Extração de Dados Não Estruturados de Promoções de Arquivamento de Inquéritos via Arquiteturas Multiagentes e LLMs

Gustavo C. M. Moreira.[1], **Victor Moreli**[1] , **Luis G. Nonato**[1]

[1]ICMC - Instituto de Ciências Matemáticas e de Computação

gucmm@hotmail.com, victormoreli@usp.br, gnonato@icmc.usp.br

Abstract. *This study analyzed the use of LLMs and multi-agent architectures to structure non-regularized police inquiry data. The results indicate that while powerful models ensure higher accuracy, high task modularization helps smaller models adapt to the required output schema. However, such task fragmentation leads to higher token consumption and reduces effectiveness in shorter texts, ultimately limiting the financial viability of the solution for large-scale applications.*

Resumo. *Este estudo tem como objetivo principal a estruturação de dados textuais de inquéritos policiais do MPSP por meio de LLMs e arquiteturas multiagentes. Para tanto, avaliaram-se diferentes modelos e graus de modularização de prompts, verificando seu impacto na precisão e no consumo de recursos. Os resultados indicam que modelos mais robustos asseguram maior precisão, ao passo que a alta modularização facilita a adesão ao esquema de saída desejado. Por outro lado, tal fragmentação resulta em maior consumo de tokens e reduz a eficácia em textos menos extensos, limitando a viabilidade financeira da solução em larga escala.*

1. Introdução

A alta criminalidade impõe severos desafios ao sistema de justiça brasileiro. Apesar do alto número homicídios desde 1980 [Cerqueira and Bueno 2024], a taxa de resolução alcançou apenas 39% em 2022 no Brasil, índice muito inferior à média global de 63% [Instituto Sou da Paz 2024]. Essa ineficiência reflete-se na formulação de políticas de segurança, que frequentemente carecem de embasamento científico [Adorno 2022]. Além disso, o volume excessivo, a omissão de informações relevantes e a despadronização dos registros policiais dificultam o uso de técnicas convencionais de análise para identificação de padrões criminais [Instituto Sou da Paz 2024].

Outro fator importante é que informações relevantes relacionadas a ocorrências criminais estão descritas de forma textuais em boletins de ocorrência e inquéritos policiais. Assim, técnicas de Processamento de Linguagem Natural (PLN), especialmente as LLMs, mostram-se promissoras devido à sua alta capacidade de representação contextual [Minaee et al. 2024]. O foco deste trabalho é estruturar dados a partir de promoções de arquivamento de inquéritos do Ministério Público do Estado de São Paulo (MPSP) fornecidos na forma de texto, viabilizando análises estatísticas sobre crimes contra a vida. Para isso, investigam-se as arquiteturas multiagentes e os modelos de linguagem mais adequados à tarefa.

Diante disso, as principais contribuições deste trabalho[1] são: (i) Estruturação de Dados Jurídicos: desenvolvimento de um pipeline para a conversão de inquéritos do MPSP; (ii) Avaliação de Arquiteturas: análise do impacto da modularização de prompts na precisão e no consumo de recursos; (iii) Benchmark de Modelos: comparação de desempenho entre LLMs (Gemini e Sabiá) quando empregadas ao português jurídico; e (iv) Identificação de Desafios Técnicos: mapeamento de obstáculos críticos na automação documental, como alucinações e conformidade com o esquema de extração.

2. Trabalhos Relacionados

A aplicação de PLN em documentos jurídicos tem avançado significativamente, impulsionada pelo uso de LLMs. Como destacado por [Zhong et al. 2020], técnicas de PLN vêm sendo utilizadas para a recuperação e análise de documentos legais, reduzindo o esforço manual em tarefas repetitivas mesmo antes do surgimento de *Deep Learning*.

No contexto das LLMs, observa-se uma rápida transição do simples uso de engenharia de *prompt* para sistemas complexos de orquestração de agentes. O framework *TabAgent* [Wu et al. 2025], por exemplo, divide a tarefa entre agentes especializados (Esquema, Extração, Semântico e Validação) para superar a dificuldade dos LLMs em seguir rigidamente restrições estruturais. Essa arquitetura mitiga problemas de alucinação e perda de contexto comuns em abordagens *zero-shot*.

Já no cenário brasileiro, o sistema de [Batitucci et al. 2026] (*Judicial Multi-Agent Metadata Extraction*) corrobora a eficácia dessa modularização ao aplicar *pipelines* multiagentes para a extração de metadados de acórdãos. O sucesso de sistemas com métricas de desempenho tão robustas no contexto nacional reforça a viabilidade de aplicar técnicas semelhantes no domínio deste artigo, que lida com linguagem e estrutura comparáveis às dos acórdãos. Ainda no contexto nacional, [Luz de Araujo et al. 2018] introduziram o LeNER-Br, um *dataset* pioneiro para o Reconhecimento de Entidades Nomeadas (NER) no domínio jurídico. Os autores anotaram manualmente documentos de diversos tribunais para identificar entidades como pessoas, organizações e dispositivos legais.

Em conjunto, os trabalhos mencionados acima evidenciam dois avanços centrais: a eficácia das arquiteturas multiagentes na conformidade estrutural das extrações [Wu et al. 2025, Batitucci et al. 2026] e a maturidade das técnicas de NER para o português jurídico [Luz de Araujo et al. 2018]. Contudo, identificam-se lacunas relevantes. Primeiro, nenhum dos estudos analisados trata de promoções de arquivamento de inquéritos policiais. Segundo, os trabalhos de [Batitucci et al. 2026] e [Wu et al. 2025] não investigam sistematicamente o impacto do grau de modularização, ou seja, quantas chamadas independentes compõem o pipeline sobre a precisão e o consumo de recursos. O presente trabalho busca preencher ambas as lacunas ao aplicar e comparar cinco arquiteturas de extração com diferentes níveis de segmentação sobre um corpus de inquéritos do MPSP, domínio ainda inexplorado na literatura de PLN jurídico brasileiro.

[1] O repositório com os códigos está disponível em: `https://github.com/vmoreli/MPSP_Extraction`

3. Metodologia

3.1. Coleta, Parsing e Filtragem dos dados

A construção do corpus ocorreu em três etapas: filtragem por metadados, extração de arquivos e controle de qualidade. Identificaram-se 117.079 processos de crimes contra a vida (homicídios, feminicídios e latrocínios) nos metadados do MPSP, obtidos do site do MPSP no período de 2018 a 2024. Desse total, 48.263 (41,2%) foram convertidos de PDF para texto; os demais foram descartados por indisponibilidade ou inconsistência com os metadados. Após isso, houve a remoção de 1.059 documentos *outliers* pelo volume de tokens, arquivos muito pequenos apenas referenciavam documentos externos, e os muito grandes continham excesso de ruído. Por fim, o corpus final consolidou-se em 47.204 processos.

3.2. Criação do esquema, Conjunto Verdade e Prompts

Para construir o conjunto verdade e validar as extrações, definiu-se um esquema em parceria com especialistas do Núcleo de Estudos da Violência (NEV). A partir de 100 documentos selecionados aleatoriamente do corpus final de 47.204 processos do MPSP, realizou-se uma extração preliminar com o modelo Sabiazinho-3.1. Essas saídas foram então revisadas por especialistas da área jurídica para consolidar a base.

Quanto à elaboração dos *prompts*, utilizou-se a atribuição de *Persona*, técnica em que se instrui o modelo a assumir um papel específico, como o de promotor de justiça, direcionando-o ao domínio legal e aplicaram-se restrições explícitas para tentar mitigar alucinações [White et al. 2023] .

3.3. Escolha dos Modelos de Linguagem de Grande Escala (LLMs) e Arquiteturas de Extração

A extração empregou os LLMs Sabiazinho 3.1, Sabia 3.1, Gemini 2.5 Flash e Gemini 2.5 Pro. A seleção justifica-se pela liderança dos modelos Gemini no *ranking* em português [Garcia 2024] e pela relevância nacional dos modelos da Maritaca AI, cujo Sabiazinho serviu como referência de custo-benefício. Custos e posições constam na Tabela 1.

Tabela 1. Estimativa de custo por modelo para processamento do corpus completo e posição no LeaderBoard

Modelo	Tokens Entrada	Tokens Saída	Custo Total R$	LLM Leaderboard
SABIAZINHO-3.1	583.736.525	117.226.886	654,79	34°
SABIA-3.1	583.736.525	117.226.886	2.863,67	8°
GEMINI-2.5-FLASH	574.774.691	114.737.347	2.526,02	2°
GEMINI-2.5-PRO	574.774.691	114.737.347	10.262,13	1°

As tarefas de extração foram segmentadas em cinco grupos de campos: Mapeamento do inquérito (M), que inclui atributos como número do processo, data de abertura e tipo de crime; Informações do inquérito (I), abrangendo local, data e hora do fato, meio utilizado e motivação do arquivamento; e identificação de Suspeitos (S), Testemunhas (T) e Vítimas (V), com campos como nome, idade, sexo e vínculo das demais partes.

As estratégias avaliadas variam de uma única chamada englobando todos os campos (MISTV) até a máxima modularização com cinco chamadas independentes (M + I +

S + T + V), com configurações intermediárias de duas (MI + STV), três (M + I + STV) e quatro chamadas (MI + S + T + V). As combinações foram baseadas na afinidade temática dos campos, agrupando, por exemplo, os sujeitos (S, T, V) pela sua natureza similar.

3.4. Métricas de Avaliação

A avaliação utiliza uma arquitetura híbrida de PLN que combina verificações exatas e semânticas. Campos de texto livre, como a razão para o arquivamento e a arma do crime, foram avaliados por meio de *embeddings* e do cálculo da similaridade de cosseno, validando a equivalência semântica entre as extrações e o conjunto verdade mesmo na presença de variações vocabulares. Para a identificação de nomes (vítimas, suspeitos e testemunhas), são aplicadas lógicas de NER e alinhamento de listas, mensuradas por métricas de Precisão, Revocação e F1 Score. Campos categóricos, como tipo de crime e sexo do indivíduo, exigem correspondência literal, enquanto a abordagem semântica garante flexibilidade na análise dos campos textuais não estruturados.

4. Experimentos e Resultados

4.1. Comparação entre os Modelos de Linguagem de Grande Escala (LLMs)

A avaliação individual dos modelos foi padronizada em cinco chamadas devido à instabilidade da Maritaca AI em prompts complexos. Esses modelos frequentemente ignoravam o esquema, exigindo reprocessamentos e gerando consumo excessivo de tokens, o que impediu a análise da amostra completa. A seguir, apresentam-se os resultados dos quatro modelos selecionados validados contra o conjunto-verdade.

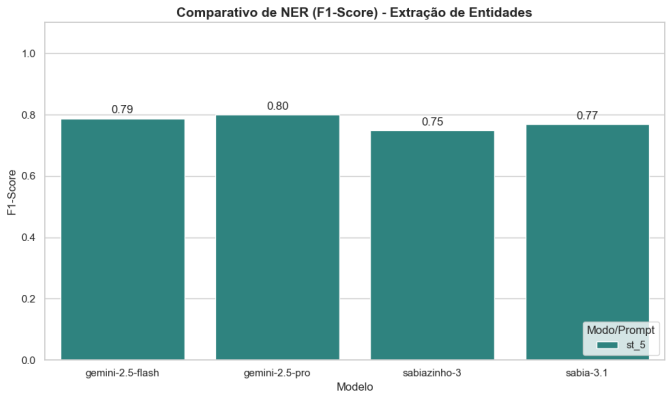

Figura 1. Resultados de F1-Score para a tarefa de Reconhecimento de Entidades Nomeadas (NER) por modelo

Os resultados apresentados na Figura 1 demonstram que os modelos do Gemini apresentaram o melhor desempenho geral na tarefa de extração. O Gemini 2.5 Pro obteve os índices de F1-score mais elevados, consolidando-se como o modelo de referência para precisão. Contudo, é importante ressaltar que essa superioridade técnica acompanha um custo financeiro significativamente maior, conforme detalhado na estimativa de gastos da Tabela 1.

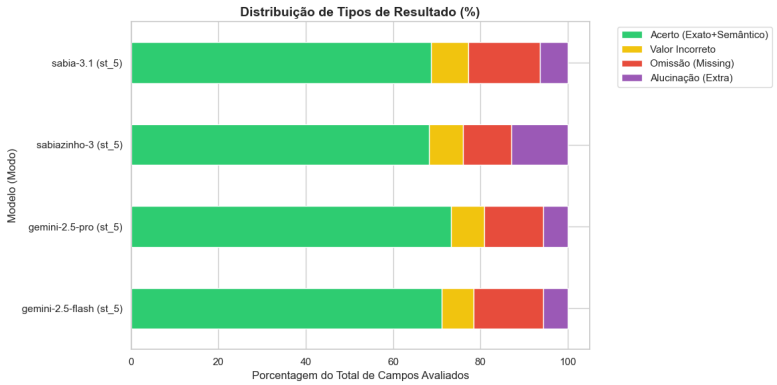

Figura 2. Distribuição de Erros dos Diferentes Modelos

Os resultados da Figura 2 corroboram o desempenho observado no F1-score, com o Gemini 2.5 Pro liderando em acertos e apresentando os menores índices de alucinação e erro. Essa hierarquia de precisão é mantida de forma decrescente pelo Gemini 2.5 Flash, seguido pelo Sabiá 3.1 e, por fim, pelo Sabiazinho 3.1. Vale notar, contudo, que o Sabiazinho 3.1 apresenta o menor índice de omissão entre todos os modelos avaliados.

4.2. Comparação entre as Arquiteturas

As avaliações das arquiteturas foram conduzidas exclusivamente com modelos da família Gemini. Como os resultados do Flash e do Pro foram muito semelhantes apesar da diferença de custo (Tabela 1), os experimentos de arquitetura foram conduzidos apenas com o Gemini 2.5 Flash.

No contexto do artigo, nós referem-se às chamadas independentes à LLM dentro do pipeline de extração. Cada nó é uma requisição separada ao modelo, responsável por extrair um subconjunto específico de campos do documento.

Diante dessa escolha, obtiveram-se os resultados apresentados nas Figuras 3, 4 e 5.

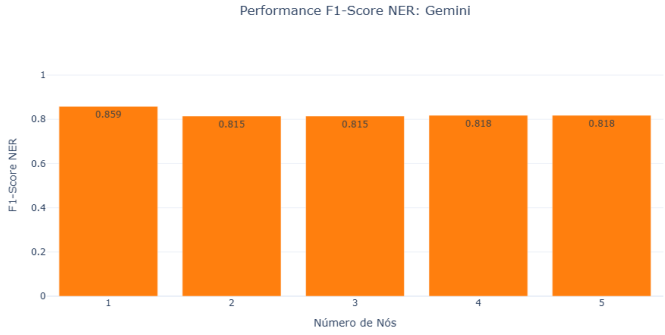

Figura 3. Resultados de F1-Score para a tarefa de Reconhecimento de Entidades Nomeadas (NER) por modelo

Vale notar que os valores de F1-Score da Figura 3 diferem dos anteriores pois, nos testes precedentes, 6 documentos com erros de extração foram excluídos (94 de 100),

enquanto aqui utilizaram-se todos os 100 registros válidos. Além disso, as arquiteturas com 2 e 3 nós, e as de 4 e 5 nós, apresentam valores idênticos pois cada par compartilha os mesmos prompts para NER, mantendo o F1-Score constante entre eles.

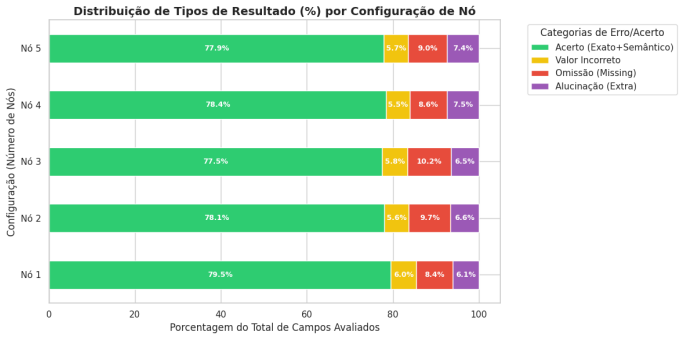

Figura 4. Distribuição de Erros das Diferentes Arquiteturas

Pela Figura 4, nota-se que as arquiteturas não possuem grande discrepância entre si, mas a configuração de 5 nós apresenta um destaque positivo, pois possui o maior volume de acertos e o menor índice de alucinação, este último sendo o tipo de erro mais crítico no contexto jurídico. Outro ponto relevante é a tendência de crescimento das alucinações conforme aumenta a modularização das tarefas.

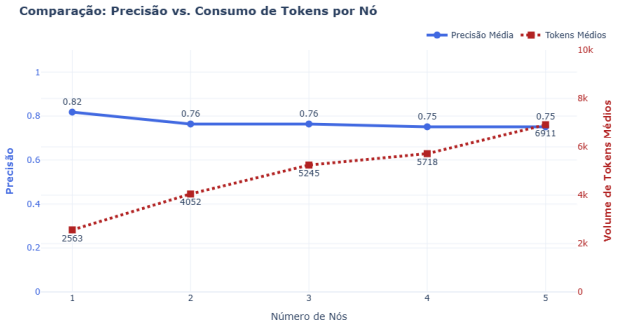

Figura 5. Consumo Médio e Precisão Média por Diferentes Arquiteturas

Além das métricas de performance, analisou-se o consumo médio de tokens por arquitetura. O gráfico da Figura 5 revela uma relação inversamente proporcional entre o volume de tokens e a precisão média: à medida que o número de nós aumenta, o consumo de recursos cresce, enquanto a precisão média apresenta um leve declínio.

5. Limitações

As limitações deste estudo incluem, principalmente, a baixa representatividade estatística e potenciais vieses, decorrentes da validação manual restrita a 100 documentos. Além disso, ruídos nos metadados oficiais retiveram casos atípicos na amostra (como mortes naturais e acidentes), gerando divergências com as regras do esquema proposto.

6. Conclusão

Este estudo indicou que é viável a extração estruturada de promoções de arquivamento de inquéritos do MPSP via LLMs. Os modelos Gemini destacaram-se, sendo o 2.5 Pro superior em precisão e o 2.5 Flash o melhor em custo-benefício. A avaliação das arquiteturas reforça que a modularização favorece a adesão ao schema, mas eleva o consumo de tokens e prejudica a precisão e o F1-Score em documentos curtos. Por fim, a metodologia proposta pode ser uma base para a automatização da análise de arquivamentos de inquéritos, abrindo caminho para ferramentas diagnósticas baseadas em evidências.

Uso de Inteligência Artificial e Origem Do Trabalho

Utilizou-se IA generativa para a revisões pontuais gramaticais do texto, mas todo o conteúdo foi desenvolvido exclusivamente pelos autores. Esse artigo teve sua origem a partir de uma Iniciação Científica financiada pela FAPESP.

Referências

Adorno, S. (2022). O fracasso do controle legal dos crimes e da violência na sociedade brasileira contemporânea. Acesso em: 8 mar. 2025.

Batitucci, L. A., Lopes, L. I., Ferreira, R., and Paraiso, E. C. (2026). Agent orchestration - LLM for legal metadata extraction: A comparative analysis of efficiency and precision. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 727–737, Salvador, Brazil. Association for Computational Linguistics.

Cerqueira, D. and Bueno, S. (2024). *Atlas da Violência 2024*. Ipea and FBSP, Brasília. Acesso em: 25 fev. 2025.

Garcia, E. A. S. (2024). Open portuguese llm leaderboard. Acesso em: nov. 2025.

Instituto Sou da Paz (2024). Onde mora a impunidade?

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil. Springer.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2311.00010*.

White, J., Fu, Q., Hays, S., Vierhauser, M., Neill, R., Wiburne, E., Dorn, B., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt.

Wu, J., Han, J., and Gao, Y. (2025). Tabagent: A multi-agent table extraction framework for unstructured documents. In *2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF)*, pages 600–607.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., and Sun, M. (2020). How does nlp benefit the legal system: A summary of legal artificial intelligence. *CoRR, abs/2004.12158*.