Fine-tuning e Análise do Aprendizado de Modelos de Linguagem Navegando por Labirintos em ASCII

Anderson N. Silva[1]

[1]Departamento Acadêmico de Informática (DAINF) –
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de Setembro, 3165 – 80230-901 – Curitiba – PR – Brasil

andersonnogueira@alunos.utfpr.edu.br

Abstract. *10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e*

1. Introdução

2. Fundamentos e Trabalhos Relacionados

Esta seção discute os trabalhos relacionados e os fundamentos teóricos principais relacionados com o tema proposto.

2.1. LLMs - Large Language Models

Os modelos de Linguagem de Grande Escala (*Large Language Models - LLMs*) são arquiteturas computacionais baseadas em redes neurais profundas, predominantemente do tipo *transformer*, desenvolvidos para processar e gerar linguagem natural. Esses modelos são treinados com grandes volumes de dados textuais, com os quais conseguem (computar/aprender) padrões estatísticos da linaguagem humana, tornando-os capazes de realizar uma ampla gama de tarefas, grande parte sem treinamento supervisionado específico (e.g. sumarização, resposta a perguntas e tradução). A característica comumente associada ao tamanho desses modelos é uma medida em bilhções de parâmetros que está diretamente associado à sua capacidade de adaptação a contextos diversos e sua generalização. A arquitetura *transformer*, introduzida por [Vaswani et al. 2017], faz parte da estrutura central da maioria dos LLMs atuais, e destaca-se pelo uso de mecanismos de atenção que possibilitam a modelagem de dependências contextuais de longo alcance.

2.2. Fine-tuning

O *fine-tunning* é uma técnica utilizada para adaptar *LLMs* a tarefas específicas, por meio de re-treinamento de parte ou totalidade de seus parâmetros com conjuntos de dados mais específicos. Em seguida da etapa inicial de pré-treinamento, na qual um *LLM* aprende as representações gerais linguagem a partir de grandes corpora, o *fine-tunning* permite especializar o comportamento do modelo para contextos mais delimitados como, por exemplo, geração de código, classificação de sentimentos ou, como no presente trabalho, navegação por labirinto em caracteres ASCII. Existem diferentes estratégias de *fine-tunning*:

- Supervised Fine-Tuning (SFT):
- Direct Preference Optimisation (DPO):
- Reinforcement Learning from Human Feedback (RLHF):
- Odds Ratio Preference Optimization (ORPO):

2.3. Explicabilidade em LLMs

Apesar de serem bem capazes em tarefas complexas, os LLMs são questionados com frequência sobre seu caráter de "caixa-preta", que é a dificuldade no entendimento sobre como tomam decisões ou gram suas respostas. Nesse contexto, é cada vez maior o interesse em técnicas de explicabilidade (*explainability*) aplicadas a redes neurais profundas para tornar seus mecanismos internos interpretáveis. A explicabilidade em LLMs tem o objetivo de revelar como os dados são processados ao longo das camadas do modelo, quais neurônios foram ativados em resposta a determinadas entradas, como essas ativações influenciam a saída final e como as representações internas evoluem durante o treinamento o *fine-tunning*. Essas técnicas são essenciais para o diagnóstico de falhas, enviesamento e limitações de raciocínio do modelo. Diversas ferramentas vêm sendo propostas para auxiliar na explicabilidade de LLMs com visualizações, utilização de grafos de conhecimento, análise de atenção e outras abordagens que ajudam a decifrar o comportamento dos modelos.

2.3.1. LLM-MRI

A biblioteca LLM-MRI (*Large Language Model - Magnetic Resonance Imaging*) é uma das ferramentas desenvolvidas com o intuito de auxiliar na explicabilidade de LLMs. Ela atua com o objetivo de simplificar o estudo de padrões de ativações em qualquer LLM baseado em trasnformer. Desenvolvida por [Costa et al. 2024], a LLM-MRI permite a coleta, organização e projeção em dimensões reduzidas dos vetores de ativação produzidos pelos modelos ao processarem diferentes entradas. Por meio de técnicas de redução de dimensionalidade, é possível visualizar como as ativações neuronais se distribuem em um espaço de menor dimensão, facilitando a identificação de padrões e também como as ativações se diferenciam após a aplicação de diferentes técnicas de fine-tunning. Tal abordagem é relevante, pois no contexto experimental deste trabalho, no qual se busca entender como LLMs se adaptam a tarefas específicas (neste caso, labirintos)

References

Costa, L., Figênio, M., Santanchè, A., and Gomes-Jr, L. (2024). LLM-MRI Python module: a brain scanner for LLMs. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 125–130, Porto Alegre, RS, Brasil. SBC.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.