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

A resolução de labirintos é uma tarefa clássica que demanda raciocínio espacial, planejamento e adaptação a ambientes estruturados. Tradicionalmente, algoritmos simbólicos e métodos baseados em busca foram empregados para solucionar esse tipo de problema. No entanto, com o avanço dos Modelos de Linguagem de Grande Escala (*Large Language Models* - LLMs), surge a oportunidade de investigar como essas arquiteturas, originalmente projetadas para linguagem natural, podem ser adaptadas para tarefas que exigem compreensão espacial e manipulação de representações simbólicas, como labirintos em ASCII.

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 e pela tomada de decisão, tornando essencial o uso de técnicas de explicabilidade para investigar as transformações ocorridas nas redes neurais durante o treinamento para tarefas específicas.

Neste contexto, este trabalho propõe investigar como LLMs abertos, com até 8 bilhões de parâmetros, aprendem a resolver labirintos representados em ASCII por meio de diferentes estratégias de *fine-tuning*. O objetivo central é analisar as mudanças nas ativações internas das redes neurais associadas ao aprendizado da tarefa, utilizando a biblioteca LLM-MRI [Costa et al. 2024] para visualizar e comparar as representações neuronais antes e depois do processo de aprendizagem. A análise busca identificar padrões emergentes, alterações estruturais e possíveis mecanismos de raciocínio espacial desenvolvidos pelos modelos.

Os principais objetivos específicos deste estudo são: (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 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.

Ao abordar essas questões, este trabalho busca contribuir tanto para o avanço prático das técnicas de especialização de LLMs quanto para o aprofundamento teórico sobre os processos de representação e raciocínio em redes neurais profundas, especialmente em contextos que exigem habilidades espaciais e interpretabilidade dos modelos.

2. Fundamentos e Trabalhos Relacionados

Esta seção apresenta os principais conceitos e trabalhos relacionados que fundamentam o desenvolvimento deste estudo, com foco em LLMs, técnicas de fine-tuning e métodos de explicabilidade aplicados à resolução de labirintos em ASCII.

2.1. LLMs - Large Language Models

Os Modelos de Linguagem de Grande Escala (*Large Language Models* - LLMs) são arquiteturas baseadas em redes neurais profundas, predominantemente do tipo *transformer*, projetadas para processar e gerar linguagem natural. Esses modelos são treinados com grandes volumes de dados textuais, aprendendo padrões estatísticos da linguagem humana e tornando-se capazes de executar uma ampla gama de tarefas, muitas vezes sem necessidade de treinamento supervisionado específico, como sumarização, resposta a perguntas e tradução. O tamanho dos LLMs, geralmente medido em bilhões de parâmetros, está diretamente associado à sua capacidade de adaptação a diferentes contextos e à generalização. A arquitetura *transformer*, introduzida por [Vaswani et al. 2023], é central para esses modelos, destacando-se pelo uso de mecanismos de atenção que permitem a modelagem de dependências contextuais de longo alcance.

2.2. Fine-tuning

O *fine-tuning* é uma técnica fundamental para adaptar LLMs a tarefas específicas, por meio do re-treinamento de parte ou da totalidade dos parâmetros do modelo com conjuntos de dados direcionados. Após o pré-treinamento, em que o LLM aprende representações gerais da linguagem, o fine-tuning permite especializar o modelo para contextos delimitados, como a navegação em labirintos ASCII, foco deste trabalho. Entre as principais estratégias de fine-tuning destacam-se:

- **Supervised Fine-Tuning (SFT)**: ajuste supervisionado com dados rotulados, comum em tarefas como classificação e tradução.
- **Direct Preference Optimisation (DPO)**: otimização baseada em preferências humanas, visando respostas mais alinhadas ao usuário.
- **Reinforcement Learning from Human Feedback (RLHF)**: aprendizado por reforço com recompensas baseadas em feedback humano.
- **Odds Ratio Preference Optimization (ORPO)**: otimização baseada na razão de chances entre respostas preferidas e não preferidas.
- **Group Relative Policy Optimization (GRPO)**: ajuste considerando preferências coletivas de grupos de usuários.

Essas estratégias são relevantes para o presente estudo, pois permitem investigar como diferentes métodos de ajuste impactam o desempenho dos LLMs em tarefas de raciocínio espacial.

2.3. Explicabilidade em LLMs

Apesar do alto desempenho, LLMs são frequentemente criticados por seu caráter de "caixa-preta", dificultando o entendimento sobre como tomam decisões. Nesse contexto, técnicas de explicabilidade (*explainability*) têm ganhado destaque, buscando tornar os mecanismos internos dos modelos mais transparentes. O objetivo dessas técnicas é revelar como os dados são processados ao longo das camadas, quais neurônios são ativados em resposta a determinadas entradas e como essas ativações influenciam as saídas e evoluem durante o treinamento ou fine-tuning.

A explicabilidade é essencial para diagnosticar falhas, enviesamentos e limitações dos modelos. Diversas ferramentas têm sido propostas, incluindo visualizações, análise de atenção e uso de grafos de conhecimento, que auxiliam na compreensão do comportamento dos LLMs.

2.3.1. LLM-MRI

A biblioteca LLM-MRI (*Large Language Model - Magnetic Resonance Imaging*) é uma ferramenta desenvolvida para facilitar a análise de padrões de ativação em LLMs baseados em *transformer*. Conforme apresentado por [Costa et al. 2024], a LLM-MRI permite coletar, organizar e projetar em dimensões reduzidas os vetores de ativação gerados pelos modelos ao processarem diferentes entradas. Técnicas de redução de dimensionalidade possibilitam visualizar como as ativações neuronais se distribuem, facilitando a identificação de padrões e a análise das mudanças provocadas por diferentes estratégias de fine-tuning.

No contexto deste trabalho, a LLM-MRI é empregada para investigar como LLMs se adaptam à tarefa de navegação em labirintos ASCII e como as estruturas internas dos modelos evoluem durante o processo de aprendizagem, contribuindo para a análise da interpretabilidade e do raciocínio espacial dos modelos.

2.4. Trabalhos Relacionados

Esta seção apresenta a revisão de literatura sobre o tema deste trabalho, com foco em LLMs, fine-tunning e explicabilidade.

2.4.1. Raciocínio Espacial em LLMs

O raciocínio espacial em LLMs tem sido explorado por diferentes abordagens, buscando adaptar esses modelos para tarefas que exigem compreensão e manipulação de ambientes estruturados, como labirintos em ASCII. O AlphaMaze [Dao and Vu 2025] propõe um processo em duas etapas: inicialmente, o modelo é ajustado via Supervised Fine-Tuning (SFT) para aprender comandos de movimentação em labirintos textuais; em seguida, utiliza-se o Group Relative Policy Optimization (GRPO) para aprimorar o raciocínio e a autocorreção. O MazeBench é empregado para avaliar o desempenho dos modelos em diferentes níveis de dificuldade, evidenciando ganhos após o GRPO. Em paralelo, [Jiang et al. 2024] questionam a vantagem de modelos multimodais em tarefas espaciais, mostrando que LLMs baseados apenas em texto podem superar alternativas multi-

modais, o que reforça a relevância de investigar estratégias específicas para LLMs textuais. Em contextos tridimensionais, [Zhang et al. 2025] identificam limitações dos LLMs, enquanto [Cheng et al. 2024] propõem integrar módulos de representação de regiões e informações de profundidade para melhorar a compreensão espacial. Outras abordagens, como o Chain-of-Symbol (COS) Prompting [Hu et al. 2023], convertem descrições em linguagem natural para representações simbólicas intermediárias, otimizando o planejamento espacial e reduzindo o uso de tokens. Já o Visualization of Thought (VoT) [Wu et al. 2024] estimula a geração de representações visuais em ASCII art, inspirando-se na cognição humana para melhorar o acompanhamento de estados e o planejamento de ações. Embora o VoT apresente avanços em determinados cenários, sua eficácia depende das habilidades emergentes dos modelos e da qualidade dos prompts, sendo limitado em tarefas mais complexas ou com modelos menos robustos.

No contexto deste trabalho, o raciocínio espacial é central, pois o objetivo é investigar como LLMs podem ser ajustados e analisados para resolver labirintos representados em ASCII. As abordagens discutidas na literatura fornecem subsídios para a escolha de técnicas de fine-tuning e avaliação, além de motivar o uso de ferramentas de explicabilidade para compreender como as representações internas dos modelos evoluem ao longo do treinamento. Dessa forma, este trabalho busca contribuir para o entendimento dos mecanismos de raciocínio espacial em LLMs, avaliando tanto o desempenho quanto a interpretabilidade dos modelos em tarefas estruturadas de navegabilidade em labirintos textuais.

2.4.2. Fine-tuning de LLMs para Tarefas Específicas

O fine-tuning de LLMs para tarefas específicas é um tema amplamente investigado, com diferentes estratégias sendo propostas para aprimorar a adaptação dos modelos. Por exemplo, [Wang et al. 2024] demonstra que o uso de SFT pode melhorar tanto as representações internas quanto a capacidade de generalização, mesmo em modelos visuais, evidenciando a importância do ajuste direcionado para tarefas de raciocínio espacial. Além disso, [Hsieh et al. 2023] propõem a utilização de rationales gerados por LLMs como supervisão adicional em um cenário multitarefa, combinando rótulos e explicações para enriquecer o sinal de treinamento. Essa abordagem, baseada em Chain-of-Thought prompting [Wei et al. 2023], permite que modelos menores alcancem desempenho superior a LLMs maiores, mesmo com menos dados, ao tornar o processo de ajuste mais eficiente e interpretável. No contexto deste trabalho, a integração de racionalizações se mostra relevante para aprimorar a eficiência e a interpretabilidade dos modelos em tarefas como a resolução de labirintos ASCII.

2.4.3. Explicabilidade em LLMs

A explicabilidade em LLMs é um campo em expansão, com abordagens que buscam tornar os modelos mais transparentes e compreensíveis. A biblioteca LLM-MRI [Costa et al. 2024] destaca-se por permitir a análise e visualização das ativações neuronais, facilitando a interpretação de como os modelos se adaptam a tarefas específicas, como a navegação em labirintos ASCII. Complementarmente, [Wu et al. 2025] propõem o con-

ceito de Usable XAI, que enfatiza a aplicação prática das explicações para diagnóstico e aprimoramento dos modelos. Entre as estratégias discutidas estão métodos de atribuição, análise das ativações internas e explicações baseadas em exemplos, todas relevantes para entender a evolução das representações internas durante o fine-tuning. A integração dessas abordagens contribui para identificar limitações, corrigir vieses e aprimorar o desempenho dos modelos em tarefas que exigem alto grau de interpretabilidade.

3. Metodologia

Esta seção apresenta o planejamento metodológico para investigar como LLMs abertos podem aprender a resolver labirintos em ASCII, analisando as mudanças nas ativações internas das redes neurais após diferentes estratégias de *fine-tuning*. O objetivo é detalhar os procedimentos que serão adotados, de modo a garantir a replicabilidade do estudo, incluindo a preparação dos dados, o ajuste dos modelos, a avaliação de desempenho e a análise das ativações neuronais.

3.1. Preparação dos Dados

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 de treinamento utilizados no trabalho AlphaMaze, uma vez que esses dados e configurações estão disponíveis publicamente. No entanto, ainda não está definido se as configurações dos labirintos e a forma de representação textual seguirão exatamente o padrão do AlphaMaze ou se será desenvolvida uma nova abordagem de geração e representação dos labirintos. Cada instância do conjunto conterá a representação textual do labirinto, a posição inicial e final, e a sequência de comandos esperada para a solução. Os dados serão divididos em conjuntos de treinamento, validação e teste, assegurando que os labirintos do conjunto de teste não sejam vistos durante o treinamento.

3.2. Modelos e Estratégias de Fine-tuning

Serão selecionados LLMs abertos com até 8 bilhões de parâmetros, compatíveis com a biblioteca LLM-MRI. Existe também a possibilidade de utilizar os mesmos modelos empregados no AlphaMaze, visto que esses modelos estão disponíveis publicamente, o que facilitaria a comparação de resultados. O processo de especialização dos modelos será realizado por meio de diferentes estratégias de *fine-tuning*, conforme discutido na literatura:

- **Supervised Fine-Tuning (SFT)**: ajuste supervisionado utilizando exemplos de labirintos e suas soluções.
- **Group Relative Policy Optimization (GRPO)**: ajuste baseado em preferências coletivas, visando aprimorar o raciocínio e a autocorreção dos modelos.

Cada modelo será treinado separadamente em cada estratégia, utilizando os mesmos dados de entrada para garantir comparabilidade.

3.3. Avaliação de Desempenho

O desempenho dos modelos será avaliado antes e após o *fine-tuning*, utilizando métricas objetivas como taxa de sucesso na resolução dos labirintos, número médio de passos até

a solução e precisão na geração dos comandos. Para garantir rigor na avaliação, os resultados serão comparados com abordagens tradicionais e com benchmarks da literatura, como o MazeBench.

3.4. Análise das Ativações Neuronais

A análise das ativações internas será conduzida com o auxílio da biblioteca LLM-MRI [Costa et al. 2024]. Para cada modelo e estratégia de ajuste, serão coletados os vetores de ativação das camadas intermediárias ao processar diferentes labirintos. Uma redução de dimensionalidade será aplicada para visualizar e comparar as distribuições das ativações antes e depois do treinamento. O objetivo será identificar padrões emergentes, alterações estruturais e possíveis mecanismos de raciocínio espacial desenvolvidos pelos modelos.

3.5. Hipóteses do Estudo

Este estudo parte das seguintes hipóteses principais:

- H1: O artigo do AlphaMaze já demonstrou que o *fine-tuning* em LLMs abertos, utilizando dados de labirintos em ASCII, resulta em melhorias mensuráveis no desempenho dos modelos na tarefa de resolução de labirintos. Este trabalho parte desse resultado, buscando investigar se tais melhorias se mantêm ou se apresentam novas características ao empregar diferentes configurações de labirintos, modelos ou estratégias de ajuste.
- H2: Estratégias distintas de *fine-tuning*, como SFT e GRPO, produzirão padrões diferentes de ativação interna, refletindo abordagens variadas de raciocínio espacial e representação da tarefa.
- H3: A análise das ativações neuronais, por meio da LLM-MRI, permitirá identificar alterações estruturais e padrões emergentes associados ao aprendizado da tarefa, contribuindo para a compreensão dos mecanismos internos dos modelos.
- H4: O uso de datasets e modelos do AlphaMaze, caso adotados, proporcionará uma base comparativa relevante, mas a adoção de novas configurações de labirintos poderá revelar limitações ou potencialidades adicionais dos LLMs.

3.6. Discussão e Interpretação dos Resultados

Os resultados quantitativos e qualitativos serão analisados de forma integrada, buscando compreender como as diferentes estratégias de *fine-tuning* influenciam tanto o desempenho quanto as representações internas dos LLMs. A discussão considerará limitações, possíveis vieses e implicações para o desenvolvimento de modelos mais interpretáveis e eficientes em tarefas que exigem raciocínio espacial.

References

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. (2024). Spatialrgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*.

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.

Dao, A. and Vu, D. B. (2025). Alphamaze: Enhancing large language models' spatial intelligence via grpo. *arXiv preprint arXiv:2502.14669*.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes.

Hu, H., Lu, H., Zhang, H., Song, Y.-Z., Lam, W., and Zhang, Y. (2023). Chain-of-symbol prompting elicits planning in large langauge models. *arXiv preprint arXiv:2305.10276*.

Jiang, X., Ge, Y., Ge, Y., Shi, D., Yuan, C., and Shan, Y. (2024). Supervised fine-tuning in turn improves visual foundation models. *arXiv preprint arXiv:2401.10222*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2023). 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.

Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., Li, S., and Joshi, N. (2024). Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., and Wei, F. (2024). 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*.

Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Hu, L., Liu, T., Zhai, X., Yao, W., Li, J., Du, M., and Liu, N. (2025). Usable xai: 10 strategies towards exploiting explainability in the llm era.

Zhang, W., Peng, R., Gao, C., Fang, J., Zeng, X., Li, K., Wang, Z., Cui, J., Wang, X., Chen, X., et al. (2025). The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? *arXiv preprint arXiv:2504.04540*.