Principais contribuições do artigo “Improving Factuality and Reasoning in Language Models through Multiagent Debate”, de Yilun Du et al

Aluno: Leandro Carísio Fernandes

- O artigo propõe, como uma forma de melhorar os resultados, que instâncias do modelo “debatam” sobre o input do usuário. Isso é feito passando o input do usuário para mais de uma instância de um LLM. Após os modelos responderem, as respostas são enviadas para os outros LLMs, que tem oportunidade de optar por atualizar a sua resposta.
- Essa estratégia pode ser feita durante um número máximo de rodadas ou pode ser feita até que os modelos convirjam.
- As vantagens da estratégia incluem: (i) os debates tendem a fazer com que a resposta final produza menos alucinações; (ii) os modelos tendem a discordar de fatos mais incertos; (iii) o debate pode trazer a resposta certa mesmo que inicialmente todas as respostas estejam erradas; (iv) é uma solução ortogonal a outras soluções de prompt, ou seja, podem ser usadas em conjunto; (v) não exige conhecimento de variáveis internas dos modelos (por exemplo, seus pesos), o que significa que pode ser usada com modelos fechados; (vi) podem ser usados com modelos diferentes (por exemplo, ChatGPT e Bard).
- A principal desvantagem é o custo. Para se obter uma única resposta, serão necessárias várias iterações, e isso aumenta consideravelmente com o número de agentes (instâncias dos modelos) envolvidos.
- Os autores testaram essa abordagem em alguns problemas e obtiveram melhorias em todos eles. Além disso, identificaram que LLMs tem problemas para tratar biografias e construíram um dataset de biografia para testar essa abordagem.