Comentário crítico: *Improving Factuality and Reasoning in Language Models through Multiagent Debate*

Aluno: Maria Fernanda Bosco
RA: 183544

O artigo propõe a interação entre múltiplos modelos de linguagem (LLMs), de forma que eles debatam entre si sobre as tarefas propostas e cheguem a um consenso sobre a resposta final. A hipótese central é que essa forma de cooperação e crítica mútua possa melhorar as capacidades de raciocínio e factualidade dos modelos, reduzindo erros lógicos e afirmações incorretas.

Na abordagem apresentada, cada agente inicialmente gera uma resposta candidata individual e, em seguida, avalia as respostas dos outros agentes, podendo usá-las para refinar ou corrigir sua própria resposta. Os autores mostram que essa técnica supera a performance de modelos de agente único, tanto em tarefas *zero-shot* quanto em configurações com *chain of thought* (CoT). Além disso, observam que o debate entre agentes reduz a incidência de informações factualmente incorretas nas respostas finais.

Para avaliação, os autores comparam o desempenho dos agentes antes e depois do processo de consenso, utilizando tarefas que exigem raciocínio e extração de informações factuais. Em raciocínio, os experimentos envolveram expressões aritméticas simples, problemas matemáticos complexos (GSM8K) e previsão de movimentos em partidas de xadrez (em notação PGN). Os resultados mostraram melhorias consistentes em todos esses domínios. Já nas tarefas factuais — que incluíram questões de biografias, conhecimento geral (MMLU) e verificação da validade de movimentos de xadrez — as melhorias foram mais modestas, aparecendo principalmente nos *prompts* que explicitamente incentivaram o debate.

Os autores também investigaram o impacto da quantidade de agentes e do número de rodadas de debate, mostrando que, até certo ponto, mais interações levam a melhores resultados, embora o ganho marginal diminua após algumas iterações. Ainda assim, não há garantia teórica de que os agentes sempre cheguem a um consenso correto — a convergência observada é apenas empírica e pode depender fortemente da tarefa e do tipo de modelo utilizado.

O trabalho apresenta resultados promissores e uma ideia conceitualmente interessante de "raciocínio coletivo" entre LLMs. Porém, é uma abordagem em que o custo computacional cresce significativamente com o número de agentes e de rodadas de interação, o que pode limitar sua aplicação prática em larga escala. Mesmo assim, o estudo contribui de forma relevante para a discussão sobre como fomentar autorreflexão, verificação cruzada e colaboração entre modelos de linguagem, abrindo caminho para futuras pesquisas sobre coordenação multiagente e raciocínio coletivo em IA.