress in individual-level Theory of Mind (ToM) reasoning, they struggle with this broader, collective task. Collective behavior emerges non-linearly from social tensions, conformity dynamics, and structural constraints, and cannot be captured by simply summing individual intentions.

- We introduce **GroupToM-Bench**, the first multimodal benchmark for group-level ToM. It traces a causal chain from micro-level BDI states (belief, desire, intention), through meso-level group tensions and structural constraints, to macro-level outcome prediction and mechanistic attribution. Using a seven-level cognitive audit framework, experiments reveal a significant gap between current models and human baselines, highlighting their inability to fully process social structures and non-linear collective dynamics.



Methodology





Overview of the dataset construction pipeline for GroupToM-Bench

Micro-Level — **Meso-Level** — **Macro-Level**

- Independent intentions derived from individual cognition
- Psychological tension arising from social pressure and collision
- Non-linear collective outcomes emerging through structural filters.

Insight

- **The "Group Cognitive Gap":** Current MLLMs exhibit a Linear Superposition Bias. While they can recover individual BDI states, they fail to model the non-linear "collapse" into group outcomes, treating social emergence as a simple sum of individual parts.
- **From Interaction to Structure:** Moving beyond "Stanford Town" (where agents use simple communication for cooperation), our findings suggest that as agent density and social complexity scale, Social Structures (e.g., Power, Norms) become the primary drivers of behavior, often overriding an agent's internal BDI logic.
- **Visual-Social Blindness:** Models struggle to integrate subtle multi-modal cues (like hesitant micro-expressions or spatial positioning) to uncover "false consensus," relying instead on surface-level textual politeness.

Experiment

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	**91.7**	76.7	70.4	63.3	**79.8**	78.9	75.1	66.0	**73.3**	65.8	58.3	66.5
	L-2	**90.5**	74.1	70.3	64.8	75.3	**77.1**	73.2	62.5	**68.8**	58.2	54.9	60.7
	L-3	**88.4**	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	**89.4**	50.5	49.4	38.0	50.3	**53.1**	50.2	**39.8**	37.3	36.4	26.2	33.1
	L-5	**90.1**	56.9	52.9	41.1	47.2	**59.7**	46.7	42.8	**47.8**	36.6	35.1	41.4
	L-6	**89.2**	45.0	42.5	32.5	**48.6**	48.3	44.1	30.1	**34.3**	31.7	17.2	26.2
	L-7	**88.1**	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

Discussion

- **Sociological Scaling:** Increasing agent counts shift behavior from individual logic to systemic Macro-constraints. AI needs a "Theory of Society" beyond simple ToM.
- **Consensus Trap:** Models suffer from "Optimistic Bias," failing to predict non-linear social collapses like Groupthink due to alignment-induced conservatism.
- **Next Frontier:** Transitioning from agent simulation to Structure-aware Intelligence, modeling how social rules (hierarchy/asymmetry) distort private intent.

ACL 2026 SAN DIEGO JULY 2–7