Atividade 02: Resumo Estendido

Informações Gerais

- **Título**: *Fine-tuning* e Análise do Aprendizado de Modelos de Linguagem Navegando por Labirintos em ASCII
- **Equipe**: Anderson Nogueira Silva, 2126516, andersonnogueira@alunos.utfpr.edu.br
- **Orientador**: Prof. Luiz Celso Gomes Junior, luizcelso@gmail.com

Resumo Estendido

Contexto e Problema:
A resolução de labirintos é uma tarefa clássica que exige raciocínio espacial, planejamento e manipulação de representações simbólicas. Com o avanço dos Modelos de Linguagem de Grande Escala (LLMs), surge a oportunidade de investigar como essas arquiteturas, originalmente voltadas para linguagem natural, podem ser adaptadas para tarefas espaciais. Apesar do sucesso dos LLMs em tarefas linguísticas, ainda há lacunas no entendimento de como esses modelos internalizam e representam informações espaciais, especialmente após processos de especialização como o fine-tuning. O caráter de "caixa-preta" dos LLMs dificulta a análise dos mecanismos internos responsáveis pelo aprendizado, tornando essencial o uso de técnicas de explicabilidade [1].

Objetivos:
O objetivo central deste trabalho é analisar as mudanças nas ativações internas de LLMs abertos, com até 8 bilhões de parâmetros, associadas ao aprendizado da tarefa de resolução de labirintos em ASCII, utilizando diferentes estratégias de fine-tuning. Busca-se identificar padrões emergentes, alterações estruturais e possíveis mecanismos de raciocínio espacial desenvolvidos pelos modelos. Os objetivos específicos incluem: (i) aplicar diferentes estratégias de fine-tuning em LLMs abertos para a tarefa de resolução de labirintos em ASCII; (ii) avaliar o desempenho dos modelos antes e após o ajuste; (iii) utilizar a biblioteca LLM-MRI para visualizar e comparar as ativações neuronais; (iv) investigar padrões emergentes e alterações nas representações internas; e (v) discutir as implicações dos resultados para o entendimento do raciocínio e da representação de tarefas específicas em LLMs.

Trabalhos Relacionados:
Diversos estudos recentes abordam o raciocínio espacial em LLMs, destacando abordagens como o AlphaMaze [2], que utiliza SFT e GRPO para aprimorar a navegação em labirintos ASCII, e o MazeBench para avaliação. [3] mostram que LLMs textuais podem superar modelos multimodais em tarefas espaciais, enquanto [4] e [5] exploram limitações e avanços em contextos tridimensionais. Estratégias como COS Prompting [6] e Visualization of Thought [7] otimizam o planejamento espacial e a representação visual. No fine-tuning, [8] e [9] evidenciam ganhos em generalização e interpretabilidade, especialmente com Chain-of-Thought [10]. Para explicabilidade, destacam-se LLM-MRI [11] e Usable XAI [1], que ampliam a compreensão das ativações e decisões dos modelos.

Metodologia:
Será utilizado um conjunto de dados composto por labirintos representados em ASCII, gerados automaticamente para garantir diversidade de estruturas e níveis de dificuldade. Existe a possibilidade de empregar os mesmos datasets e modelos do AlphaMaze, já que estão disponíveis publicamente, mas ainda não está definido se as configurações dos labirintos seguirão exatamente o padrão do AlphaMaze ou se será desenvolvida uma nova abordagem. Os dados serão divididos em conjuntos de treinamento, validação e teste. Serão selecionados LLMs abertos compatíveis com a LLM-MRI, e o fine-tuning será realizado por meio de SFT e GRPO. O desempenho será avaliado por métricas como taxa de sucesso, número médio de passos e precisão dos comandos, comparando-se com benchmarks da literatura, como o MazeBench [2]. A análise das ativações internas será conduzida com a LLM-MRI, aplicando técnicas de redução de dimensionalidade para identificar padrões emergentes e alterações estruturais.

Resultados Esperados:
Espera-se observar melhorias no desempenho dos modelos após o fine-tuning, bem como identificar padrões distintos de ativação interna associados a cada estratégia de ajuste. A análise das ativações deve revelar mecanismos de raciocínio espacial e possíveis limitações dos modelos. Resultados preliminares, visualizações das ativações e comparações quantitativas serão apresentados como evidências.

Avaliação:
Os dados e modelos utilizados são públicos, dispensando a necessidade de coleta própria ou aprovação ética. A avaliação será feita por meio de métricas objetivas e análise qualitativa das ativações. O ambiente experimental será composto por recursos computacionais já disponíveis, sem necessidade de aporte financeiro adicional ou aquisição de hardware específico.

Viabilidade:
O trabalho é viável dentro do tempo previsto para o TCC, considerando a disponibilidade dos dados, modelos e ferramentas. Os principais riscos envolvem desafios técnicos no ajuste dos modelos e na análise das ativações.

Impactos:
O projeto tem potencial impacto científico ao aprofundar o entendimento sobre a representação e o raciocínio espacial em LLMs, além de contribuir para o desenvolvimento de modelos mais interpretáveis. Pode gerar impactos educacionais ao fornecer exemplos e ferramentas para o ensino de IA explicável, e tecnológicos ao aprimorar métodos de análise de modelos de linguagem.

Entregáveis e Evidências:
No TCC1, serão entregues: revisão da literatura, definição dos requisitos experimentais, preparação dos dados, seleção dos modelos, implementação inicial dos pipelines de treinamento e análise, além de resultados preliminares e visualizações das ativações. Essas etapas facilitarão a continuidade e o aprofundamento do trabalho no TCC2.

Referências

[1] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Lijie Hu, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. Usable xai: 10 strategies towards exploiting explainability in the llm era, 2025.

[2] Alan Dao and Dinh Bach Vu. Alphamaze: Enhancing large language models' spatial intelligence via grpo. *arXiv preprint arXiv:2502.14669*, 2025.

[3] Xiaohu Jiang, Yixiao Ge, Yuying Ge, Dachuan Shi, Chun Yuan, and Ying Shan. Supervised fine-tuning in turn improves visual foundation models. *arXiv preprint arXiv:2401.10222*, 2024.

[4] Weichen Zhang, Ruiying Peng, Chen Gao, Jianjie Fang, Xin Zeng, Kaiyuan Li, Ziyou Wang, Jinqiang Cui, Xin Wang, Xinlei Chen, et al. The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? *arXiv preprint arXiv:2504.04540*, 2025.

[5] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*, 2024.

[6] Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. Chain-of-symbol prompting elicits planning in large langauge models. *arXiv preprint arXiv:2305.10276*, 2023.

[7] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[8] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024.

[9] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[11] Luiz Costa, Mateus Figênio, André Santanchè, and Luiz Gomes-Jr. 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, 2024. SBC.