Da criação de um Corpus ao treinamento de um grande modelo de linguagem: O que pode dar errado em uma IC?

Lucas B. Bulcão Mota[1], **Aline Athaydes**[1], **Babacar Mane**[1], **Daniela Barreiro Claro**[1], **Marlo Souza**[1], **Fernando Humberto** [1]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos, s/n - Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{lucasbulcao,alineathaydes,babacarm,dclaro,msouza1,}@ufba.br

{fernando.humberto}@ufba.br

Abstract. *This paper presents the experience of a scientific initiation project focused on the development of a chatbot specialized in Consumer Law. One of the main challenges faced was the creation of a synthetic dataset to enable the fine-tuning of a language model. Throughout the process, several technical and methodological difficulties were identified, ranging from data collection and structuring to model training. The objective of this work is to report these challenges, highlighting the importance of error as part of the scientific learning process and reflecting on the lessons learned in the development of AI-based legal systems.*

Resumo. *Este artigo apresenta a experiência de uma iniciação científica voltada ao desenvolvimento de um chatbot especializado em direito do consumidor. Um dos principais desafios enfrentados foi a criação de um conjunto de dados sintético para permitir o ajuste fino de um modelo de linguagem. Ao longo do processo, diversas dificuldades técnicas e metodológicas foram identificadas, desde a coleta e estruturação dos dados até o treinamento do modelo. O objetivo deste trabalho é relatar essas dificuldades, destacando a importância do erro como parte do processo de aprendizagem científica e refletindo sobre os aprendizados obtidos na construção de sistemas jurídicos baseados em IA.*

1. Introdução

Com o avanço das tecnologias de inteligência artificial, especialmente dos grandes modelos de linguagem (LLMs), tem-se observado um crescente interesse na aplicação dessas ferramentas em áreas tradicionalmente complexas como o direito. Estudos recentes apontam que o uso de modelos de linguagem podem contribuir para a democratização do conhecimento jurídico [Malaquias Junior et al. 2024]. No entanto, a adoção dessas tecnologias exige cuidados metodológicos específicos, sobretudo no que diz respeito à qualidade dos dados utilizados e à responsabilidade na geração de conteúdo legal.

A subárea do jurídico, direito do consumidor tem ganhado relevância significativa nos últimos anos, refletindo o crescente interesse da população na garantia de seus direitos. De acordo com dados apresentados por [ConJur 2023], entre 2018 e 2022, uma em cada quatro ações distribuídas nas Justiças estadual e federal tratava de questões relacionadas ao direito do consumidor.

Neste contexto, minha iniciação científica (IC) está inserida em um projeto de pesquisa que busca desenvolver um chatbot voltado para o domínio do direito do consumidor, com foco em usuários leigos. Desde o início da pesquisa, fiquei responsável por uma das etapas centrais

do projeto: a construção de um conjunto de dados sintético para permitir o ajuste fino (fine-tuning) de um modelo de linguagem de grande porte. No entanto, ao longo do processo,foi encontrada uma série de dificuldades, desde a elaboração do dataset até o treinamento do modelo. Essas experiências revelaram não apenas a complexidade técnica da tarefa, mas também a importância do erro como parte fundamental do processo de aprendizagem na pesquisa científica.

Este artigo tem como objetivo relatar as principais dificuldades enfrentadas ao longo dessa trajetória, destacando os aspectos técnicos e metodológicos envolvidos no ajuste de um modelo de linguagem para o domínio de defesa do consumidor.

2. Metodologia

Este projeto de pesquisa tem como objetivo desenvolver um chatbot capaz de responder a perguntas no domínio do direito do consumidor. Para isso, foi elaborada uma metodologia que inclui a construção de um conjunto de dados representativo, a avaliação de uma amostra por uma especialista jurídica e o posterior treinamento de um modelo de linguagem apto a compreender o contexto jurídico e gerar respostas adequadas ao público-alvo.

2.1. Desenvolvimento do Dataset Sintético

A construção do conjunto de dados representou uma das etapas mais complexas da pesquisa, exigindo diversas decisões metodológicas e técnicas, como mostrada na figura 1.

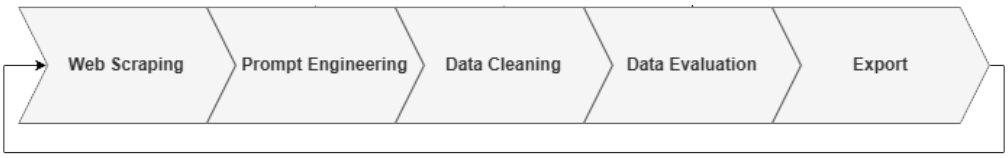

Figura 1. Pipeline de elaboração do Dataset

O primeiro passo consistiu na obtenção de dados jurídicos de referência, que seriam fornecidos como contexto ao modelo GPT-4o mini [OpenAI 2024]. Utilizamos o framework em python *Selenium* para realizar o *web scraping*, extraindo os conteúdos diretamente relacionados à defesa do consumidor, incluindo artigos do Código de Defesa do Consumidor (CDC), súmulas e acórdãos do site do Supremo Tribunal de Justiça (STJ). Entretanto, essa escolha não foi trivial, considerando a dificuldade em encontrar dados limpos e estruturados, que garantem diversidade temática para o domínio de defesa do consumidor.

A escolha adequada das fontes é fundamental para assegurar a relevância e a consistência do conteúdo gerado. Após duas gerações do dataset, constatamos que as três fontes iniciais poderiam não ser suficientes para capturar a diversidade temática necessária à tarefa. Assim, consideramos a inclusão de materiais adicionais, como livros especializados em direitos do consumidor e seções de perguntas frequentes (FAQs), como alternativas para enriquecer o contexto fornecido ao modelo.

Com os dados de referência devidamente preparados, elaborou-se um prompt destinado a orientar o modelo na geração de pares de perguntas e respostas. O prompt foi cuidadosamente projetado para promover diversidade temática e assegurar a fundamentação jurídica com base nos documentos extraídos. Posteriormente, os conjuntos de perguntas e respostas foram estruturados no formato de chat template apresentado na listagem abaixo

```
{
  "messages": [
    {"role": "system", "content": "instrução"},
    {"role": "user", "content": "Pergunta"},
    {"role": "assistant", "content": "Resposta + Contexto"}
  ]
}
```

Listing 1: Chat template em formato JSON.

A geração automática do dataset resultou em um conjunto de dados brutos, contendo inconsistências típicas de modelos generativos, como campos vazios, perguntas em outras línguas, erros ortográficos e alucinações de modo geral. Isso demandou uma etapa subsequente de limpeza e curadoria dos dados.

A limpeza foi realizada por meio de duas estratégias: manual e automática. A abordagem manual consistiu na identificação e exclusão de pares com erros evidentes ou conteúdo irrelevante. Já a limpeza automática envolveu a contagem de tokens por par pergunta-resposta, eliminando os conjuntos com menos de 50 tokens, e menos de 10 tokens por campo (pergunta, resposta ou contexto) além da remoção de mais de 3100 duplicatas. Gerando um conjunto de dados limpo composto por 61870 pares de perguntas e respostas.

Após essa etapa, foi necessário avaliar a qualidade do conjunto de dados. Dado que os textos foram gerados artificialmente, essa avaliação se mostrou desafiadora. Selecionamos, então, uma amostra aleatória com 101 pares pergunta–resposta, que passou por uma análise qualitativa conduzida por uma especialista na área jurídica. A avaliação seguiu os princípios da Análise de Conteúdo proposta por [Bardin 2011]. Dos 101 pares analisados, 76 foram considerados satisfatórios, enquanto os demais apresentaram algum tipo de inconsistência.

Por fim, após a avaliação, o conjunto de dados foi exportado em formato CSV e disponibilizado na plataforma Hugging Face [Hugging Face], que hospeda tanto datasets quanto modelos de linguagem.

Todo esse processo resultou em um dataset limpo e parcialmente validado, pronto para ser utilizado no fine-tuning do modelo. A construção cuidadosa da base de dados é um dos elementos centrais para garantir o bom desempenho de modelos de linguagem: dados de qualidade são fundamentais para resultados satisfatórios.

2.2. Treinamento do Modelo de Linguagem

Após a criação do conjunto de dados de treino, a próxima etapa foi definir uma metodologia para treinar um LLM. A Figura 2 mostra o pipeline de elaboração do modelo.

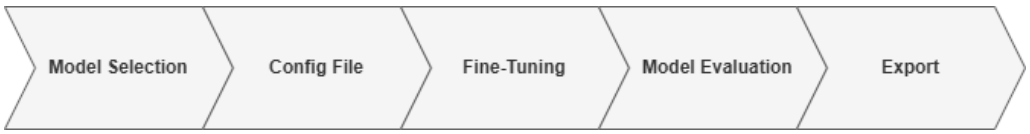

Figura 2. Pipeline de elaboração do Modelo

As maiores dificuldades nesse processo se concentraram na escolha do modelo base e na elaboração do arquivo de configuração para o treinamento.

Diante da vasta gama de modelos abertos disponíveis, enfrentamos dificuldades em compreender as diferenças entre eles. Inicialmente consideramos modelos menores, mas observamos limitações quanto ao desempenho e à capacidade de generalização. A escolha pelo Qwen3-8B [Qwen Team 2025] foi influenciada por testes preliminares positivos, pela documentação acessível e, principalmente, pela possibilidade de utilizar um modelo relativamente grande graças à infraestrutura disponibilizada pelo Sistema de Computação de Alto Desempenho (SMCAD) da universidade.

A metodologia de ajuste fino do modelo de linguagem foi conduzida utilizando o framework Axolotl [Ebrahimi et al. 2024]. A abordagem inicial, que consistia em um ajuste fino completo (Full Fine-Tuning), demonstrou ser impraticável por duas razões principais. Primeiramente, o modelo exibiu um claro comportamento de superajuste (overfitting), gerando respostas com baixa variabilidade e excessivamente aderentes ao conjunto de treinamento. Em segundo lugar, o processo demandou um custo computacional elevado, dificultando sua execução com os recursos de hardware disponíveis. Para superar esses desafios, a estratégia foi redefinida com a adoção da técnica de Adaptação de Baixo Ranque, ou Low-Rank Adaptation (LoRA) [Hu et al. 2021], uma abordagem de ajuste fino eficiente em parâmetros (Parameter-Efficient Fine-Tuning, PEFT). A LoRA foi escolhida especificamente por mitigar o risco de overfitting e reduzir drasticamente a carga computacional, uma vez que congela os pesos do modelo pré-treinado e treina apenas um número reduzido de parâmetros em matrizes de baixo ranque injetadas em sua arquitetura.

Após um processo de experimentação iterativa para a otimização dos hiperparâmetros, a configuração final do treinamento com LoRA foi estabelecida para maximizar a capacidade de generalização do modelo.definiu-se o ranque $r = 16$, valor que representa um balanço entre a capacidade de aprender as nuances do domínio jurídico e a simplicidade necessária para evitar o superajuste. Para reforçar a regularização, foi aplicado um dropout de 0,05 nas matrizes LoRA e limitou-se o treinamento a três épocas, evitando memorização excessiva. O modelo resultante apresentou melhor desempenho em dados não vistos, gerando respostas mais diversificadas e semanticamente alinhadas ao domínio jurídico, o que confirma a eficácia da abordagem LoRA para a tarefa.

3. Conclusão e Trabalhos Futuros

O desenvolvimento deste projeto de pesquisa proporcionou uma valiosa oportunidade de aprendizado prático sobre o ciclo completo de construção de um sistema baseado em modelos de linguagem, desde a elaboração do conjunto de dados até o treinamento de um modelo generativo.

Ao longo da trajetória, os maiores aprendizados vieram justamente das escolhas tomadas, ainda que nem sempre tenham sido as melhores. Cada etapa, da seleção das fontes de dados à definição dos hiperparâmetros de treinamento apresentou desafios que exigiram estudo, tentativa e ajuste. A experiência mostrou que ajustes são sempre necessários e fundamentais para o amadurecimento científico.

Como continuidade deste trabalho, pretende-se avançar no desenvolvimento do *chatbot* explorando novas abordagens que permitam ampliar a capacidade de compreensão e contextualização do modelo. Além disso, será avaliada a aplicação da estratégia de *Chain-of-Thought prompting* [Wei et al. 2022], que estimula o raciocínio passo a passo por parte do modelo, com potencial para melhorar a coerência e a explicabilidade das respostas. Essas abordagens visam não apenas aprimorar a qualidade das interações, mas também tornar o sistema mais robusto e confiável para usuários leigos.

Agradecimentos

Agradecimentos ao Escavador e a FAPESB por meio dos projetos TIC 0002/2015, CCE 0022/2023 e INCITE PIE0002/2022

Referências

Bardin, L. (2011). *Análise de conteúdo*. Edições 70, São Paulo, 1 edition. Traduzido por Luís Antero Reto e Augusto Pinheiro.

ConJur (2023). Cada ações judiciais estaduais e federais sobre consumo.

Ebrahimi, S., Chen, K., Asudeh, A., Das, G., and Koudas, N. (2024). Axolotl: Fairness through assisted self-debiasing of large language model outputs. *arXiv preprint arXiv:2403.00198*. Disponível em: https://arxiv.org/abs/2403.00198.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models.

Hugging Face. Hugging face: The ai community building the future. https://huggingface.co. Acesso em: jun. 2025.

Malaquias Junior, R., Pires, R., Romero, R., and Nogueira, R. (2024). Juru: Legal brazilian large language model from reputable sources. *arXiv preprint arXiv:2403.18140*. Disponível em: https://arxiv.org/abs/2403.18140.

OpenAI (2024). Gpt-4o mini: advancing cost-efficient intelligence. Online. Disponível via anúncio oficial da OpenAI; modelo lançado em 18 de julho de 2024.

Qwen Team (2025). Qwen3 technical report. Technical report, Qwen Research. Relatório técnico sobre a série de modelos Qwen3, incluindo Qwen3-8B.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.