Elton Cardoso do Nascimento - 233840

IA024 - Redes Neurais Profundas para Processamento de Linguagem Natural - 1s2024

Leitura do Artigo "Improving Factuality and Reasoning in Language Models through Multiagent Debate" (Du et al.)

O trabalho propõe uma nova técnica para melhorar a performance de LLMs na resposta de problemas, baseada na ideia de vários agentes debatendo qual deveria ser a solução correta. O processo se inicia pedindo para cada agente uma resposta para uma questão, e em seguida as respostas de outros agentes são apresentadas e é solicitado uma nova solução considerando essas e a própria resposta. O processo é repetido várias vezes, obtendo a resposta ao final. Este processo permite reduzir a quantidade de respostas incorretas e o impacto de alucinações, melhorando a performance em diversos datasets.

Entre outras descobertas do trabalho, temos:

- Processo funciona mesmo com vários agentes utilizando o mesmo modelo, sendo possível variar o modelo entre eles.
- Processo converge na maioria dos casos.
- Permite corrigir alguns casos em que as respostas iniciais dos modelos estão todas incorretas
- É possível controlar a duração do debate alterando o prompt de debate, fazendo o modelo duvidar mais ou menos da própria respostas
    - Fazer o modelo duvidar menos da própria resposta gera debates mais longos, porém que chegam a soluções melhores
    - Modelos tendem a concordar com outros neste processo, possivelmente resultado do processo de treino
- Aumento das rodadas de debate gera melhoria monotonicamente crescente na performance, mas saturando com poucos acréscimos
- Aumento da quantidade de agentes participando no debate gera melhoria monotonicamente crescente na performance
- Sumarizar as respostas dos outros agentes se faz necessário com o aumento de rodadas e agentes
    - Sumarizar gera melhoria na performance
    - Agente tende a não considerar toda a informação em debates mais longos, se focando nas informações mais recentes

Embora tenha alto custo computacional, a proposta dos desenvolvedores é que a técnica seja utilizada futuramente para gerar dados de treinos adicionais para outros modelos. O artigo também cria um novo dataset e benchmark para avaliação de acurácia factual utilizando biografias de cientistas da computação famosos.