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

## Proposta de Tese de Doutorado

**Área: Computação Inteligente**

**Título: Utilização de Modelos de Linguagem de Larga Escala para Interpretação de Bases de Dados Relacionais e Geração de Documentação Semântica.**

**Orientador: Alexandre Magno Andrade Maciel (alexandre.maciel@upe.br)**

**Descrição:**

Ao analisar a situação das organizações atuais, é possível evidenciar a geração de enormes volumes de dados a uma velocidade sem precedentes. Esses dados vêm de uma variedade de fontes em diversos domínios, com sua geração impulsionada por várias tendências tecnológicas como os dispositivos inteligentes, a internet das coisas e a computação em nuvem [1]. Para que os dados sejam verdadeiramente valiosos, eles devem ser acompanhados por metadados significativos, descrevendo não apenas a lista de atributos e suas tipagens, mas também uma descrição semântica que entregue o conhecimento de domínio necessário para o consumo e análise de tais dados [2].

Segundo Paulus et al. [3] existem dois processos fundamentais para a eficácia e adoção da Gestão de Dados Semânticos: o desenvolvimento e a manutenção de conceitualizações, e o enriquecimento semântico das fontes de dados. O enriquecimento semântico dos metadados básicos transforma a maneira como os usuários interagem com os dados, facilitando uma compreensão mais abrangente e precisa. Esse processo de vinculação entre os esquemas de origem e as conceitualizações semânticas não apenas melhora a descoberta e a interpretação dos dados, mas também promove uma utilização mais eficiente e eficaz dos dados [2].

Modelos de Linguagem de Larga Escala (*Large Language Models* - LLM) são essencialmente modelos de aprendizagem profunda que seguem a arquitetura transformer e pré-treinados com grandes volumes de dados não estruturados, no formato de texto [4]. Este tipo de modelo pode ter um uso relevante na identificação semântica, pois diferente das redes neurais tradicionais, são capazes de realizar a extração de informação de um conjunto de entidades e relacionamentos, transformando dados não estruturados de textos em dados estruturados, o que permite identificar a relação semântica entre duas entidades em sequência. [5]

O objetivo deste projeto é desenvolver um sistema baseado em algoritmos de LLM para interpretar bases de dados relacionais e gerar documentação semântica automaticamente. A documentação semântica descreve o significado, a estrutura e os relacionamentos entre os dados, facilitando a compreensão e o uso dos bancos de dados por parte de desenvolvedores, analistas de dados e outros stakeholders.

**Referências Bibliográficas**

[1] - Oussous, A. et al. (2018) ‘Big Data Technologies: A survey’, Journal of King Saud University - Computer and Information Sciences, 30(4), pp. 431–448.
[2] - Hoseini, S., Theissen-Lipp, J. and Quix, C. (2024) ‘A survey on semantic data management as intersection of ontology-based data access, semantic modeling and Data Lakes’, Journal of Web Semantics, 81, p. 100819.
[3] - Paulus, A. et al. (2021) ‘Recent advances and future challenges of Semantic Modeling’, 2021 IEEE 15th International Conference on Semantic Computing (ICSC)
[4] - Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[5] - Kalyan, K.S. (2023) ‘A survey of GPT-3 family large language models including chatgpt and GPT-4’, SSRN Electronic Journal.