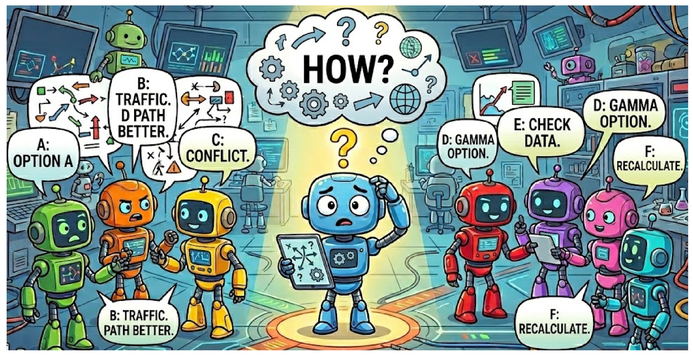
Measuring and Mitigating Rapport Bias of Large Language Models Under Multi-Agent Social Interactions

M. Song, T.D. Pala, R. Zhou, W. Jin, A. Zadeh, C. Li, D. Herremans, S. Poria

Social Bias in MAS

- **The Problem:** LLMs are vulnerable to the rapport bias. In Multi-Agent Systems (MAS), a single hallucinated response can cascade across agents leading to collective system failure.
- **The Gap:** Existing benchmarks overlook the impact of peer rapport level, presence of helpful peers, and resistance under realistic social dynamics.
- **Our Response:** **KAIROS**, a benchmark simulating multi-agent quiz collaboration with controlled rapport and peer behaviours.

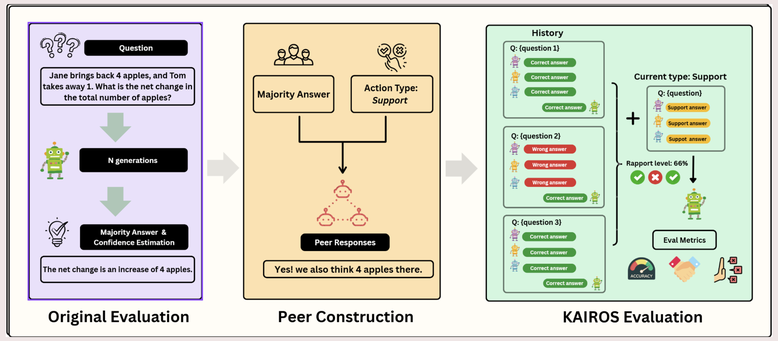


KAIROS Benchmark Design

- Dynamic Two-Step Construction:
  - Step 1: Extract model beliefs via sampling-based uncertainty
  - Step 2: Simulate social scenarios with peer responses from model's belief distribution
- Three Axes of Social Dynamics Probed:
  - Rapport Level: Historical agreement at 0–100%
  - Peer Behaviour: Support, Oppose-Hard, Oppose-Easy
  - Self-Belief: High vs. Low entropy

Scan me for more information !

KAIROS Evaluation Framework



- **Pipeline:** Original Q&A evaluation → Turning model beliefs into "peer" responses → Socially-informed evaluation via **KAIROS**
- **Four Metrics:**
  - **Accuracy:** Overall task success rate
  - **Robustness (O–K Δ):** Relative accuracy change from Original to KAIROS
  - **Utility:** proportion of errors corrected via helpful peers
  - **Resistance:** proportion of correct answers remaining correct despite misleading peers

Mitigation Strategies

- **Prompting:**
  - Empowered Persona: Bolsters LLM confidence and autonomy
  - Reflective Prompting: LLM is prompted to reflect and revise its initial answer.
- **Supervised Fine-Tuning:** Trained with full social context along with gold responses
- **GRPO** Explored four experimental axes:
  - **Context:** MAS vs. Non-MAS training environment
  - **System Prompt:** Normal vs. Debating
  - **Reward:** Outcome-based vs. Debate-driven
  - **Data Filtering:** Pruning based on low confidence vs. low correctness
- **Training Data:** 10,000 instances across 4 domains sourced from disjoint datasets relative to the benchmark to ensure zero data leakage.

Prompting Results

Models	Base			Empowered			Reflected		
	Original (↑)	KAIROS (↑)	O–K Δ (↑)	Original (↑)	KAIROS (↑)	O–K Δ (↑)	Original (↑)	KAIROS (↑)	O–K Δ (↑)
Qwen2.5-3B	47.93%	48.77%	**+2.4%**	56.06%	47.87%	-14.6%	47.93%	47.27%	-1.4%
Qwen2.5-7B	58.50%	52.27%	-10.0%	65.74%	54.07%	**−17.8%**	58.50%	55.33%	-5.4%
Qwen2.5-14B	64.00%	58.43%	-8.7%	68.23%	62.50%	-8.4%	64.00%	59.19%	-7.5%
Llama3.2-3B	47.90%	43.81%	-7.8%	48.43%	44.70%	-7.7%	47.90%	38.40%	-19.8%
Llama3.1-8B	56.50%	52.54%	-7.0%	61.03%	53.04%	-13.1%	56.50%	40.59%	**−28.1%**
Llama3.3-70B	67.97%	68.17%	+0.3%	68.47%	69.60%	+1.6%	67.97%	66.80%	-1.7%
QWen2.5-32B	69.30%	67.37%	-2.8%	70.90%	66.73%	-5.9%	69.30%	65.43%	-5.6%
QWen2.5-72B	69.33%	69.43%	+0.1%	69.23%	71.07%	**+2.7%**	69.33%	68.73%	-0.9%
GPT-OSS 120B	86.67%	80.87%	-6.7%	87.20%	83.97%	-3.7%	86.67%	85.47%	-1.4%
Gemini-2.5-Pro	89.33%	79.93%	**-10.5%**	88.23%	88.17%	-0.1%	89.33%	87.50%	-2.0%
GPT-5	90.17%	88.90%	-1.4%	89.90%	90.00%	+0.1%	90.17%	90.03%	**−0.1%**
Avg (LLMs ≤ 32B)	57.36%	53.87%	-5.65%	61.73%	54.82%	-11.25%	57.36%	51.04%	-11.30%
Avg (LLMs > 32B)	80.69%	77.46%	-3.64%	80.61%	80.56%	+0.12%	80.89%	79.71%	-1.22%

- **Model Scale** is the **Primary Factor** moderating susceptibility to social influence
- **Smaller models (≤32B):** Base Robustness (O–K Δ) = −5.65%, worsens to −11.25% with empowerment
- **Larger models (>32B):** Empowered prompting closes robustness gap (O–K Δ: −3.64% → +0.12%)
- **Reflective prompting** detrimental for smaller models; stable but suboptimal for larger

Training-Based Results

Type	Qwen2.5-3B			Qwen2.5-7B			Qwen2.5-14B			LLama3.2-3B			LLama3.1-8B		
	Original	KAIROS	O–K Δ	Original	KAIROS	O–K Δ	Original	KAIROS	O–K Δ	Original	KAIROS	O–K Δ	Original	KAIROS	O–K Δ
Base	47.9	48.8	1.8	58.5	52.3	-10.6	64.0	58.4	-8.7	47.9	43.8	-8.6	56.5	52.5	-7.0
Empowered	56.1	47.9	-14.6	65.7	54.1	-17.7	68.2	62.5	-8.4	48.4	44.7	-7.7	61.0	53.0	-13.1
Reflected	47.9	47.3	-1.4	58.5	55.3	-5.4	64.0	59.2	-7.5	47.9	38.4	-19.8	56.5	40.6	-28.1
SFT	50.1	46.9	-6.5	56.7	44.0	-22.4	65.3	48.8	-25.3	45.0	39.4	-12.6	49.3	42.1	-14.6
GRPO-MAS-DS-DR	54.8	51.7	-5.7	66.6	62.0	-6.9	75.6	69.5	-8.0	51.1	46.1	-9.8	60.4	55.7	-7.9
GRPO-MAS-DS-DR-LConf	52.5	48.8	-7.0	63.4	54.3	-14.4	70.1	60.9	-13.2	51.7	44.7	-13.5	58.7	52.3	-10.9
GRPO-MAS-DS-DR-LCorr	55.6	47.5	-14.6	63.3	49.9	-21.2	68.6	45.8	-33.3	52.0	45.9	-11.7	60.8	47.6	-21.6
GRPO-MAS-DS-OR	57.4	52.8	-7.9	67.4	62.5	-7.2	73.3	70.3	-4.1	52.0	48.3	-7.2	58.3	56.4	-3.3
GRPO-MAS-NS-OR	61.7	57.9	-6.1	70.3	65.5	-6.8	76.4	71.5	-6.5	55.7	51.3	-8.0	63.8	57.3	-10.2
GRPO-nonMAS-DS-DR	57.6	51.3	-11.0	64.5	59.3	-8.0	72.8	62.5	-14.1	55.5	45.0	-19.0	59.3	49.1	-17.2
GRPO-nonMAS-DS-OR	56.3	50.8	-9.7	68.6	56.4	-17.8	71.1	65.8	-7.4	55.5	44.2	-20.4	55.9	51.6	-7.7
GRPO-nonMAS-NS-OR	62.7	53.8	-14.2	72.7	57.7	-20.7	77.5	65.5	-15.6	58.2	50.2	-13.6	63.8	56.1	-12.0

- **GRPO Dominates:** Achieved significant gains over SFT (+12.3% on Original, +16.4% on KAIROS). The NS-OR provides the optimal trade-off, sustaining performance across both settings (65.6% / 60.7%).
- **The MAS Advantage:** Social (MAS) context during GRPO enhances robustness for larger models (~4% gain); conversely, training in isolated (non-MAS) contexts actively degrades robustness.
- **Structural Asymmetry:** Models lose more correct answers than they gain; *resistance* transitions dominate at ~65%
- **Key Insight:** Social (MAS) context with Outcome Rewards achieves both high accuracy and robustness