# Extração de Informação Jurídica com LLMs e RAG: um Estudo sobre Casos de Feminicídio

**Barbara Martínez de Lucena**[1], **Carina F. Dorneles**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brazil

barbara.lucena@posgrad.ufsc.br

carina.dorneles@ufsc.br

***Abstract.*** *Legal documents are long, complex, and written in technical language, which makes automatic information extraction challenging. This work investigates the use of Large Language Models (LLMs) to extract structured data from legal documents related to femicide cases. We compare two strategies: direct extraction with LLMs and a pipeline combining LLMs, Retrieval-Augmented Generation (RAG), and zero-shot prompts. The experiments use Qwen 2.5, Llama 3.3, and Gemma 3 to extract attributes related to the crime, investigation, and judicial process. The outputs are compared against a manually annotated dataset using accuracy by model and attribute. Preliminary results indicate that the RAG-based approach improves performance compared to direct extraction, contributing to the transformation of unstructured legal documents into structured data for femicide studies.*

***Resumo.*** *Documentos jurídicos são longos, complexos e escritos em linguagem técnica, o que dificulta a extração automática de informações. Este trabalho investiga o uso de Modelos de Linguagem de Grande Escala (LLMs) para extrair dados estruturados de documentos jurídicos sobre feminicídio. Comparamos duas estratégias: extração direta com LLMs e um pipeline que combina LLMs, Retrieval-Augmented Generation (RAG) e prompts zero-shot. Os experimentos utilizam Qwen 2.5, Llama 3.3 e Gemma 3 para extrair atributos relacionados ao crime, à investigação e ao processo judicial. Os resultados são comparados com uma base anotada manualmente por meio da acurácia por modelo e atributo. Os resultados preliminares indicam que a abordagem com RAG melhora o desempenho em relação à extração direta, contribuindo para transformar documentos jurídicos não estruturados em dados estruturados para estudos sobre feminicídio.*

**Informações Gerais. Nível:** Mestrado. **Aluno:** Barbara Rarine Nery Martínez de Lucena. **Orientador:** Carina Friedrich Dorneles. **Programa:** Programa de Pós-Graduação em Ciências de Computação da Universidade Federal de Santa Catarina. **Admissão:** Agosto 2025. **Exame de qualificação:** Dezembro 2026. **Defesa prevista:** Agosto 2027. **Etapas concluídas:** Créditos obrigatórios em disciplinas, revisão bibliográfica, definição do problema.

## 1. Introdução

Pesquisas em Grandes Modelos de Linguagem (do inglês, Large Language Models - LLMs) avançaram significativamente nos últimos anos, possibilitando a extração, organização e transformação de informações provenientes de grandes volumes de textos não estruturados em dados estruturados e consultáveis [Lai et al. 2024]. Esses modelos têm demonstrado forte capacidade em tarefas de extração de informação, interpretação semântica e identificação de atributos em fontes textuais complexas [Lewis et al. 2020]. Tais avanços têm apoiado aplicações em diversos domínios, incluindo análise de literatura biomédica, relatórios financeiros, automação de atendimento ao cliente e processamento de documentos jurídicos. No entanto, apesar de sua crescente adoção, o uso de LLMs em contextos jurídicos ainda apresenta desafios importantes, especialmente quando os documentos são longos, heterogêneos e escritos em linguagem técnica.

Documentos jurídicos são frequentemente densos, formais e semanticamente complexos, contendo terminologia específica do domínio e informações distribuídas em diferentes seções ou documentos [Ashley 2017, Chalkidis et al. 2020]. Em processos criminais, e particularmente em casos de feminicídio, dados relevantes sobre vítimas, autores, circunstâncias do crime, investigações e tramitação judicial podem estar dispersos em boletins de ocorrência, inquéritos policiais, denúncias, decisões judiciais e outros documentos processuais, tornando a extração manual custosa e sujeita a inconsistências. Embora LLMs tenham sido exploradas para o processamento de textos jurídicos, a extração de informação permanece desafiadora devido a contextos extensos, fundamentação factual inconsistente e à necessidade de interpretação específica do domínio [Chalkidis et al. 2022, El Hamdani et al. 2024]. Além disso, o uso direto de prompts pode produzir saídas incompletas ou inconsistentes quando as informações relevantes estão dispersas em documentos longos [Liu et al. 2024]. Nesse contexto, Retrieval-Augmented Generation (RAG) apresenta-se como uma estratégia promissora ao recuperar passagens relevantes antes da geração, melhorando a fundamentação contextual das saídas dos modelos [Lewis et al. 2020].

Neste trabalho, investigamos o uso de LLMs para extrair automaticamente informações estruturadas de documentos jurídicos relacionados a casos de feminicídio em um estado brasileiro. Inicialmente, avaliamos o uso direto de diferentes LLMs, incluindo Gemma, Llama e Qwen, para extrair atributos predefinidos de documentos jurídicos utilizando prompts zero-shot. Com base nas limitações observadas nessa primeira abordagem, propomos um pipeline que combina LLMs, RAG e prompts zero-shot projetados para produzir saídas estruturadas em JSON para tarefas específicas de extração. O pipeline inclui extração de texto de arquivos PDF, segmentação dos documentos, indexação vetorial, recuperação de passagens relevantes e extração baseada em prompts de atributos relacionados a circunstâncias do crime, detalhes da investigação e informações do processo judicial.

Os dados extraídos são comparados com uma base de referência anotada manualmente, a fim de avaliar acurácia, valores ausentes e inconsistências entre modelos e abordagens. A principal contribuição deste trabalho é uma metodologia para transformar documentos jurídicos não estruturados em dados estruturados, apoiando estudos empíricos sobre feminicídio e contribuindo para pesquisas em bancos de dados e extração de informação.

Este artigo está organizado da seguinte forma. A Seção 2 apresenta conceitos fundamentais e trabalhos relacionados sobre LLMs, processamento de textos jurídicos e RAG. A Seção 3 descreve a metodologia de extração proposta e a configuração experimental. A Seção 4 apresenta e discute a avaliação e os resultados preliminares. Por fim, a Seção 5 resume os principais achados e aponta direções para trabalhos futuros.

## 2. Trabalhos relacionados

Nesta seção, apresentamos uma breve visão geral dos avanços recentes na interseção entre Processamento de Linguagem Natural (PLN), Modelos de Linguagem de Grande Escala (LLMs) e o domínio jurídico, com atenção especial ao processamento de textos jurídicos e à extração de informações de documentos jurídicos não estruturados. Nesse contexto, a pesquisa sobre LLMs e Direito refere-se a como esses modelos podem ser adaptados, avaliados e aplicados ao processamento de estatutos, jurisprudência, contratos, regulamentos, decisões judiciais, registros policiais e outras formas de linguagem jurídica, ao mesmo tempo em que examina suas capacidades e limitações em fluxos de trabalho jurídicos reais.

Estudos recentes têm mostrado a crescente relevância das LLMs para o processamento de textos jurídicos. Sun et al. [Sun 2023] documentam a expansão do uso de LLMs em tarefas como predição de julgamentos, raciocínio estatutário, redação jurídica, tutoria jurídica e assistência consultiva automatizada. No entanto, os autores também destacam desafios persistentes relacionados à privacidade, propriedade intelectual, equidade e vieses em sistemas jurídicos automatizados. Essas preocupações são especialmente relevantes ao lidar com dados jurídicos e criminais sensíveis, nos quais erros ou inferências sem suporte podem afetar a confiabilidade e a interpretabilidade das informações extraídas.

Complementando essa perspectiva, Katz et al. [Katz et al. 2023] apresentam um mapeamento em larga escala do panorama de pesquisa em PLN jurídico, documentando o rápido crescimento da área e sua transição de métodos baseados em regras e estatísticos para modelos neurais e baseados em transformers. Seus resultados revelam desafios estruturais persistentes, incluindo disponibilidade limitada de conjuntos de dados, barreiras de reprodutibilidade e complexidade linguística, que restringem a aplicação de ferramentas de PLN em fluxos de trabalho jurídicos reais. Essas limitações estão diretamente relacionadas ao problema abordado neste trabalho, uma vez que a construção de bases de dados estruturadas a partir de documentos jurídicos depende tanto de métodos de extração precisos quanto de uma avaliação confiável em relação a referências anotadas manualmente.

Zhong et al. [Zhong et al. 2020] contextualizam ainda mais a área ao distinguir entre abordagens baseadas em símbolos, que enfatizam conhecimento jurídico explícito e interpretabilidade, e abordagens baseadas em embeddings, que utilizam representações orientadas por dados para apoiar tarefas como correspondência de casos, predição de julgamentos, extração de informação e resposta a perguntas jurídicas. Essa distinção é relevante para o presente estudo porque o pipeline proposto combina métodos de recuperação orientados por dados com regras de extração baseadas em prompts, buscando preservar algum grau de controle sobre as informações extraídas dos documentos jurídicos.

Além disso, Siino et al. [Siino et al. 2025] revisam aplicações de LLMs em diversas tarefas de PLN jurídico, incluindo busca jurídica, revisão de documentos jurídicos e predição jurídica. O trabalho destaca avanços recentes em recuperação em documentos

longos, análise de contratos, inferência entre casos e conjuntos de dados jurídicos multilíngues. Esses desenvolvimentos são particularmente importantes para a extração de informação jurídica, uma vez que evidências ou atributos relevantes podem aparecer apenas em pequenas porções de documentos extensos. Nesse cenário, Retrieval-Augmented Generation (RAG) surgiu como uma estratégia promissora, pois recupera passagens relevantes antes da geração e pode reduzir a dependência do conhecimento interno do modelo ou de contextos de entrada excessivamente longos.

Em conjunto, esses estudos mostram que as LLMs estão reformulando a forma como textos jurídicos são buscados, analisados, resumidos e estruturados. Ao mesmo tempo, eles enfatizam desafios em aberto relacionados à qualidade dos dados, fundamentação factual, interpretabilidade, reprodutibilidade e uso responsável. Este trabalho situa-se nesse contexto ao investigar a extração de informações estruturadas de documentos jurídicos relacionados a casos de feminicídio. Em vez de focar em raciocínio jurídico geral ou predição, o objetivo é avaliar como LLMs, combinadas com RAG e prompts zero-shot com saída JSON, podem apoiar a transformação de documentos jurídicos não estruturados em dados estruturados para análise empírica.

## 3. Metodologia

Para investigar como diferentes estratégias de extração afetam o desempenho de LLMs na extração de informação jurídica, este estudo foi organizado em duas etapas experimentais. Os experimentos focaram em três LLMs de pesos abertos, pertencentes a diferentes famílias de modelos e escalas de parâmetros: Qwen 2.5 (72B), Llama 3.3 (70B) e Gemma 3 (27B). Esses modelos foram selecionados por representarem alternativas recentes de alta capacidade e pesos abertos que podem ser executadas em um ambiente computacional local, um requisito importante ao lidar com documentos jurídicos sensíveis e dados de acesso restrito.

O conjunto de dados inicial consistiu em 55 casos relacionados a feminicídio do ano de 2021. Oito atributos foram selecionados para extração: *Data do fato*, *Município*, *Meio empregado*, *Local*, *Horário*, *Preventiva*, *Flagrante* e *Diligências*. Esses atributos abrangem informações sobre o fato criminoso, sua localização, o meio empregado e elementos da investigação policial. Os atributos *Preventiva*, *Flagrante* e *Diligências* foram modelados como campos booleanos, com valores possíveis restritos a SIM ou NÃO. Esses atributos foram selecionados por representarem informações centrais para a caracterização inicial dos casos, abrangendo dimensões temporais, geográficas, circunstanciais e investigativas do processo.

A Figura 1 apresenta uma visão geral da metodologia adotada neste estudo. Na primeira etapa (V1), conduzida no final de 2025, as três LLMs foram avaliadas sem RAG. Nesse cenário, o conteúdo textual extraído dos documentos jurídicos foi fornecido diretamente aos modelos, juntamente com prompts zero-shot projetados para extrair os atributos predefinidos. Os modelos não foram ajustados por fine-tuning e não receberam exemplos anotados nos prompts; em vez disso, os prompts descreviam a tarefa de extração, as regras de extração e o formato estruturado esperado para a saída.

Na segunda etapa (V2), implementamos um pipeline de extração baseado em RAG. Nessa abordagem, os documentos em PDF são convertidos em texto e segmentados em chunks. Como os documentos jurídicos são extensos, utilizamos chunks de

500 caracteres com sobreposição de 300 caracteres. Essa configuração foi adotada para preservar o contexto local entre segmentos adjacentes, ao mesmo tempo em que limita a quantidade de informação irrelevante fornecida aos modelos. Os chunks foram indexados para recuperação e posteriormente selecionados por meio de consultas associadas a cada tarefa de extração. Para cada extração, as passagens recuperadas foram fornecidas à LLM juntamente com um prompt zero-shot instruindo o modelo a retornar a informação extraída.

A mesma estratégia geral de prompting foi utilizada nas duas etapas. As saídas dos modelos foram comparadas com uma base de dados anotada manualmente. Antes da avaliação, tanto os valores de referência quanto os valores preditos foram normalizados para reduzir diferenças superficiais de formatação.

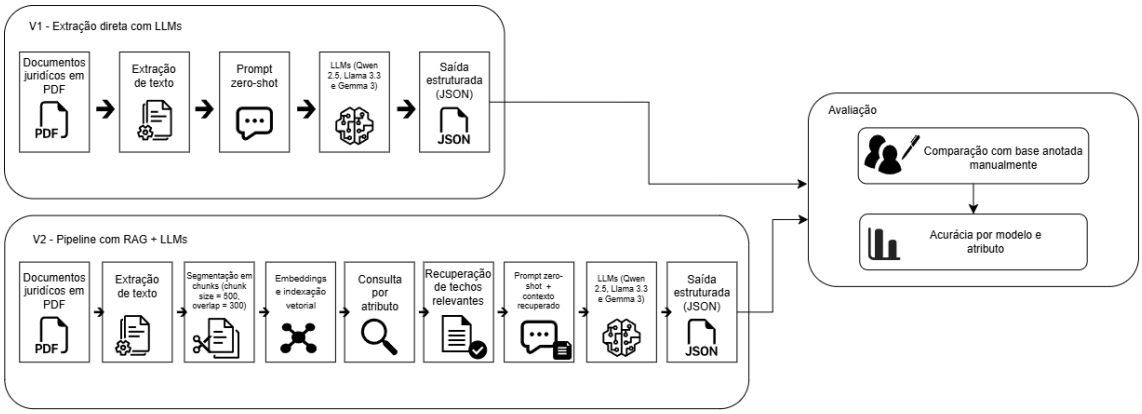


**Figura 1. Visão geral do pipeline de extração de informações em documentos jurídicos de feminicídio.**

## 4. Resultados preliminares

Devido à sensibilidade dos documentos jurídicos e dos dados pessoais envolvidos, trechos reais dos documentos não são reproduzidos integralmente neste trabalho. Assim, os resultados são apresentados de forma agregada, por meio de métricas por modelo e atributo, evitando a exposição de informações identificáveis dos casos analisados. Os resultados preliminares mostram que a segunda versão do pipeline de extração melhorou substancialmente o desempenho dos três modelos de linguagem quando comparada à primeira versão. Essa melhoria é especialmente visível em atributos como *Data do fato*, *Horário*, *Flagrante* e *Preventiva*, sugerindo que os prompts revisados e as regras de normalização contribuem para uma extração mais consistente.

Entre os três modelos, Qwen e Llama alcançaram os melhores resultados gerais na V2. O Qwen obteve a maior acurácia em atributos como *Município*, *Horário*, *Flagrante* e *Diligências*. O Llama alcançou o melhor resultado para *Local*, além de apresentar um desempenho geral forte nos demais atributos.

A Figura 2 resume a acurácia obtida pelos três modelos de linguagem nos atributos avaliados, comparando a primeira e a segunda versões do pipeline de extração.

De modo geral, os resultados indicam que a V2 tornou a extração mais robusta e consistente em comparação à abordagem direta com LLMs. A melhora observada nos três

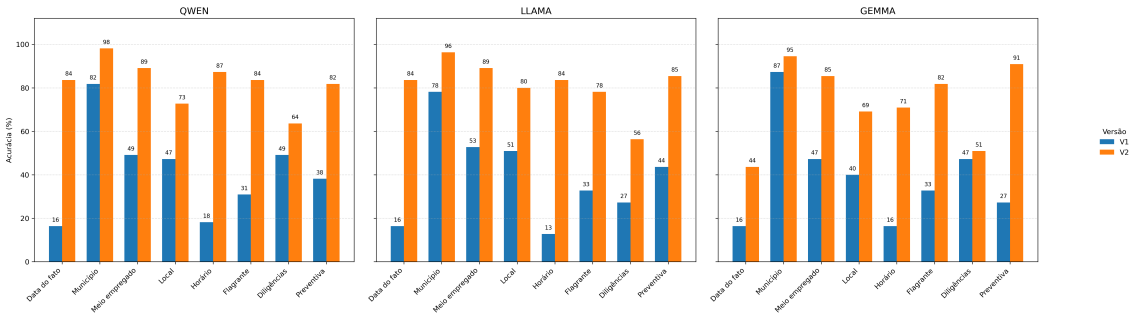


**Figura 2. Comparação de acurácia entre V1 e V2 nos modelos de linguagem e atributos avaliados.**

modelos sugere que a recuperação de trechos relevantes antes da geração contribuiu para orientar melhor a extração dos atributos. No entanto, campos que dependem de maior interpretação contextual ou de informações distribuídas nos documentos, como *Local* e *Diligências*, ainda apresentam maior variação entre os modelos. Esses achados indicam que a combinação entre RAG, prompts zero-shot e normalização melhora a extração automática, mas ainda requer refinamentos adicionais.

## 5. Conclusões e trabalhos futuros

Este trabalho investigou o uso de Modelos de Linguagem de Grande Escala para extração automática de informações estruturadas a partir de documentos jurídicos relacionados a casos de feminicídio. A pesquisa foi organizada em duas etapas: a primeira baseada na extração direta com LLMs, e a segunda baseada em um pipeline que combina LLMs, RAG e prompts zero-shot com saída estruturada. Os resultados preliminares indicam que a segunda versão do pipeline apresentou melhora consistente em relação à primeira.

De modo geral, os resultados sugerem que a recuperação de trechos relevantes antes da geração contribui para tornar a extração mais robusta em documentos jurídicos longos e não padronizados. Além disso, a comparação entre os modelos mostrou que Qwen e Llama apresentaram os melhores desempenhos gerais na V2, enquanto Gemma obteve ganhos relevantes em alguns atributos. Esses achados reforçam o potencial da combinação entre LLMs, RAG, prompts zero-shot e normalização para apoiar a transformação de documentos jurídicos não estruturados em dados estruturados.

Como trabalho futuro, pretendemos expandir as abordagens atuais de RAG Hiperbólico baseadas no Disco de Poincaré ao migrar para o modelo de Lorentz com curvatura variável para analisar documentos de feminicídio. Em vez de adotar uma curvatura fixa e global, avaliaremos o impacto da variação do parâmetro $K$ na recuperação de trechos relevantes em documentos jurídicos sobre feminicídio. Essa abordagem busca ajustar a representação geométrica conforme a complexidade do texto, utilizando curvaturas mais acentuadas para trechos com maior densidade hierárquica e curvaturas mais brandas para relatos narrativos ou cronológicos. Também será analisado se a estabilidade numérica da métrica de Lorentz contribui para melhorar a recuperação de informação e o contexto fornecido às LLMs. Por fim, serão incluídos exemplos sintéticos da pipeline, preservando a confidencialidade dos dados.

## Referências

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330. Association for Computational Linguistics.

El Hamdani, R., Bonald, T., Malliaros, F., Holzenberger, N., and Suchanek, F. (2024). The factuality of large language models in the legal domain. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery.

Katz, D. M., Hartung, D., Gerlach, L., Jana, A., and Bommarito, M. J. (2023). Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Lai, J., Gan, W., Wu, J., Qi, Z., and Yu, P. S. (2024). Large language models in law: A survey. *AI Open*, 5:181–196.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Siino, M., Falco, M., Croce, D., and Rosso, P. (2025). Exploring LLMs applications in law: A literature review on current legal NLP approaches. *IEEE Access*, 13:18253–18276.

Sun, Z. (2023). A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.