Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Computação Inteligente

Título: Avalição de Estratégias para Garantia de Respostas Precisas em Modelos de Linguagem

Orientador – Leandro Honorato de Souza Silva (leandro.ssilva@upe.br)
Coorientador – Bruno José Torres Fernandes (bjtf@ecomp.poli.br)

Descrição – O desenvolvimento de algoritmos de Inteligência Artificial (IA) capazes de compreender e dominar uma linguagem é um desafio significativo na área de Processamento de Linguagem Natural (NLP) [1]. Recentemente, os *Large Language Models* (LLMs), como o ChatGPT, têm ganhado destaque na academia, indústria e sociedade devido aos resultados notáveis [2].

A área jurídica oferece oportunidades para aplicação dos modelos de linguagem. Uma análise de mais de seiscentos artigos relacionados a NLP & Law revelou um aumento no número de trabalhos sobre o tema, bem como a diversificação das tarefas e línguas abordadas [3]. No entanto, o uso de modelos como o ChatGPT também apresenta problemas conhecidos, como a alucinação, em que as respostas geradas são coerentes, mas falsas [1].

Nesse sentido, foi proposto o TruthfulQA, um benchmark composto de 817 perguntas que abrangem 38 categorias, incluindo saúde, direito, finanças e política [4]. As perguntas do TruthfulQA foram elaboradas de forma que alguns seres humanos responderiam falsamente devido a crenças errôneas ou conceitos equivocados. Estratégias de treinamento podem ser exploradas para reduzir a alucinação e aumentar a veracidade das respostas dos LLMs [5]. Por exemplo, em uma versão do LLM LLaMA chamado Alpaca, a estratégia ITI (*Improving Truthfulness Induction*) melhora sua veracidade de 32,5% para 65,1% [5].

Outro desafio é a dependência de modelos e bases de dados proprietários, além dos altos custos computacionais para treinamento [6]. Portanto, pesquisas estão sendo realizadas para tornar o treinamento desses modelos mais acessível e desenvolver modelos e bases de dados de código aberto [7].

O objetivo deste trabalho é investigar o efeito de estratégias de treinamento, pós processamento e baseadas em agentes (*Agentic* LLM [8]) na veracidade das respostas de LLMs no âmbito jurídico. Além disso, propõe-se a criação de um benchmark análogo ao TruthfulQA, porém em língua portuguesa.

Referências Bibliográficas

[1] W. X. Zhao *et al.*, “A Survey of Large Language Models”, p. 1–58, mar. 2023, [Online]. Disponível em: http://arxiv.org/abs/2303.18223
[2] S. R. Bowman, “Eight Things to Know about Large Language Models”, 2023.
[3] D. M. Katz, D. Hartung, L. Gerlach, A. Jana, e M. J. Bommarito, “Natural Language Processing in the Legal Domain”, *SSRN Electronic Journal*, p. 1–13, 2023, doi: 10.2139/ssrn.4336224.
[4] S. Lin, J. Hilton, e O. Evans, “TruthfulQA: Measuring How Models Mimic Human Falsehoods”, *Proceedings of the Annual Meeting of the Association for*

Computational Linguistics, vol. 1, p. 3214–3252, 2022, doi: 10.18653/v1/2022.acl-long.229.

[5] K. Li, O. Patel, F. Viégas, H. Pfister, e M. Wattenberg, "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model", em *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, p. 41451–41530. [Online]. Disponível em: http://arxiv.org/abs/2306.03341

[6] A. Köpf *et al.*, "OpenAssistant Conversations -- Democratizing Large Language Model Alignment", em *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, abr. 2023, p. 47669–4768. [Online]. Disponível em: http://arxiv.org/abs/2304.07327

[7] T. Dettmers, A. Pagnoni, A. Holtzman, e L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs", em *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, maio 2023, p. 10088–10115. [Online]. Disponível em: http://arxiv.org/abs/2305.14314

[8] M. Sudarshan *et al.*, "Agentic LLM Workflows for Generating Patient-Friendly Medical Reports", ago. 2024, [Online]. Disponível em: http://arxiv.org/abs/2408.01112