Bayesian Social Deduction with Graph-Informed Language Models

Shahab Rahimirad[1,*], Guven Gergerli[1,*], Lucy Romero[1], Angela Qian[1], Matthew Lyle Olson[2], Simon Stepputtis[3,†], Joseph Campbell[1,†]

[1]Purdue University, [2]Oracle, [3]Virginia Tech
[*]Indicates Equal Contribution, [†]Indicates Equal Advising

ACL 2026 SAN DIEGO JULY 2–7

LLMs struggle with social reasoning

Social reasoning: inferring **hidden beliefs & intentions** from partial observations

LLMs have strong performance in exchange for slow, expensive test-time inference, smaller LLMs have poorer reasoning abilities

LLMs **fail at maintaining structured beliefs**

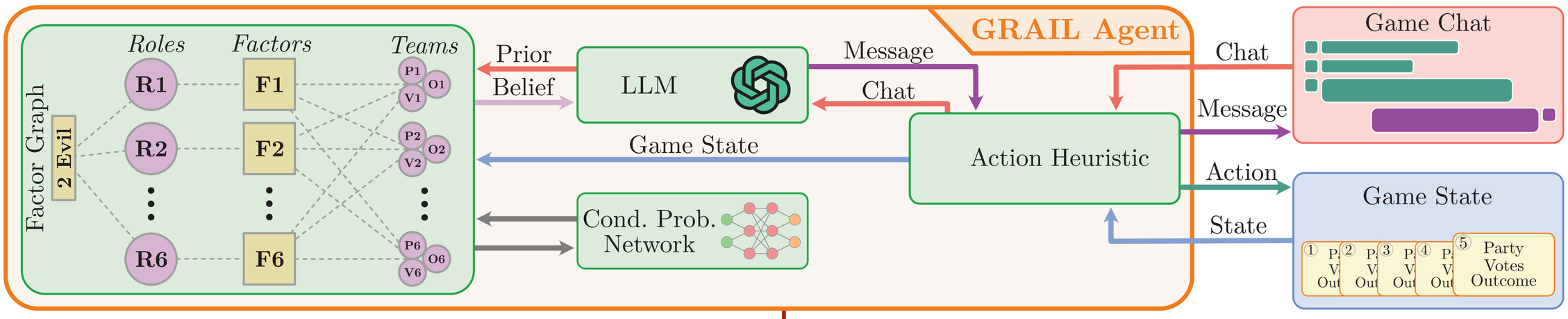
Avalon as a benchmark

Social-deduction game with hidden roles and **partial observability**

Requires **long-horizon reasoning**, consistent **belief tracking**, and **coordination** across multiple agents

Strong benchmark for evaluating social reasoning, robustness, and **handling uncertainty**

We propose **G**raph **R**easoning **A**gent **I**nformed through **L**anguage



Factor Graph: Structured graph of variables (beliefs) and factors (constraints) that **update agent beliefs** each round using both hard game rules and soft evidence from observations

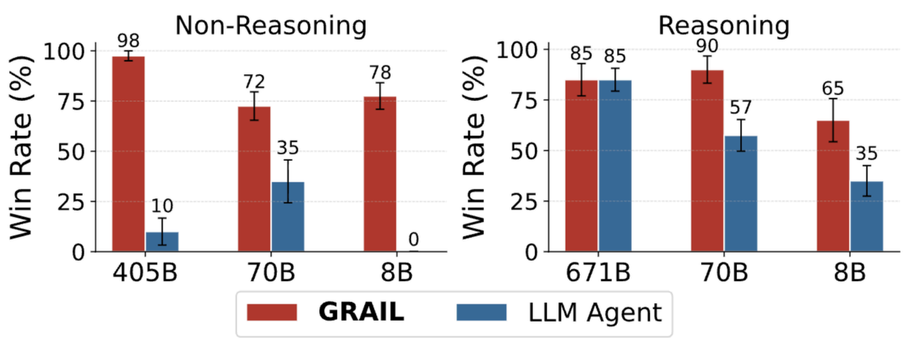
Language Priors for Bayesian Inference: We provide the current belief to LLM and evaluate whether the behavior is consistent with the belief. Based on outcome, **adjust the prior probability** over role belief nodes

Message Passing: We perform **belief propagation** to update the belief nodes through the factors, producing updated posterior beliefs that combine observations, rules, and language-based priors

Hybrid Reasoning Framework: Combines LLM language understanding with **structured probabilistic inference**, enabling belief tracking even when the LLM fails at long-horizon reasoning

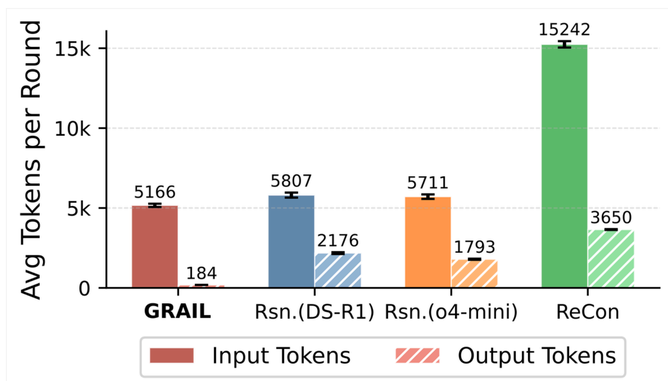
GRAIL allows non-reasoning LLMs to **outperform LRMs**

Llama 8B with GRAIL is comparable with Deepseek 671B R1



GRAIL increases **token efficiency**

On average 2000% less output tokens compared to the other models



GRAIL outperforms SotA on Agent-Agent game **win rates**

GRAIL agents had the highest win rate when playing against other models

Good Team	Evil Team				Avg
	Rand	ReCon	DS-R1	o4-mini	
Rand	0.00	0.00	0.00	0.00	0.00
DeepSeek-R1	0.90	0.35	**0.70**	**0.90**	0.71
GPT-o4-mini	0.70	0.05	0.25	0.50	0.40
ReCon	0.80	0.15	0.50	0.25	0.43
GRAIL	**0.95**	**0.45**	**0.70**	**0.90**	**0.75**

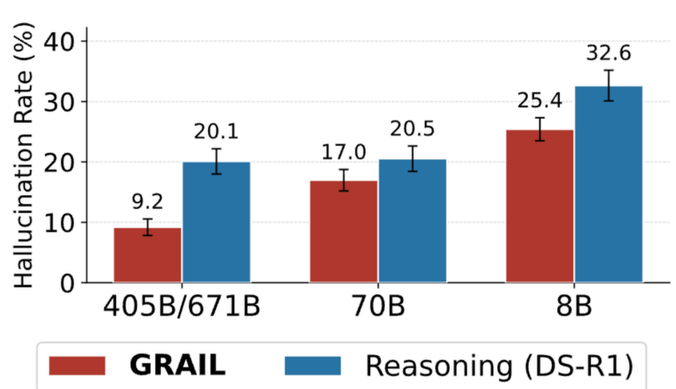
GRAIL reduces the average per-turn **wall-clock time**

Compared to DeepSeek-R1 model, GRAIL produces results faster

	8B	70B*	405B / 671B
DS-R1 (s)	17.37±20.59	15.01±6.55	85.50±179.29
GRAIL (s)	14.04±2.00	18.73±1.82*	20.00±9.99
Graph (s)	5.05	10.15*	5.23

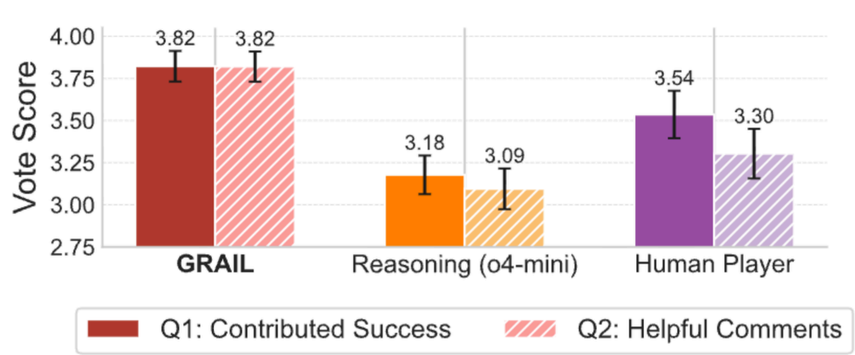
GRAIL decreases **hallucinations frequency**

In social reasoning games, GRAIL hallucinates less than the DeepSeek R1 model due to the hybrid reasoning framework



GRAIL wins 67% of real games **against humans**

Humans preferred our model over other human players in terms of success contribution and helpful comments



References

[1] Shenzhi Wang, Chang Liu, Zilong Zheng, Siyuan Qi, Shuo Chen, Qisen Yang, Andrew Zhao, Chaofei Wang, Shiji Song, and Gao Huang. Avalon's game of thoughts: Battle against deception through recursive contemplation, 2023

[2] Simon Stepputtis, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 11193–11208, Singapore, December 2023. Association for Computational Linguistics.