

Trabalho Final de PLN

Objetivo: Este trabalho visa desenvolver uma LLM (Large Language Model) capaz de responder perguntas sobre a legislação acadêmica de Graduação da Universidade Federal do Amazonas (UFAM), disponível [neste link](https://proeg.ufam.edu.br/normas-academicas/57-proeg/146-legislacao-e-normas.html).

<https://proeg.ufam.edu.br/normas-academicas/57-proeg/146-legislacao-e-normas.html>

Estrutura do Trabalho:

- Alunos de pós-graduação: Trabalho individual
- Alunos de graduação: Trabalho em trios

Etapas do Trabalho

1. Download e Pré-processamento da Legislação:

- Acesse o link fornecido e faça o download de todas as resoluções e normas presentes.
- Alguns documentos estão em formato PDF escaneado. Utilize ferramentas apropriadas para extrair o texto desses PDFs, garantindo a integridade e a precisão das informações extraídas.
- Realize o pré-processamento do texto, incluindo limpeza, normalização e estruturação dos dados para facilitar as etapas subsequentes.

2. Geração de Base de Dados Sintética de Instruções:

- Com base nas legislações pré-processadas, gere uma base de dados sintética de instruções contendo 1000 exemplos de perguntas e respostas.
- As perguntas devem cobrir uma ampla gama de tópicos e detalhes presentes nas legislações, garantindo diversidade e relevância.
- Utilize técnicas de geração de texto automatizado para criar exemplos coerentes e variados, baseando-se no conteúdo das normas acadêmicas.

3. Treinamento do Modelo de Linguagem com LoRA/QLoRA:

- Utilize a técnica de Low-Rank Adaptation (LoRA) ou Quantized LoRA (QLoRA) para fazer o tuning de instruções treinando um modelo de linguagem a sua escolha.
- O modelo deve ser treinado utilizando a base de dados sintética gerada, focando em otimizar a capacidade do modelo em compreender e responder perguntas relacionadas à legislação acadêmica.

4. Implementação de RAG (Retrieval-Augmented Generation):

- Indexe todo o conteúdo da legislação utilizando uma ferramenta de busca eficiente.
- Desenvolva um sistema de RAG que, ao receber uma pergunta, recupere trechos relevantes da legislação e gere uma resposta baseada tanto no conteúdo recuperado quanto no conhecimento do modelo de linguagem treinado.
- O sistema deve ser capaz de fornecer respostas precisas e contextualmente apropriadas, utilizando os trechos relevantes como suporte.

Entregáveis

1. **Relatório de Pré-processamento:**
 - Descrição detalhada das etapas de download, extração e pré-processamento dos textos das legislações.
 - Ferramentas utilizadas e desafios enfrentados durante o processo.
 - Base de dados.
2. **Base de Dados Sintética:**
 - Arquivo contendo os 1000 exemplos de perguntas e respostas gerados.
 - Metodologia utilizada para a geração dos exemplos.
3. **Modelo Treinado:**
 - Código fonte utilizado para o treinamento do modelo de linguagem com LoRA/QLoRA.
 - Relatório de desempenho do modelo, incluindo métricas de avaliação e análise de resultados.
4. **Sistema de RAG Implementado:**
 - Código fonte do sistema de RAG.
 - Demonstração de funcionamento do sistema com exemplos de perguntas e respostas.
 - Relatório de avaliação da eficácia do sistema, incluindo exemplos de consultas e as respostas geradas.

Critérios de Avaliação

- **Completeness e Precisão:** Qualidade do pré-processamento dos textos e integridade dos dados extraídos.
- **Criatividade e Variedade:** Diversidade e relevância dos exemplos de perguntas e respostas gerados.
- **Qualidade do Modelo:** Desempenho do modelo de linguagem treinado, avaliado por métricas apropriadas.
- **Eficiência do Sistema de RAG:** Capacidade do sistema em recuperar trechos relevantes e gerar respostas precisas.

Prazo de Entrega

- **Data de Entrega:** 5/8

Nota: Certifique-se de documentar todas as etapas do trabalho, incluindo código, relatórios e quaisquer ferramentas utilizadas. A clareza e a organização da documentação serão consideradas na avaliação final.