USO DE LLMS PARA AUXILIAR O PROVISIONAMENTO DE INFRAESTRUTURAS VIRTUAIS

Lívia Mayumi Kawasaki Alves, Guilherme Piêgas Koslovski

INTRODUÇÃO

Large Language Models (LLMs) representam um salto significativo em sistemas computacionais capazes de compreender e gerar textos em linguagem humana. O rápido desenvolvimento dos LLMs estimulou inovações nas estratégias e nas técnicas de *fine-tuning*. Especificamente, o *fine-tuning* usa um modelo pré-treinado como base e envolve um novo treinamento em um conjunto de dados menor e específico, melhorando o desempenho em tarefas específicas com dados e requisitos computacionais reduzidos (PARTHASARATHY et al., 2024). O presente trabalho investiga a aplicação de LLMs como ferramenta auxiliar no escalonamento de tarefas, que se refere a ação de alocar servidores para atender as solicitações dos usuários, considerando vários parâmetros como tempo de execução, memória e interdependências de tarefas, definindo o sequenciamento da execução (BRUCKER, 2006). A escolha do algoritmo de escalonamento impacta diretamente no tempo de espera e no uso dos servidores. Por isso, a otimização dessa tarefa é computacionalmente desafiadora, pois os métodos tradicionais apresentam limitações em termos de escalabilidade, motivando, dessa forma, o estudo de alternativas utilizando técnicas de inteligência artificial (ABGARYAN et al., 2024). Nesse contexto, o objetivo deste trabalho é propor a integração de *fine-tuning* nos modelos pré-treinados para otimizar o processo de escalonamento de tarefas a fim de analisar e comparar os resultados.

DESENVOLVIMENTO

Como o estudo foca na resposta gerada pelos modelos para decidir o escalonamento, tais modelos devem seguir a estrutura *Generative Pre-trained Transformer* (GPT) que apresenta três componentes principais: (i) generativo, eles aprendem as relações entre as variáveis de um conjunto de dados para gerar novos dados; (ii) pré-treinado, ou seja, são treinados usando uma grande base de dados, podendo assim economizar tempo e melhorar o desempenho; e (iii) *transformer*, uma arquitetura de rede neural artificial capaz de lidar com dados sequenciais (YENDURI et al., 2024). A entrada provida aos modelos nesse estudo foi um arquivo que simula a alocação de tarefas em um *data center* (DC) composto por servidores, tarefas agendadas e tarefas na fila, incluindo as informações dos recursos de cada um deles. Inicialmente, os modelos assimilam os padrões do comportamento e determinam uma política de escalonamento, gerando uma saída da ordem de execução. A partir da saída, são recortadas as informações sobre quais tarefas foram escalonadas e como foram escalonadas. Esses resultados são então passados para um simulador que gerencia a alocação das tarefas em servidores ao longo do tempo, registrando as métricas de desempenho. Para realizar os testes experimentais, iniciou-se o treinamento de um modelo GPT próprio usando um fragmento da entrada dada aos modelos já treinados para alimentar o modelo. Além desse modelo, os testes foram executados com outros já consolidados e publicamente disponibilizados no Hugging Face, plataforma que fornece ferramentas de inteligência artificial. Foi possível realizar o teste para os modelos descritos: gpt2-large, EleutherAI/gpt-neo-1.3B, meta-llama/Llama-3.2-1B-Instruct, microsoft/phi-1.5 e deepseek-ai/deepseek-coder-1.3b-instruct.

RESULTADOS

Os resultados estão ilustrados pelas Figuras 1 e 2, nas quais as linhas vermelhas (LLM) representam o modelo treinado pelos autores. A Figura 1 apresenta no eixo y o valor médio do

bounded slowdown (bsld, uma métrica que relaciona o tempo total de espera e processamento de cada tarefa) pelo tempo em segundos no eixo x. Notou-se que o LLM se saiu ligeiramente melhor com bsld menor comparado aos outros modelos. A Figura 2 mostra no eixo x a quantidade de tarefas na fila e no eixo y o *Cumulative Distribution Function* (CDF) que representa a porcentagem de valores que estão inseridos nos intervalos específicos de x. Nessa métrica, o LLM se mostrou melhor com menores filas, e com os valores de CDF crescendo mais rapidamente. Em geral, o modelo treinado apresentou resultados relativamente melhores que os outros modelos pré-treinados. Porém, é conhecido que estes foram treinados em dados de uso geral, portanto, eles podem não ter sido expostos, durante o pré-treinamento, a dados semelhantes ao contexto apresentado, e por isso ter tido dificuldades em gerar resultados significativos e diferentes entre si. Em vista disso, propõe-se a implementação da etapa de *fine-tuning* no estudo. A implementação dessa técnica mostra melhoria no desempenho dos LLMs efetivamente nos problemas de escalonamento, podendo até superar as abordagens tradicionais (ABGARYAN et al., 2024).

CONSIDERAÇÕES FINAIS

Este trabalho contribuiu para a exploração da aplicação de LLMs no escalonamento de tarefas comunicantes. Analisou-se que o treinamento de um modelo próprio para aplicações na área de escalonamento pode superar, em alguns casos, as limitações de outros modelos pré-treinados. Ademais, propôs-se a integração da técnica de *fine-tuning* para adaptar os modelos para contextos específicos e melhorar seu desempenho. Os trabalhos futuros envolvem a implementação da solução proposta, para o estudo dos resultados com o intuito de realizar uma análise comparativa. Espera-se, dessa forma, resultados que possam contribuir na busca por soluções eficientes para escalonar tarefas em ambientes de alto desempenho.

Palavras-chave: *Large Language Models*; Escalonamento; *Fine-tuning.*

ILUSTRAÇÕES

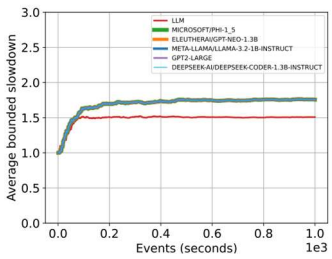

Figura 1. *bsld time consolidation.*

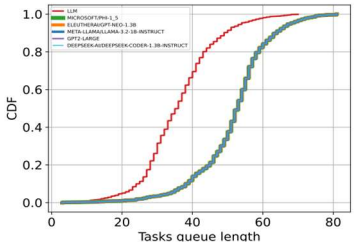

Figura 2. *tasks queue consolidation.*

REFERÊNCIAS BIBLIOGRÁFICAS

[1] BRUCKER, Peter. Scheduling Algorithms. Berlin: Springer, 2006.

[2] PARTHASARATHY, Venkatesh Balavadhani et al. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities. 2024. Disponível em: https://arxiv.org/abs/2408.13296. Acesso em: 31 jul. 2025.

[3] ABGARYAN, Henrik; HARUTYUNYAN, Ararat; CAZENAVE, Tristan. LLMs can schedule. arXiv preprint arXiv:2408.06993, 2024. Disponível em: https://arxiv.org/abs/2408.06993. Acesso em: 31 jul. 2025.

[4] YENDURI, Gokul et al. GPT (Generative Pre-Trained Transformer)—a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access, v. 12, p. 54608-54649, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3389497.

DADOS CADASTRAIS

BOLSISTA: Lívia Mayumi Kawasaki Alves
MODALIDADE DE BOLSA: PIBIC/CNPq
VIGÊNCIA: 02/08/2024 a 31/08/2025 – Total: 12 meses
ORIENTADOR(A): Guilherme Piêgas Koslovski
CENTRO DE ENSINO: CCT
DEPARTAMENTO: Departamento de Ciência da Computação
ÁREAS DE CONHECIMENTO: Ciências Exatas e da Terra / Ciência da Computação
TÍTULO DO PROJETO DE PESQUISA: Mecanismos para Alocação de Infraestruturas Virtuais baseados em Aprendizado de Máquina e Acelerados por GPUs Parte 3
Nº PROTOCOLO DO PROJETO DE PESQUISA: NPP3275-2023