Comentário crítico: *Language Models are Few-Shot Learners*

Aluno: Maria Fernanda Bosco
RA: 183544

O artigo apresenta o GPT-3, um modelo de linguagem autoregressivo com 175 bilhões de parâmetros, mais de dez vezes maior que qualquer modelo anterior. A principal aposta em relação ao GPT-2 foi o aumento de escala, tanto em número de parâmetros (de 1,5B para 175B), quanto em volume de dados de treino. Enquanto o GPT-2 utilizava apenas o *WebText* (40GB), o GPT-3 foi treinado em cerca de 570GB (após filtragem), combinando *Common Crawl* (60%), *WebText2* (22%), *Books1 e 2* (16%) e *Wikipedia* (3%), com maior cuidado na limpeza e deduplicação.

A inovação metodológica central está na forma de avaliação: em vez de *fine-tuning* para tarefas específicas, o GPT-3 é testado diretamente via *prompting*, explorando cenários *zero-shot*, *one-shot* e *few-shot*. Cada tarefa recebeu um template de prompt adequado, com exemplos limitados ao contexto de 2048 tokens. O objetivo foi mostrar que apenas com o aumento massivo de escala, o modelo adquire fortes capacidades de generalização, sem precisar de ajustes adicionais.

Nos resultados, o GPT-3 se destacou em *language modeling* e demonstrou que modelos maiores se beneficiam mais do *few-shot learning*. Em tradução, foi competitivo, chegando próximo de sistemas especializados, e surpreendeu com bom desempenho em tradução *zero-shot*. Em tarefas de perguntas e respostas obteve resultados sólidos, ainda que inferiores a alguns modelos dedicados. Destacou-se no benchmark LAMBADA, além de revelar capacidades inesperadas, como resolver operações matemáticas simples, completar analogias, lidar com raciocínio simbólico e até gerar código, notícias, poemas e diálogos coerentes.

Os autores também reconhecem várias limitações. O modelo ainda produz repetições, incoerências ou contradições, e tem dificuldade em lidar com noções básicas de “física do senso comum”. Sua eficiência de aprendizado é baixa, exigindo quantidades enormes de dados, o que levanta dúvidas sobre a viabilidade de continuar ampliando apenas pela escala. Há também forte sensibilidade ao prompt e incerteza sobre se o desempenho em *few-shot* decorre de aprendizado real ou simples memorização. Outras preocupações incluem viéses sociais, risco de uso malicioso e o alto custo energético do treinamento.

O trabalho demonstra que a simples ampliação de escala — em parâmetros e dados — pode gerar capacidades emergentes notáveis em modelos de linguagem. Porém, também evidencia que a escalada não resolve problemas fundamentais, como raciocínio robusto, eficiência de aprendizado, controle de vieses e sustentabilidade. O GPT-3 marca um avanço histórico, mas aponta para os desafios éticos, técnicos e práticos que acompanharão o desenvolvimento de modelos ainda maiores.