Evaluating LLMs for Healthcare-Related Named Entity Recognition in Brazilian Judicial Decisions

Elias Jacob de Menezes-Neto
Metropole Digital Institute
Federal University of Rio Grande do Norte
Natal, Brazil
elias.jacob@ufrn.br

Fabio Luiz de Oliveira Bezerra
Department of Private Law / 7th Federal Court
Federal University of Rio Grande do Norte / 5th Regional Federal Court
Natal, Brazil
fabiobezerra@jfrn.jus.br

Marco Bruno Miranda Clementino
Department of Private Law / 6th Federal Court
Federal University of Rio Grande do Norte / 5th Regional Federal Court
Natal, Brazil
marcobruno@jfrn.jus.br

***Abstract*— The increasing reliance on the judiciary to secure access to healthcare in Brazil has created a vast corpus of legal rulings that contain insights into healthcare delivery. However, extracting actionable information from these unstructured texts remains a significant challenge. This study evaluates the performance of state-of-the-art large language models (LLMs) for healthcare-related named entity recognition (NER) in Brazilian judicial decisions. Using our release dataset, LexCare.BR, a manually annotated gold-standard dataset of 1,200 legal rulings, we assessed 10 predefined healthcare entities across multiple LLMs, including both open-source and closed-source models. Our results show that larger models, such as GPT-4o and Llama 3.1, achieved the highest overall F1-scores (0.739 and 0.694, respectively), demonstrating robust capabilities in extracting clinically and policy-relevant entities. However, we observed significant variations in performance across entity types, with higher precision for standardized codes, such as ICD-10, but lower recall for context-dependent categories like dietary supplements. These findings highlight the potential of LLMs to automate the extraction of structured healthcare data from judicial texts, thereby enabling real-time monitoring of healthcare judicialization and informing targeted policy interventions. Data is available at https://github.com/eliasjacob/lexcare.br**

Keywords—Large Language Models, Named Entity Recognition, Healthcare, Brazilian Judicial Decisions, Cross-Domain Benchmark, Public Health Analytics, Legal Informatics, Weak Supervision, Function Calling, Entity Extraction.

I. Introduction

Brazil's constitution establishes health as a fundamental right, obligating the federal, state, and local governments to guarantee comprehensive healthcare for every person. Rooted in the principles of universality and full protection, this mandate covers medical services, hospital procedures, and the provision of essential medicines and supplies. This policy affects over 212 million residents and will cost more than 35 billion dollars in public health expenditures in 2025.

This constitutional guarantee has contributed to the "judicialization of health care." Structural inefficiencies and rising public expectations have driven individuals to seek court intervention for medical treatments and pharmaceuticals, even when these interventions fall outside the parameters of public policy. In 2024 alone, 657,473 new health-related cases were filed, with 584,264 resolved within the year. In 2023, court rulings impacted 232 million dollars in pharmaceutical spending, representing 13.2% of the total allocated for medication purchases in the national health budget. Meanwhile, several challenges persist, such as increased costs, opportunistic litigation, forum shopping, and the imposition of treatments not yet integrated into health protocols. The issue is so complex and consequential that Brazil's Supreme Court has refrained from issuing a definitive ruling. Instead, it endorsed an intergovernmental agreement establishing criteria for court jurisdiction and the granting of health-related claims. Yet, despite this framework, several uncertainties persist, and the Supreme Court itself has rendered inconsistent decisions in subsequent cases.

In this context, the ability to extract healthcare related data from judicial proceedings (e.g., identifying patterns in medical conditions, surgical procedures, treatments, and pharmaceutical claims) can provide valuable insights for both judicial decision-making and the optimization of health policy administration. Manual extraction from hundreds of millions of unstructured legal texts is prohibitively labor-intensive, limiting the ability to identify spatial-temporal trends or recurring gaps in resource allocation. Advances in artificial intelligence, particularly large language models (LLMs), offer an interesting opportunity to automate this process.

This study systematically compares state-of-the-art LLMs in extracting ten predefined healthcare entities from Brazilian court rulings, using a manually annotated gold-standard dataset. Our analysis identifies the LLM architectures best suited to convert unstructured judicial data into actionable insights that support evidence-based policymaking in Brazil's public health system.

II. Related works

Many studies in the legal field address the judicialization of healthcare; however, they seldom analyze the specific types of medications and procedures being litigated, primarily because such information is not systematically documented within legal procedures [1].

Few studies have had access to the primary documents from the involved parties, and those that have are inherently limited and partial. The application of natural language

processing models to judicial decisions may offer a more comprehensive mapping of healthcare-related demands. Detailed analysis has typically been possible only when data are obtained directly from the involved parties, who must comply with the court ruling. For instance, an investigation based on documents from the Attorney's Office of São Paulo, a prominent litigant in healthcare-related cases, revealed that the most frequently requested medications in 2022 were cannabidiol, nintedanib, dupilumab, and insulin degludec [2]. Moreover, the study identified insulin-dependent diabetes mellitus, epilepsy, atopic dermatitis, interstitial lung diseases, and global developmental disorders as the most cited conditions.

The application of LLMs has drastically changed NER tasks in legal and health domains, though challenges persist. In healthcare, LLMs like GPT-4 face obstacles due to the scarcity of domain-specific data in their pre-training corpora, the complexity of medical terminology, and the need for nuanced, context-aware understanding [3], [4], [5], [6], [7]. Early studies showed that LLMs, especially in few-shot learning settings, lacked the specialized knowledge required for biomedical NER [8]. However, recent advancements, including models like GPT-4 and knowledge distillation, have demonstrated improved performance on biomedical NER tasks [4], [9], [10], [11], [12]. Techniques such as targeted distillation, where smaller models are trained using larger models, have also shown promise in enhancing NER capabilities [10], [13], [14], [15]. \

In legal texts, NER is increasingly important for legal research, document analysis, and process automation [16], [17], [18], [19], [20], [21]. However, legal NER presents unique challenges, as it often requires extracting specialized entities that differ from those in general NER. While LLMs have gained traction in legal applications like judgment prediction [22], [23], [24], their use for NER in legal texts, particularly at the intersection of law and healthcare, remains underexplored. Initial studies highlight both opportunities and limitations, with generic legal entities providing foundational insights but often failing to capture substantive information for cross-disciplinary analysis [18].

The choice between closed-source, commercially available LLMs (e.g., GPT-4) and open-source alternatives remains a key consideration, balancing performance, cost, privacy, and control. Together, these developments underscore the potential of LLMs to advance NER in specialized domains, though further research is needed to address domain-specific challenges and improve model adaptability.

III. Methodology

A. Gold Dataset Construction

This study used publicly available court rulings from all federal courts within the 5th Regional Federal Court (TRF5, the acronym in the Portuguese language) one of the six regional federal courts in Brazil. As per Brazilian law, these documents are publicly accessible through the PJe case management system unless the case was sealed by a judge, in which case the document becomes unavailable online.

We obtained 50,945 court rulings from cases that were classified as health-related by the plaintiff's lawyer. We divided each document into manageable segments using LangChain's recursive character text splitter, with a maximum of 1,024 tokens per segment. We randomly selected 1,200 documents from the dataset for annotation. We calculated that these documents averaged 709.57 tokens (σ = 266.99), which offered a diverse range of text lengths for analysis. We release the annotated dataset under the name *LexCare.BR*.

B. Annotation Process

Each of the 1,200 selected texts was independently annotated by four distinct annotators to ensure comprehensiveness and diverse perspectives. Two annotators which were experienced in handling legal cases with healthcare implications, annotated every document in our dataset.

In addition, a group of 14 judicial clerks and one federal judge, all of whom were appointed by the Federal Courts' Judicial Center of Intelligence and have extensive experience working in federal courts, provided other two annotations for each instance in our dataset. Prior to annotation, all annotators participated in a two-hour training session and received detailed written guidelines to standardize their use of the labeling tool [25]. On average, each annotator from this pool annotated 161.27 documents (σ = 32.61), with individual contributions ranging from 71 to 206 document annotations. Once a document received four independent annotations, it was removed from the annotation pool.

C. Labeling Schema

The annotation process included 10 predefined labels, selected to capture key elements relevant to healthcare and legal contexts: DISEASE_ICD_CODE: ICD identification for disease (e.g., E11.9); DISEASE_NAME: Disease nomenclature (e.g., Type 2 Diabetes Mellitus); MEDICAL_SUPPLY_NAME: Healthcare delivery tools (e.g., syringes, MRI machines); PHARMACEUTICAL_BRAND_NAME: Proprietary medication names (e.g., Advil); ACTIVE_PHARMACEUTICAL_INGREDIENT: Biologically active drug components (e.g., metformin); SURGICAL_PROCEDURE_NAME: Standardized surgical interventions (e.g., laparoscopic cholecystectomy); DIAGNOSTIC_PROCEDURE_NAME: Disease identification methods (e.g., PCR tests); THERAPEUTIC_INTERVENTION_NAME: Non-surgical treatments (e.g., radiation therapy); DIETARY_SUPPLEMENT_BRAND_NAME: Commercial supplement identifiers (e.g., IsoSource 1.5); DIETARY_SUPPLEMENT_ACTIVE_COMPONENT: Bioactive substances (e.g., calcium).

These labels address the key challenges of healthcare delivery and management in Brazil. They are aligned with our goal to provide a rich, structured dataset for the formulation of policies, resource allocation, and targeted public health interventions.

D. Annotator Reliability

To evaluate the reliability of individual human annotators on the task, we performed pairwise comparisons of annotations across all text instances. The reliability of each

annotator was computed by aggregating the pairwise agreement scores, which were based on both the spatial overlap of annotation spans and the agreement in assigned labels.

Let each annotation be represented as a span, with a start index s, an end index e, and a corresponding label. Given two annotations, a_1 and a_2, we first compute the span overlap using an Intersection over Union (IoU) metric. If annotation a_1 has span $[s_1, e_1]$ and annotation a_2 has span $[s_2, e_2]$ the overlap is defined as

$$Overlap(a_1, a_2) = \frac{\min(e_1,e_2)-\max(s_1,s_2)}{\max(e_1,e_2)-\min(s_1,s_2)} \quad (1)$$

We use the convention that if the spans do not overlap (i.e., if $e_1 < s_2$ or $e_2 < s_1$), the overlap is set to 0. This measure returns a value in the range $[0,1]$ where 0 indicates no overlap and 1 indicates identical spans. Once we compute the span overlap, we perform the comparison of the labels like this:

$$\delta(a_1, a_2) = \begin{cases} 1, & if\ the\ labels\ of\ a_1\ and\ a_2\ match \\ 0, & otherwise \end{cases} \quad (2)$$

An overall agreement score for the annotation pair is then computed as follows. First, if the overlap is significant, i.e., if $Overlap(a_1, a_2) \geq \tau$, where $\tau = 0.5$ is a user-defined threshold, the overall agreement score $S(a_1, a_2)$ is given by

$$S(a_1, a_2) = \frac{Overlap(a_1,a_2)+\delta(a_1,a_2)}{2} \quad (3)$$

If the overlap does not meet the threshold (i.e., if $Overlap(a_1, a_2) < \tau$), the pair is considered to have no significant agreement, and the score is set to 0.

E. Label Aggregation

For each text instance in our dataset, annotations from different annotators were compared pairwise. Let each instance i be associated with a set of annotators $\mathcal{A}_i$. For every unique pair(j, k), with $j, k \in \mathcal{A}_i$ and $j < k$, the pairwise score $S(a_j, a_k)$ was computed as described above. For each annotator, we then aggregated the scores over all instances in which they participated by maintaining a running sum of the agreement scores and a count of comparisons. For an annotator j, let the total score be as

$$total\_score_j = \sum_{comparisons\ involving\ j} S(a_j, a_k) \quad (4)$$

and let N_j denote the number of such pairwise comparisons. The final reliability score for annotator $j$is computed as

$$Reliability(j) = \frac{total\_score}{N_j} \quad (5)$$

This statistic, bounded between 0 and 1, represents the average level of agreement between annotators j and their peers. In line with weakly supervised learning approaches [26], we treat each human annotator as a labeling function and aggregate their outputs using a generative model based on a Hidden Markov Model (HMM). In this framework, the hidden states represent the true, latent labels, while the observed states correspond to the labels provided by the annotators. We adopted this approach because, while still relatively straightforward to implement, the HMM-based aggregation outperforms traditional majority voting methods [27], [28], [29].

Additionally, we incorporated the computed annotator reliability scores as weights during the aggregation process, so that annotations from more reliable annotators have greater influence on the final label determination. Finally, we estimated label quality issues for each sentence [30] and manually inspected and corrected any annotation problems.

In the final dataset, the number of labeled spans for each category was as follows: DISEASE_NAME had 1,595 spans; ACTIVE_PHARMACEUTICAL_INGREDIENT presented 1,400 spans; PHARMACEUTICAL_BRAND_NAME included 873 spans; MEDICAL_SUPPLY_NAME comprised 437 spans; DISEASE_ICD_CODE registered 373 spans; SURGICAL_PROCEDURE_NAME contained 318 spans; THERAPEUTIC_INTERVENTION_NAME had 243 spans; DIETARY_SUPPLEMENT_BRAND_NAME showed 212 spans; DIAGNOSTIC_PROCEDURE_NAME accounted for 176 spans; and finally, DIETARY_SUPPLEMENT_ACTIVE_COMPONENT was identified in 33 spans.

F. Generation of NER labels using LLMs

Large Language Models can interact with external systems using a task called "function calling," where they interpret declarative instructions and generate outputs that mimic a structured function call (e.g., in JSON). This process involves fine-tuning the model to add such capabilities and designing prompts that clearly define the function's schema and parameters.

In our experiments, we limited the scope to models that support function calling, are multilingual with native Portuguese support, and have publicly available weights compatible with our infrastructure, which comprises 8xH-100 GPUs. For benchmarking purposes, we also included GPT-4o results to serve as a performance reference for closed-source LLMs. We used LangChain structured outputs and Pydantic to define the data schema to be extracted and queried open models through Ollama. The list of models included: Athene-V2 [31], Aya Expanse [32], Command-R [33], Firefunction-V2 [34], GPT-4o [35], Granite 3.0 and 3.1 [36], Hermes 3 [37], Llama 3 [38], Mistral [39], Mixtral [40], Nemotron [41], and Qwen 2.5 [42]. We excluded models like DeepSeek-R1 and Phi4 due to their lack of built-in function calling support.

We presented the task as a few-shot function calling problem, where each request included 9 manually curated examples within the prompt. These 9 examples were not part of our gold dataset. This approach achieved higher performance compared to zero-shot methods, as demonstrated in prior research [43], [44], highlighting the effectiveness of example-based guidance for improving model accuracy without requiring additional fine-tuning of the model weights.

We evaluated each model using precision, recall, and F1-score for every entity type, employing a strict exact-match criterion, which considers a predicted entity correct only if

both its starting and ending boundaries are perfectly aligned with those in our gold dataset. The overall performance was then calculated using a micro-averaging strategy, which aggregates the true positives, false positives, and false negatives across all entity classes to provide a measure of each model's performance. This approach, which combines strict boundary matching with a micro-averaged overall metric, allows a fair assessment of the models' abilities to accurately capture the targeted entities.

IV. Results

This section highlights the overall effectiveness of each LLM, pinpointing top-performing models and identifying specific entity categories that presented greater challenges for accurate extraction. The aim is to provide a broad picture of the capabilities of these models in converting unstructured legal texts into structured data for evidence-based health policy. Table I describes the overall metrics for each LLM. Please refer to the supplementary material for the table containing the full results with precision, recall and F1-Score for each entity category.

TABLE I. Overall metrics for each LLM

Name	*Size*	*Quantization*	*F1-Score*	*Precision*	*Recall*
Athene-v2	72B	Q4_K_M	0,672	0,759	<u>0,603</u>
Aya Expanse	32B	Q4_K_M	0,507	0,623	0,427
	8B	Q4_K_M	0,487	0,645	0,391
Command-R Plus	104B	Q4_0	0,424	0,657	0,313
Command-R	35B	Q4_0	0,618	0,795	0,505
Firefunction-v2	70B	Q4_0	0,593	0,760	0,487
GPT-4o	N/A[a]	N/A	**0,739**	0,829	**0,667**
Granite 3 - Dense	8B	Q4_K_M	0,218	0,556	0,136
Granite 3.1 - Dense	8B	Q4_K_M	0,421	0,586	0,328
Granite 3.1 - MoE	3B	Q4_K_M	0,321	0,555	0,226
Hermes 3	405B	Q4_0	0,664	0,819	0,558
	70B	Q4_0	0,558	0,687	0,470
	8B	Q4_0	0,505	0,664	0,408
Llama 3.1	405B	Q4_K_M	0,672	0,822	0,568
		Q8	<u>0,694</u>	<u>**0,840**</u>	0,591
	70B	Q4_K_M	0,552	0,702	0,455
	8B	Q4_K_M	0,455	0,597	0,367
Llama 3.2	3B	Q4_K_M	0,384	0,571	0,289
Llama 3.3	70B	Q4_K_M	0,681	0,806	0,589
		FP16	0,683	0,812	0,589
Mistral-Small	24B	Q4_K_M	0,666	0,767	0,588
Mixtral - MoE	8x22B	Q4_0	0,519	0,676	0,421
Nemotron	70B	Q4_K_M	0,659	0,806	0,558
Qwen 2.5-Coder	32B	Q4_K_M	0,613	0,732	0,527
Qwen 2.5	72B	Q4_K_M	0,663	0,744	0,597

[a] **Bold** values represent the best results. <u>Underlined</u> values represent the best results for open-weights models.

The results reveal substantial differences in performance across the expanded set of models, with clear trends emerging based on model size, architecture, and quantization. Larger models, such as GPT-4o, Hermes 3 (405B), and Llama 3.3 (70B), consistently achieved the highest F1-scores, demonstrating superior entity extraction capabilities. These models excelled across most categories, with F1-scores frequently exceeding 0.75, particularly for entities like ACTIVE_PHARMACEUTICAL_INGREDIENT and DISEASE_NAME. In contrast, smaller models, including Granite 3-Dense (8B) and Aya Expanse (8B), struggled significantly, often scoring below 0.50 on the overall metric. This suggests that model size remains a critical determinant of performance, with larger models benefiting from greater representational capacity.

Entity-specific analysis uncovered varying degrees of difficulty across categories. For example, ACTIVE_PHARMACEUTICAL_INGREDIENT and DISEASE_NAME were relatively well-handled by most models, with many achieving F1-scores. However, categories like DIETARY_SUPPLEMENT_ACTIVE_COMPONENT and DIETARY_SUPPLEMENT_BRAND_NAME posed significant challenges, particularly for smaller models, which often scored near zero. This disparity likely reflects differences in category complexity and the availability of training data. Larger models, while still not excelling in these categories, performed notably better, suggesting that increased model capacity can mitigate some of the challenges posed by sparse or ambiguous data.

The analysis also revealed that many models prioritized precision over recall, particularly in challenging categories. For instance, Athene-v2 achieved high precision (≈0.910) but lower recall (≈0.667) for the ACTIVE_PHARMACEUTICAL_INGREDIENT category, resulting in an F1-score of 0.812. GPT-4o, while maintaining high precision (≈0.924), demonstrated better recall (≈0.724) for the same category, illustrating the trade-offs between precision and recall. However, certain categories, such as DIAGNOSTIC_PROCEDURE_NAME, saw near-zero recall across multiple models, indicating systematic challenges in identifying these entities despite accurate predictions when they were detected.

In conclusion, the results underscore the dominance of larger models like GPT-4o and Llama 3.3 (70B) in few-shot entity extraction tasks. While these models achieved competitive F1-scores across most categories, smaller models struggled, particularly with recall in challenging categories. This highlights the importance of model capacity and the need for continued improvements in smaller, more efficient architectures.

V. Future works

There's an interesting opportunity here for future work to leverage LLMs as labeling functions within a weak supervision approach. In this scenario, individual LLMs can

be treated as noisy annotators whose outputs are aggregated and denoised to produce a single, reliable label for each instance. This technique could have several benefits:

- Expanding labeled data: LLMs can be used as labeling functions, generating large amounts of annotated data cheaply. This approach circumvents the costly and time-consuming process of manual annotation, particularly in specialized domains such as legal healthcare texts. The scalability of LLM-driven labeling can foster the creation of huge datasets that cover a diverse range of scenarios and linguistic variations
- Weak supervision approaches: These methods would allow the individual outputs of several LLMs to be combined into one robust probabilistic label per instance. By modeling the uncertainty and potential biases of each LLM as a labeling function, the aggregated labels could better reflect the underlying true labels. This strategy also opens possibilities for refining labeling reliability by incorporating domain expertise as prior knowledge or by weighting each model's contribution based on historical performance.
- Distillation into smaller models: Once high-quality probabilistic labels are established, one could use these large, weakly labeled datasets to train smaller, more computationally efficient models. Knowledge distillation methods [10], [13], [14], [15] could transfer the capabilities of larger models into these smaller ones, allowing inference to become cheaper while maintaining robust performance. Such models would be particularly well-suited for deployment in real-world settings where computational resources or latency needs to be considered.

VI. Conclusion

This study provides the first systematic evaluation of large language models (LLMs) for healthcare-related named entity recognition (NER) in Brazilian judicial decisions, addressing a critical gap at the intersection of legal informatics and health analytics. With our newly released LexCare.BR dataset—a manually annotated gold-standard corpus of 1,200 legal rulings—we showed that LLMs can be used to convert unstructured judicial texts into structured, actionable, healthcare data.

GPT-4o achieved the highest overall F1-score (0.739) among tested models, closely followed by open-weights models like Llama 3.1 (0.694) and Athene-v2 (0.672). These results highlight the feasibility of automating the extraction of clinically and policy-relevant entities, such as diseases, pharmaceuticals, and medical procedures, from complex legal documents, despite the challenges posed by linguistic heterogeneity, structural irregularity, and semantic ambiguity in Brazilian court rulings.

The analysis revealed two main insights for cross-domain NER in resource-constrained settings. First, model performance varied substantially across entity types, with high precision for standardized codes (e.g., ICD-10: F1 ≈ 0.937) but lower recall for context-dependent categories like dietary supplements active components (F1 < 0.5). This precision-recall asymmetry suggests that while LLMs can reliably detect unambiguous entities, they remain conservative in recognizing less structured medical concepts—a limitation with direct implications for public health surveillance. Second, model scale emerged as a key determinant of robustness, with larger models (≥70B parameters) showing better balance across metrics compared to smaller counterparts, though at higher computational costs.

From a policy perspective, these technical capabilities, validated on the LexCare.BR dataset, could transform how Brazil monitors healthcare judicialization. Automated extraction from court rulings of entities involving denied treatments or recurring medication shortages would enable real-time identification of systemic gaps in the SUS, thus allowing for more targeted resource allocation and informed policy interventions.

Acknowledgment

This research was supported by the High-Performance Computing Center at UFRN (NPAD/UFRN). Additionally, this study was partially funded by the Brazilian National Council for Scientific and Technological Development (CNPq) through a research productivity scholarship awarded to E.J. Menezes-Neto (302582/2023-1).

We gratefully acknowledge the support of the Intelligence Center of Federal Courts in the state of Rio Grande do Norte for the invaluable contributions of their staff, particularly A.P.B.A. Oliveira, C.D. Fonseca, C.G.B. Lima, C.F.S. Carvalho, D.M.S. Vieira, E.S. Medeiros, I.S.R.A. Dantas, I.C.F.C. Augusto, J.M.R. Vasconcelos, L.L.M. Mota, L.C.F.A. Aquino, N.E.C.P. Salustino, R.H.V.P. Pinto and others, in data labeling.

AI Usage Disclosure: In preparing this manuscript, we used OpenAI's o1 model solely for improving readability and clarity. We employed the model to refine sentence structure, thereby enhancing comprehension, without altering the scientific content or research findings. All AI-assisted revisions were thoroughly reviewed and approved by the authors, who remain fully responsible for the final version of the text.

References

[1] L. B. D. Sá, Y. H. Bezerra, and I. M. G. D. Silva, "Judicialização da saúde e o fornecimento de fármacos não constantes na RENAME," *Rease*, vol. 8, no. 5, pp. 675–693, May 2022, doi: 10.51891/rease.v8i5.5529.

[2] G. M. da Silva, A. S. Gabriel, P. K. Psanquevich, and B. A. B. Neto, "Retrato da judicialização de medicamentos no município de São Paulo em 2022: o impacto dos pareceres técnicos da Secretaria Municipal de Saúde nas demandas judiciais," *Revista de Administração em Saúde*, vol. 23, no. 93, Art. no. 93, Dec. 2023, doi: 10.23973/ras.93.365.

[3] J. Bian, J. Zheng, Y. Zhang, H. Zhou, and S. Zhu, "One-shot Biomedical Named Entity Recognition via Knowledge-Inspired Large Language Model," in *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, in BCB '24. New York, NY, USA: Association for Computing Machinery, Dec. 2024, pp. 1–10. doi: 10.1145/3698587.3701356.

[4] J. Biana, W. Zhai, X. Huang, J. Zheng, and S. Zhu, "VANER: Leveraging Large Language Model for Versatile and Adaptive Biomedical

Named Entity Recognition," Apr. 27, 2024, *arXiv*: arXiv:2404.17835. doi: 10.48550/arXiv.2404.17835.
[5] X. Li, K. Chen, Y. Long, and M. Zhang, "LLM with Relation Classifier for Document-Level Relation Extraction," Dec. 07, 2024, *arXiv*: arXiv:2408.13889. doi: 10.48550/arXiv.2408.13889.
[6] Z. Zhu *et al.*, "Comparative Analysis of Large Language Models in Chinese Medical Named Entity Recognition," *Bioengineering*, vol. 11, no. 10, Art. no. 10, Oct. 2024, doi: 10.3390/bioengineering11100982.
[7] Z. Tan *et al.*, "Large Language Models for Data Annotation: A Survey," Feb. 20, 2024, *arXiv*: arXiv:2402.13446. Accessed: Mar. 28, 2024. [Online]. Available: http://arxiv.org/abs/2402.13446
[8] M. Moradi, K. Blagec, F. Haberl, and M. Samwald, "GPT-3 Models are Poor Few-Shot Learners in the Biomedical Domain," Jun. 01, 2022, *arXiv*: arXiv:2109.02555. doi: 10.48550/arXiv.2109.02555.
[9] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
[10] K. S. Vedula, A. Gupta, A. Swaminathan, I. Lopez, S. Bedi, and N. H. Shah, "Distilling Large Language Models for Efficient Clinical Information Extraction," Dec. 21, 2024, *arXiv*: arXiv:2501.00031. doi: 10.48550/arXiv.2501.00031.
[11] G. Peikos, P. Kasela, and G. Pasi, "Leveraging Large Language Models for Medical Information Extraction and Query Generation," Oct. 31, 2024, *arXiv*: arXiv:2410.23851. doi: 10.48550/arXiv.2410.23851.
[12] V. K. Keloth *et al.*, "Advancing entity recognition in biomedicine via instruction tuning of large language models," *Bioinformatics*, vol. 40, no. 4, p. btae163, Mar. 2024, doi: 10.1093/bioinformatics/btae163.
[13] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *Int J Comput Vis*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021, doi: 10.1007/s11263-021-01453-z.
[14] W. Zhou, S. Zhang, Y. Gu, M. Chen, and H. Poon, "UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition," Jan. 18, 2024, *arXiv*: arXiv:2308.03279. doi: 10.48550/arXiv.2308.03279.
[15] X. Xu *et al.*, "A Survey on Knowledge Distillation of Large Language Models," Mar. 08, 2024, *arXiv*: arXiv:2402.13116. Accessed: Aug. 02, 2024. [Online]. Available: http://arxiv.org/abs/2402.13116
[16] F. X. B. Da Silva *et al.*, "Named Entity Recognition Approaches Applied to Legal Document Segmentation," in *Anais do X Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2022)*, Brasil: Sociedade Brasileira de Computação - SBC, Nov. 2022, pp. 210–217. doi: 10.5753/kdmile.2022.227949.
[17] V. Naik, P. Patel, and R. Kannan, "Legal Entity Extraction: An Experimental Study of NER Approach for Legal Documents," *IJACSA*, vol. 14, no. 3, 2023, doi: 10.14569/IJACSA.2023.0140389.
[18] P. H. Luz de Araujo, T. E. de Campos, R. R. R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo, "LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text," in *Computational Processing of the Portuguese Language*, vol. 11122, A. Villavicencio, V. Moreira, A. Abad, H. Caseli, P. Gamallo, C. Ramisch, H. Gonçalo Oliveira, and G. H. Paetzold, Eds., in Lecture Notes in Computer Science, vol. 11122. , Cham: Springer International Publishing, 2018, pp. 313–323. doi: 10.1007/978-3-319-99722-3_32.
[19] T. W. T. Au, I. J. Cox, and V. Lampos, "E-NER -- An Annotated Named Entity Recognition Corpus of Legal Text," Dec. 19, 2022, *arXiv*: arXiv:2212.09306. doi: 10.48550/arXiv.2212.09306.
[20] D. Bernsohn *et al.*, "LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text," Feb. 06, 2024, *arXiv*: arXiv:2402.04335. doi: 10.48550/arXiv.2402.04335.
[21] J. Breton, M. B. Billami, M. Chevalier, and C. Trojahn, "Leveraging Semantic Model and LLM for Bootstrapping a Legal Entity Extraction: An Industrial Use Case," in *Studies on the Semantic Web*, A. Salatino, M. Alam, F. Ongenae, S. Vahdati, A.-L. Gentile, T. Pellegrini, and S. Jiang, Eds., IOS Press, 2024. doi: 10.3233/SSW240004.
[22] J. Cui, X. Shen, and S. Wen, "A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges," *IEEE Access*, vol. 11, pp. 102050–102071, 2023, doi: 10.1109/ACCESS.2023.3317083.
[23] M. Medvedeva and P. Mcbride, "Legal Judgment Prediction: If You Are Going to Do It, Do It Right," in *Proceedings of the Natural Legal Language Processing Workshop 2023*, D. Preo\textcommabelowtiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. (Jerry) Spanakis, and N. Aletras, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 73–84. doi: 10.18653/v1/2023.nllp-1.9.
[24] E. Jacob de Menezes-Neto and M. B. M. Clementino, "Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts," *PLOS ONE*, vol. 17, no. 7, p. e0272287, Jul. 2022, doi: 10.1371/journal.pone.0272287.
[25] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software." 2020. [Online]. Available: https://github.com/HumanSignal/label-studio
[26] J. Zhang, C.-Y. Hsieh, Y. Yu, C. Zhang, and A. Ratner, "A Survey on Programmatic Weak Supervision," Feb. 14, 2022, *arXiv*: arXiv:2202.05433. Accessed: Dec. 31, 2022. [Online]. Available: http://arxiv.org/abs/2202.05433
[27] P. Lison, J. Barnes, and A. Hubin, "skweak: Weak Supervision Made Easy for NLP," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 2021, pp. 337–346. doi: 10.18653/v1/2021.acl-demo.40.
[28] Y. Li, P. Shetty, L. Liu, C. Zhang, and L. Song, "BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition," May 30, 2021, *arXiv*: arXiv:2105.12848. doi: 10.48550/arXiv.2105.12848.
[29] Y. Li, L. Song, and C. Zhang, "Sparse Conditional Hidden Markov Model for Weakly Supervised Named Entity Recognition," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Aug. 2022, pp. 978–988. doi: 10.1145/3534678.3539247.
[30] W.-C. Wang and J. Mueller, "Detecting Label Errors in Token Classification Data," Oct. 08, 2022, *arXiv*: arXiv:2210.03920. Accessed: Apr. 22, 2024. [Online]. Available: http://arxiv.org/abs/2210.03920
[31] "Nexusflow.ai | Blog :: Introducing Athene-V2: Advancing Beyond the Limits of Scaling with Targeted Post-training." Accessed: Feb. 10, 2025. [Online]. Available: https://nexusflow.ai/blogs/athene-v2
[32] J. Dang *et al.*, "Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier," Dec. 05, 2024, *arXiv*: arXiv:2412.04261. doi: 10.48550/arXiv.2412.04261.
[33] "Command R: RAG at Production Scale," Cohere. Accessed: Feb. 16, 2025. [Online]. Available: https://cohere.com/blog/command-r
[34] "Firefunction-v2: Function calling capability on par with GPT4o at 2.5x the speed and 10% of the cost," Firefunction-v2: Function calling capability on par with GPT4o at 2.5x the speed and 10% of the cost. Accessed: Feb. 10, 2025. [Online]. Available: https://fireworks.ai/blog/firefunction-v2-launch-post
[35] OpenAI *et al.*, "GPT-4o System Card," Oct. 25, 2024, *arXiv*: arXiv:2410.21276. doi: 10.48550/arXiv.2410.21276.
[36] M. Mishra *et al.*, "Granite Code Models: A Family of Open Foundation Models for Code Intelligence," May 07, 2024, *arXiv*: arXiv:2405.04324. doi: 10.48550/arXiv.2405.04324.
[37] R. Teknium, J. Quesnelle, and C. Guang, "Hermes 3 Technical Report," Aug. 15, 2024, *arXiv*: arXiv:2408.11857. doi: 10.48550/arXiv.2408.11857.
[38] "The Llama 3 Herd of Models | Research - AI at Meta." Accessed: Feb. 10, 2025. [Online]. Available: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/
[39] A. Q. Jiang *et al.*, "Mistral 7B," Oct. 10, 2023, *arXiv*: arXiv:2310.06825. doi: 10.48550/arXiv.2310.06825.
[40] A. Q. Jiang *et al.*, "Mixtral of Experts," Jan. 08, 2024, *arXiv*: arXiv:2401.04088. doi: 10.48550/arXiv.2401.04088.
[41] Z. Wang *et al.*, "HelpSteer2-Preference: Complementing Ratings with Preferences," Oct. 02, 2024, *arXiv*: arXiv:2410.01257. doi: 10.48550/arXiv.2410.01257.
[42] A. Yang *et al.*, "Qwen2 Technical Report," Sep. 10, 2024, *arXiv*: arXiv:2407.10671. doi: 10.48550/arXiv.2407.10671.
[43] R. Agarwal *et al.*, "Many-Shot In-Context Learning," Oct. 17, 2024, *arXiv*: arXiv:2404.11018. doi: 10.48550/arXiv.2404.11018.
[44] H. Zeghidi and L. Moncla, "Evaluating Named Entity Recognition Using Few-Shot Prompting with Large Language Models," Sep. 04, 2024, *arXiv*: arXiv:2408.15796. doi: 10.48550/arXiv.2408.15796.