Aplicação de Large Language Models na Análise e Síntese de Documentos Jurídicos: Uma Revisão de Literatura

Matheus Belarmino[1], **Rackel Coelho**[1], **Roberto Lotufo**[2], **Jayr Pereira**[1]

[1]Universidade Federal do Cariri (UFCA)
Juazeiro do Norte – CE – Brasil

[2]NeuralMind.ai
Campinas – SP – Brasil

matheus.pereira@aluno.ufca.edu.br, jayr.pereira@ufca.edu.br

Abstract. *Large Language Models (LLMs) have been increasingly used to optimize the analysis and synthesis of legal documents, enabling the automation of tasks such as summarization, classification, and retrieval of legal information. This study aims to conduct a systematic literature review to identify the state of the art in prompt engineering applied to LLMs in the legal context. The results indicate that models such as GPT-4, BERT, Llama 2, and Legal-Pegasus are widely employed in the legal field, and techniques such as Few-shot Learning, Zero-shot Learning, and Chain-of-Thought prompting have proven effective in improving the interpretation of legal texts. However, challenges such as biases in models and hallucinations still hinder their large-scale implementation. It is concluded that, despite the great potential of LLMs for the legal field, there is a need to improve prompt engineering strategies to ensure greater accuracy and reliability in the generated results.*

Resumo. *Os Modelos de Linguagem de Grande Escala (LLMs) têm sido cada vez mais utilizados para otimizar a análise e síntese de documentos jurídicos, possibilitando a automatização de tarefas como sumarização, classificação e recuperação de informações legais. Este estudo tem como objetivo realizar uma revisão sistemática da literatura para identificar o estado da arte em engenharia de prompts aplicada a LLMs no contexto jurídico. Os resultados indicam que modelos como GPT-4, BERT, Llama 2 e Legal-Pegasus são amplamente empregados na área jurídica, e técnicas como Few-shot Learning, Zero-shot Learning e Chain-of-Thought prompting têm se mostrado eficazes para melhorar a interpretação de textos legais. No entanto, desafios como a presença de vieses nos modelos e a ocorrência de alucinações ainda dificultam sua implementação em larga escala. Conclui-se que, apesar do grande potencial dos LLMs para o direito, há a necessidade de aprimoramento das estratégias de engenharia de prompts para garantir maior precisão e confiabilidade nos resultados gerados.*

1. Introdução

Os avanços em inteligência artificial e aprendizado de máquina têm revolucionado diversas áreas do conhecimento, incluindo o direito. A crescente digitalização dos processos jurídicos gerou um volume exponencial de documentos legais, tornando cada vez mais desafiador para advogados, juízes e pesquisadores processar, interpretar e extrair

informações relevantes de maneira eficiente. Nesse contexto, os modelos de linguagem de grande escala (LLMs, do inglês, Large Language Models) surgem como uma tecnologia promissora para auxiliar na análise e síntese de documentos jurídicos [SILVA 2023].

Os LLMs são modelos de aprendizado de máquina treinados em vastos conjuntos de dados textuais para compreender e gerar linguagem natural. Esses modelos são baseados na arquitetura de Transformers [Vaswani et al. 2017], que permite que eles aprendam diferentes tarefas a partir de grandes volumes de texto por meio de mecanismos de autoatenção. Dentre os LLMs mais conhecidos estão o GPT-4, Gemini, LLaMa e Sabiá, um modelo voltado para o português brasileiro [Abonizio et al. 2025]. Esses modelos são capazes de realizar tarefas como responder perguntas, traduzir textos, resumir documentos, classificar informações e até mesmo gerar conteúdos com coerência contextual.

A aplicação de LLMs no contexto legal tem o potencial de otimizar significativamente processos como revisão contratual, predição de decisões judiciais, recuperação de jurisprudência e geração automática de resumos legais. No entanto, apesar de seus avanços, esses modelos ainda enfrentam desafios como alucinações (geração de informações incorretas), viés nos dados de treinamento e dificuldades na interpretação de contextos jurídicos específicos. Assim, um dos principais focos de pesquisa atualmente é a utilização de técnicas de engenharia de *prompts* para melhorar a precisão e relevância das respostas dos LLMs em contextos jurídicos [Peixoto and Bonat 2023].

A engenharia de *prompts* refere-se ao processo de formular comandos ou instruções otimizadas para obter as melhores respostas possíveis dos LLMs. Técnicas como *Zero-shot Learning* (quando o modelo responde sem exemplos), *Few-shot Learning* (quando é treinado com poucos exemplos) e *Chain-of-Thought* (quando se estimula um raciocínio passo a passo) têm sido amplamente estudadas para aprimorar a performance dos LLMs em tarefas jurídicas complexas. Pesquisadores vêm explorando essas estratégias para reduzir erros e viabilizar o uso confiável desses modelos no setor jurídico [Moura and Carvalho 2023].

Apesar da ampla gama de pesquisas sobre LLMs aplicados ao direito, os estudos existentes estão dispersos e não sistematizados em uma única fonte. Isso dificulta a compreensão do estado da arte e a identificação de lacunas e oportunidades de aprimoramento na aplicação dessas tecnologias no contexto jurídico. Atualmente, não há uma revisão consolidada na literatura que reúna e analise criticamente os avanços recentes na engenharia de *prompts* aplicada aos LLMs para documentos jurídicos. Diante dessa lacuna, o objetivo principal desta pesquisa é realizar uma revisão da literatura para identificar o estado da arte em técnicas de engenharia de *prompts* aplicadas a LLMs na análise e síntese de documentos jurídicos complexos em Língua Portuguesa.

2. Metodologia

2.1. Definição das Perguntas de Pesquisa

A pesquisa foi conduzida com o objetivo de compreender como os Modelos de Linguagem de Grande Escala (LLMs) são aplicados na análise de documentos jurídicos. Para isso, foram formuladas perguntas de pesquisa principais e secundárias, conforme apresentado na Tabela 1.

Esta pesquisa busca compreender de forma ampla o impacto dos LLMs na análise

Table 1. Perguntas de Pesquisa e Facetas

ID	Pergunta de Pesquisa	Faceta
RQ1	Qual o modelo de linguagem utilizado?	Modelo
RQ2	Quais as técnicas de engenharia de prompt utilizadas?	Técnica
RQ3	Como a solução é avaliada?	Avaliação
RQ4	Quais as métricas usadas na avaliação?	Métrica
RQ5	Qual o desempenho dos LLMs na tarefa?	Desempenho
RQ6	Qual o domínio de aplicação?	Domínio

de documentos jurídicos. As perguntas secundárias foram elaboradas para permitir uma avaliação detalhada dos aspectos técnicos, metodológicos e aplicacionais dos modelos de linguagem.

2.2. Bases de Dados e Estratégia de Busca

Para garantir um levantamento abrangente da literatura científica sobre o tema, a pesquisa foi realizada nas bases de dados SCOPUS e Springer, que são amplamente reconhecidas pela indexação de estudos relevantes na área de inteligência artificial e direito.

A estratégia de busca foi definida para recuperar artigos relacionados à aplicação de engenharia de *prompts* e modelos de linguagem na análise de documentos jurídicos. A busca foi limitada ao período de 2020 a 2024 para garantir a inclusão de estudos recentes e relevantes. A query utilizada para a busca dos estudos foi:

(“Prompt Engineering” OR “Prompt Optimization” OR “Large Language Models” OR “LLMs” OR “generative artificial intelligence”) AND (“Legal Document” OR “Judicial Processes” OR “Legal Cases” OR “Court Documents”)

2.3. Seleção de Estudos

Os estudos foram selecionados com base em critérios de inclusão e exclusão previamente estabelecidos, conforme apresentado na Tabela 2. Para garantir uma triagem eficiente e sistemática das publicações, utilizou-se o ChatGPT como suporte na organização e análise preliminar dos artigos. Foram utilizados *prompt*s específicos, como ’Liste os principais métodos utilizados neste estudo’ e ’Qual o modelo de linguagem empregado neste artigo?’, para auxiliar na categorização e extração dos dados. O modelo de linguagem foi empregado para identificar padrões nos dados, sugerir categorizações e auxiliar na extração de informações relevantes dos textos científicos, tornando o processo mais ágil e estruturado.

No entanto, todas as decisões finais sobre a inclusão ou exclusão dos estudos foram realizadas manualmente, garantindo rigor metodológico e alinhamento com os objetivos da pesquisa. Dessa forma, a ferramenta foi utilizada como um recurso complementar para otimizar a análise, sem comprometer a integridade dos resultados.

2.4. Extração de Dados

A extração de dados foi realizada de maneira sistemática para garantir uma análise comparativa entre os estudos selecionados. A Tabela 3 apresenta os principais elementos extraídos de cada estudo.

Table 2. Critérios de Inclusão e Exclusão

Tipo	Critério
Inclusão	IC1: O estudo aborda a aplicação de técnicas de engenharia de *prompts* na análise de documentos jurídicos.
	IC2: O estudo explora o uso de modelos de linguagem (LLMs) na interpretação de processos judiciais ou outros documentos legais.
Exclusão	EC1: O estudo está escrito em uma língua diferente de português ou inglês.
	EC2: O estudo é um editorial, keynote, biografia, opinião, tutorial, workshop, resumo de relatório, poster, tese, dissertação, capítulo de livro ou painel.
	EC3: O estudo não aborda a análise ou síntese de documentos jurídicos.
	EC4: O estudo não utiliza técnicas de engenharia de *prompts* ou Large Language Models (LLMs).
	EC5: O foco do estudo é em outra área do direito que não envolve processos ou documentos jurídicos complexos.

Table 3. Elementos Extraídos dos Estudos

Elemento	Descrição
Referência do estudo	Autor, ano, título e fonte do estudo.
Modelos de linguagem utilizados	LLMs empregados na análise de documentos jurídicos.
Técnicas de engenharia de prompt	Métodos utilizados para otimizar a interação com os LLMs.
Métodos de avaliação	Procedimentos aplicados para validar a eficácia das soluções.
Métricas de desempenho	Indicadores utilizados para medir a qualidade das respostas.
Principais achados e limitações	Resultados obtidos e desafios identificados nos estudos.

3. Resultados

Na Tabela 4, apresenta-se a revisão sistemática de artigos científicos sobre o uso de LLMs na área jurídica, explorando aspectos fundamentais como os modelos utilizados, as técnicas de engenharia de prompt, os métodos de avaliação das soluções, as métricas empregadas, o desempenho dos modelos e os domínios de aplicação. A partir dessa análise, é possível identificar tendências, desafios e avanços na aplicação dessas tecnologias para a compreensão e síntese de documentos jurídicos.

Table 4. Resultados

Artigo	RQ1: Modelo	RQ2: Prompt	RQ3: Avaliação	RQ4: Métricas	RQ5: Desempenho	RQ6: Domínio
[Deroy et al. 2024]	GPT-4 e 3.5, LLama-2, Legal-Pegasus, Legal-LED	TL;DR	Métricas automáticas e avaliação com estudantes de Direito	Rouge, METEOR, BERTScore, SummaC, NumPrec	LLMs superaram métodos extrativos	Sumarização de decisões judiciais do Reino Unido e Índia
[Ghosh et al. 2023]	BART (Encoder-Decoder)	Mascaramento seletivo de spans	Testado em 13 datasets de 6 tarefas	F1, Precisão, Diversidade de Tokens	Melhorou o desempenho em 1%-50%	Aumento de dados para NLP jurídico
[Wu et al. 2023]	ChatGPT (GPT-3.5)	Reorganização de fatos jurídicos	Experimentos com CAIL2018 e CJO22	Acurácia, Precisão, Recall, F1	Proposta superou baseline	Predição de julgamentos na China
[Venkatakrishnan et al. 2024]	GPT-3.5, Mistral 8*7B	Construção de grafos	Avaliação qualitativa dos grafos	Não especificado	Geração de grafos com alta precisão	Análise de dados de imigração
[Deroy et al. 2023]	Text-Davinci-003, GPT-3.5, modelos ajustados	Tl;Dr, divisão em blocos de 1024 palavras	Similaridade entre resumos gerados e padrões	ROUGE, METEOR, BLEU, SummaC, NEPrec, NumPrec	Modelos jurídicos específicos tiveram melhor desempenho	Sumarização de decisões judiciais indianas
[Wu et al. 2024]	D3LM fine-tune do LLaMa-2	Perguntas diagnósticas	avaliação automática e humana	ROUGE, BLEU, fluência, precisão	Superou GPT-4 em métricas específicas	Análise de casos criminais nos EUA
[Zhou et al. 2023]	GPT-3.5-turbo, BERT	Sumarização extrativa	Experimentos com modelos de busca	P@5, P@10, MAP, NDCG	Melhor que baselines	Recuperação de casos jurídicos
[Kang et al. 2023]	ChatGPT	Decomposição de perguntas, in-context learning, CoT	Avaliação por estudantes e especialistas jurídicos	Precisão, recall, F1-score, Cohen's Kappa	Desempenho misto, F1 médio de 0.49	Direito contratual e familiar (Malásia e Austrália)
[Prasad et al. 2024]	GPT-Neo, GPT-J, LEGAL-BERT, InLegalBERT	Segmentação em chunks, embeddings de última camada	Testes em datasets jurídicos	Micro-F1, Macro-F1, acurácia	MESc teve desempenho superior aos métodos anteriores	Classificação de documentos jurídicos (Índia, UE, EUA)
[Nunes et al. 2024]	Sabiá (Llama 7B ajustado para português)	In-Context Learning com heurísticas K similares	Testes em corpora brasileiros (LeNER-Br, UlyssesNER-Br)	F1-micro, precisão, recall	Melhor desempenho com K similares, F1 de 51%	NER em textos jurídicos brasileiros
[Hijazi et al. 2024]	GPT-4, Jais, Llama 3, Claude-3-opus	In-Context Learning, Chain of Thought, Few-shot e Zero-shot	Testes em MCQs, Q&A e traduções	F1-score, ROUGE, métrica de similaridade GPT-4	GPT-4 foi o melhor geral, Jais destacou-se no direito saudita	Direito saudita e árabe
[Chen et al. 2024]	Qwen-14B, LLaMA, GPT-3.5, GPT-4	Ajuste fino em duas etapas, R-Drop	Testes com corpus de e-commerce	SacreBLEU, ROUGE	G2ST superou modelos de última geração	Tradução de textos de e-commerce
[Coelho et al. 2024]	ChatGPT, Doc2Vec, C-LSTM, BERT	Prompts específicos com justificativa	Testes em pareceres jurídicos brasileiros	Precisão, recall, F1-score, acurácia, RMSE	Desempenho competitivo	Extração de informações jurídicas
[Choi 2024]	GPT-4 e 3.5 Turbo, RandomForest, SVM	Zero-shot, Few-shot, Chain-of-Thought	Comparação com humanos	Acurácia, precisão, recall, F1-score	GPT-4 comparável a humanos (F1=0.89)	Classificação de opiniões

O avanço dos LLMs tem proporcionado mudanças significativas no campo da inteligência artificial aplicada ao direito. Com a crescente digitalização dos processos jurídicos e a expansão do uso de documentos eletrônicos, os LLMs apresentam-se como uma ferramenta essencial para análise, classificação e geração de conteúdo legal. Neste contexto, diversos estudos têm sido conduzidos para avaliar o desempenho desses modelos na análise de documentos jurídicos. A partir das questões de pesquisa fundamentais (RQs), este trabalho busca explorar como os LLMs são empregados, suas limitações e potencialidades dentro do campo jurídico.

Modelos de Linguagem Utilizados (RQ1). Os modelos de linguagem empregados variam conforme a abordagem do estudo e o objetivo específico de cada análise. Muitos estudos utilizam modelos de uso genérico como GPT-3.5, GPT-4 e LLaMa 2 [Deroy et al. 2024, Wu et al. 2023, Venkatakrishnan et al. 2024, Deroy et al. 2023, Zhou et al. 2023, Kang et al. 2023, Hijazi et al. 2024, Chen et al. 2024, Choi 2024, Nunes et al. 2024]. Alguns estudos optam por usar também por modelos especializados, como Legal-Pegasus e o Legal-LED [Deroy et al. 2023, Deroy et al. 2024]. Enquanto outros optaram pelo BERT adaptados ao domínio jurídico [Zhou et al. 2023, Coelho et al. 2024]. Os estudos demonstram, que a escolha do modelo influencia diretamente a capacidade do sistema de interpretar textos legais, pois modelos treinados em domínios específicos demonstram melhor compreensão dos conceitos jurídicos e maior precisão na classificação de documentos e extração de informações.

Técnicas de Engenharia de Prompt Utilizadas (RQ2). Para otimizar o desempenho dos modelos de linguagem, diferentes técnicas de engenharia de prompt são utilizadas. As estratégias incluem zero-shot prompting, no qual o modelo é solicitado a realizar uma tarefa sem exemplos prévios, e few-shot prompting, que fornece alguns exemplos para guiar a resposta do modelo. O Chain of Thought (CoT) [Wei et al. 2022], por sua vez, incentiva o raciocínio passo a passo antes de fornecer uma resposta final. Outros métodos, como prompts adaptativos e reformulação dinâmica de consultas, também são utilizados para aprimorar a precisão das respostas, principalmente em contextos complexos como a extração de informações de pareceres jurídicos.

Avaliação da Solução (RQ3). A avaliação do desempenho dos modelos de linguagem no domínio jurídico ocorre por meio de diferentes abordagens. Algumas pesquisas utilizam métricas automáticas, como a similaridade entre resumos gerados e padrões de referência, enquanto outras adotam avaliação qualitativa conduzida por especialistas da área jurídica [Deroy et al. 2024, Wu et al. 2024]. Em estudos de classificação de documentos, por exemplo, os modelos são comparados com abordagens tradicionais, como aprendizado supervisionado e sistemas baseados em regras. Em determinados casos, os resultados são revisados por estudantes de direito e advogados para avaliar a coerência e precisão das respostas geradas.

Métricas Utilizadas na Avaliação (RQ4). A qualidade das saídas geradas pelos LLMs é medida através de um conjunto variado de métricas, dependendo da tarefa em questão. Para sumarização de textos, são amplamente empregadas métricas tradicionais para sumarização de texto, como ROUGE, METEOR e BLEU [Deroy et al. 2024, Deroy et al. 2023, Wu et al. 2024, Hijazi et al. 2024, Chen et al. 2024], que avaliam a similaridade entre os resumos automáticos e os produzidos por humanos. No caso de classificação de textos e extração de entidades nomeadas, as métricas mais utilizadas

são F1-score, Precisão e Recall [Coelho et al. 2024, Ghosh et al. 2023, Wu et al. 2023, Hijazi et al. 2024, Choi 2024]. Quando se trata de recuperação de informações, se observa o uso de métricas como MAP (Mean Average Precision) e NDCG (Normalized Discounted Cumulative Gain), que medem a relevância dos documentos retornados para uma consulta [Zhou et al. 2023].

Desempenho dos Modelos de Linguagem na Tarefa (RQ5). O desempenho dos LLMs varia significativamente de acordo com o modelo utilizado e a complexidade da tarefa. Em geral, modelos jurídicos especializados, como Legal-BERT e Sabiá, tendem a apresentar melhores resultados do que modelos genéricos como GPT-3.5 e GPT-4, especialmente em tarefas que envolvem terminologia técnica. No entanto, mesmo os modelos mais avançados enfrentam desafios relacionados às chamadas ”alucinações” — respostas que parecem plausíveis, mas contêm informações erradas. Para mitigar esse problema, algumas abordagens combinam aprendizado supervisionado com interação humana para refinar os resultados.

Domínio de Aplicação (RQ6). Os modelos de linguagem são aplicados a diversas áreas do direito, variando de acordo com a necessidade específica do estudo. Alguns artigos se concentram na sumarização de decisões judiciais em sistemas legais da Índia e do Reino Unido [Deroy et al. 2024], enquanto outros exploram a recuperação de casos jurídicos no contexto chinês [Zhou et al. 2023]. Há também aplicações voltadas para o reconhecimento de entidades nomeadas em textos legislativos brasileiros e a extração de informações de processos judiciais americanos. A versatilidade dos LLMs permite que sejam empregados em múltiplos contextos, desde a automação de tarefas rotineiras até análises complexas de cenários jurídicos.

4. Conclusões

O avanço dos Modelos de Linguagem de Grande Escala (LLMs) tem proporcionado impactos significativos na análise e síntese de documentos jurídicos. A crescente digitalização dos processos legais exige soluções cada vez mais eficazes para lidar com a complexidade e o volume crescente de informações. Nesse contexto, os LLMs se apresentam como ferramentas promissoras para otimizar a extração de informações, a sumarização de decisões judiciais e a automação de tarefas repetitivas no direito.

A revisão sistemática realizada neste estudo demonstrou que diferentes modelos de LLMs têm sido aplicados na área jurídica, desde os modelos generalistas, como GPT-4, Llama 2 e BERT, até aqueles mais especializados no contexto legal, como Legal-Pegasus e Sabiá. A escolha do modelo influencia diretamente o desempenho nas tarefas propostas, com os modelos treinados especificamente para o domínio jurídico apresentando maior precisão na classificação de textos, recuperação de jurisprudência e análise de contratos. Outro aspecto essencial abordado foi a engenharia de prompts, que se mostra crucial para maximizar o desempenho dos LLMs. Técnicas como Zero-shot Learning, Few-shot Learning e Chain-of-Thought prompting foram amplamente utilizadas para aprimorar a precisão das respostas geradas pelos modelos. Os estudos analisados demonstraram que a formulação adequada de prompts pode reduzir significativamente as alucinações e viéses, tornando os modelos mais confiáveis para aplicações jurídicas.

A avaliação das soluções apresentou grande diversidade de metodologias. Alguns estudos adotaram métricas quantitativas, como ROUGE, METEOR, BLEU e F1-score,

enquanto outros utilizaram avaliações qualitativas, com a participação de especialistas jurídicos para verificar a coerência e confiabilidade das respostas dos modelos. Os resultados indicam que, embora os LLMs tenham mostrado grande potencial na automação de tarefas jurídicas, ainda existem limitações que precisam ser superadas, como a necessidade de maior contextualização e a mitigação de erros sistemáticos.

No que diz respeito ao desempenho dos LLMs, observou-se que os modelos especializados na área jurídica tendem a superar os modelos generalistas em tarefas específicas. No entanto, ainda há desafios relacionados à interpretabilidade dos resultados e à dificuldade dos modelos em lidar com textos jurídicos altamente complexos e técnicos. Para que os LLMs sejam amplamente adotados no setor jurídico, é fundamental investir em treinamentos mais refinados e na criação de datasets jurídicos robustos que possam aprimorar a adaptação desses modelos ao domínio legal.

Por fim, a análise do domínio de aplicação dos LLMs revelou que essas tecnologias estão sendo empregadas em diversas frentes do direito, incluindo sumarização de decisões judiciais, extração de entidades nomeadas, recuperação de informações em bancos de dados jurídicos e predição de resultados de processos. No entanto, sua aplicação em larga escala ainda exige estudos mais aprofundados, especialmente para garantir a transparência, confiabilidade e conformidade com as normativas jurídicas vigentes.

Em conclusão, este estudo destacou que os LLMs representam uma ferramenta poderosa para transformar o setor jurídico, trazendo eficiência e acessibilidade ao processamento de informações legais. No entanto, seu uso ainda requer avanços na engenharia de prompts, no controle de viés e na confiabilidade dos modelos. Pesquisas futuras devem explorar estratégias de ajuste fino (fine-tuning) mais eficazes, além do desenvolvimento de modelos jurídicos mais especializados para diferentes sistemas legais e línguas. Dessa forma, será possível maximizar o impacto positivo dessas tecnologias no campo do direito, garantindo que os LLMs sejam aplicados de forma ética, precisa e eficiente

References

Abonizio, H., Almeida, T. S., Laitz, T., Junior, R. M., Bonás, G. K., Nogueira, R., and Pires, R. (2025). Sabiá-3 technical report.

Chen, K., Chen, B., Gao, D., Dai, H., Jiang, W., Ning, W., Yu, S., Yang, L., and Cai, X. (2024). General2specialized llms translation for e-commerce. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 670–673, New York, NY, USA. Association for Computing Machinery.

Choi, J. H. (2024). How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics (JITE)*, 180(2):214–233.

Coelho, G., Celecia, A., de Sousa, J., Lemos, M., Lima, M., Mangeth, A., Frajhof, I., and Casanova, M. (2024). Information extraction in the legal domain: Traditional supervised learning vs. chatgpt. In *Proceedings of the 26th International Conference on Enterprise Information Systems*, volume 1, pages 579–586.

Deroy, A., Ghosh, K., and Ghosh, S. (2023). How ready are pre-trained abstractive models and llms for legal case judgement summarization?

Deroy, A., Ghosh, K., and Ghosh, S. (2024). Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, pages 1–44.

Ghosh, S., Evuru, C. K., Kumar, S., Ramaneswaran, S., Sakshi, S., Tyagi, U., and Manocha, D. (2023). Dale: Generative data augmentation for low-resource legal nlp.

Hijazi, F., AlHarbi, S., AlHussein, A., Shairah, H. A., AlZahrani, R., AlShamlan, H., Knio, O., and Turkiyyah, G. (2024). Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. *arXiv preprint arXiv:2408.07983*.

Kang, X., Qu, L., Soon, L.-K., Trakic, A., Zhuo, T., Emerton, P., and Grant, G. (2023). Can chatgpt perform reasoning using the irac method in analyzing legal scenarios like a lawyer? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 13900–13923. Association for Computational Linguistics.

Moura, A. and Carvalho, A. A. (2023). Literacia de prompts para potenciar o uso da inteligência artificial na educação. *RE@ D-Revista de Educação a Distância e Elearning*, 6(2):e202308–e202308.

Nunes, R. O., Spritzer, A. S., Freitas, C. M. D. S., and Balreira, D. G. (2024). Out of sesame street: a study of portuguese legal named entity recognition through in-context learning. In *Proceedings of the 26th International Conference on Enterprise Information Systems*.

Peixoto, F. H. and Bonat, D. (2023). Gpts e direito: impactos prováveis das ias generativas nas atividades jurídicas brasileiras. *Sequência (Florianópolis)*, 44(93):e94238.

Prasad, N., Boughanem, M., and Dkaki, T. (2024). Exploring large language models and hierarchical frameworks for classification of large unstructured legal documents. In Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., and Ounis, I., editors, *Advances in Information Retrieval*, pages 221–237, Cham. Springer Nature Switzerland.

SILVA, D. H. S. d. (2023). Análise de sentimentos em decisões e entendimentos do supremo tribunal de justiça utilizando large language models.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Venkatakrishnan, R., Tanyildizi, E., and Canbaz, M. A. (2024). Semantic interlinking of immigration data using llms for knowledge graph construction. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 605–608, New York, NY, USA. Association for Computing Machinery.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wu, Y., Wang, C., Gumusel, E., and Liu, X. (2024). Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning.

Wu, Y., Zhou, S., Liu, Y., Lu, W., Liu, X., Zhang, Y., Sun, C., Wu, F., and Kuang, K. (2023). Precedent-enhanced legal judgment prediction with llm and domain-model collaboration.

Zhou, Y., Huang, H., and Wu, Z. (2023). Boosting legal case retrieval by query content selection with large language models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 176–184, New York, NY, USA. Association for Computing Machinery.