Sep 22nd 2024

LBM

Level up your Agents with LBM

- Ajay
- Adrian Lam
- Benedict Neo
- Tan Wei Chun

Problem Statement

- LLMs fail to generalize for specific expertise
- Hallucinations and lack of creativity
- Fixed by large scale agent testing with human in the loop

Our Solution

- Generate multiple agents using various divergent styles
- Evaluate using an LLM
- Align the evaluations with human feedback

Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi

Modern large language models (LLMs) like ChatGPT have shown remarkable performance on general language tasks but still struggle on complex reasoning tasks, which drives the research on cognitive behaviors of LLMs to explore human-like problem-solving strategies. Along this direction, one representative strategy is self-reflection, which asks an LLM to refine the solution with the feedback generated by itself iteratively. However, our study shows that such reflection-style methods suffer from the Degeneration-of-Thought (DoT) problem: once the LLM has
gh
Multi-
state of
r MAD
equire
nse
ness of
e modest
find that
his https

Improving Factuality and Reasoning in Language Models through Multiagent Debate

Yilun Du[1], Shuang Li[1], Antonio Torralba[1], Joshua B. Tenenbaum[1], Igor Mordatch[2]

[1] MIT [2] Google Brain

ICML 2024

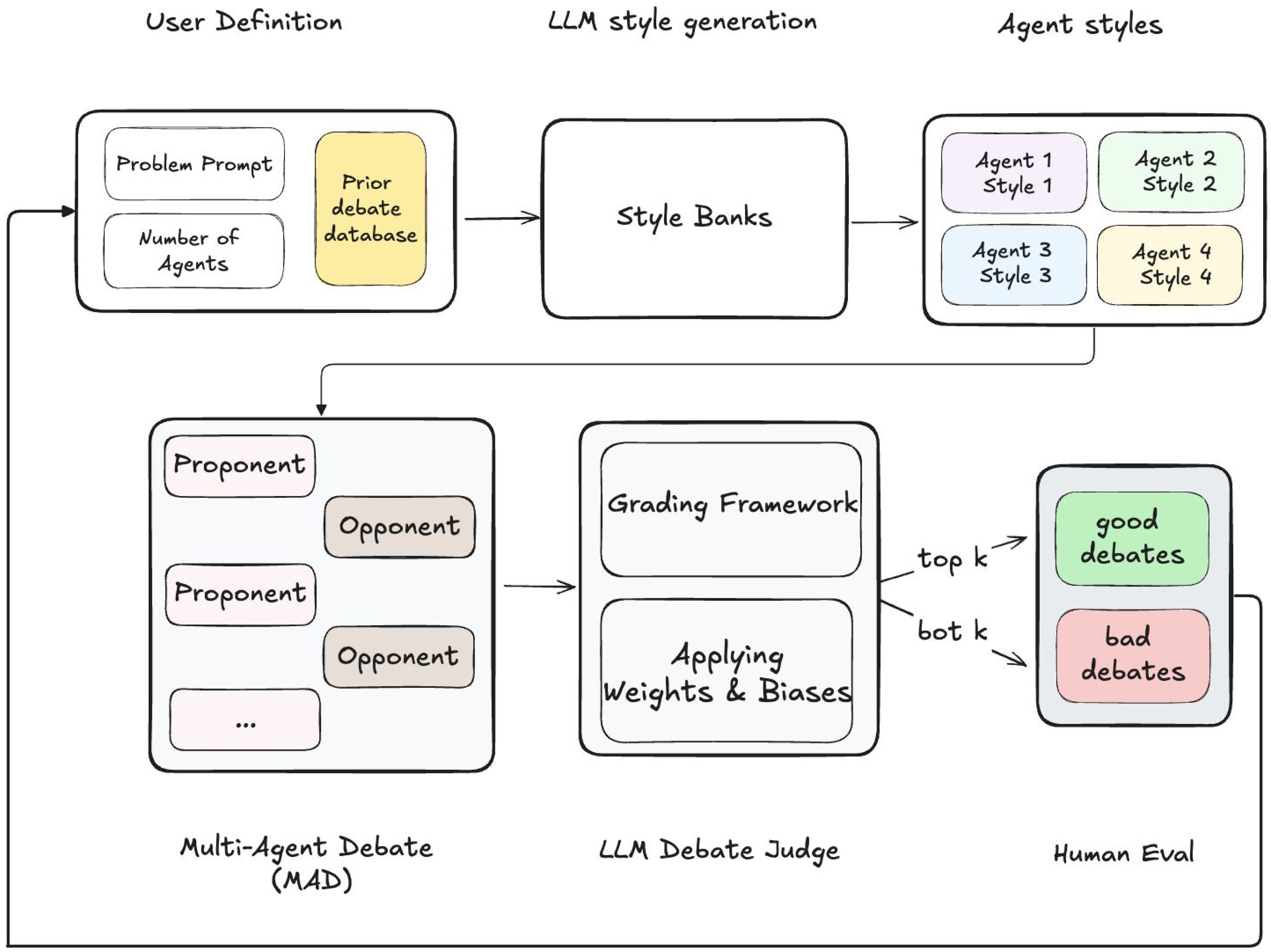
Paper Code

[Submitted on 23 Feb 2024]

Fine-Grained Self-Endorsement Improves Factuality and Reasoning

Ante Wang, Linfeng Song, Baolin Peng, Ye Tian, Lifeng Jin, Haitao Mi, Jinsong Su, Dong Yu

This work studies improving large language model (LLM) generations at inference time by mitigating fact-conflicting hallucinatic
endorsement framework that leverages the fine-grained fact-level comparisons across multiple sampled responses. Compared v
2022;Chen et al., 2023)) that perform response-level selection, our approach can better alleviate hallucinations, especially for lo
can broadly benefit smaller and open-source LLMs as it mainly conducts simple content-based comparisons. Experiments on Bi
effectively improve the factuality of generations with simple and intuitive prompts across different scales of LLMs. Besides, comp
GSM8K demonstrate the potential of self-endorsement for broader application.



User Definition
LLM style generation
Agent styles
Problem Prompt
Number of Agents
Prior debate database
Style Banks
Agent 1 Style 1
Agent 2 Style 2
Agent 3 Style 3
Agent 4 Style 4
Proponent
Opponent
Proponent
Opponent
...
Grading Framework
Applying Weights & Biases
top k
bot k
good debates
bad debates
Multi-Agent Debate (MAD)
LLM Debate Judge
Human Eval

LLM Debating Styles

The **optimistic** debating style presents a positive outlook and focuses on potential benefits and opportunities. This style is characterized by a **focus on human potential, resilience, and the capacity for positive change**...

The **economic** debating style frames arguments in terms of cost-benefit analysis, resource allocation, and economic impact. This style is characterized by a **focus on incentives, trade-offs, and opportunity costs**....

The **technological** debating style focuses on the role of innovation, digital transformation, and scientific advancements in shaping solutions. This style is characterized by a **focus on disruption, efficiency gains, and the transformative power of technology**...

The **contrarian** debating style consistently takes positions opposite to the mainstream view or prevailing wisdom. This style is characterized by a **willingness to stand apart from the crowd, a skepticism towards widely accepted ideas...**

LLM Evaluation

- Specify the grading notes to judge responses
- Each number provided matches a list of criteria
- Fine tuned based on human evaluation of scoring quality

Assistant LLM	Judge Method				
	Human	GPT-4	GPT-4 + Grading Notes	GPT-4-Turbo	GPT-4-Turbo + Grading Notes
	Positive Label Rate by Judge				
Llama3-70b	71.9%	96.9%	73.1%	83.1%	65.6%
GPT-4o	79.4%	98.1%	81.3%	91.9%	68.8%
	Alignment Rate with Human Judge				
Llama3-70b	-	74.7%	**96.3%**	76.3%	91.3%
GPT-4o	-	78.8%	**93.1%**	77.5%	84.4%

Respect	1-5
Accurate Information	1-5
Relevance	1-5
Argument Quality	1-5
Critical Thinking	1-5
Organization	1-5
Preparation	1-5

Demo time!

TODAY'S DATE

Next steps: Fine-tuning Styles

With our validated response dataset, the model can fine tune for the next iteration

Get Highest Score Agents

Integrate Past Responses

Update Style Generator LLM

PROJECT NAME

Ask your
questions

TODAY'S DATE

PROJECT NAME

Questions?