Um Pipeline Data-Centric para Construção de Corpus de Reconhecimento de Entidades Nomeadas para o Domínio Trabalhista Brasileiro

Emerson Diego da Costa Araujo[1,2], **Diego Ernesto Rosa Pessoa**[1], **Hidelberg Oliveira Albuquerque**[3]

[1]Instituto Federal da Paraíba (IFPB) – João Pessoa – PB – Brasil

[2]Divisão de Inteligência Artificial (DIA)
Tribunal Regional do Trabalho da 13ª Região (TRT-13) – João Pessoa – PB – Brasil

[3]Laboratório de Estudos em Informática Aplicada (LEIA)
Universidade Federal Rural de Pernambuco (UAST/UFRPE) – Serra Talhada – PE – Brasil

emerson.diego@academico.ifpb.edu.br, diego.pessoa@ifpb.edu.br

hidelberg.albuquerque@ufrpe.br

Resumo. *Este trabalho apresenta um pipeline data-centric para construção e curadoria de corpora de Reconhecimento de Entidades Nomeadas na Justiça do Trabalho brasileira. A abordagem integra geração sintética, supervisão fraca, aprendizado ativo e validação humana, resultando em dois recursos: um corpus de treinamento com 53.324 segmentos anotados em 32 classes e um conjunto Gold Standard independente. A qualidade do corpus foi validada por ajuste fino do BERTimbau, que atingiu 98,15% de micro-F1 no Gold Standard. Em comparação com Qwen-2.5-72B, a solução especializada reduziu a latência de inferência em aproximadamente 1.500× (31 s para 0,02 s por segmento), evidenciando viabilidade operacional em infraestrutura institucional local.*

Abstract. *This work presents a data-centric pipeline for building and curating Named Entity Recognition corpora in the Brazilian Labor Court domain. The approach integrates synthetic generation, weak supervision, active learning, and human validation, resulting in two resources: a training corpus with 53,324 segments annotated into 32 classes and an independent Gold Standard set (100 segments). Corpus quality was validated by fine-tuning BERTimbau, which achieved 98.15% micro-F1 on the Gold Standard. Compared to Qwen-2.5-72B, the specialized solution reduced inference latency by approximately 1,500× (from 31 s to 0.02 s per segment), demonstrating operational viability and technological sovereignty in local institutional infrastructure.*

1. Introdução

Órgãos do Judiciário operam diariamente com um volume elevado de documentos jurídicos (p. ex., sentenças, petições iniciais, despachos, além de ofícios e portarias), produzidos e armazenados em sistemas eletrônicos. Embora a digitalização tenha ampliado a disponibilidade desse acervo, informações críticas para a gestão institucional — como identificação de partes, números de processo, datas, valores e referências normativas — permanecem distribuídas em texto livre, dificultando automação, auditoria e recuperação de informação.

Nesse cenário, o Reconhecimento de Entidades Nomeadas (REN) é essencial para transformar linguagem natural em estrutura, permitindo localizar e classificar entidades relevantes, além de apoiar a detecção e o tratamento de dados pessoais e sensíveis conforme a LGPD [Brasil 2018]. O domínio institucional, porém, impõe alta variabilidade textual e restrições normativas, restringindo a exposição de dados institucionais a terceiros e exigindo soluções locais com soberania tecnológica e confidencialidade.

Apesar do avanço das arquiteturas Transformer, o uso direto de modelos generalistas em ambientes institucionais segue limitado por custo computacional, requisitos de execução local e necessidade de previsibilidade. Em tarefas de extração estruturada em nível de *token*, a especialização de modelos *encoder-only* via ajuste fino supervisionado tende a oferecer maior estabilidade e desempenho [Nunes et al. 2024], desde que sustentada por um processo de construção e curadoria de dados aderente ao domínio [Nunes 2025].

Surge o desafio de construir corpora de alta qualidade para extração automática de entidades em documentos institucionais, de forma sistemática e reprodutível, preservando soberania tecnológica e proteção de dados. Sob a perspectiva centrada em dados [Zha et al. 2025], o ganho desloca-se do modelo para o dataset. Logo, em cenários com infraestrutura local restrita, o sucesso da extração depende menos do porte da arquitetura e mais da curadoria rigorosa dos dados voltada ao domínio-alvo. Este trabalho propõe um pipeline data-centric para REN na Justiça do Trabalho, integrando geração sintética, supervisão fraca, aprendizado ativo e validação humana para superar gargalos de rotulagem e sustentar ciclos iterativos de melhoria contínua.

O objetivo central é avaliar a viabilidade técnica da construção de corpora especializados para REN em ambiente regulado, usando o desempenho do modelo como evidência da qualidade do dado curado. As principais contribuições deste trabalho são:

- A concepção e implementação de um pipeline *data-centric* para curadoria contínua de corpora em ambiente institucional, com protocolo de qualidade baseado em pré-rotulagem, revisão seletiva e ciclos orientados por incerteza;
- A definição de uma taxonomia granular (32 classes) e a construção de dois recursos complementares: corpus de treinamento com 53.324 segmentos (reais e sintéticos) e Gold Standard independente com validação humana integral;
- A validação empírica do *corpus* via especialização de um modelo *encoder-only*, alcançando F1 micro de 98,15% no *Gold Standard* e latência de 0,02 s por segmento.

Para fins de reprodutibilidade e transparência, os artefatos metodológicos e de dados utilizados neste estudo estão publicamente disponíveis em um repositório suplementar[1].

2. Trabalhos Relacionados

O Reconhecimento de Entidades Nomeadas no domínio jurídico é central para estruturação de conhecimento e conformidade institucional, exigindo adaptação a vocabulários e padrões complexos [Jurafsky and Martin 2024, Zhong et al. 2020, Ariai et al. 2025]. Esta seção revisa a evolução das arquiteturas de REN no domínio jurídico, os principais corpora nacionais e as metodologias *data-centric* que fundamentam a proposta.

Embora o uso de Grandes Modelos de Linguagem (LLMs) via *prompting* tenha se popularizado, abordagens puramente generativas ainda apresentam limitações de con-

[1] https://github.com/emerson-diego/DataCentric-LegalNER-BR

sistência e eficiência em tarefas de anotação em nível de *token* [Vithanage et al. 2025]. Para extração estruturada de entidades, a especialização de arquiteturas *encoder-only*, como o BERTimbau [Souza et al. 2020], por ajuste fino supervisionado (*fine-tuning*) tende a oferecer maior estabilidade e previsibilidade operacional do que modelos generalistas [Nunes et al. 2024, Mosbach et al. 2023]. Essa vantagem, contudo, depende de construção e curadoria rigorosas, aderentes ao domínio e capazes de mitigar riscos de contaminação [Nunes 2025].

A literatura nacional sobre REN jurídico evoluiu de redes recorrentes para arquiteturas Transformer. O trabalho pioneiro LeNER-Br [Araujo et al. 2018] estabeleceu as bases para legislação e jurisprudência, introduzindo entidades do domínio legal brasileiro. Posteriormente, iniciativas como o UlyssesNER-Br [Albuquerque et al. 2022] ampliaram a cobertura do domínio legislativo ao disponibilizar um *corpus* anotado voltado à análise de projetos de lei, recentemente estendido pelo UlyssesLegalNER-Br [Albuquerque et al. 2026], que unifica projetos de lei, jurisprudência e leis sob taxonomia jurídica mais ampla. O CDJUR-BR [Brito et al. 2023] introduziu uma coleção de documentos jurídicos com anotação mais granular, contribuindo para o avanço de tarefas de extração de informação no contexto da Justiça brasileira.

Apesar desses avanços, persiste uma lacuna no domínio jurídico: a complexidade terminológica e as exigências da LGPD demandam taxonomias mais granulares. O *DataSense NERCorpus* [Dias et al. 2020] abordou dados sensíveis em português, mas com foco em documentos corporativos e administrativos (como contratos e currículos). No contexto da Justiça do Trabalho, Castro [Castro 2019] explorou modelos ELMo com anotação semissupervisionada e, até onde sabemos, constitui o único trabalho de REN voltado a esse subdomínio. Ainda assim, o cenário atual requer arquiteturas Transformer e estratégias de curadoria mais escaláveis.

A construção de corpora anotados para REN em domínios regulados beneficia-se da integração de *supervisão fraca* (pré-rotulagem automática para expandir a cobertura do acervo anotado) e *aprendizado ativo* (priorização de exemplos incertos para validação humana). Em conjunto, essas estratégias viabilizam ciclos iterativos de curadoria e ajuste fino com elevada eficiência [Nunes et al. 2024].

A proposta deste trabalho diferencia-se por operacionalizar essas técnicas em arquitetura *data-centric* unificada. As iniciativas voltadas à Justiça do Trabalho ainda carecem de maior granularidade taxonômica, sobretudo para dados sensíveis definidos no art. 5º, inciso II, e submetidos ao rigor do art. 11 da LGPD [Brasil 2018]. Adota-se, portanto, 32 classes de entidades, desenhadas para lidar com variabilidade textual, identificadores estruturais e recorrência de categorias sensíveis, com execução local em ambiente institucional. A Tabela 1 resume os trabalhos relacionados no domínio jurídico.

Tabela 1. Comparativo de Trabalhos Relacionados no Domínio Jurídico

Trabalho	Domínio	Ents.	Foco LGPD	Estratégia
LeNER-Br [Araujo et al. 2018]	Legis./Juris.	6	Baixa	Manual
UlyssesNER-Br [Albuquerque et al. 2022]	Legislativo (PLs)	18	Baixa	Manual
UlyssesLegalNER-Br [Albuquerque et al. 2026]	Jurídico legislativo	23	Baixa	Híbrida
Castro [Castro 2019]	Justiça do Trabalho	11	Baixa	Semi-supervisionada
DataSense [Dias et al. 2020]	Genérico (Currículos/Contratos)	18	Média	Híbrida
CDJUR-BR [Brito et al. 2023]	Peças Processuais (Judicial)	21	Média	Manual
Nunes et al. [Nunes et al. 2024]	Legislativo (PLs)	7	Baixa	*Self-Learning*
Proposta deste trabalho	***Justiça do Trabalho***	***32***	***Alta***	***Data-Centric (Híbrida)***

3. Metodologia de Construção e Curadoria do Corpus

3.1. Pipeline de Geração do Corpus

A construção do *corpus* foi estruturada como um fluxo contínuo de ingestão e enriquecimento, com *pipeline* reprodutível de engenharia de dados. Documentos em PDF são extraídos diariamente dos sistemas institucionais Processo Judicial Eletrônico (PJe) e Sistema de Processos Administrativos e Ouvidoria (PROAD), cobrindo processos administrativos e judiciais no recorte de 2021 a 2026. Os dados são armazenados em *Parquet* em um *Data Lake* (*MinIO*) com orquestração via *Apache Airflow*. A base inicial de treinamento foi formada por amostragem aleatória desse acervo para capturar diversidade documental e mitigar vieses de seleção.

Conforme ilustrado na Figura 1, o *pipeline* transforma documentos brutos em unidades textuais anotáveis, viabilizando a construção progressiva de um *corpus* supervisionado com dados reais e sintéticos. PromptNER [Andrade et al. 2023] mostrou a viabilidade de usar LLMs na rotulagem automática de dados públicos para treinar modelos menores que, posteriormente, podem processar dados sensíveis localmente. Em linha com essa estratégia, modelos generativos externos (como Gemini 2.5 Flash) restringem-se à etapa inicial de geração sintética/pública, sem exposição de documentos sensíveis reais.

Figura 1. Fluxo conceitual de geração do corpus

O processo é composto pelas seguintes etapas principais:

1. **Extração e Segmentação**: Textos do *Data Lake* são segmentados em *chunks* semânticos com *NLTK* para a detecção de fronteiras linguísticas de sentenças, evitando sua fragmentação. As sentenças são agrupadas sequencialmente até 512 *tokens* pelo tokenizador do modelo final. Sentenças isoladas que excedem 512 *tokens* são descartadas ($<$1% do total), pois ultrapassar esse limite em uma única sentença é raro neste domínio.
2. **Enriquecimento do Corpus**: Esta etapa operacionaliza o paradigma de *Teacher-Student Learning* [Ding et al. 2024], onde modelos generativos (*Teachers*) transferem conhecimento para o modelo especializado (*Student*) por meio de:
    - **Geração Sintética**: Uso de Gemini 2.5 Flash (única etapa externa) para mitigar o *cold start* e o desbalanceamento de classes. Por operar sobre dados sintéticos, essa etapa envia apenas taxonomia e cenários fictícios, sem peças do PJe ou PROAD, permitindo o uso de modelos generativos externos sem comprometer o sigilo de documentos reais. A geração estruturou-se

em duas trilhas de *prompt*: entidades administrativas e estruturadas (ênfase em classes sub-representadas) e categorias sensíveis (LGPD), compensando o desbalanceamento entre dados sensíveis raros e entidades estruturadas abundantes no acervo real. Exemplos dos *prompts*, pós-processamento e revisão constam no repositório suplementar.

- **Pré-rotulagem Automática**: o acervo real foi pré-rotulado com o modelo *Qwen-2.5-72B* (*on-premise*, sobre o acervo real) e adotado como *baseline* generativo. A escolha decorreu de exploração qualitativa entre LLMs abertos no Tribunal. Em amostra inspecionada, o Qwen apresentou menor incidência de erros críticos, ainda com limitações de fronteira e rótulos fora do *schema*. Trata-se de seleção para o *cold start*, não de *benchmark* exaustivo (detalhes no repositório suplementar). À medida que o modelo especializado ganhou confiabilidade, passou a pré-rotular o acervo no lugar do Qwen, caracterizando um processo de autoespecialização.

3.2. Curadoria Humana (Human-in-the-Loop)

A curadoria manual foi operacionalizada por uma ferramenta de rotulagem desenvolvida para o fluxo de revisão proposto. A interface (Figura 2) facilita a inspeção e correção das entidades pré-rotuladas, com destaque por cores conforme a classe de entidade, reduzindo a carga cognitiva dos revisores e acelerando a validação.

Figura 2. Interface da ferramenta de curadoria humana utilizada para revisão

- **Diretrizes de Anotação:** Elaborou-se um manual técnico definindo categorias taxonômicas, critérios de fronteira e resoluções de ambiguidades contextuais, disponível no repositório suplementar.
- **Ciclos de Alinhamento:** Cinco revisores realizaram reuniões semanais. Dúvidas de marcação eram consolidadas, discutidas e resolvidas por consenso, com atualização versionada das diretrizes no manual.

Dada a complexidade terminológica, adotou-se anotação por consenso contínuo. Exemplares impactados por mudanças no manual eram revistos antes de reingresso no *corpus*. Complementarmente, um piloto de concordância interanotadores (IAA) em 200 segmentos (dois revisores revisando de forma independente a mesma pré-rotulagem do modelo especializado) obteve $\kappa = 0{,}95$ de Cohen [Cohen 1960] no *Exact Match* por entidade.

3.3. Modelo de Dados e Rastreabilidade

Para garantir governança e reprodutibilidade ao longo do ciclo de curadoria, o *corpus* foi estruturado em JSONL. Embora a extração inicial consuma dados brutos em *Parquet* no *Data Lake*, a transformação para um modelo semiestruturado na fase de anotação responde à necessidade de flexibilidade de esquema (*schema evolution*). Em *pipelines* iterativos, nos quais diretrizes e metadados evoluem continuamente, cada registro JSONL atua como documento desnormalizado e autossuficiente. Além do texto e das entidades anotadas, inclui metadados de proveniência (origem e metadados de produção) e de auditoria (histórico de intervenções humanas), preservando a rastreabilidade temporal em um único objeto. A especificação completa consta no repositório suplementar. O *corpus* final utilizado totaliza 107 MB.

3.4. Composição do Corpus

O *corpus* resultante compreende 53.324 exemplos anotados em 32 classes de entidades, integrando dados reais e sintéticos em um fluxo de evolução contínua. A composição híbrida equilibra a representatividade do domínio jurídico com a ampliação da cobertura semântica em categorias de baixa frequência natural.

O conjunto de treinamento é predominantemente composto por dados reais (≈70,2%), extraídos de 16.832 documentos únicos, variando de peças públicas a expedientes de tramitação restrita. Do total de 37.447 exemplos reais, 1.573 são oriundos de arquivos sigilosos. Complementarmente, 15.877 instâncias sintéticas (≈29,8%) foram incorporadas para reforçar a detecção de entidades menos representadas.

A estratégia de controle de qualidade priorizou a fidelidade dos dados: 20.362 exemplos (38,2% do total) passaram por revisão humana direta, correspondendo a 68,4% do bloco sintético (10.857) e a 25,4% do bloco real (9.505), enquanto a maior parte dos reais remanescentes entrou via *Self-Learning* controlado por limiares de confiança.

A taxonomia final contempla 32 classes com distribuição heterogênea. Em frequência, destacaram-se `ORGANIZACAO` (15,48%), `NUMERO_GENERICO` (15,24%), `NOME_PESSOA` (12,15%), `PERIODO_TEMPO` (10,65%) e `LEGISLACAO` (9,51%). Na cauda da distribuição, predominam categorias sensíveis (LGPD), como `ORIGEM_RACIAL_ETNICA` (0,70%), `CONVICCAO_RELIGIOSA` (0,89%), `OPINIAO_POLITICA` (1,15%) e `DADO_SAUDE_VIDA_SEXUAL` (2,49%), reforçadas por segmentos sintéticos.

No *Gold Standard* (100 segmentos; 2.605 entidades), reporta-se o suporte: OPINIAO_POLITICA, 2 instâncias; CONVICCAO_RELIGIOSA, 1; suporte mínimo de 1 instância (p.ex., DADO_BANCARIO), conforme repositório suplementar. Suporte unitário ou binário, sobretudo em categorias sensíveis à LGPD, compromete a confiabilidade da avaliação por classe na avaliação cega. Prevê-se ampliar o *Gold Standard* por amostragem dirigida a essas categorias.

O detalhamento das 32 classes, com definições semânticas, exemplos, frequências absolutas e F1 individual no teste reservado (10%), encontra-se no repositório suplementar.

3.5. Evolução Arquitetural e Aprendizado Contínuo

Como a curadoria humana se mostrou custosa e era necessário incorporar mais exemplos reais em escala, o *pipeline* evoluiu de uma rotulagem assistida para um fluxo mais auto-

matizado. Na fase de *cold start*, a identificação primária de dados sensíveis dependeu de geração sintética e pré-rotulagem via LLMs (Seção 3.1). Com a maturidade preditiva e a consistência taxonômica do modelo especializado, ele próprio passou a assumir o motor de inferência. A partir daí, adotou-se uma triagem baseada em confiança, dividida em quatro zonas: Zona LGPD (revisão prioritária), Alta Confiança (auto-incorporação por *Self-Learning*), Incerteza (aprendizado ativo) e Descarte (supressão de ruído).

A Figura 3 resume ingestão diária, inferência, roteamento às quatro zonas e retorno ao treino ou à revisão.

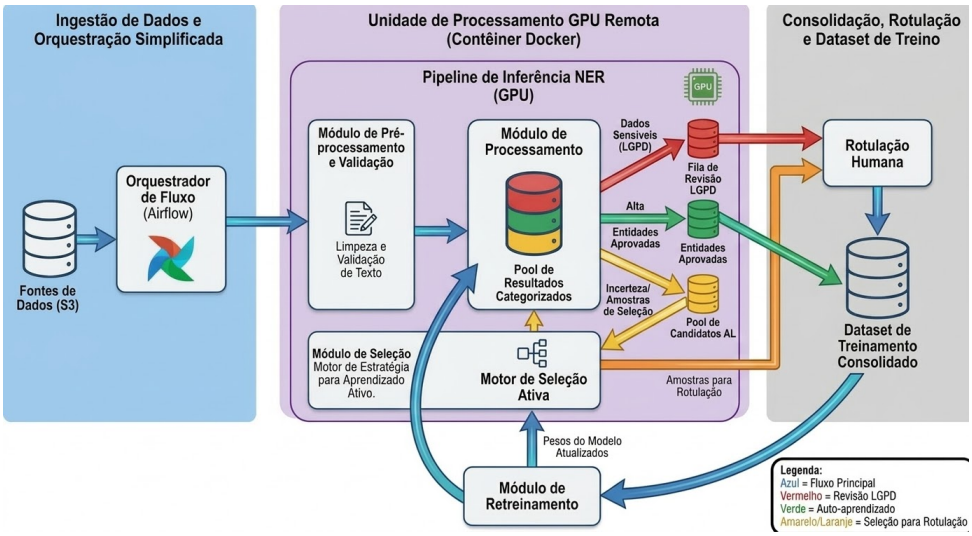

Figura 3. Pipeline com seleção ativa e retreinamento contínuo.

Diariamente, o *pipeline* executa ingestão, segmentação, inferência, roteamento às zonas (Tabela 2), revisão no rotulador de entidades sensíveis e candidatos de aprendizado ativo, e auto-incorporação por *Self-Learning*. As etapas automatizadas, incluída a filtragem ativa, concluem-se em janelas de minutos. Retreinamento do *encoder* e recalibração de T_{ent} ocorrem em ciclos separados da ingestão diária, quando há acúmulo de segmentos revisados ou auto-incorporados. O diagrama é esquemático: encadeia numa única figura etapas de operação diária e de ciclos de curadoria, embora não ocorram no mesmo ritmo. Mesmo nessa visão unificada, a Zona LGPD e a triagem ativa exigem revisão humana.

Tabela 2. Mapeamento de Zonas Operacionais, Limiares e Ações do Sistema

Zona Operacional	Critério / Limiar de Probabilidade (P)	Ação do Sistema
Zona LGPD	Ao menos uma entidade sensível (qualquer P)	Fila de Revisão Humana (Enriquecimento)
Alta Confiança	*Chunk* aprovado no limiar (todas as entidades com $P \geq T_{\text{ent}}$)	Auto-incorporação (*Self-Learning*)
Incerteza	*Chunk* não aprovado no limiar e com ao menos uma entidade em $0{,}50 \leq P < (T_{\text{ent}} - 0{,}10)$	Envio para o *Pool* de Candidatos AL
Descarte	*Chunk* não aprovado no limiar e sem entidades na faixa de incerteza (caso residual)	Descarte de predições ruidosas

O fluxo de decisão prioriza o risco de negócio (Zona LGPD), seguido pela auto-incorporação de predições seguras (Alta Confiança) e, por fim, a separação de amostras para o aprendizado ativo (Incerteza). A distribuição média de roteamento dos *chunks* reflete o rigor conservador do sistema: Zona LGPD ($\approx 3\%$), Alta Confiança ($\approx 7\%$), Incerteza ($\approx 16\%$) e Descarte residual ($\approx 74\%$).

Os limiares dinâmicos (T_{ent}), calculados a partir do desempenho individual de cada classe, buscaram equilibrar o controle de ruído com um custo de revisão viável. A faixa

de incerteza estabelece uma margem de segurança abaixo do limiar de auto-incorporação por classe, enquanto o descarte atua como filtro residual. Em ciclos pilotos de calibração, limiares mais permissivos aumentaram falsos positivos no *Self-Learning*, e limiares mais rígidos reduziram a recuperação de classes raras; os valores adotados apresentaram melhor compromisso entre precisão, ganho informacional e esforço humano.

Na triagem ativa, seja c um *chunk* e $\hat{\mathcal{E}}_c = \{(\hat{e}_i, \hat{y}_i)\}_i$ o conjunto de entidades previstas, com probabilidade $P(\hat{e}_i \mid c)$ e limiar $T_{\text{ent}}(\hat{y}_i)$ por classe (Tabela 2). *Chunks* com entidade sensível (LGPD) são desviados à revisão humana antes da auto-aprovação, qualquer que seja P. Um *chunk* é auto-aprovado quando $P(\hat{e}_i \mid c) \geq T_{\text{ent}}(\hat{y}_i)$ para toda entidade prevista. Caso contrário, integra a zona de incerteza se existir entidade na faixa

$$0{,}50 \leq P(\hat{e}_i \mid c) < T_{\text{ent}}(\hat{y}_i) - 0{,}10. \tag{1}$$

Denotamos por $\mathcal{Z}_{\text{inc}}$ o conjunto de *chunks* não auto-aprovados com ao menos uma entidade na faixa da Eq. (1). Candidatos roteados à Zona LGPD permanecem excluídos do *pool* de aprendizado ativo. Cada $c \in \mathcal{Z}_{\text{inc}}$ passa por deduplicação ($h(c) = \text{SHA256}(\text{id}(c) \| \text{texto}(c))$; *id* ou h já no acervo são ignorados) e filtragem semântica no lote diário, com *encoder* `gte-multilingual-base` (distinto do BERTimbau), vetor e_c L2-normalizado e similaridade de cosseno:

$$\mathcal{P}'_{\text{AL}} = \big\{ c \in \mathcal{Z}_{\text{inc}} : \max_{t \in \mathcal{D}_{\text{treino}}} \cos(e_c, e_t) < \tau_{\text{ds}} \big\}, \quad \tau_{\text{ds}} = 0{,}82, \tag{2}$$

$$\mathcal{P}_{\text{AL}} = \text{GreedyDedup}\big(\mathcal{P}'_{\text{AL}};\, \tau_{\text{int}} = 0{,}85\big), \tag{3}$$

em que GreedyDedup mantém c_i e descarta c_j ($j > i$) se $\cos(e_{c_i}, e_{c_j}) \geq \tau_{\text{int}}$. Seja $f(y)$ a frequência da classe y no treinamento, N o total de entidades anotadas e $|C| = 32$. Para $c \in \mathcal{P}_{\text{AL}}$, com $\mathcal{C}(c)$ o conjunto de classes em c e $n_{\text{incertas}}(c)$ entidades na faixa da Eq. (1), define-se

$$\text{freq}_{\text{rel}}(y) = \frac{f(y) + 1}{N + |C|}, \qquad \text{score}_{\text{base}}(c) = \sum_{y \in \mathcal{C}(c)} \frac{1}{\text{freq}_{\text{rel}}(y)}, \tag{4}$$

$$\text{score}(c) = \text{score}_{\text{base}}(c) \cdot \left(1 + 0{,}1\, n_{\text{incertas}}(c)\right). \tag{5}$$

Selecionam-se $\max(1, \lfloor 0{,}10\, |\mathcal{P}_{\text{AL}}| \rfloor)$ candidatos com maior $\text{score}(c)$ para revisão humana prioritária. Segmentos revisados ou auto-incorporados por *Self-Learning* retornam ao *corpus* em ciclos de retreinamento.

4. Especialização do Modelo e Configuração Experimental

Esta seção detalha o processo de especialização do modelo, a configuração de treino e o protocolo de partição de dados, visando a validade estatística dos resultados.

4.1. Modelo Base

A especialização partiu do BERTimbau Large [Souza et al. 2020], arquitetura *encoder-only* monolíngue em português brasileiro, alinhada ao acervo jurídico-trabalhista do estudo. Optou-se por pré-treino nacional em detrimento de *encoders* multilíngues generalistas, dado o escopo monolíngue da taxonomia. O BERTimbau Base foi avaliado no mesmo protocolo da Tabela 3: o Large superou o Base em média cerca de 0,5 e 0,6 ponto percentual em F1 micro e macro, respectivamente, com superioridade em todos os dez *folds* no F1 macro. O incremento é modesto frente à variabilidade entre *folds*, porém consistente. Métricas agregadas e valores por *fold* constam no repositório suplementar.

4.2. Configuração de Treinamento

O treinamento foi realizado por meio de ajuste fino supervisionado (*fine-tuning*) sobre o *corpus* descrito na seção anterior, de forma incremental ao longo dos ciclos de curadoria.

A seleção de hiperparâmetros seguiu heurísticas recomendadas para BERTimbau [Souza et al. 2020]. A Tabela 3 resume a configuração adotada na validação cruzada. Em cada *fold*, o melhor *checkpoint* foi selecionado por `eval_f1_micro` na partição de validação. Ao término da validação cruzada, o *checkpoint* do melhor *fold* foi avaliado no teste reservado (10%).

Tabela 3. Hiperparâmetros de ajuste fino.

Parâmetro	**Valor**
Taxa de aprendizado	2×10^{-5}
Otimizador	AdamW (`adamw_torch`, Hugging Face `Trainer`)
Weight decay	0,03
Scheduler	*Warmup* 500 passos; decaimento cosseno
Épocas (máx.)	10
Batch / lote efetivo	8 por GPU; acumulação 8 (64)
Early stopping	Paciência 2; $\Delta loss \geq 0{,}001$
Hardware	$1\times$ NVIDIA L40S (46 GB VRAM)
Validação cruzada / semente	Estratificada, $k = 10$; semente 42

4.3. Protocolo Experimental e *Gold Standard*

O acervo de treino compreende 53.324 segmentos (média de 297 *tokens*). O ajuste fino foi acompanhado por validação interna (validação cruzada em dez *folds* e teste reservado de 10%) e por avaliação cega no *Gold Standard*.

Na validação interna do ajuste fino, a unidade de partição é o *chunk*: 10% destinados ao teste reservado (5.332 exemplos) e 90% submetidos à validação cruzada estratificada (43.193 treino e 4.799 validação por *fold*). A estratificação preservou perfis de entidades entre treino e validação, condição necessária dado o desbalanceamento do domínio. Esse arranjo prioriza volume estatístico e equilíbrio taxonômico. Como os segmentos derivam de 16.832 documentos segmentados em múltiplos *chunks*, trechos adjacentes do mesmo processo ou peça podem, excepcionalmente, ocupar partições distintas.

Na avaliação cega, a restrição é documental. O *Gold Standard* reúne 100 segmentos reais do PJe e do PROAD (79 do PJe e 21 do PROAD), anotados integralmente à luz das diretrizes. Os documentos de origem não integram o acervo de treino e foram avaliados apenas na etapa cega, reduzindo vazamento direto (*data leakage*). São 2.605 entidades no total (média de 26,0 por segmento; *tokens* entre 64 e 512, média de 468).

A seleção buscou refletir classes frequentes do domínio e incluir categorias raras e sensíveis. As cinco mais frequentes concentram 41,9% das 2.605 entidades, e sete categorias têm no máximo dez instâncias, cobrindo 31 das 32. A classe `IP` não ocorre no *Gold Standard*, pois os documentos reais elegíveis não continham instâncias anotáveis, enquanto no acervo de treino ela provém sobretudo de segmentos sintéticos. A distribuição por classe e uma amostra reformulada e anonimizada constam no repositório suplementar.

4.4. Métricas de Avaliação

O desempenho foi mensurado por Precisão, Revocação e F1. Adotou-se F1 micro para o desempenho global e F1 macro para equidade entre as 32 classes, evitando que categorias críticas e minoritárias fossem mascaradas pelas majoritárias. Toda a avaliação seguiu *Exact Match* em nível de entidade, exigindo correspondência exata de fronteiras e classe.

5. Resultados

5.1. Análise de Eficácia e Baseline de Literatura

Na validação interna do ajuste fino, a validação cruzada (dez *folds*) registrou F1 micro médio de 97,76% ($\pm$ 0,08%), com pico de 97,92% (melhor *fold*) e 97,61% (pior *fold*), evidenciando estabilidade do aprendizado. No teste reservado (10% do acervo de treino), o modelo alcançou F1 micro de 97,87% e F1 macro de 96,65%. A estreita margem entre micro e macro indica mitigação do viés de classes majoritárias, com bom desempenho também nas categorias de baixa frequência.

Na avaliação cega, o modelo alcançou Precisão de 98,21%, Revocação de 98,09%, F1 micro de 98,15% e F1 macro de 98,67%. A estratificação do *Gold Standard* buscou cobertura mínima de classes raras e sensíveis e altera a composição em relação ao acervo de treino. `ORGANIZACAO`, por exemplo, passa de 15,48% das entidades no treino a 10,4% na amostra, e `DADO_SAUDE_VIDA_SEXUAL` totaliza 14 instâncias. O F1 macro superior ao micro neste conjunto reflete essa composição amostral, e não desempenho superior nas classes raras, cujas métricas individuais permanecem frágeis pelo baixo suporte.

A análise por classe revela desempenho heterogêneo. Identificadores estruturados (`CPF`, `CNPJ`, `OAB`) atingiram F1 quase perfeito, favorecidos pela baixa ambiguidade morfológica. As maiores dificuldades concentraram-se em `OPINIAO_POLITICA` e `CONVICCAO_RELIGIOSA`, classes mais subjetivas e dependentes de contexto mais amplo. Na inspeção qualitativa, predominaram erros de fronteira (*boundary errors*): o modelo acerta o núcleo da entidade, mas diverge na inclusão de preposições ou artigos adjacentes.

A Tabela 4 consolida eficácia, eficiência e requisitos de hardware. Os F1 marcados com † vêm dos trabalhos originais, em condições distintas, e servem apenas como referência contextual. A comparação sob protocolo unificado (‡) restringe-se ao Qwen-2.5-72B frente ao BERTimbau especializado.

Quanto ao Qwen-2.5-72B (F1 micro 63,83%, ‡), o desempenho reflete limitações de *in-context learning* para REN. Para preservar comparabilidade, a avaliação seguiu o protocolo da Seção 4.3 no *Gold Standard* (32 classes). O *prompt* utilizado, disponível no repositório suplementar, instrui a extração em JSON com texto literal e classe. A partir disso, um script localizou cada string no documento original e obteve os índices de início e fim. Na análise qualitativa, predominaram erros de fronteira (p. ex., inclusão de pontuação ou espaços, como em "Tribunal Regional do Trabalho."), além de alucinação taxonômica com classes fora do *schema* de 32 categorias. Em linha com [Mosbach et al. 2023], esses resultados confirmam as limitações da inferência via *prompt* frente ao ajuste fino. Paralelamente, [Tran et al. 2024] corrobora a competitividade de *encoders* em NER jurídico. Maffeo *et al.* [Maffeo et al. 2026] reportam micro-F1 de 57,9% com Qwen-2.5:32B no LeNER-Br [Araujo et al. 2018], patamar da mesma ordem de grandeza em extração via *prompting* (*corpus*, taxonomia e porte do modelo distintos).

5.2. Análise de Eficiência e Viabilidade Operacional

A implantação em ambientes regulados depende da eficiência computacional e da sustentabilidade da infraestrutura local. As medições usaram a mesma GPU (NVIDIA L40S, 46 GB) e segmentos do *Gold Standard* (média de 468 *tokens*), reproduzindo o fluxo segmento a segmento. O especializado inferiu em *fp16* (Hugging Face Transformers), e o Qwen-2.5-72B em quantização 4-bit (Ollama). Reporta-se latência média *wall-clock* por segmento e, no generativo, tempo ponta a ponta com decodificação autoregressiva, além de vazão como razão entre *tokens* totais e tempo agregado. As precisões distintas refletem arranjos viáveis *on-premise*, sem equiparação numérica estrita.

O modelo generativo apresentou latência média de 31,08 s por segmento, consumindo cerca de 40 GB de VRAM. Em contrapartida, o modelo especializado (BERTimbau *fine-tuned*) processou segmentos equivalentes em 0,02 s, com vazão de aproximadamente 23.400 *tokens*/segundo e consumo inferior a 4 GB de VRAM.

Essa redução da latência em aproximadamente 1.500$\times$ aliada aos baixos requisitos de hardware viabiliza o processamento em larga escala em infraestruturas modestas. Tal desempenho assegura a soberania tecnológica e o cumprimento de diretrizes de privacidade sem comprometer a celeridade da tramitação documental na Justiça do Trabalho.

Tabela 4. Comparativo contextual de eficácia, eficiência e infraestrutura.

Modelo / Trabalho	F1 Micro	Ents.	Latência / Seg.	Vazão (Tk/s)	VRAM (GB)	Hardware Mínimo
DataSense [Dias et al. 2020]	83,01%†	*	–	–	–	–
LeNER-Br [Araujo et al. 2018]	92,53%†	6	–	–	–	–
Castro [Castro 2019]	93,81%†	11	–	–	–	–
Qwen-2.5-72B (4-bit)	$\approx$ 64,00%‡	32	31,08 s	$\approx$ 16	$\approx$ 40	1x GPU 48 GB
BERTimbau *fine-tuned*	**98,15%**	**32**	**0,02 s**	**$\approx$ 23.400**	**< 4**	**1x GPU 16 GB**

†F1 do trabalho original (domínio e protocolo distintos).
‡Reavaliado neste estudo (protocolo unificado).

6. Discussão

O F1 micro de 98,15% sugere que a especialização de *encoders*, retroalimentada por um pipeline *data-centric* iterativo, constitui estratégia eficaz para extração de entidades institucionais. Apesar do bom desempenho global, a análise qualitativa expôs maior incidência de ambiguidades de fronteira em trechos de tabelas achatadas ou cabeçalhos desestruturados em PDFs, além de viés de padronização na escrita sintética inicial, mitigado nos ciclos subsequentes de aprendizado ativo com exemplos reais do PJe e PROAD.

O tratamento de dados sensíveis (LGPD) revelou-se o principal desafio, pois a rotulagem depende fortemente do contexto. Termos como "atestado" podem ser apenas jargões administrativos ou revelar condição de saúde do titular. Essa distinção semântica fina eleva o custo da curadoria humana.

Sob a perspectiva ética, a geração sintética restringiu-se ao *cold start*, sem envio de documentos judiciais a modelos externos; textos reais permanecem em processamento local. Segmentos sintéticos em categorias LGPD passaram por revisão humana intensiva (68,4% revisados), com o objetivo de evitar estereótipos e padrões identificáveis. Sobre dados reais sensíveis, a Zona LGPD exige revisão manual antes da incorporação ao *corpus*.

Por fim, o Judiciário Trabalhista acumula terabytes de textos não estruturados e incorpora milhares de novos documentos diariamente. Nesse cenário, velocidade de processamento é requisito basilar. Conforme a Tabela 4, o modelo especializado combina alta vazão com baixo consumo de memória, viabilizando processamento sustentável em infraestrutura modesta. Mais do que ganho de engenharia, essa eficiência reforça a soberania tecnológica: ao permitir execução local, evita a exposição de dados judiciais sigilosos a APIs externas e fortalece a governança institucional.

6.1. Limitações do Estudo

Apesar dos resultados, três limitações devem ser consideradas: (i) o sigilo judicial restringe a publicação integral do *corpus* (mitigada por amostras anonimizadas e diretrizes); (ii) o *Gold Standard* (100 segmentos) apresenta baixo suporte em classes raras, comprometendo a avaliação por classe na avaliação cega; (iii) dados sintéticos podem induzir viés de padronização linguística, mitigado por exemplos reais e aprendizado ativo.

Do ponto de vista experimental, o estudo avalia o *pipeline* integrado sem ablação que isole cada etapa (geração sintética, supervisão fraca, aprendizado ativo e *Self-Learning*), nem comparação quantitativa sintético *versus* real em treinos controlados. *Baselines* de *encoder* adicionais além de BERTimbau Base/Large e Qwen também não foram realizados. Esses desdobramentos permanecem como trabalho futuro. Reconhece-se, ainda, que a mesma equipe conduziu a curadoria do acervo de treino e a anotação do *Gold Standard*.

7. Conclusão

Este trabalho apresentou um pipeline *data-centric* para construção e curadoria contínua de corpora de REN no domínio jurídico trabalhista brasileiro. Ao integrar geração sintética, supervisão fraca, aprendizado ativo e revisão humana seletiva, consolidou-se um *corpus* robusto, com mais de 53 mil exemplos em 32 classes.

A validação interna e a avaliação cega mostraram que a especialização orientada por dados de qualidade superou o paradigma generativo avaliado, com ganho adicional de eficiência computacional. Em termos práticos, isso sustenta a implantação local com preservação do sigilo no processamento documental.

Como trabalhos futuros, vislumbra-se a operacionalização do modelo em ferramentas institucionais, com destaque para: (i) leitores estruturados, convertendo peças processuais em bases analíticas; e (ii) sistemas de varredura em larga escala para detectar dados sensíveis em documentos disponibilizados publicamente, com auditoria contínua de conformidade LGPD. Prevê-se a publicação dos pesos do modelo para fomentar soluções abertas. Por fim, o pipeline e a taxonomia propostos oferecem base replicável para outros contextos institucionais. A extensão a outros ramos do Judiciário e da administração pública poderá requerer validação externa, curadoria local e *Gold Standards* independentes, a partir do recorte consolidado no TRT-13.

Declaração de uso de IA generativa: Claude 4.6 Sonnet e Gemini 3.1 apoiaram código e revisão do manuscrito; Gemini 2.5 Flash, a geração sintética (Seção 3.1). Todas as saídas foram validadas pelos autores.

Agradecimentos

Agradecemos ao TRT-13, à Divisão de Inteligência Artificial e ao IFPB pelo apoio institucional. Agradecemos ainda às nossas famílias pelo apoio pessoal.

Referências

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). UlyssesNER-Br: A corpus of Brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language (PROPOR 2022)*, pages 3–14, Fortaleza, Brazil. Springer.

Albuquerque, H. O., Souza, E., Lucena, D. C. G., Albuquerque, H. J. O., Silva, N. F. F. d., Dias, M. d. S., Nunes, R. O., Oliveira, A. L. I., and Carvalho, A. C. P. L. F. d. (2026). UlyssesLegalNER-Br: from legislative to legal, a comprehensive corpus of Brazilian legal documents for named entity recognition. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 331–341, Salvador, Brazil. Association for Computational Linguistics.

Andrade, C. M. V. d., França, C., Belém, F., Jallais, G., Ganem, M. A. S., Teixeira, G., Laender, A. H. F., and Gonçalves, M. A. (2023). PromptNER: Uma abordagem para reconhecimento de entidades nomeadas em dados sensíveis a partir de instâncias rotuladas automaticamente. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados (SBBD 2023)*, pages 269–281, Belo Horizonte, MG, Brasil. SBC.

Araujo, P. H. L. d., Campos, T. E. d., Oliveira, R. R. R. d., Stauffer, M., Couto, S., and Bermejo, P. (2018). LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *Computational Processing of the Portuguese Language (PROPOR 2018)*, pages 313–323. Springer.

Ariai, F., Mackenzie, J., and Demartini, G. (2025). Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*, 58(6):1–37.

Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018 (Lei Geral de Proteção de Dados Pessoais).

Brito, A. M., Pinheiro, V., Furtado, V., Monteiro Neto, J. A., Bomfim, F. d. C. J., da Costa, A. C. F., Silveira, R., and Aragão, N. (2023). CDJUR-BR — Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas. In *Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology (STIL 2023)*, pages 176–195, Belo Horizonte, Brazil. Association for Computational Linguistics.

Castro, P. V. Q. d. (2019). Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico. Master's thesis, Universidade Federal de Goiás, Goiânia.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Dias, M., Boné, J., Ferreira, J. C., Ribeiro, R., and Maia, R. (2020). Named entity recognition for sensitive data discovery in Portuguese. *Applied Sciences*, 10(7):2303.

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., and Joty, S. (2024). Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand. Association for Computational Linguistics.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing*. Pearson, 3rd ed. draft edition.

Maffeo, G., Silva, C., and Oliveira, H. G. (2026). Prompt engineering for named entity extraction from Portuguese legal documents. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 1092–1097, Salvador, Brazil. Association for Computational Linguistics.

Mosbach, M., Pimentel, T., Ravfogel, S., Klakow, D., and Cotterell, R. (2023). Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314. Association for Computational Linguistics.

Nunes, R. O. (2025). Data contamination in specialized named entity recognition corpora. Master's thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Nunes, R. O., Balreira, D. G., Spritzer, A. S., and Freitas, C. M. D. S. (2024). A named entity recognition approach for Portuguese legislative texts using self-learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR 2024)*, pages 290–300. Springer.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference (BRACIS 2020)*, pages 403–417. Springer.

Tran, H. T. H., Chatterjee, N., Pollak, S., and Doucet, A. (2024). DeBERTa beats behemoths: A comparative analysis of fine-tuning, prompting, and PEFT approaches on LegalLensNER. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 371–380, Miami, FL, USA. Association for Computational Linguistics.

Vithanage, D., Yu, P., Xie, Q., Xu, H., Wang, L., and Deng, C. (2025). A comprehensive evaluation of large language models for information extraction from unstructured electronic health records in residential aged care. *Computers in Biology and Medicine*, 197:111013.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. (2025). Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5).

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. *arXiv preprint arXiv:2004.12158*.