# Atividade 01

## Definição do Orientador, Equipe, Tema

- **Orientador**: Prof. Luiz Celso Gomes Junior, luizcelso@gmail.com
- **Equipe**: Anderson Nogueira Silva, 2126516, andersonnogueira@alunos.utfpr.edu.br
- **Título (pode alterar)**: Fine-tuning e Análise de LLMs em Tarefas Não Linguísticas: Navegando por Labirintos ASCII
- **Descrição**: Este trabalho tem como cerne investigar os mecanismos internos de aprendizaado de LLMs a partir de uma tarefa específica e formalmente simples: resolução de labirintos representados em caracteres ASCII. O trabalho envolve o uso de modelos abertos, que serão ajustados por meio de diferentes estratégias de *fine-tuning* para aprenderem a realizar essa tarefa. Além do treinamento em si, o trabalho deve buscar compreender o que muda no interior das redes neurais ao longo do processo de aprendizado. Para isso, serão utilizados bibliotecas de visualização com a LLM-MRI [1], que permitem observar as ativações neuronais e suas representações dimensionais reduzidas. Ao comparar essas ativações antes e depois do aprendizado da tarefa, pretende-se obter insights sobre as estruturas e padrões internos que emergem nos modelos ao aprenderem a navegar nos labirintos. Este trabalho, portanto, se propõe a contribuir tanto para a prática do *fine-tuning* quanto para a compreensão teórica dos processos de representação e raciocíciodos LLMS.
- **Objetivo Geral**: Investigar como modelos de linguagem de grande escala (LLMs) aprendem a realizar a tarefa de resolução de labirintos representados em ASCII, por meio de técnicas de fine-tuning, e analisar as mudanças nas ativações internas das redes neurais associadas ao aprendizado da tarefa.
- **Objetivos Específicos**:
  - Aplicar diferentes estratégias de fine-tuning em LLMs abertos com até 8B de parâmetros para que aprendam a resolver labirintos em ASCII.
  - Avaliar o desempenho dos modelos na tarefa proposta antes e após o fine-tuning.
  - Utilizar a biblioteca LLM-MRI para visualizar e comparar as ativações neuronais das redes antes e depois do processo de aprendizagem.
  - Investigar padrões emergentes e alterações estruturais nas representações internas dos modelos a partir da nova tarefa aprendida.
  - Discutir as implicações dos resultados observados para o entendimento do raciocínio e da representação de tarefas específicas em LLMs.

# Referências

[1] Luiz Costa, Mateus Figênio, André Santanchè, and Luiz Gomes-Jr. LLM-MRI Python module: a brain scanner for LLMs. In *Anais do 39º Simpósio Brasileiro de Banco de Dados (SBBD) - Demonstrações e Aplicações*, pages 125–130, Florianópolis, SC, 2024. Sociedade Brasileira de Computação.