

# Universidade de Pernambuco

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

### Proposta de Dissertação de Mestrado

Área: Inteligência Computacional

**Título: Uso de Large Language Models para rotulação de bases de dados para treinamento supervisionado de modelos em tarefas de extração de entidades e relações em documentos**

**Orientador – Byron Leite Dantas Bezerra ([byron.leite@upe.br](mailto:byron.leite@upe.br))**

#### Descrição

A utilização de modelos de Inteligência Artificial em Processamento de Linguagem Natural (*Natural Language Processing, NLP*) possibilitou avanços na extração de informações a partir de texto estruturado [1]. Mais especificamente, uma das tarefas de interesse neste projeto é a identificação de entidades em um documento e as relações destas entidades umas com as outras, e seus atributos [2]. Por exemplo, em um contrato de aluguel de um imóvel, as entidades seriam o locatário, o locador, o imóvel, entre outros. Neste caso, podemos imaginar como atributos do locatário e do vendedor: nome da pessoa, cpf, e data de nascimento, no caso de pessoa física, ou nome da empresa, cnpj, dados dos sócios, no caso de pessoa jurídica. Por fim, pode ser de interesse a obtenção de relações entre as entidades identificadas no documento.

Tradicionalmente, estes tipos de tarefas alcançaram melhores desempenho com arquiteturas de redes neurais baseadas em Transformers [1], com treinamento supervisionado a partir de bases de dados rotuladas. Para tanto, um desafio que surge é ter a disposição a base de dados com quantidade e representatividade suficientes, além de completamente rotulada para treinamento do modelo. Mais recentemente, com o surgimento dos modelos de linguagens largos (Large Language Models, LLMs) [3], a exemplo do ChatGPT, vimos uma explosão de aplicações de NLP alcançando desempenho surpreendente, algumas vezes superando até o desempenho humano. Dessa forma, surge a questão que se objetiva responder neste projeto: qual o desempenho dos LLMs na tarefa de rotulação automática de bases de dados para tarefas de extração de entidades e relações? Como lidar com alucinações dos LLMs e aumentar a confiança destes modelos para minimizar o esforço humano neste tipo de tarefa? Assim, o escopo deste projeto de mestrado compreende o estudo comparativo de LLMs no cenário anteriormente descrito, a fim de investigar até que ponto esses modelos alcançam resultados satisfatórios e sobre quais condições.

A proposta envolve uma equipe multidisciplinar e faz parte do projeto de pesquisa e inovação “*Algoritmos e Modelos de Inteligência Artificial e Visão Computacional para Processamento Inteligente de Documentos*” fomentado pelo CNPQ, e em parceria com a empresa Di2Win ([www.di2win.com](http://www.di2win.com)). Para conhecer mais sobre o orientador e seus temas de pesquisa, convido a assistir a entrevista [aqui](#).

#### Referências Bibliográficas

1. NASAR, Zara; JAFFRY, Syed Waqar; MALIK, Muhammad Kamran. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, v. 54, n. 1, p. 1-39, 2021.
2. SMALL, Sharon Gower; MEDSKER, Larry. Review of information extraction technologies and applications. *Neural computing and applications*, v. 25, p. 533-548, 2014.
3. Wadhwa, S., Amir, S., & Wallace, B. C. (2023). Revisiting Relation Extraction in the era of Large Language Models. *ArXiv*. /abs/2305.05003.