UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS – CECH
DEPARTAMENTO DE LETRAS
BACHARELADO EM LINGUÍSTICA

BIANCA COUTO VINCENZI

INVESTIGAÇÃO DE TÉCNICAS DE IN-CONTEXT LEARNING PARA RECONHECIMENTO DE ENTIDADES NOMEADAS PARA PORTUGUÊS

SÃO CARLOS - SP
2025

BIANCA COUTO VINCENZI

INVESTIGAÇÃO DE TÉCNICAS DE IN-CONTEXT LEARNING PARA RECONHECIMENTO DE ENTIDADES NOMEADAS PARA PORTUGUÊS

Monografia apresentada ao Departamento de Letras da Universidade Federal de São Carlos como Trabalho de Conclusão de Curso para obtenção do título de Bacharel em Linguística.

Orientador: Prof.a Dra. Ariani Di Felippo.

SÃO CARLOS/SP
2025

Agradecimentos

À toda a minha família e principalmente aos meus pais, Leandro e Elizangela, meu alicerce e meu porto seguro. Agradeço imensamente por todo o suporte que me permitiu chegar até aqui e por terem permanecido ao meu lado em cada etapa desta formação, sempre com compreensão, fé e amor incondicional. Tudo o que sou e cada passo que conquistei devo ao esforço e à dedicação de vocês. À minha irmã, Alice, por todos os momentos de companheirismo e risadas, os quais foram minha fonte de força nos momentos de dificuldade.

À minha namorada, Marcia, agradeço pelo amor e pela parceria diária inabalável. Ter você ao meu lado, dividindo os sonhos e trazendo leveza aos dias difíceis, tornou tudo especial. Levo desta fase o crescimento acadêmico e pessoal, mas, principalmente, a sorte de ter encontrado você.

Às minhas amigas de vida e de curso, Anaterra e Pamela, minha gratidão pela troca enriquecedora de experiências e pelo apoio nos momentos de exaustão. A amizade de vocês foi um refúgio e é um presente que levarei para sempre.

À minha professora e orientadora, Ariani, expresso minha profunda admiração e gratidão. Obrigada pela generosidade no compartilhar do saber, pela paciência incansável e pela condução segura e brilhante deste trabalho. Agradeço pela confiança depositada em mim.

Por fim, agradeço à UFSCar e a todas as pessoas que fizeram parte desse ciclo.

RESUMO

No campo do Processamento de Língua Natural (PLN), o Reconhecimento de Entidades Nomeadas (REN) é a tarefa responsável por identificar as entidades nomeadas (ENs) nos textos e classificá-las de acordo com categorias pré-definidas. Essa tarefa permite a extração automatizada de informações-chave, como nomes de pessoas, organizações, locais, datas e outras, contribuindo para a análise conveica textual. Diante da eficácia dos grandes modelos de linguagem (*Large Language Models* - LLM) para a tarefa de REN e outras no PLN, este trabalho investigou, de modo incipiente, algumas abordagens de Engenharia de *Prompt* para a tarefa de REN em textos escritos em português padrão ou formal. Especificamente, exploraram-se as técnicas de *In-Context Learning* (ICL) *zero-shot*, em que a tarefa é realizada apenas com base no conhecimento prévio do LLM, e *few-shot,* por meio da qual os LLMs aprendem a realizar uma tarefa a partir de alguns exemplos fornecidos no *prompt*. Cada uma delas foi empregada em conjunto a outras técnicas de *prompting,* sendo duas baseadas em instruções (*instruction-based*) (no caso, *persona prompting* e *rule-based prompting*) e outra de formatação estruturada de saída (*structured output*). Para tanto, utilizou-se uma amostra do *corpus* Porttinari-base, cujas 471 sentenças foram manualmente anotadas com ENs. Tal amostra serviu como *gold-standard* para a avaliação da tarefa de REN baseada em LLM com técnicas de ICL. Como resultado, verificou-se um salto de desempenho com a abordagem *few-shot* (medida-F1 de 66%) em comparação ao *zero-shot* (medida-F1 de 37,7%). Apesar das limitações amostrais, conclui-se que a Engenharia de *Prompt* oferece uma alternativa promissora e ágil para a tarefa de REN frente aos modelos treinados.

Palavras-chave: Reconhecimento de Entidades Nomeadas. *In-Context Learning*. Engenharia de *Prompt*.

ABSTRACT

In the field of Natural Language Processing (NLP), Named Entity Recognition (NER) is the task responsible for identifying named entities (NEs) in texts and classifying them according to predefined categories. This task enables the automated extraction of key information—such as names of people, organizations, locations, dates, and other elements—thus contributing to semantic text analysis. Given the effectiveness of Large Language Models (LLMs) for NER and other NLP tasks, this study conducted an initial investigation of several Prompt Engineering approaches for NER in texts written in standard or formal Portuguese. Specifically, we explored the In-Context Learning (ICL) strategy known as zero-shot prompting, in which the task is performed solely based on the LLM's prior knowledge, and few-shot prompting, through which LLMs learn to perform a task from a few examples provided directly in the prompt. Each strategy was employed in combination with additional prompting techniques, including two *instruction*-based approaches (i.e., *persona prompting* and *rule-based prompting*) and one type of *output formatting.* To this end, a sample from the Porttinari-base corpus was used, consisting of 471 sentences manually annotated with NEs. This sample served as the gold standard for evaluating NER based on LLM with ICL techniques. Results show that a substantial performance improvement with the few-shot approach (F1-Score of 66.0%) compared to the zero-shot scenario (F1-Scores of 37.7%). Despite the sample limitations, it is concluded that Prompt Engineering offers a promising and agile alternative to the completion of NER to fine-tuned models.

Keywords: Named Entity Recognition. *In-Context Learning*. *Prompt* Engineering.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

PLN	Processamento de Língua Natural
REN	Reconhecimento de Entidade Nomeada
LLM	*Large Language Model*
ICL	*In Context-Learning*
HAREM	Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas
EN	Entidade Nomeada
CD	Coleção Dourada
IA	Inteligência Artificial
AM	Aprendizado de Máquina
UD	*Universal Dependencies*

SUMÁRIO

1. Introdução

O Processamento de Língua Natural (PLN) é um campo que tem como objetivo investigar e propor métodos e sistemas de processamento computacional da linguagem humana (Caseli *et al*., 2023). Além disso, é um campo que busca automatizar a interpretação e/ou a geração de língua natural predominantemente escrita, abrangendo tarefas tradicionais como tradução automática e outras mais recentes como análise de sentimento e detecção de *fake news* (Jurafsky; Martins, 2025).

O Reconhecimento de Entidades Nomeadas (REN) (em inglês, *Named Entity Recognition* - NER) é uma tarefa que pode compor a fase de interpretação de um sistema de PLN. Essa tarefa engloba duas etapas: (i) identificação das Entidades Nomeadas (ENs) e (ii) classificação das ENs de acordo com o contexto em categorias pré-definidas (Freitas, 2022). Em (1), por exemplo, há 2 ENs de 2 categorias genéricas bastante difundidas (PESSOA e LOCAL). Assim, o REN organiza elementos dispersos ao classificá-los em categorias formais, convertendo referentes centrais em unidades semanticamente tipificadas. Isso permite reconhecer "que tipo de coisa" está sendo mencionada no texto, aprimorando a interpretação.

(1) [PESSOA Dra. Ana] voou para [LOCAL São Paulo].

Desde seu surgimento na década de 1990 como um segmento da Extração de Informação (Grishman; Sundheim, 1996), a tarefa de REN ganhou espaço no PLN. A Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas, conhecida como HAREM, foi a primeira campanha de avaliação de REN para o português, organizada pela Linguateca[1] em meados dos anos 2000. O objetivo da iniciativa era criar um padrão de avaliação, um conjunto de categorias e *corpora* de referência que permitissem comparar sistemas de forma sistemática. Tal avaliação teve 2 edições: Primeiro HAREM (2007) e o Segundo HAREM (2008).

O estado-da-arte atual de REN é o método supervisionado baseado em *fine-tuning* de Wang et al. (2020), no qual a versão *large* do BERT[2] é combinada a um mecanismo de concatenação de *embeddings* (ACE) e a um decodificador CRF (*Conditional Random Fields*). No CoNLL-2003 (Sang; De Meulder, 2003), que é um dos *benchmarks*[3] mais tradicionais para

[1] A Linguateca é um consórcio e portal que disponibiliza recursos, avaliações, ferramentas e documentação para o PLN em português, sendo um dos pilares da pesquisa na área desde os anos 2000 (https://www.linguateca.pt/).
[2] Segundo Devlin *et al.* (2018), BERT é um modelo de linguagem baseado em *transformers* que aprende representações contextuais bidirecionais das palavras, permitindo capturar o significado considerando tanto o contexto anterior quanto o posterior, o que o torna extremamente eficaz em diversas tarefas de PLN.
[3] *Benchmark* é um conjunto padronizado de dados rotulados (comumente *corpora* manuais) e métricas usado para avaliar e comparar o desempenho de diferentes modelos ou métodos em uma tarefa específica.

REN em inglês, composto por ~20 mil sentenças (~300 mil *tokens*) de textos jornalísticos anotados com quatro tipos de entidades (*pessoa, organização, local e miscelânea*), o sistema de Wang et al. alcança medida-F1[4] de 94,6%.

Abordagens baseadas em modelos pré-treinados (como o BERT) usados em *fine-tuning* (ajuste fino) supervisionado são muito eficazes em *benchmarks*, mas apresentam limitações, como: (i) necessidade de grandes quantidades de dados anotados, cuja produção é lenta e cara; (ii) alto custo computacional, especialmente para treinar ou ajustar modelos grandes; e (iii) baixa portabilidade entre domínios, pois o modelo tende a degradar rapidamente quando é aplicado fora do contexto em que foi treinado.

Buscando formas alternativas para superar esses problemas, tem-se visto um novo paradigma com o uso dos grandes modelos de linguagem (em inglês, *Large Language Models* - LLMs) e técnicas de Aprendizado em Contexto (em inglês, *In-Context Learning* - ICL) (Brown *et al.,* 2020). Essa combinação dispensa ajuste de parâmetros e funciona só com exemplos fornecidos no *prompt* (Ashok; Lipton, 2023).

Nesse cenário, destaca-se o emprego da técnica *few-shot* ICL, que é uma técnica de aprendizado em contexto que fornece ao modelo exemplos para que ele aprenda a realizar determinada tarefa. Alguns trabalhos nessa linha empregam diferentes técnicas de *prompting* (como instrução, exemplos, raciocínio explicativo e controle/formato de saída) em conjunto (Ashok; Lipton, 2023) ou técnicas adicionais como *entity-level embedding* para selecionar exemplos relevantes (do conjunto de treinamento) e *self-verification* para validar as entidades geradas (Wang *et al.,* 2023).

O método *few-shot* de Ashok e Lipton (2023) (PromptNER), que se baseia no GPT4 e em várias técnicas de *prompting* combinadas, alcançou medida-F1 de 83,48% no CoNLL-2003. O método *few-shot* de Wang et al. (2023) (GPT-Ner), que se baseia no GPT3 (Brown *et al.,* 2020) e nas estratégias *entity-level embedding* e *self-verification*, alcançou medida-F1 de 90,91% para o CoNLL-2003. Vale ressaltar, no entanto, que a replicação dos resultados pode variar, uma vez que pequenas alterações no *prompt*, no LLM usado, no pós-processamento ou no conjunto de exemplos, podem afetar o desempenho.

Esses resultados recentes evidenciam que é válido usar LLMs com *prompting* do tipo ICL (sem treinamento) para realizar REN em *corpora* padrão-ouro como o CoNLL-2003,

[4] A medida-F1 é uma métrica que combina precisão e revocação por meio de sua média harmônica, resumindo em um único valor o equilíbrio entre acertos e cobertura do sistema (Jurafsky; Martin, 20025).

obtendo resultados considerados competitivos frente ao estado-da-arte dos métodos supervisionados tradicionais. Esse avanço não se deve apenas à substituição de modelos como o BERT por LLMs, mas à combinação de *prompts* bem estruturados, recuperação inteligente de exemplos relevantes e mecanismos adicionais de verificação e pós-processamento. Com isso, tais métodos tornam-se uma alternativa prática ao *fine-tuning* intensivo.

Além dos resultados competitivos, esses métodos se destacam pela flexibilidade e generalidade, qualidades valorizadas no ecossistema contemporâneo do PLN. Cada vez mais, a área também tem priorizado critérios de usabilidade e robustez (especialmente em cenários de poucos dados e domínios variados), o que tem impulsionado o desenvolvimento de abordagens capazes de equilibrar desempenho, adaptação e custo.

Para textos formais em português, o estado-da-arte é, segundo Vieira e Silva (2023), o BERTimbau (Souza *et al.* 2019), que é um modelo pré-treinado do tipo BERT desenvolvido para o português. Quando ajustado à tarefa de REN, Souza *et al.* (2019) reportam que ele obteve 78,6% de medida-F1 para as 10 categorias gerais (PESSOA, LOCAL, ORGANIZAÇÃO, TEMPO, OBRA, EVENTO, ABSTRAÇÃO, COISA, MODO E ACONTECIMENTO) do Primeiro HAREM.

No que tange ao uso de LLMs com técnicas de ICL (sem treinamento) para REN em português, destacam-se os trabalhos que focam em domínios especializados, como o (i) jurídico, utilizando algum membro da família GPT (Oliveira *et al.,* 2024, Coelho *et al.,* 2024) ou modelos de língua em português (Nunes *et al.,* 2024), como o Sabiá (Pires *et al.,* 2023), e (ii) em textos históricos (Sarcinelli, 2025), utilizando modelo de pesos abertos[5] como o LLaMA (Grattafiore *et al.,* 2024) e outros.

Diante disso, este Trabalho de Conclusão de Curso (TCC) investigou o emprego de LLMs e algumas técnicas de ICL para a tarefa de REN em textos jornalísticos em português, considerando, dado o escopo do trabalho, apenas entidades das categorias mais difundidas na literatura, que são PESSOA, LOCAL E ORGANIZAÇÃO (as chamadas PLO).

Para apresentação do trabalho, este texto está organizado em 6 seções. Na seção 2, apresenta-se uma revisão da literatura sobre a atividade de REN, destacando noções ou conceitos importantes, métricas de avaliação, recursos e métodos para REN em português. Na seção 3, serão descritas as metodologias utilizadas para a realização da presente investigação,

[5] O termo "modelos de pesos abertos" (do inglês *open-weights models*) refere-se a modelos cujos parâmetros treinados (pesos) são disponibilizados publicamente, permitindo que sejam utilizados em infraestrutura própria (Jurafsky; Martin, 20025).

tal como o *corpus*, as diretrizes de anotação das Entidades Nomeadas no corpus Porttinari-base e explicitação da metodologia do *In-Context Learning*. Na seção 4, apresenta-se o trabalho de anotação das ENs realizado no corpus Porttinari-base e a aplicação do modelo LLM Llama-3.1 em conjunto com os *prompts* para a tarefa de REN nos dois experimentos que serão realizados. Na seção 5, serão apresentadas discussões sobre os resultados obtidos, que serão discutidos a partir de métricas de análises e gráficos de matriz de confusão. E para finalizar, a seção 6 conta com as considerações finais do presente trabalho.

2. Revisão da literatura

2.1 Entidade Nomeada: definição, segmentação e classificação

O conceito de entidade nomeada (EN) varia bastante na literatura (Piai, 2025). As definições, segundo Marrero et al. (2013), pautam-se principalmente nas noções de (i) categoria gramatical (nome próprio), (ii) identificador único ou (iii) domínio ou propósito de aplicação, as quais nem sempre são empregadas de forma excludente.

As definições desse termo no PLN são comumente determinadas em eventos de avaliação conjunta (em inglês, *shared task*) (Santos; Freitas, 2024). Mais especificamente, as *shared tasks* fornecem um cenário padronizado para pesquisa ao disponibilizar definições claras das tarefas, conjuntos de entidades, *corpora* anotados, manuais de anotação detalhados e métricas de avaliação unificadas, permitindo que diferentes grupos desenvolvam, comparem e aprimorem seus métodos sob as mesmas condições experimentais. Nesses desafios, equipes de pesquisa trabalham sobre o mesmo problema e os mesmos dados, utilizando protocolos de avaliação comuns, o que possibilita mensurar de forma justa avanços metodológicos e estabelecer ou superar o estado da arte. Dessa forma, as *shared tasks* desempenham papel crucial na consolidação de taxonomias, práticas de anotação e *benchmarks* robustos para o desenvolvimento contínuo do PLN. (Santos; Freitas, 2024).

Quanto à tarefa de REN, destacam-se as iniciativas MUC (*Message Understanding Conference*) Grishman e Sundheim (1996), ACE (*Automatic Content Extraction*) (Doddington *et al.,* 2004), o CoNLL (*Conference on Computational Natural Language Learning*) (Sang, De Meulder, 2003), e HAREM (Santos; Cardoso, 2007, Mota; Santos, 2008).

As edições MUC-6 (1995) e MUC-7 (1997) iniciaram o conceito moderno de REN. Nesses eventos, as expressões que designam ENs foram definidas como "identificadores únicos". Essa definição diz respeito particularmente aos nomes próprios das categorias PESSOA (PER) (p.ex.: "Barack Obama"), LOCAL (LOC) (p.ex.: "Maracanã") e ORGANIZAÇÃO (ORG) (p.ex.: "TAM"), classificados como ENAMEX (em inglês, *entity named expression*) e amplamente tratados como o foco da tarefa de REN. O MUC também estava interessado no reconhecimento de expressões temporais (como datas e horas) e numéricas (como porcentagens e valores). Tais expressões foram classificadas como TIMEX (do inglês, *temporal expression*) e NUMEX (do inglês, *numerical expression*), respectivamente.

Adotando uma concepção fixa e unívoca de EN, a categoria de uma entidade no MUC não muda conforme o contexto e as categorias PER, LOC e ORG não podem coincidir. Isso quer dizer que, no MUC, o conjunto de categorias de ENs foi definido de cima para baixo (*top-down*), pois os organizadores estabeleceram previamente quais categorias existiriam (PER, ORG, LOC, etc.), tratadas como tipos fixos, estáveis e não ambíguos, que não mudam conforme o contexto. Assim, a classe de uma entidade é decidida antes da análise dos textos, e o anotador humano (ou automático) deve encaixar cada menção em uma dessas classes rígidas, mesmo quando o contexto permitir outras interpretações. Tal concepção de EN recebeu várias críticas, pois ignora a polissemia e o uso metonímico dos nomes próprios.

As edições do ACE de 2002 a 2005 expandiram a tarefa. Ao buscar avançar a extração de informações para segurança, inteligência e análise de eventos (como conflitos, ataques terroristas e segurança nacional), a iniciativa considerou, além de PER, LOC e ORG, outras 4 categorias: GEO-POLITICAL ENTITY (GPE) (entidades geo-políticas), FACILITY (FAC) (instalações como prédios, rodovias, pontes, etc.), VEHICLE (VEH) (veículos como carros, caminhões, aviões, etc.) e WEAPON (WEA) (armas ou armamento como revólveres, mísseis, armas biológicas, etc.). Além disso, esses eventos não se limitaram à identificação de nomes próprios, mas sim ao reconhecimento de tudo o que fosse pessoa, local, organização, local, etc., sem restrições de forma (podiam ser realizados linguisticamente como substantivo, pronome, nome próprio, sintagma nominal, etc.). Ademais, ao contrário do MUC, as categorias eram definidas com base no contexto, levando-se em conta a polissemia.

As conferências CoNLL de 2002 e 2003 consolidaram *benchmarks* amplamente usados para avaliação em REN, sobretudo o CoNLL-2003. Mais precisamente, essa iniciativa adotou as categorias do MUC (PER, LOC e ORG) e acrescentou mais uma, MISC, que engloba evento

(p.ex.: Segunda Guerra Mundial, Olimpíadas de 2024, etc.), nacionalidade (p.ex: brasileiro, catalão, etc.), afiliação religiosa/política (p.ex.: católico, muçulmano, etc.) e obra cultural (p.ex.: Harry Potter, Mona Lisa, etc.). Com isso, o CoNLL estendeu o conceito de EN para incluir também nomes próprios e comuns que expressam conceitos considerados relevantes para a extração de informação e que não se encaixavam nas categorias anteriores.

No HAREM (2007-2008), por exemplo, adotou-se a terminologia "entidade mencionada", pois o objetivo central da avaliação não era apenas testar reconhecimento de entidades convencionais, mas avaliar a capacidade dos sistemas de identificar qualquer expressão referencial que exercesse papel de entidade no discurso (Vieira e Silva, 2023). Além das 3 categorias mais tradicionais (PESSOA, LOCAL e ORGANIZAÇÃO), o evento considerou mais 7 categorias (DATA, VALOR, ABSTRAÇÃO, OBRA, EVENTO, COISA e OUTRO), pautando a identificação de uma EN pela ocorrência de ao menos uma letra maiúscula (ou algarismo) (p.ex.: "**B**arack Obama", "presidente da **R**epública", "R$ **200**", etc.). Ademais, o HAREM, em conformidade com o ACE, adotou a anotação *bottom-up*, na qual as categorias são atribuídas após análise contextual, permitindo polissemia, metonímia e vagueza.

Outros autores, como Freitas e Souza et al. (2023), trabalham com a definição de EN em função do domínio do *corpus* ou da aplicação de REN. Ao anotar um *corpus* composto por boletins e relatórios técnicos do domínio do gás e petróleo (PetroNer), os autores consideram que as ENs são elementos nominais (próprios e comuns) e adjetivais relevantes para a área/aplicação, como "bioturbação", "Bacia do Espírito Santo", "estratigráfico", etc.

Em suma, a concepção de EN no PLN evoluiu para além da noção estrita de nome próprio. A concepção mais funcional do termo abrange não apenas entidades *per se*, mas também elementos cruciais para a análise de um domínio, como datas, expressões numéricas e conceitos nominais específicos. E quando se diz "concepção funcional", enfatiza-se que definir o que é uma entidade em um *corpus* está ligado muito mais ao propósito da aplicação e das especificidades do domínio do que a uma definição gramatical consensual.

Além da definição do que anotar ou reconhecer como EN, a tarefa de REN impõe desafios de segmentação e classificação. Uma EN pode ser segmentada de diferentes maneiras, a depender da granularidade empregada. As segmentações em (2), cujos exemplos foram elaborados com base em Freitas (2022), são igualmente aceitáveis. A diferença entre elas é que a primeira (2a) considera a expressão maior, enquanto as demais (2b,c) consideram entidades mais granulares, que correspondem a uma leitura mais composicional. O reconhecimento em

(2b,c) considera o que se denomina ENs aninhadas, quando duas ou mais entidades compõe outra maior.

(2) a. [Secretaria de Educação de São Carlos] **LOC** ou **ORG**

b. [Secretaria de Educação]**ORG** de [São Carlos] **LOC** ou ORG

c. [Secretaria] **ORG** de [Educação] **SABER** de [São Carlos] **LOC** ou **ORG**

Um aspecto que pode influenciar a identificação/segmentação é a irregularidade no uso das maiúsculas em expressões compostas. Expressões compostas como "Porto de Santos" e "porto de Santos" refletem o uso pouco sistemático da maiúscula, sendo, por isso, expressões linguísticas distintas de uma mesma EN. Sintagmas nominais como "casa de João" (em "A casa de João é grande.") e "relógio de Sol" (em "O relógio de Sol da praça.") demandam tratamento diferente. No primeiro caso, apenas "João" é uma EN da classe PESSOA, posto que a expressão completa ("casa de João") não designa uma classe ou objeto. Esse já não é o caso de "relógio de Sol", pois a expressão completa se refere a uma EN composta.

Expressões de valor também impõem desafios à segmentação. Intervalos de valor como "entre 3 e 4%" podem ser anotados de forma mais restrita, segundo a qual apenas os números e símbolos ("3", "4%") são marcados como ENs, enquanto preposições ("entre", "e") não fazem parte da entidade. O mesmo ocorre com expressões quantificadas como "quase 10 kg", cuja segmentação pode ou não incluir o quantificador ("quase") como parte integrante da EN. As expressões de tempo também impõem desafios semelhantes, merecendo, aliás, uma documentação específica no Segundo HAREM (Hagège, 2008).

No que tange à classificação, a polissemia é um fenômeno presente na tarefa de REN. Nos exemplos em (2c), indicou-se, por exemplo, que "São Carlos" pode ser LOC ou ORG a depender de quando é utilizado para fazer referência à cidade enquanto espaço geográfico ou à administração pública. Outros fenômenos semânticos envolvidos são a metonímia e vagueza. A metonímia ocorre quando um termo é usado para se referir a uma entidade de maneira indireta ou simbólica, em vez de mencionar literalmente o nome da entidade. Em "fontes próximas ao Palácio do Planalto", por exemplo, "Palácio do Planalto" não se refere literalmente a um local (edifício), mas ao "presidente da república", sendo, portanto, classificado como PESSOA (do tipo cargo). A vagueza ocorre quando uma EN pode ser simultaneamente interpretada em categorias diferentes. "Immanuel Kant" em "um filósofo seguidor de Immanuel Kant", por exemplo, pode ser classificado como PESSOA ou ABSTRAÇÃO (*obra*). Nesses casos em que

o contexto não permite estabelecer com certeza a categoria, o HAREM, por exemplo, considera aceitável anotar múltiplas classes.

2.2. Reconhecimento de Entidade Nomeada: caracterização da tarefa

Como mencionado, a tarefa de REN surgiu na década de 90, mais especificamente, na 6ª edição do evento *Message Understanding Conference* (MUC), concebida como um tipo de Extração de Informação (Grishman; Sundheim, 1996). Em (1), há um exemplo de identificação/segmentação e classificação de REN. Nele, há 2 menções, sendo cada uma delas pertencente a uma categoria genérica distinta, no caso, PESSOA ("Dra. Ana") e LOCAL ("São Paulo"). Tais menções são do tipo *single*, isto é, *tokens* unitários. Caso a sentença fosse "Dra. Ana voou para São Paulo em 30 de novembro", ter-se-ia mais uma EN, no caso, "30 de novembro, pertencente à categoria TEMPO e do tipo multipalavra.

(1) [PESSOA Dra. Ana] voou para [LOCAL São Paulo].

Como mencionado, o REN é responsável por identificar e classificar todos os elementos pertencentes a uma EN. Abordagem clássica para esse problema é a classificação sequencial, isto é, *token a token.* Segundo autores como Vieira e Silva (2023) e Jurafsky e Martin (2025), existem diversos esquemas de anotação que funcionam como um sistema de marcação posicional, identificando quais termos compõem uma entidade e suas delimitações de início e fim dentro de uma sequência de texto, a saber: IO (*Inside-Outside*) (Pirovani; Oliveira, 2018), BIO (*Begin-Inside-Outside*) (Ramshaw; Marcus, 1999), BIOES (*Beginning, Inside, Outside, End, Single*) (Ratinov; Roth, 2009) e BILOU (*Begin-Inside-Last-Outside-Unit*) (Amaral; Vieira, 2014). Entre eles, o formato BIO é o mais difundido.

Na Figura 1, tem-se uma comparação bastante ilustrativa entre os formatos IO, BIO e BIOES, extraída de Jurafsky e Martin (2025). Vale destacar que, em qualquer um desses esquemas, as etiquetas de delimitação da sequência de *tokens* (ou *span*) que representa uma EN são subespecificadas pela categoria da EN em questão, como I-LOC para "Chicago".

No esquema BIO, todo *token* que inicia de uma sequência de interesse é anotado com a etiqueta "B-" (*Begin*), os demais *tokens* que ocorrem no interior do *span* são anotados com "I-" (*Inside*) e os *tokens* fora do *span* são anotados com "O" (*Outside*). Dessa forma, esse tipo de anotação representa exatamente as mesmas informações que a anotação entre colchetes ilustrada no exemplo (1), com a vantagem de ser mais explícita no tocante à delimitação das ENs, uma vez que envolve sequências de *tokens*.

Comparativamente ao formato BIO, o esquema IO é o mais simples, pois distingue apenas dois estados: *tokens* dentro de uma entidade (“I”) e *tokens* fora de qualquer entidade (“O”). Por não marcar o início de um *span*, o formato IO perde informação estrutural importante, já que sequências consecutivas de “I” não permitem saber onde uma entidade começa ou termina, nem diferenciar entidades adjacentes de mesmo tipo.

O formato BIO resolve esse problema ao introduzir a etiqueta B (*Begin*), que marca explicitamente o primeiro *token* de uma entidade. Dessa forma, o modelo consegue identificar fronteiras de entidades e distinguir entidades vizinhas.

O BIOES é uma extensão do formato BIO, acrescentando (i) “E” (*End*), para indicar o último *token* de uma entidade multi-token e (ii) “S” (*Single*), para entidades compostas por um único *token*. Essas duas novas etiquetas tornam a estrutura interna dos *spans* mais explícita, facilitando o aprendizado da tarefa. Entretanto, essa granularidade aumenta a complexidade da anotação e do processamento, pois o modelo precisa aprender não apenas o início e o interior dos *spans* (como em BIO), mas também seus finais e a identificação de entidades unitárias.

Figura 1. Diferentes formatos de anotação de entidade nomeada.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Fonte: Jurafsky e Martin (2025).

Outro aspecto importante sobre REN é conjunto de categorias de entidades nomeadas (também chamado de *tagset*), os quais variam amplamente na literatura, sobretudo pela dependência do domínio e da aplicação de REN. No entanto, algumas taxonomias se tornaram especialmente influentes e recorrentes.

O *benchmark* CoNLL-2003 popularizou um conjunto enxuto de 4 categorias (PESSOA, ORGANIZAÇÃO, LOCAL e MISC), privilegiando universalidade e fácil portabilidade entre domínios. Já o OntoNotes 5.0 (Weischedel *et al.,* 2013), amplamente utilizado em sistemas

supervisionados e empregado no CoNLL-2012, adota um *tagset* substancialmente mais amplo do que os esquemas clássicos (como MUC ou CoNLL-2003). Tal taxonomia inclui várias categorias de ENs, ultrapassando uma dezena de classes, entre elas: PERSON, ORGANIZATION, LOCATION/GPE (local geopolítico), WORK-OF-ART (obra), PRODUCT (produto), LAW (lei), EVENT (evento), LANGUAGE (língua), NORP (nacionalidade/grupo religioso ou político), FACILITY (instalação), QUANTITY (quantidade), etc.

Em português, o HAREM estabeleceu uma das taxonomias mais detalhadas, com 10 categorias principais (ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO), 41 tipos e alguns subtipos, tornando-se referência para múltiplos projetos de anotação e avaliação. Outras taxonomias ainda mais extensas aparecem no contexto biomédico, jurídico e financeiro, em que categorias altamente especializadas (como gene, doença, fármaco, jurisprudência, contrato etc.) são necessárias. Em síntese, a literatura oferece desde *tagsets* mínimos e altamente generalizáveis até taxonomias extensas e orientadas a domínios, e a escolha depende do equilíbrio desejado entre cobertura semântica, consistência de anotação e viabilidade de treinamento.

2.3 Avaliação do desempenho de métodos de REN

Independentemente da abordagem utilizada, qualquer método ou modelo de PLN precisa ser avaliado, uma vez que somente por meio disso é possível compreender em quais situações ele funciona, por que funciona e como pode ser aprimorado (Madureira, 2024). Essa exigência sustenta a prática de comparar um sistema ao estado-da-arte, entendido como o patamar mais avançado de desenvolvimento e conhecimento disponível sobre determinado tópico.

Existem diferentes maneiras de se contabilizar as predições corretas de ENs no esquema de classificação sequencial. Segundo Sarcinelli (2025), abordagens mais restritas, as quais são mais comuns, não consideram predições parciais (isto é, de limites desalinhados) como corretas. Isso significa que um sistema que reconhece apenas “Ana” e não “Dr. Ana” (em (1)) não prediz corretamente a EN. Contudo, há abordagens mais flexíveis, que consideram diferentes tipos de acerto por parte dos métodos ou sistemas, incluindo predições parciais, como é o caso do HAREM (Santos *et al,* 2006).

Seguindo a lógica da classificação sequencial dos *tokens* que compõem uma sentença, a avaliação do desempenho de sistemas de REN é feita com as clássicas medidas de precisão

(*precision*) (P), revocação (R) (*recall*) e medida-F1 (F1-*score*), as quais se aplicam a cada entidade. Essas medidas podem ser calculadas da seguinte forma (3):

(3)
$$P = \frac{VP}{VP + FP}$$
$$R = \frac{VP}{VP + FN}$$
$$F1 = \frac{2 \times P \times R}{P + R}$$

Nas fórmulas em (3), VP (verdadeiro positivo) é o número de ENs corretamente reconhecidas, FP (falso positivo) é o de entidades erroneamente reconhecidas, e FN (falso negativo) é o de ENs não reconhecidas pelo método. A medida-F1 é a média harmônica de P e R.

Quando há mais de uma categoria, o cálculo da F1 pode variar. Na F1 *micro*, VP, FP e FN de todas as categorias são computados conjuntamente. Na F1 *macro*, calculam-se P e R por categoria e, depois, faz-se a média dos valores. A F1 *micro* é mais adequada em cenários com desbalanceamento entre categorias, pois leva em conta a quantidade de exemplos de cada classe; já a F1 *macro* atribui o mesmo peso a todas elas.

2.4 Recursos linguísticos e computacionais para REN em português

Para o desenvolvimento de sistemas de PLN, os *corpora* anotados, isto é, coleções de textos autênticos com informações linguísticas explícitas (Sinclair, 2005; McEnery *et al.,* 2006), são recursos linguísticos (textuais) essenciais, especialmente os de referência (*gold standard*), pela consistência das anotações (Duran; Pardo, 2024). Quando as anotações são inteiramente manuais ou automáticas com posterior revisão humana (semiautomáticas), diz-se que elas são do tipo "padrão ouro" ou "douradas". Esses recursos são essenciais porque atuam tanto no desenvolvimento de sistemas simbólicos, fornecendo conhecimento explícito para a formulação de regras linguísticas, quanto em sistemas baseados em Aprendizado de Máquina (AM), servindo como base para o aprendizado supervisionado. Além disso, são fundamentais para a avaliação desses métodos e modelos.

Para investigar a tarefa de REN em portugês, os pesquisadores do PLN contam principalmente com os *corpora*, *tagsets* e diretrizes de anotação do HAREM.

A duas edições desse evento produziram as Coleções Douradas (CDs), chamadas Primeiro HAREM e Segundo HAREM, além do Mini HAREM (organizado para uma edição especial do evento). O Quadro 1 sistematiza as estatísticas das CDs principais. Os *corpora* do HAREM foram elaborados com o objetivo de abranger variantes do português de diferentes regiões, reunindo materiais provenientes na sua maioria de Portugal e Brasil, mas também de Angola, Moçambique, Macau, Índia, Timor-Leste e Cabo Verde.

A primeira edição do HAREM (Santos, Cardoso, 2007) foi dividida em dois subeventos: Primeiro HAREM e Mini HAREM. Realizada em 2005, essa edição contou com 10 equipes de seis países. A CD do Primeiro HAREM reúne 129 textos anotados manualmente de diferentes gêneros, incluindo conteúdo extraído da *web*, notícias jornalísticas, transcrições de entrevistas, relatórios técnicos online e textos políticos. No total, essa coleção reúne aproximadamente 80.000 mil palavras e 5.132 entidades distribuídas em 10 categorias e 41 tipos (e alguns subtipos) (cf. Seção 3.2). Essa organização hierárquica entre categorias e tipos foi adotada em todas as CDs subsequentes. O Mini HAREM ocorreu em 2006, quando cinco equipes da edição anterior reenviaram seus sistemas. Para essa avaliação, produziu-se uma nova CD com 128 textos do mesmo domínio do Primeiro HAREM e 3.758 entidades anotadas.

Em 2008, organizou-se a segunda edição do HAREM (Mota; Santos, 2008), ampliando a diversidade textual ao incluir *blogs* e artigos da Wikipédia. A nova CD reúne 129 documentos e 7.847 entidades categorizadas. Quanto à *taxonomia*, a principal diferença em relação à edição anterior foi a troca da categoria "Variado" por "Outro". Além disso, ressalta-se que diversos tipos foram renomeados ou ajustados, assim como algumas diretrizes de anotação (segmentação e classificação) foram modificadas.

Quadro 1. Os principais *corpora* em português com anotação dourada de ENs.

Corpus	**Palavras**	**Entidades**	**Domínio**	**Anotação**
Primeiro HAREM	80.000	5.132	Geral	Dourada
Segundo HAREM	100.000	7.847	Geral	Dourada

Fonte: Vieira e Silva (2023).

A tarefa de REN, sobretudo os métodos desenvolvidos segundo o paradigma simbólico (cf. Seção 2.6), também pode ser auxiliada por recursos lexicais do tipo *gazetteers*. No geral, os *gazetteers* são coleções de nomes próprios. Para o português, tem-se o REPENTINO (acrônimo

para REPositório para reconhecimento de ENTIdades NOmeadas para o português) (Sarmento *et al.* 2006) para o português. O REPENTINO contém mais de 45.000 nomes próprios extraídos do *corpus* WPT03 (isto é, coleção da Web portuguesa de 2003) e classificados com base em um sistema amplo e detalhado, composto por 11 categorias e 97 subcategorias.

Vale ressaltar que o repositório NILC-*Embeddings*[6] também é fonte importante de recursos lexicais, ao fornecer conjunto de *word embeddings* estáticos (como Word2Vec, FastText e GloVe) pré-treinados especificamente para o português. O treinamento foi realizado sobre um grande *corpus* multigênero, com aproximadamente 1,39 bilhão de *tokens*, compilado a partir de 17 fontes textuais diferentes, o que garante diversidade linguística e boa cobertura lexical. Por serem *embeddings* estáticos, cada palavra recebe um único vetor independente do contexto, diferentemente de modelos contextuais como BERT. Esses recursos são amplamente usados como representações semânticas em tarefas de PLN para o português.

No que tange aos expedientes computacionais para pesquisas sobre REN em português, estes podem ser organizados em três grupos: (i) ferramentas ou *toolkits* de análise linguística, spaCy[7], Polyglot[8], FreeLing[9] e NLPyPort[10], que fornecem *pipelines* prontos para uso, englobando modelos pré-treinados e módulos linguísticos específicos (ii) *frameworks* como Keras[11], PyTorch[12] e TensorFlow[13], que fornecem infraestrutura para desenvolver e treinar modelos neurais, isto é, para construir modelo de AM (necessitando de *corpus*, arquitetura e treinamento fornecidos pelo usuário), e (iii) plataformas que fornecem modelos pré-treinados (especialmente, *transformers*), isto é, *hubs* de modelos modernos reutilizáveis. O HuggingFace[14] é a mais difundida das plataformas descritas em (c), disponibilizando modelos de redes neurais pré-treinados em dados do português (especificamente, no brWaC (*Brazilian Web-as-Corpus*) (cf. Wagner Filho *et al.,* 2018)), como o BERTimbau.

[6] https://huggingface.co/collections/nilc-nlp/nilc-embeddings
[7] https://spacy.io/
[8] https://draquet.github.io/PolyGlot/
[9] https://nlp.lsi.upc.edu/freeling/node/1
[10] https://github.com/NLP-CISUC/NLPyPort
[11] https://keras.io/
[12] https://pytorch.org/
[13] https://www.tensorflow.org/?hl=pt-br
[14] https://huggingface.co/

2.5 Grandes Modelos de Linguagem (LLMs) e Engenharia de Prompt

O termo "LLM" recobre modelos fundacionais de IA baseados em redes neurais profundas (tipicamente *transformers*), geralmente com bilhões de parâmetros, que aprendem distribuições probabilísticas de *tokens* (palavras) por meio de um pré-treinamento massivo em grandes *corpora* não rotulados, sendo capazes de compreender, gerar e manipular língua natural (Brown *et al.,* 2020).

Nesse cenário, destacam-se os LLMs generativos que, embora sejam construídos a partir de *transformers*, diferenciam-se dos demais porque são pré-treinados para prever apenas o próximo *token* em uma sequência, com acesso apenas aos *tokens* anteriores (Zhao *et al.,* 2023). Essa abordagem se diferencia dos modelos bidirecionais como o BERT, que utilizam máscaras para prever *tokens* em qualquer posição da sentença, tendo acesso ao contexto completo (*tokens* antes e depois do *token* alvo). Além disso, os LLMs normalmente passam por uma segunda etapa de treinamento especializada em *instruction following*, na qual aprendem a seguir comandos e responder a *prompts* de forma mais alinhada com as intenções humanas (Sarcinelli, 2025).

As particularidades do treinamento de LLMs, aliadas à escala muito maior desses modelos em termos de parâmetros treináveis, conferem a eles habilidades consideradas "emergentes" (Wei *et al.,* 2022), isto é, capacidades que não se manifestam em modelos menores. Entre essas habilidades, destaca-se a possibilidade de executar tarefas sem treinamento explícito para cada problema (Zhao *et al.,* 2023). Em vez de alterar a arquitetura ou ajustar pesos para cada nova tarefa, basta interagir com o modelo por meio de sua interface textual (o *prompt*), formulando instruções adequadas. Esse modo de uso, que é denominado ICL, dispensa qualquer etapa adicional de treinamento. Além disso, o LLM pode adaptar-se a uma tarefa a partir de exemplos incluídos no próprio *prompt*.

A seguir, no Quadro 2, sistematizam-se as principais categorias de técnicas de ICL, também chamadas *tuning-free prompts.* Vale ressaltar que não há uma tipologia de técnicas já estabelecida e as listadas no Quadro 2 são fruto da leitura dos vários trabalhos que compõem a revisão da literatura deste trabalho.

Quadro 2. Principais técnicas de ICL.

Técnica	Definição breve
Demonstration-based	Fornece ao modelo nenhum (*zero-shot)*, um (*one-shot*) ou poucos exemplos (*few-shot*) da tarefa diretamente no prompt para que ele aprenda o padrão e o reproduza
Chain-of-Thought prompting	Técnica em que o prompt incentiva o modelo a explicitar passo a passo o raciocínio, melhorando tarefas complexas e lógicas.
Self-consistency	Gera várias cadeias de raciocínio (múltiplos CoTs) e escolhe a resposta mais frequente ou consistente para aumentar a precisão.
RAG / retrieval-based prompting	Combina *prompting* com recuperação de documentos relevantes (via busca) que são incluídos no contexto para melhorar precisão e factualidade.
Self-verification	O próprio modelo revisa, verifica ou corrige sua resposta inicial, muitas vezes com um segundo passo de julgamento.
Instruction-following	O prompt é construído como instruções claras e diretas, otimizando a obediência do modelo a comandos explícitos.
Constraint-based prompting	Impõe regras, limites ou proibições ("não faça X", "use apenas Y"), garantindo que a saída respeite restrições específicas.
Structured output prompting	Orienta o modelo a produzir saídas em formato rígido (JSON, tabela, XML, lista numerada), útil em IE, NER e extração estruturada.
Exemplars selection	Seleção automática ou inteligente de exemplos relevantes (via *embeddings* ou similaridade) para compor o *few-shot*, aumentando qualidade do ICL.

Fonte: A autora (2025).

Resumidamente, as abordagens de ICL vêm sendo amplamente exploradas no PLN devido a alguns fatores (Dong *et al.,* 2024): (i) formulação do *prompt* via língua natural, (ii) simulação do aprendizado por analogias, semelhante ao processo de decisão realizado por humanos, e (iii) aprendizado sem custos computacionais extras e sem o esforço de um retreinamento.

Das listadas no Quadro 2, as técnicas "baseadas em demonstração" (em inglês, *Demonstration-based*) do tipo *zero-shot* e *few-shot prompting* são as mais difundidas. Mais especificamente, o modelo aprende a realizar uma tarefa específica na abordagem *zero-shot prompting* somente com base em instruções fornecidas no *prompt*, mas sem exemplos sobre ela. Já na *few-shot prompting,* o modelo aprende a realizar determinada tarefa a partir de um

conjunto de exemplos. É comum utilizar a nomenclatura *k-shot* para se referenciar a essa abordagem, em que *k* é a quantidade de exemplos empregados via *prompt* (Jurafsky; Martin, 2025). Todas as categorias do Quadro 2 possuem vários tipos, que não serão listados aqui por questão do escopo do trabalho.

2.6 Métodos de REN e o estado-da-arte

As investigações sobre REN podem ser classificados de acordo com os paradigmas de Inteligência Artificial (IA)/PLN empregados. Assim, tem-se métodos do paradigma (i) simbólico, (ii) estatístico (IA clássica ou conexionista "leve"), (iii) neural profundo (ou IA conexionista), (iv) neural baseado em *transformers* (ou IA profunda contextual) e (v) baseados em LLMs (ou IA generativa).

No paradigma simbólico, a tarefa de REN é realizada por meio de regras linguísticas extraídas de *corpus,* as quais podem ser baseadas, por exemplo, em padrões sintático-semânticos, frequência de termos e similaridade com palavras em um dicionário de referência (como os *gazetteers*) (Vieira e Silva, 2023). Esse método foi o mais explorado no início do desenvolvimento de sistemas para REN, destacam-se o PALAVRAS-NER (Bick, 2006), que obteve um resultado de 58,29% de medida-F, e o PAMPO (Rocha *et al.,* 2016), que atingiu 73,36% de medida-F, ambos avaliados na CD do Primeiro HAREM.

No paradigma estatístico, REN é tratado como uma tarefa de classificação multiclasse baseada em AM, típico da IA estatística. O modelo aprende a partir de *corpus* anotado (AM supervisionado), utilizando *features* manuais para representar o contexto linguístico. Após o treinamento, o sistema é capaz de generalizar para novos textos e prever automaticamente o rótulo de cada *token* ou sequência. Modelos amplamente usados incluem HMMs (*Hidden Markov Models*), CRFs e SVMs (*Support Vector Machine*), que tradicionalmente realizam identificação e classificação de entidades de forma integrada. Esse paradigma dominou o REN entre 2000 e 2015, sendo empregado em trabalhos para o português como os de Solorio (2007), Amaral e Vieira (2014), Júnior et al. (2015) e Lopes *et al* (2019).

No paradigma da IA conexionista moderna, caracterizado pelo uso de AM profundo, a tarefa de REN passa a ser conduzida por modelos capazes de aprender automaticamente as próprias características, dispensando a engenharia manual de *features*. Esses sistemas utilizam representações distribuídas, como *embeddings* de palavras e subpalavras, e exploram arquiteturas neurais especializadas, incluindo redes neurais recorrentes (*Recurrent Neural*

Network – RNN) (como BiLSTM e modelos LSTM (*Long Short Term Memory*)+CRF), redes convolucionais e mecanismos de atenção neural, além de combinações híbridas que integram *embeddings* com camadas de CRF. Entre 2015 e 2018, esse paradigma marcou a transição dos métodos estatísticos tradicionais para modelos que capturam dependências contextuais profundas e relações sequenciais de maneira mais robusta e flexível. Nesse paradigma, destacam-se para o português os trabalhos de Santos e Guimarães (2015), Castro *et al.* (2018), Santos *et al.* (2019) e outros.

No paradigma dos métodos baseados em *transformers*, o REN passa a ser realizado por modelos profundamente contextuais, fundamentados em pré-treinamento em larga escala e no uso de *embeddings* contextuais, que variam conforme o ambiente linguístico em que cada palavra aparece. Nessa abordagem, arquiteturas como BERT e RoBERTa e suas variações específicas de língua (como mBERT e BERTimbau para o português), servem como base sobre a qual são adicionadas camadas finais de classificação para rotular entidades. Entre 2018 e 2023, esse paradigma tornou-se dominante por permitir que modelos pré-treinados generalizassem bem com pouco ajuste fino, capturando dependências contextuais complexas de forma muito mais eficaz que redes recorrentes ou convolucionais.

E é nesse paradigma que se encontra o estado-da-arte atual de REN para o inglês, com método supervisionado baseado em *fine-tuning* de Wang et al. (2020), no qual a versão *large* do BERT é combinada a um mecanismo de concatenação de *embeddings* (ACE) e a um decodificador CRF (*Conditional Random Fields*). No CoNLL-2003 (Sang; De Meulder, 2003), que é um dos *benchmarks* mais tradicionais para REN em inglês, composto por ~20 mil sentenças (~300 mil *tokens*) de textos jornalísticos anotados com quatro tipos de entidades (*pessoa, organização, local e miscelânea*), o sistema de Wang et al. alcança medida-F1 de 94,6%. Segundo Vieira e Silva (2023), o estado-da-arte do REN para o português (formal e de domínio geral) é o BERTimbau. De acordo com Souza *et al.* (2019), o Bertimbau foi ajustado (*fine-tuning*) para a tarefa REN utilizando a CD do Primeiro HAREM. Para as 10 categorias da taxonomia do HAREM, ele alcançou uma medida-F de 78,6%.

Os métodos baseados em LLMs (como os da família GPT, LLaMA, Gemma e outros) representam o paradigma da IA generativa. Destacam-se os trabalhos que empregam LLMs e técnicas de ICL, que dispensam ajuste de parâmetros e funcionam só com exemplos no *prompt* (Ashok; Lipton, 2023). Neles, a língua natural torna-se a interface: o usuário especifica instruções, formatos de saída ou exemplos, e o modelo realiza o reconhecimento de entidades

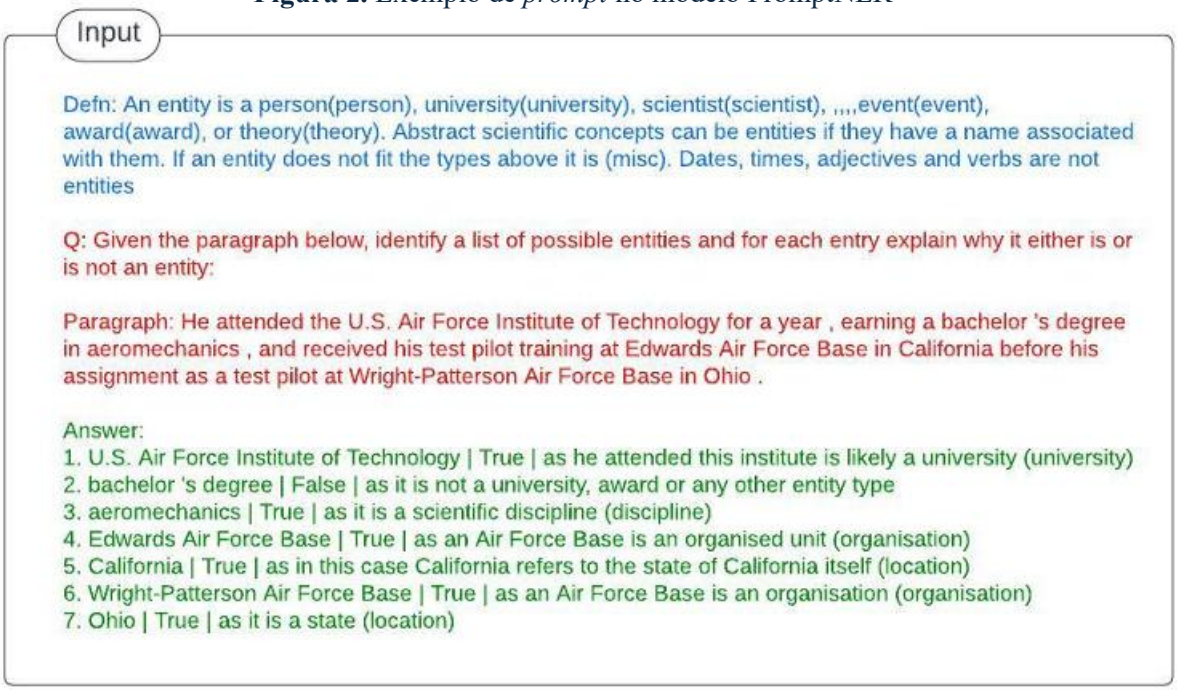
diretamente. O emprego de LLMs e técnicas de ICL deslocam o foco do ajuste de parâmetros para o *design* de *prompts* – termo utilizado neste trabalho como sinônimo de Engenharia de Prompt, sob a perspectiva de que o design, assim como a engenharia, ocupa-se da criação de soluções estratégicas para resolver problemas de desempenho e otimização de tarefas - explorando capacidades emergentes de raciocínio contextual e instrução e, com isso, evitando a necessidade de *corpora* grandes e poder computacional caro e buscando robustez e flexibilidade

Trabalhos iniciais nessa linha trataram REN como preenchimento de lacunas (*cloze*), usando *templates* rígidos, sendo, por isso, bastante dependentes do *template*, da ordem dos exemplos e do tamanho do contexto (p.ex.: Lee *et al.,* 2021, Cui et al, 2021, Wei *et al.,* 2021).

Trabalhos mais recentes, como o de Ashok e Lipton (2023), adotam LLM e combinação de diferentes estratégias de *prompting.* Em Wang *et al*. (2023), os autores propõem o GPT-Ner, um modelo baseado no GPT-3 (Brown *et al.,* 2020) combinado com duas estratégias adicionais*: entity-level embedding*, que reforça a representação das entidades no texto, e *self-verification*, na qual o próprio modelo revisa e valida suas previsões. Essa combinação aumenta a robustez do REN e permitiu que o modelo obtivesse 90,91% de F1 no *benchmark* CoNLL-2003.

Com base na Figura 2, observa-se que a técnica *Instruction* (cf. Quadro 2) está presente no composto **Defn** (em azul), que guia o modelo na tarefa de REN, fornecendo os tipos de entidades que devem ser reconhecidas, e no componente **Q(uestion)-Paragraph** (em vermelho), que apresenta a tarefa propriamente dita, solicitando que o modelo identifique ENs no parágrafo fornecido e explique, para cada entrada, se é ou não uma entidade. O componente **Answer** (em verde), por sua vez, instancia as estratégias *Structured output*, *Demonstration-based e Chain-of-Thought* (CoT), uma vez que, respectivamente, fornece ao LLM o formato exato da saída, exemplos de como as respostas devem ser (*few-shot*) e o raciocínio passo a passo que justifica cada decisão do modelo. O PromptNER alcançou medida-F1 de 83,48% no CoNLL-2003. Uma versão traduzida do prompt se encontra no Apêndice B.

Figura 2. Exemplo de *prompt* no modelo PromptNER



Fonte: Ashok e Lipton (2023).

Em Wang et al. (2023), os autores propõem o GPT-Ner, um modelo baseado no GPT-3 (Brown *et al.,* 2020) combinado com duas estratégias adicionais*: entity-level embedding*, que reforça a representação das entidades no texto, e *self-verification*, na qual o próprio modelo revisa e valida suas previsões. Essa combinação aumentou a robustez do REN e permitiu que o modelo obtivesse 90,91% de F1 no *benchmark* CoNLL-2003.

Esses resultados mostram que métodos ICL-*based* podem realizar REN de forma competitiva, desde que combinados a *prompts* estruturados, seleção inteligente de exemplos e mecanismos de verificação, tornando-se uma alternativa prática aos modelos BERT-*like*.

Quanto ao uso de LLMs com técnicas *few-shot* ICL para REN em português, destacam-se os trabalhos que focam em domínios especializados, como o (i) jurídico, utilizando algum membro da família GPT (Coelho *et al.,* 2024) ou modelos de língua em português (Nunes *et al.* 2024), como o Sabiá (Pires *et al.,* 2023), e (ii) em textos históricos (Sarcinelli, 2025), utilizando modelo de pesos abertos como o LLaMA (Grattafiore *et al.,* 2024) e outros.

3. Metodologia

3.1. Seleção do corpus

Para investigar o emprego de LLMs e técnicas de ICL na tarefa de REN em *corpora* com escrita formal em português, selecionou-se um dos 3 *subcorpora* jornalísticos que compõem o *treebank* Porttinari (Duran *et al.,* 2023). Esse *treebank* foi desenvolvido pelo projeto POeTiSA (*POrtuguese processing - Towards Syntactic Analysis and parsing*) no âmbito da frente NLP2 (*Natural Language Processing for Portuguese*) do Centro de Inteligência Artificial (C4AI) da Universidade de São Paulo (USP). O objetivo do POeTiSA é avançar o estado-da-arte da análise sintática do português, desenvolvendo recursos de base, como *corpora* anotados (multigênero), e métodos ou modelos de *parsing*.

No que diz respeito à parcela do gênero jornalístico, o Porttinari engloba três *subcorpora*, que possuem diferentes dimensões e funcionalidades, a saber: (i) Porttinari-base, *corpus* revisado para servir como o padrão-ouro, (ii) Porttinari-check, um pequeno *corpus*, com uma estrutura similar ao Porttinari-base, que tem como função demonstrar os testes e diferenças entre as anotações manuais e automáticas e (iii) Porttinari-automatic, um grande *corpus* que foi anotado automaticamente com um parser treinado no Porttinari-base (Duran *et al.,* 2023). Como pode visto na Figura 3, o Porttinari-base é composto por 8.418 sentenças, as quais são retiradas do jornal Folha de São Paulo, totalizando 168.000 *tokens*. Ele está subdividido em três *subsets*: treinamento, desenvolvimento e teste. A quantidade de sentenças, de *tokens* e a média de *tokens* por sentença de cada *subset* podem ser observadas na Figura 3.

Figura 3. Dados estatísticos sobre subcorpora do Treebank Porttinari.

		Number of sentences	Number of tokens	Avg number of tokens per sentence
Porttinari-base	*Training (70%)*	5,893	117,972	19.97
	Development (10%)	842	16,528	
	Test (20%)	1,683	33,580	
	Overall	8,418	168,080	
Porttinari-check		1,685	33,576	19.93
Porttinari-automatic		3,954,218	94,444,424	23.88

Fonte: https://sites.google.com/icmc.usp.br/poetisa/porttinari-2-1.

O Portinari-base, segundo Duran *et al*. (2023), foi construído a partir dos 5.000 documentos iniciais que compõem o *dataset* Folha-Kaggle. Esse conjunto original de notícias da Folha de São Paulo, disponível no repositório Kaggle, possui temas diversos como economia, educação e política global. Para compor o Portinari-base, evitou-se selecionar (i) documentos inteiros, buscando garantir maior diversidade de autores e tópicos, e (ii) sentenças muito pequenas, uma vez que elas podem ter estrutura sintática muito simples; assim, privilegiou-se a seleção de sentenças com tamanho entre 10 e 40 *tokens*.

Para fomentar as pesquisas sobre análise sintática automática, o Porttinari-base possui anotação em nível morfológico e sintático (relações de dependências) segundo o modelo gramatical *Universal Dependencies* (UD) (Nivre *et al.,* 2020). Como mencionado, o Porttinari-base é um *corpus* padrão-ouro, o que significa que a anotação UD de todos os níveis foi manualmente revisada. Na Figura 4, ilustra-se a anotação-UD da sentença "O jornalista viajou a convite do Festival do Rio." do Porttinari-base*,* a qual está no formato CoNLL-U, constituído de 10 colunas, sendo cada uma delas destinada a uma informação específica[15].

Figura 4. Exemplo de anotação-UD de uma sentença do Porttinari-base.

Id	Form	Lemma	Upos Tag	Xpos Tag	Feats	Head	DepRel	Deps	Misc
1	O	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	2	det	_	_
2	jornalista	jornalista	NOUN	_	Number=Sing	3	nsubj	_	_
3	viajou	viajar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	_
4	a	a	ADP	_	_	5	case	_	_
5	convite	convite	NOUN	_	Gender=Masc|Number=Sing	3	obl	_	_
6-7	do	_	_	_	_	_	_	_	_
6	de	de	ADP	_	_	8	case	_	_
7	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	8	det	_	_
8	Festival	Festival	PROPN	_	_	5	nmod	_	_
9-10	do	_	_	_	_	_	_	_	_
9	de	de	ADP	_	_	11	case	_	_
10	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	11	det	_	_
11	Rio	Rio	PROPN	_	_	8	nmod	_	_
12	.	.	PUNCT	_	_	3	punct	_	_

Fonte: A autora (2025).

Cada linha corresponde a uma palavra ou *token* da sentença. A coluna 1 indica a ordem de ocorrência do *token* na sentença (ID) e a coluna 2 exibe o próprio *token* (FORM). As informações morfológicas são descritas nas colunas 3, 4, 5 e 6, sendo, respectivamente, lemas (LEMMA), etiquetas universais de *part-of-speech* (PoS) (classe de palavra) (UPOS), etiquetas

[15] https://universaldependencies.org/format.html

PoS específicas da língua[16] (XPOS) e traços gramaticais (FEATS). As informações sintáticas são indicadas nas colunas 7 e 8. A coluna 7 indica o *head* (cabeça) da relação (HEAD), sendo esta, de fato, rotulada na coluna 8 (DEPREL). A coluna 9 fica destinada às chamadas relações *enhanced* (ou enriquecidas)[17] (DEPS), que basicamente buscam explicitar certas relações implícitas entre as palavras. Por fim, a coluna 10 (MISC) pode exibir informações adicionais sobre os *tokens*.

3.2. Criação de uma anotação gold-standard

Após a seleção do *corpus* Porttinari-base, realizou-se uma anotação *gold-standard,* que serviu de referência para a avaliação de desempenho do emprego de LLM e técnicas de ICL. Para tanto, selecionaram-se de forma aleatória 471 sentenças (9.470 *tokens*) que compõem o Porttinari-base. Tal seleção foi feita em função do escopo e tempo de execução do TCC.

A anotação manual das ENs nessa parcela do *corpus* englobou apenas as categorias de PESSOA, LOCALIÇÃO e ORGANIZAÇÃO (PLO) da taxonomia do Segundo HAREM (Mota; Santos, 2008) (Figura 5), excluindo os tipos, e foi feita por apenas um anotador. A adoção das categorias PLO se deve ao fato de serem amplamente reconhecidas na literatura de REN e, por sua natureza abrangente e consolidada, contribuírem para reduzir tanto o escopo quanto a complexidade da anotação.

A anotação manual das ENs PLO na parcela de estudo do Porttinari-base foi feita basicamente de acordo com as diretrizes do Segundo HAREM (Mota; Santos, 2008). As diretrizes neste trabalho foram divididas em gerais e específicas.

A regra geral para identificação das ENs apoiou-se em critérios básicos de anotação: (i) a entidade deve incluir ao menos uma letra maiúscula, (ii) ser reconhecida preliminarmente pela *tag* PoS PROPN (nome próprio) da anotação UD e (iii) ser anotada em sua sequência mais ampla de *tokens*, desconsiderando entidades aninhadas (*nested entities*). Já para a classificação das ENs, este trabalho adotou diretrizes gerais que contemplam: (i) o uso de um esquema *single-label*, atribuindo uma única categoria por entidade; (ii) a consideração de fenômenos linguísticos como *polissemia*; e (iii) a atenção a casos de *metonímia*. Esses elementos orientaram a escolha da categoria mais adequada a cada ocorrência, sempre com base no

[16] Essas etiquetas não foram contempladas na anotação conduzida no POeTiSA.
[17] A anotação desse tipo de relação de dependência no Porttinati-base está em andamento.

contexto discursivo, assegurando assim uma classificação consistente e alinhada ao sentido efetivamente expresso no texto. As diretrizes específicas de anotação apresentam os critérios que qualificam uma entidade para cada categoria PLO. Estas estão detalhadas e exemplificadas no manual de anotação que foi seguido neste trabalho (Apêndice A).

Figura 5. Categorias do Segundo HAREM.

Categoria	Tipos	Subtipos
ABSTRAÇÃO (5)	DISCIPLINA	—
	ESTADO	—
	IDEIA	—
	NOME	—
	OUTRO	—
ACONTECIMENTO (4)	EFEMÉRIDE	—
	EVENTO	—
	ORGANIZADO	—
	OUTRO	—
COISA (5)	CLASSE	—
	MEMBROCLASSE	—
	OBJECTO	—
	SUBSTÂNCIA	—
	OUTRO	—
LOCAL (4)	FÍSICO (7)	ILHA, AGUACURSO, PLANETA, REGIÃO, RELEVO, AGUAMASSA, OUTRO
	HUMANO (6)	RUA, PAÍS, DIVISÃO, REGIÃO, CONSTRUÇÃO, OUTRO
	VIRTUAL (4)	COMSOCIAL, SÍTIO, OBRA, OUTRO
	OUTRO	—
OBRA (4)	ARTE	—
	PLANO	—
	REPRODUZIDA	—
	OUTRO	—
ORGANIZAÇÃO (4)	ADMINISTRAÇÃO	—
	EMPRESA	—
	INSTITUIÇÃO	—
	OUTRO	—
PESSOA (8)	CARGO	—
	GRUPOCARGO	—
	GRUPOIND	—
	GRUPOMEMBRO	—
	INDIVIDUAL	—
	MEMBRO	—
	POVO	—
	OUTRO	—
TEMPO (5)	DURAÇÃO	—
	FREQUÊNCIA	—
	GENÉRICO	—
	TEMPO_CALEND (4)	HORA, INTERVALO, DATA, OUTRO
	OUTRO	—
VALOR (4)	CLASSIFICAÇÃO	—
	MOEDA	—
	QUANTIDADE	—
	OUTRO	—
OUTRO (1)	OUTRO	—

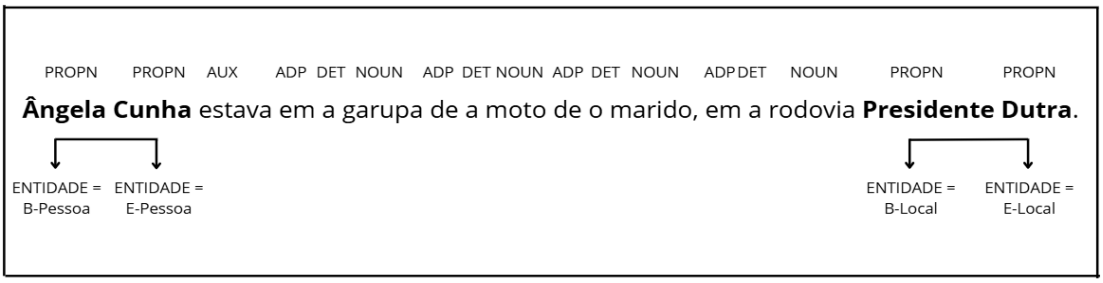
Fonte: Mota e Santos (2008).

Quanto ao esquema de anotação, adotou-se o BIOES. As Figuras 6 e 7 ilustram a anotação de ENs *single* e multipalavra nesse esquema.

Figura 6. Exemplo de anotaçao de ENs simples no modelo BIOES.

Fonte: A autora (2025).

Figura 7. Exemplo de anotação de ENs multipalavra no modelo BIOES.



Fonte: A autora (2025).

Ademais, destaca-se que anotação das ENs no *corpus* Porttinari-base foi feita nos arquivos no formato CoNLL-U, resultantes da anotação do *corpus* segundo o modelo gramatical UD. Para este trabalho, selecionou-se, como se observa na Figura 8, a 5ª coluna desse arquivo para a inserção das anotações de ENs, uma vez que a informação originalmente prevista para esta coluna (XPoS, isto é, etiquetas morfossintáticas específicas de uma língua) não está sendo usado pelo projeto POeTiSa, responsável pela anotação UD do *corpus*.

Figura 8. Exemplo da anotação de REN dentro do modelo UD no formato CoNLL-U.

# sent_id = FOLHA_DOC003104_SENT003									
# text = Em comunicado, o Hamas, que governa a Faixa de Gaza, afirmou que aceitou condições exigidas pelo Fatah para que um acordo entre os grupos se concretize.									
1	Em	em	ADP	_	_	2	mark	_	_
2	comunicado	comunicado	NOUN	_	Gender=Masc|Number=Sing	14	obl	_	SpaceAfter=No
3	,	,	PUNCT	_	_	2	punct	_	_
4	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	5	det	_	_
5	Hamas	Hamas	PROPN	S-Pessoa	_	14	nsubj	_	SpaceAfter=No
6	,	,	PUNCT	_	_	8	punct	_	_
7	que	que	PRON	_	PronType=Rel	8	nsubj	_	_
8	governa	governar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fir	5	acl:relcl	_	_
9	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	10	det	_	_
10	Faixa	Faixa	PROPN	B-Local	_	8	obj	_	_
11	de	de	ADP	I-Local	_	12	case	_	Proper=Yes
12	Gaza	Gaza	PROPN	E-Local	_	10	flat:name	_	SpaceAfter=No
13	,	,	PUNCT	_	_	8	punct	_	_
14	afirmou	afirmar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fir	0	root	_	_
15	que	que	SCONJ	_	_	16	mark	_	_
16	aceitou	aceitar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fir	14	ccomp	_	_
17	condições	condição	NOUN	_	Gender=Fem|Number=Plur	16	obj	_	_
18	exigidas	exigir	VERB	_	Gender=Fem|Number=Plur|VerbForm=Part|Voice=Pass	17	acl	_	_
19-20	pelo	_	_	_	_	_	_	_	_
19	por	por	ADP	_	_	21	case	_	_
20	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	21	det	_	_
21	Fatah	Fatah	PROPN	S-Organizaçã	_	18	obl:agent	_	_
22	para	para	ADP	_	_	30	mark	_	_
23	que	que	SCONJ	_	_	22	fixed	_	_
24	um	um	DET	_	Definite=Ind|Gender=Masc|Number=Sing|PronType=Art	25	det	_	_
25	acordo	acordo	NOUN	_	Gender=Masc|Number=Sing	30	nsubj	_	_
26	entre	entre	ADP	_	_	28	case	_	_
27	os	o	DET	_	Definite=Def|Gender=Masc|Number=Plur|PronType=Art	28	det	_	_
28	grupos	grupo	NOUN	_	Gender=Masc|Number=Plur	25	nmod	_	_
29	se	se	PRON	_	Case=Nom|Person=3|PronType=Prs	30	expl	_	_
30	concretize	concretizar	VERB	_	Mood=Sub|Number=Sing|Person=3|Tense=Pres|VerbForm=Fi	16	advcl	_	SpaceAfter=No
31	.	.	PUNCT	_	_	14	punct	_	SpaceAfter=No

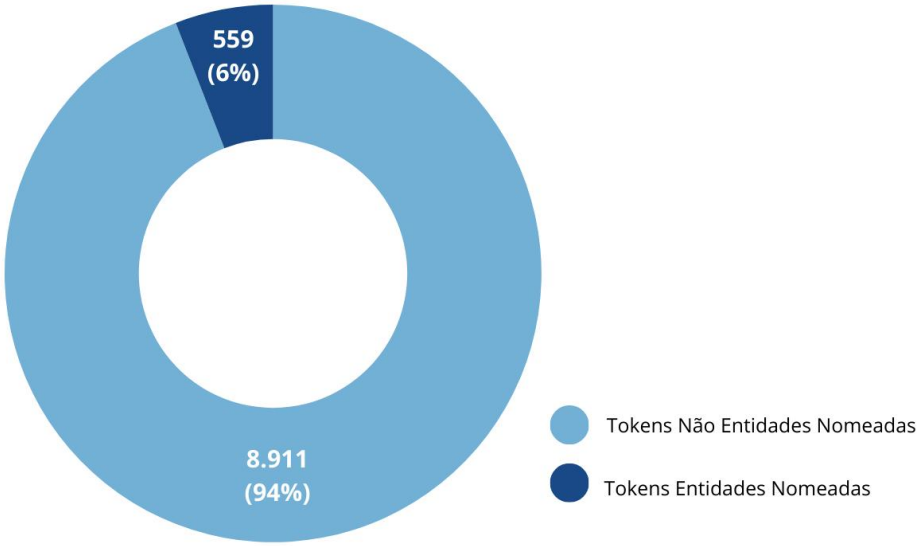
Fonte: A autora (2025).

O processo de anotação manual foi feito através de uma ferramenta chamada Visual Studio Code (VS Code[18]), selecionada por suportar extensões que otimizam a manipulação de arquivos CoNLL-U. Seus recursos conferem destaque de sintaxe e uma estruturação visual clara, assegurando a legibilidade e a precisão necessárias para a validação das etiquetas. A anotação das ENs se deu, portanto, considerando as diretrizes gerais e específicas e a posição da 5ª coluna de cada sentença do arquivo CONLL-U.

Sobre a anotação manual, destaca-se que, do total de 9.470 *tokens* que compõem a parcela de 471 sentenças, 559 deles são *tokens* classificados como ENs (6% do total) e 8,911 não correspondem a ENs (isto é, 94%), assim como ilustrado da Figura 9. Essa distribuição é esperada em tarefas como a de REN, em que a maior parte dos *tokens* do *corpus* não pertence a nenhuma categoria de entidade. Entre os 559 *tokens* de entidade, foram classificados 290 *tokens* pertencentes à categoria de PESSOA (52% de 559), 137 *tokens* pertencentes à categoria de ORGANIZAÇÃO (24,5% do total) e 132 *tokens* pertencentes à categoria de LOCAL (23,6%) (Figura 10).

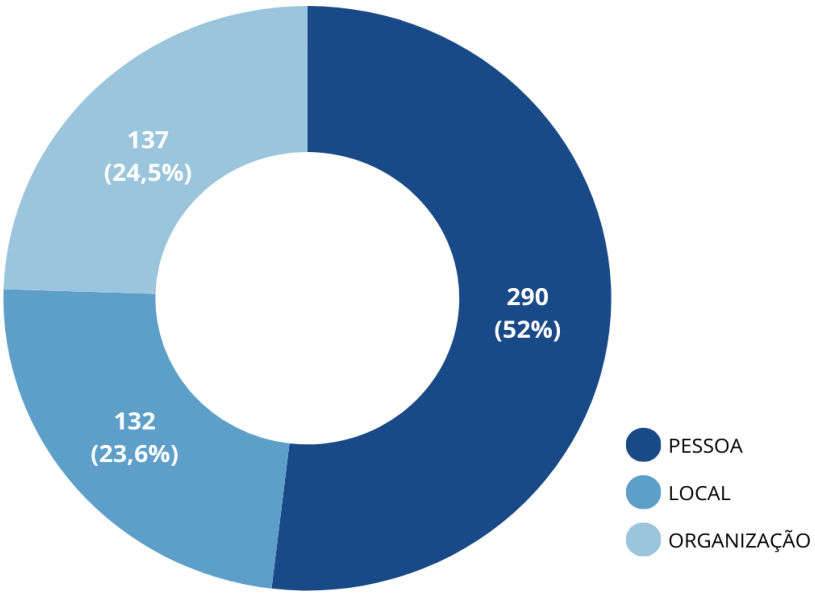
[18] https://code.visualstudio.com/

Figura 9. Estatística geral da classificação de tokens na anotação manual.



Fonte: A autora (2025)

Figura 10. Estatística da classificação de tokens por categoria na anotação manual.



Fonte: A autora (2025)

3.3. Seleção do LLM e das técnicas de prompting

O LLM selecionado como ferramenta principal para esta pesquisa foi o Llama 3.1, especialmente sua variante llama-3.1-8b-instant, modelo lançado pela Meta em julho de 2024.

A escolha desse LLM pautou-se em alguns critérios metodológicos fundamentais. Primeiramente, seu status como um modelo de código aberto (do inglês, *open source*) garante a transparência e reprodutibilidade, tendo em vista seu fácil acesso. Em segundo lugar, o Llama 3.1 é reconhecido por suas capacidades avançadas de raciocínio, que o habilitam a seguir e cumprir com uma grande margem de sucesso tarefas complexas. Por fim, o modelo possui um suporte multilíngue robusto, que o classifica como uma ferramenta adequada para o processamento de um *corpus* em língua portuguesa.

O modelo Llama 3.1[19] (llama-3.1-8b-instant) de 8 bilhões de parâmetros, apesar de ser o menor da família, mostrou-se o mais adequado para tarefa por se tratar de um modelo otimizado para inferência de alta velocidade e baixa latência, por isso, o sufixo *-instant* no nome.

A abordagem adotada para uso do modelo foi através de uma API (Interface de Programação de Aplicações, do inglês *Application Programming Interface*), que funciona como uma interface de comunicação padronizada. Ela permite que um *software* local envie requisições, os *prompts*, pela internet para um servidor remoto, que hospeda e executa o LLM, esse servidor processa a requisição e retorna a resposta, o texto gerado. O uso de uma API garante a completude do experimento sem a necessidade de uma infraestrutura computacional local e garante a validade dos resultados, pois as chamadas feitas para o modelo são executadas na versão oficial e otimizada.

Entre as plataformas de API disponíveis, selecionou-se a plataforma Groq[20], que disponibiliza o acesso ao Llama 3.1 8B de maneira gratuita, além de oferecer um limite de requisição extenso o suficiente para a realização de pesquisas aprofundadas. Além desses benefícios, a adoção da Groq permitiu que o modelo obtivesse uma excelente performance graças ao seu *hardware* proprietário (LPU), que oferece uma velocidade altíssima de processamento e geração de *tokens*.

Na sequência, outra importante definição metodológica foi a escolha das técnicas de ICL. Para especializar o modelo à tarefa de REN para um *corpus* do gênero jornalístico em língua portuguesa, selecionaram-se as duas técnicas de ICL mais difundidas no PLN e em REN, a saber: z*ero-shot* e *few-shot prompting*. Cada uma delas foi empregada em conjunto a outras técnicas, diferentes das já empregadas na literatura, a saber: (i) "*prompting* baseado em

[19] https://ai.meta.com/blog/meta-llama-3-1/
[20] https://groq.com/

persona" (em inglês, *persona* ou *role-playing prompting*), (ii) "*prompting* baseado em regra" (do inglês, *rule-based prompting*) e (iii) "formatação estruturada de saída" (inglês, *structured output* ou *output formatting*). As traduções dos termos em inglês estão entre aspas porque ainda não estão consolidadas na literatura, sendo, portando, traduções livres.

A técnica *persona prompting* (Shanahan *et al.,* 2023) é da categoria *Instruction-based*., consistindo em instruir o modelo a assumir um papel, ativando o conhecimento latente do modelo que será relevante para a tarefa e melhorando a aderência às instruções. Outra técnica baseada em instruções, denominada *rule-based prompting* (Zhang *et al.,* 2025), consiste em construir conjuntos explícitos de regras e diretrizes que orientam o comportamento do modelo durante a realização da tarefa. Essa abordagem busca reduzir alucinações e assegurar que as respostas do modelo permaneçam alinhadas à lógica e às restrições específicas do problema.

Já a técnica *structured output* (Cheng *et al.,* 2024) constitui uma forma específica de restrição que define rigorosamente a estrutura que a resposta do modelo deve seguir. Essa técnica é essencial para garantir que o modelo apresente os resultados no formato desejado, evitando variações ou organizações indesejadas na saída.

4. Exploração de estratégias de ICL

O emprego do Llama 3.1 e das técnicas de ICL combinadas foi realizado em dois cenários: *zero-shot* e *few-shot*. Para tanto, fez-se necessário inicialmente elaborar ou construir o *prompt* para cada um dos cenários.

4.1. Zero-shot com "persona+rules+output"

A técnica *zero-shot* foi aplicada em conjunto com *persona prompting*, *rule-based prompting* e *structured output formatting*, de modo que a tarefa de REN é executada exclusivamente a partir do conjunto de instruções fornecidas no *prompt*, sem o uso de exemplos adicionais. A Figura 11 ilustra o *prompt zero-shot* com as demais técnicas de ICL. Nele, vê-se 4 componentes, cada um deles orientando o modelo de forma específica:

1. *Persona prompting* (azul): o modelo assume a persona de um especialista em REN, direcionando quais conteúdos e informações do pré-treinamento devem ser mobilizados para gerar respostas alinhadas à tarefa.

2. *Output Formatting* (primeira parte) (vermelho): essa instrução inicial de formatação estabelece restrições globais sobre a saída antes da aplicação das regras detalhadas, garantindo que o modelo compreenda desde o início o formato esperado das respostas.
3. *Rule-based prompting* (verde): inclui diretrizes explícitas que delimitam os aspectos centrais da anotação; as regras definem a taxonomia de ENs (PESSOA, LOCAL e ORGANIZAÇÃO), o esquema de anotação (BIOES), a estrutura da resposta (uso de hífen e proibição de abreviação das categorias) e o formato de saída (uso exclusivo de aspas retas e correspondência entre tuplas de saída e entrada); essas regras mantêm consistência na anotação e evitam problemas no pós-processamento e na avaliação.
4. *Output Formatting* (segunda parte) (vermelho): funciona como gatilho de execução, sinalizando ao modelo a conclusão da geração do texto; essa segmentação final organiza e estrutura a resposta dentro dos parâmetros especificados, aumentando a precisão e a confiabilidade dos resultados.

Figura 11. *Prompt zero-shot* com técnicas *persona, ruled-based* e *output formatting*.

Persona prompting
Você é um especialista em linguística computacional e precisa realizar a tarefa de REN (reconhecimento de entidades nomeadas) em um corpus de língua portuguesa.

Output Formatting
Sua saída deve ser exclusivamente uma lista de tuplas Python no formato [('token', 'tag')]. Não inclua nenhuma outra explicação ou texto.

Rule-based prompting
As entidades devem ser anotadas para as categorias de Pessoa, Local e Organização, seguindo rigorosamente estas diretrizes:
1. **Esquema de Anotação:** Use o padrão BIOES (`B-`, `I-`, `E-`, `S-`). O separador deve ser exclusivamente um hífen (`-`).
2. **Tokens Não-Entidade:** Para tokens que não são entidades, use a string `'_'`.
3. **Formato da String:** Use apenas aspas retas (`'`) dentro das tuplas.
4. **Rótulos Proibidos:** Nunca use as abreviações em inglês `PER`, `LOC`, ou `ORG`. Use sempre os nomes completos em português: `Pessoa`, `Local`, `Organização`.
5. **Restrição:** A sua resposta DEVE conter exatamente o mesmo número de tuplas que o número de tokens na lista **Tokens de Entrada**. Você não deve inserir tokens que não estavam na entrada, nem omitir tokens que estavam nela. A lista de tokens da sua saída deve ser idêntica à lista de Tokens de Entrada.

Output Formatting
TAREFA:
Tokens de Entrada: {lista_de_tokens_a_ser_analisada}
Saída:

Fonte: A autora (2025).

4.2. Few-shot com "persona+rules+output"

Além dos componentes instrucionais do *zero-shot,* esse *prompt* contém outro componente com exemplos de anotação de REN, os quais permitem que o modelo aprenda a realizar a tarefa a partir de analogias. A Figura 12 ilustra o *prompt* em questão, especialmente o componente *few-*

shot prompting (roxo). Para a confecção desse *prompt,* selecionaram-se 4 sentenças provenientes da parcela de 471 sentenças anotadas do Porttinari-base. A seleção ocorreu de forma estratégica, garantindo que as sentenças-exemplo incluíssem ENs das três categorias (PESSOA, LOCAL e ORGANIZAÇÃO) e contemplassem as diferentes expressões linguísticas (*single* e multipalavras), de forma que o aprendizado não ocorresse de maneira enviesada.

Figura 12. *Prompt few-shot* com técnicas *persona, ruled-based* e *output formatting*.

Persona prompting
Você é um especialista em linguística computacional e precisa realizar a tarefa de REN (reconhecimento de entidades nomeadas) em um corpus de língua portuguesa.

Output Formatting
Sua saída deve ser exclusivamente uma lista de tuplas Python no formato [('token', 'tag')]. Não inclua nenhuma outra explicação ou texto.

Rule-based prompting
As entidades devem ser anotadas para as categorias de Pessoa, Local e Organização, seguindo rigorosamente estas diretrizes:
1. **Esquema de Anotação:** Use o padrão BIOES (`B-`, `I-`, `E-`, `S-`). O separador deve ser exclusivamente um hífen (`-`).
2. **Tokens Não-Entidade:** Para tokens que não são entidades, use a string `'_'`.
3. **Formato da String:** Use apenas aspas retas (`'`) dentro das tuplas.
4. **Rótulos Proibidos:** Nunca use as abreviações em inglês `PER`, `LOC`, ou `ORG`. Use sempre os nomes completos em português: `Pessoa`, `Local`, `Organização`.
5. **Restrição:** A sua resposta DEVE conter exatamente o mesmo número de tuplas que o número de tokens na lista **Tokens de Entrada**. Você não deve inserir tokens que não estavam na entrada, nem omitir tokens que estavam nela. A lista de tokens da sua saída deve ser idêntica à lista de Tokens de Entrada.

Few-shot prompting
Observe os exemplos perfeitos abaixo:
Exemplo 1:
Tokens de Entrada: ['Em', 'comunicado', ',', 'o', 'Hamas', ',', 'que', 'governa', 'a', 'Faixa', 'de', 'Gaza', ',', 'afirmou', 'que', 'aceitou', 'condições', 'exigidas', 'por', 'o', 'Fatah', 'para', 'que', 'um', 'acordo', 'entre', 'os', 'grupos', 'se', 'concretize', '.']
Saída: [('Em', '_'), ('comunicado', '_'), (',', '_'), ('o', '_'), ('Hamas', 'S-Pessoa'), (',', '_'), ('que', '_'), ('governa', '_'), ('a', '_'), ('Faixa', 'B-Local'), ('de', 'I-Local'), ('Gaza', 'E-Local'), (',', '_'), ('afirmou', '_'), ('que', '_'), ('aceitou', '_'), ('condições', '_'), ('exigidas', '_'), ('por', '_'), ('o', '_'), ('Fatah', 'S-Organização'), ('para', '_'), ('que', '_'), ('um', '_'), ('acordo', '_'), ('entre', '_'), ('os', '_'), ('grupos', '_'), ('se', '_'), ('concretize', '_'), ('.', '_')]

Exemplo 2:
Tokens de Entrada: ['Em', 'Miami', ',', 'a', 'brasileira', 'Miriam', 'Friedman', ',', '65', 'diz', 'estar', 'acostumada', 'com', 'a', 'temporada', 'de', 'os', 'furacões', '.']
Saída: [('Em', '_'), ('Miami', 'S-Local'), (',', '_'), ('a', '_'), ('brasileira', '_'), ('Miriam', 'B-Pessoa'), ('Friedman', 'E-Pessoa'), (',', '_'), ('65', '_'), ('diz', '_'), ('estar', '_'), ('acostumada', '_'), ('com', '_'), ('a', '_'), ('temporada', '_'), ('de', '_'), ('os', '_'), ('furacões', '_'), ('.', '_')]

Exemplo 3:
Tokens de Entrada: ['A', 'Polícia', 'Federal', 'monitorou', 'em', 'maio', 'um', 'encontro', 'de', 'ambos', 'com', 'Francisco', 'Assis', 'e', 'Silva', ',', 'advogado', 'e', 'delator', 'de', 'a', 'empresa', '.']
Saída: [('A', '_'), ('Polícia', 'B-Organização'), ('Federal', 'E-Organização'), ('monitorou', '_'), ('em', '_'), ('maio', '_'), ('um', '_'), ('encontro', '_'), ('de', '_'), ('ambos', '_'), ('com', '_'), ('Francisco', 'B-Pessoa'), ('Assis', 'I-Pessoa'), ('e', 'I-Pessoa'), ('Silva', 'E-Pessoa'), (',', '_'), ('advogado', '_'), ('e', '_'), ('delator', '_'), ('de', '_'), ('a', '_'), ('empresa', '_'), ('.', '_')]

Exemplo 4:
Tokens de Entrada: ['Em', 'a', 'vizinhança', ',', 'ficam', 'também', 'a', 'Catedral', 'de', 'St.', 'Patrick', 'e', 'o', 'Rockfeller', 'Center', '.']
Saída: [('Em', '_'), ('a', '_'), ('vizinhança', '_'), (',', '_'), ('ficam', '_'), ('também', '_'), ('a', '_'), ('Catedral', 'B-Local'), ('de', 'I-Local'), ('St.', 'I-Local'), ('Patrick', 'E-Local'), ('e', '_'), ('o', '_'), ('Rockfeller', 'B-Local'), ('Center', 'E-Local'), ('.', '_')]

Output Formatting
TAREFA:
Tokens de Entrada: {lista_de_tokens_a_ser_analisada}
Saída:

Fonte: A autora (2025).

Com base na Figura 12, observa-se que o componente *"few-shot prompting*" foi inserido após as diretrizes gerais (*persona prompting*, *output formatting* (primeira parte) e *rule-based prompting*). Essa decisão foi tomada pautando-se na hipótese de que essa ordem faz o modelo "ancorar" o comportamento nas regras explícitas (o que anotar, rótulos permitidos, formato

BIOES etc.) e só então interpretar os exemplos como demonstrações dessas regras em uso, e não como o único padrão possível. Assim, o LLM aprende o procedimento a partir das instruções e o concretiza com os exemplos, o que tende a gerar uma aprendizagem em contexto mais estável e robusta para REN.

4.3. Fluxo de anotação de REN com LLM e técnicas de ICL

Uma vez que os *prompts* haviam sido construídos, o processo de aplicação foi organizado em um fluxo de processamento em Python, no ambiente Google Colab, composto por quatro etapas: (i) preparação dos dados do corpus; (ii) submissão das sentenças via prompt — que, no cenário *few-shot*, incluía exemplos provenientes da anotação manual do Porttinari-base; (iii) armazenamento e pós-processamento das respostas do modelo; e (iv) cálculo das métricas e geração dos gráficos de avaliação.

Na etapa (i), elaborou-se um *script* em Python responsável por extrair e formatar as informações pertinentes do corpus anotado, disponibilizado no formato CoNLL-U. As 471 sentenças anotadas manualmente tiveram seus *tokens* (segunda coluna dos arquivos CoNLL-U) coletados e organizados em listas representando as sentenças de entrada. Cada lista foi vinculada à variável do *prompt* {lista_de_*tokens*_a_ser_analisado}.

Na etapa (ii), um laço de repetição automatizado atualizava os *prompts zero-shot* e *few-shot* com a lista de *tokens* correspondente a cada sentença. O *prompt* completo era então submetido ao modelo, via API, produzindo dois conjuntos de saídas, um para cada cenário experimental. Testes preliminares evidenciaram limitações da API quando o processamento ocorria de forma contínua. Por isso, implementaram-se mecanismos de robustez, como o processamento em lotes de 25 sentenças, com salvamento automático das respostas, e um sistema de espera em caso de falha, garantindo a estabilidade do fluxo.

Na etapa (iii), as respostas retornadas pela API (listas de tuplas no formato especificado) passaram por um *parser* de conversão. Esse processamento resultou em arquivos CoNLL-U, análogos aos originais do *corpus*, contendo as etiquetas de REN previstas pelo modelo.

Por fim, na etapa (iv), os arquivos gerados pelos *prompts* nos dois cenários foram comparados às anotações manuais, o que permitiu calcular as métricas e produzir as matrizes de confusão para análise do desempenho dos modelos na tarefa de REN.

As métricas de avaliação utilizadas foram precisão, revocação e medida-F1, complementadas pela matriz de confusão, que permite visualizar, de forma detalhada, a distribuição dos acertos e dos erros do modelo ao comparar as categorias previstas com as categorias reais. O cálculo dessas métricas foi realizado com a biblioteca seqeval, em Python, utilizando o esquema BIOES. Como o corpus Porttinari-base emprega apenas o rótulo "_" para marcar tokens que não correspondem a entidades, foi necessário aplicar uma normalização prévia, garantindo a compatibilidade entre os formatos e evitando inconsistências no cálculo das métricas e na geração dos gráficos.

5. Resultados e discussão

Após descrição metodológica e detalhamento do fluxo de processamento nas seções anteriores, esta seção é dedicada à apresentação dos resultados obtidos nos cenários *zero-shot* e *few-shot,* ambos em combinação com as outras técnicas de ICL já mencionadas. A seguir, procede-se à análise dos dados quantitativos, com base nas métricas de desempenho (precisão, revocação e medida-F1), e dos dados qualitativos, por meio da inspeção das matrizes de confusão.

O Quadro 3 apresenta as medidas de desempenho no cenário *zero-shot*, acompanhadas de suas médias ponderadas. A média de 37,7% de medida-F1 — derivada de 27,4% de precisão e 62,9% de revocação — evidencia que o Llama 3.1 apresenta desempenho limitado para REN nesse cenário, mesmo quando combinado às técnicas de ICL adotadas. Esse resultado contrasta fortemente com o estado-da-arte para o português, como o BERTimbau (78,6% de F1), e com métodos baseados em LLM aliados a técnicas mais avançadas de ICL para o inglês, como PromptNER (83,48%) e GPT-NER (90,91%).

Quadro 3. Métricas de desempenho do Llama no cenário *zero-shot* com técnicas de ICL.

Categoria	Precisão	Revocação	Medida-F1
PESSOA	30,6%	79,0%	44,1%
LOCAL	28,6%	42,4%	34,1%
ORGANIZAÇÃO	19,5%	48,9%	27,6%
Média	27,4%	62,9%	37,7%

Fonte: A autora (2025).

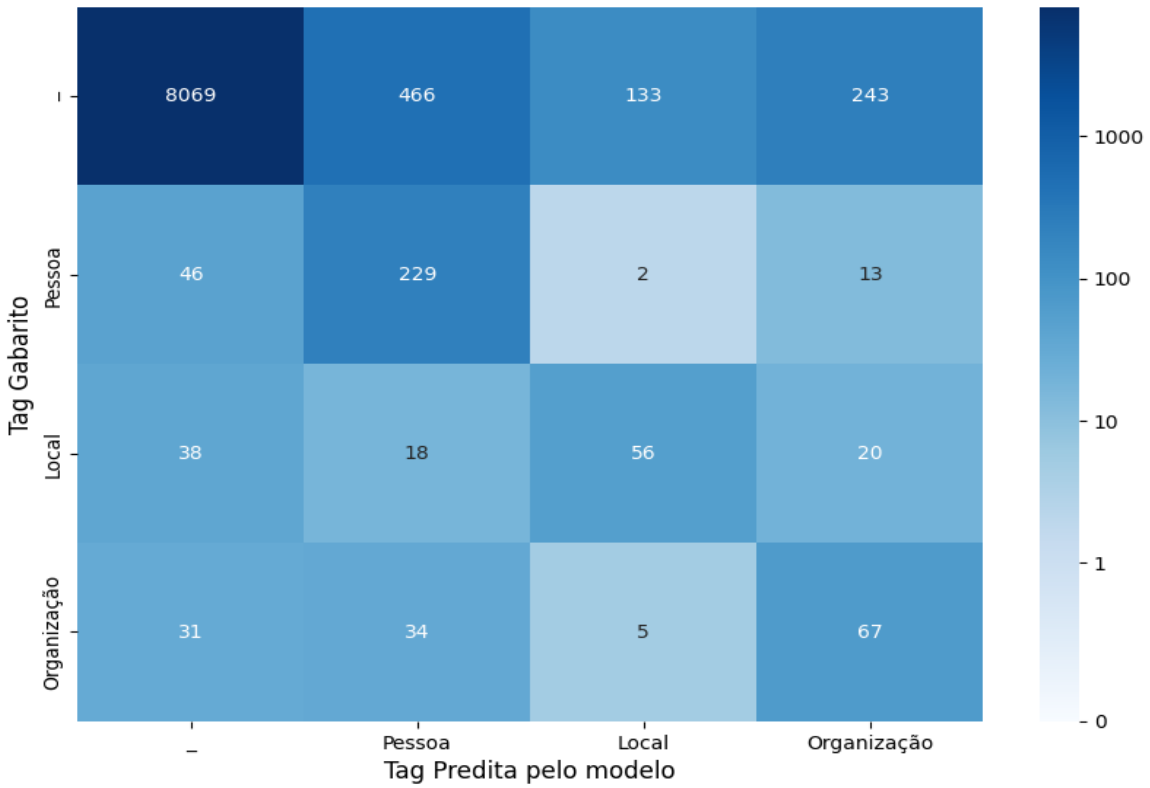
Analisando os resultados por categoria, vê-se que PESSOA apresentou o melhor desempenho, alcançando medida-F de 44,1%. Em seguida, aparece LOCAL, com 34,1%, e ORGANIZAÇÃO, com 27,6%. A categoria PESSOA apresenta a maior discrepância entre

precisão (30,6%) e revocação (79,0%). Esse comportamento sugere que o modelo tende a superclassificar *tokens* como PESSOA, aplicando o rótulo com baixa criticidade e pouca ancoragem contextual, resultando em uma identificação pouco seletiva dessa categoria.

A Figura 13 apresenta a matriz de confusão resultante da comparação entre as *tags* de entidades nomeadas da anotação de referência e as *tags* atribuídas pelo modelo. As categorias avaliadas incluem PESSOA, LOCAL, ORGANIZAÇÃO e não-entidade (rótulo "–").

Observa-se com base na primeira linha da Figura 13 que, dos 8,911 *tokens* que não são ENs (cf. Figura 9), o modelo identifica corretamente 8,069 (isto é, "–" → "–"), o que equivale a 90,5% do total. Assim, a matriz evidencia que o modelo identifica com alta precisão *tokens* que não são ENs. Isso pode ser explicado pelo fato de que, como a maior parte dos *tokens* em textos naturais não pertence a categorias de EN, o modelo tende a aprender padrões fortes e estáveis associados à classe "não-EN".

Figura 13. Matriz de confusão do cenário *zero-shot* com técnicas de ICL.



Fonte: A autora (2025).

A maior confusão se dá com a categoria PESSOA, posto que 466 *tokens* que não correspondem a ENs foram classificados como tal, frente a 243 como ORGANIZAÇÃO e 133 como LOCAL. A maior confusão com PESSOA pode estar associada a fenômenos linguísticos frequentes do domínio jornalístico, como a metonímia e a personificação de instituições. Esses sinais levam o modelo a superestimar a presença de entidades do tipo PESSOA, gerando mais falsos positivos do que nas demais categorias.

No que se refere à categoria PESSOA, o modelo apresenta 229 acertos (isto é, PESSOA → PESSOA) de um total de 290 ocorrências da anotação manual. Apresenta poucos casos de confusão com as outras duas categorias. A confusão mais frequente é com a categoria de "não-entidade" (PESSOA → "–", com 46 ocorrências), enquanto confusões com ORGANIZAÇÃO e LOCAL são mais raras, com 13 e 2 casos, respectivamente.

No que se refere à categoria LOCAL, o modelo apresenta 56 acertos (LOCAL → LOCAL) de um total de 132 casos da anotação manual. Observa-se uma dispersão mais ampla de erros, com confusões envolvendo "não-entidade" (38 ocorrências), ORGANIZAÇÃO (20) e PESSOA (18). Esse comportamento indica que o modelo encontra maior dificuldade em delinear fronteiras precisas entre referências a locais e outras categorias de entidades.

Quanto à ORGANIZAÇÃO, o modelo apresenta 67 acertos (ORGANIZAÇÃO → ORGANIZAÇÃO) de um total de 137 ocorrências da anotação manual. Essa categoria apresenta taxas relativamente mais altas de confusão, sobretudo com PESSOA (34 ocorrências) e "não entidade" (31 ocorrências), além de alguns casos com LOCAL (5). Tal comportamento pode estar associado à heterogeneidade lexical das entidades nomeadas da categoria ORGANIZAÇÃO. Diferentemente das categorias de PESSOA e LOCAL, que possuem padrões linguísticos mais regulares, as ENs de ORGANIZAÇÃO manifestam-se através de formas variadas, desde siglas até sintagmas nominais complexos e extensos. Essa variabilidade estrutural impede que o modelo aprenda com clareza as regras de delimitação necessárias para classificar corretamente a entidade

A baixa performance do modelo Llama 3.1 no cenário *zero-shot* com as técnicas de ICL selecionadas não constitui surpresa. Diz-se isso porque não se espera que um LLM alcance bons resultados em REN sem qualquer exemplo. Em tarefas sensíveis ao contexto, como REN, os LLMs dependem de pistas explícitas sobre o tipo de entidade e os limites de anotação. A ausência de exemplos reduz drasticamente a capacidade do modelo, mesmo quando o ICL é

empregado com instruções e formatação de saída, que, ao que parece, não fornecem informações suficientes para alinhar o comportamento do LLM ao domínio-alvo.

Apesar da baixa performance já esperada, a investigação de REN no cenário *zero-shot* é relevante, pois: (i) estabelece um *baseline* para quantificar o impacto das técnicas de ICL no cenário *few-shot*; (ii) permite avaliar a capacidade instrucional "pura" do modelo, ou seja, aquilo que ele é capaz de realizar sem exemplos; (iii) evidencia o grau de dependência da tarefa em relação ao domínio, auxiliando na escolha de estratégias de adaptação; e (iv) reflete situações práticas em que não há exemplos anotados disponíveis, verificando a viabilidade do uso direto do LLM.

Em seguida, apresenta-se o Quadro 4 que sintetiza as medidas de precisão, revocação e medida-F1 para cada categoria de entidade no cenário *few-shot*, além da média ponderada de cada uma delas. A média ponderada de 66% de medida-F1 no cenário *few-shot* representa um avanço expressivo em relação ao desempenho *zero-shot* (37,7%). Embora não alcance o estado-da-arte para o português nem mesmo as abordagens baseadas em LLM combinadas com técnicas mais sofisticadas de ICL, o valor de 66% é claramente mais competitivo, colocando o modelo em um patamar intermediário dentro do cenário de REN.

Assim como no cenário *zero-shot*, a hierarquia de desempenho entre as categorias se mantém: PESSOA apresenta a maior medida-F, seguida de perto por LOCAL, e, por fim, ORGANIZAÇÃO. Além disso, vale ressaltar que a média 66% de medida-F pode ser considerada um valor competitivo ao do estado-da-arte, com a vantagem de ter sido atingido sem pré-treinamento em larga escala.

Quadro 4. Métricas de desempenho do Llama no cenário *few-shot* com técnicas de ICL.

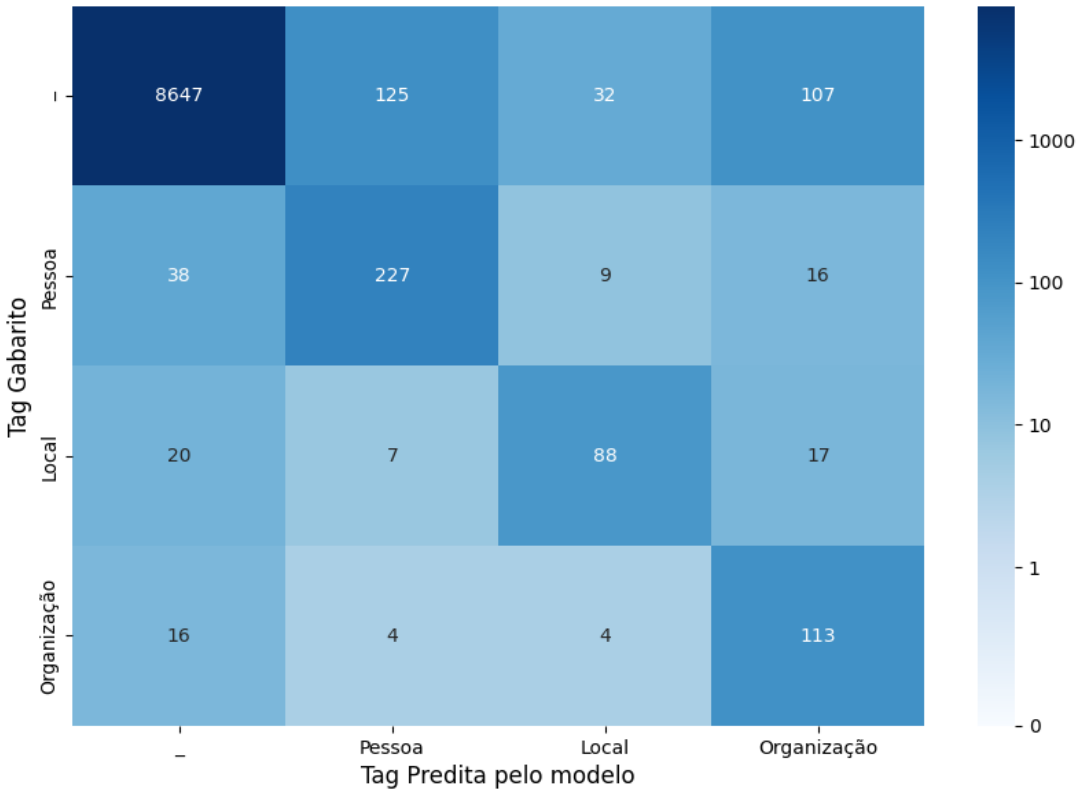
Categoria	Precisão	Revocação	Medida-F1
PESSOA	62,5%	78,3%	69,5%
LOCAL	66,2%	66,7%	66,4%
ORGANIZAÇÃO	44,7%	82,5%	57,9%
Média	59,0%	76,5%	66,0%

Fonte: A autora (2025).

Na Figura 14, tem-se a matriz de confusão do cenário *few-shot* com técnicas de ICL. Nela, destaca-se a redução de *tokens* "não-entidades" incorretamente classificados como ENs em relação ao cenário anterior. Sobre isso, observa-se que a quantidade de *tokens* não-entidades

("–") classificados como PESSOA caiu de 466 para 125, enquanto as classificações incorretas para ORGANIZAÇÃO e LOCAL caíram de 243 para 107 e de 133 para 32, respectivamente.

Figura 14. Matriz de confusão no cenário *few-shot* com técnicas de ICL.



Fonte: A autora (2025).

No que se refere à categoria PESSOA, destaca-se a medida-F1 de 69,5% frente a de 44,1% do cenário *zero-shot.* Esse valor mais elevado da medida-F é principalmente resultante da melhora na precisão por meio da qual o LLM identifica as ENs da categoria PESSOA. A precisão de PESSOA aumentou de 30,6% para 62,5% porque, no cenário *few-shot,* o modelo reduziu de forma significativa o número de falsos positivos atribuídos à categoria, sobretudo aqueles provenientes de *tokens* não-entidade. Como os verdadeiros positivos permanecem praticamente constantes entre os dois cenários, a melhora na precisão decorre essencialmente dessa queda acentuada nas classificações indevidas como PESSOA.

Quanto à categoria LOCAL, observa-se o melhor equilíbrio entre precisão (66,2%) e revocação (66,7%) das três categorias, resultando em uma medida-F de 66,4%, bem acima do valor de 34,1% do cenário *zero-shot*. A categoria LOCAL, aliás, apresenta o maior aumento absoluto de medida-F entre os cenários *zero-* e *few-shot*, com um ganho de 32,3%, superando

tanto PESSOA quanto ORGANIZAÇÃO. Os valores referentes a LOCAL nesse segundo cenário mostram que o modelo consolidou critérios estáveis para a classificação dessa categoria, evitando a atribuição excessiva de etiquetas LOCAL a *tokens* que não pertencem a essa classe (32 casos) e aprimorando sua capacidade de recuperar ocorrências efetivamente anotadas como LOCAL na referência (88 de 132). Os dados da matriz de confusão (Figura 14) revelam, entretanto, que a distinção entre LOCAL e ORGANIZAÇÃO continua sendo um ponto sensível para o modelo, embora tenha havido melhora no cenário *few-shot*, já que a confusão passou de 20 para 17 casos.

Sobre ORGANIZAÇÃO, destaca-se que também houve uma melhora significativa de mais de 30% do primeiro cenário para o segundo, pois a média obtida para a medida-F1 passou de 27,5% para 57,9%. Entretanto, o desempenho do modelo para essa categoria demonstra a maior assimetria entre precisão (44,7%) e revocação (82,5%). A revocação alta mostra que ele recupera a maior parte das instâncias anotadas manualmente (113 de 137 casos), demonstrando boa sensibilidade à categoria. No entanto, a precisão relativamente baixa revela que muitos *tokens* foram rotulados como ORGANIZAÇÃO sem que de fato pertençam à categoria, totalizando 140 predições de ORGANIZAÇÃO, das quais apenas 113 eram corretas.

Ressalta-se, ao final, que a melhora no desempenho do Llama 3.1 (com as referidas técnicas de ICL) para as tarefas de identificação e classificação das ENs na parcela do *corpus* Porttinari-base parece ser resultante do emprego da técnica *few-shot prompting*, de modo que as sentenças-exemplo foram importantes para o aumento dos valores atingidos por cada categoria.

A análise comparativa entre os cenários *zero-shot* e *few-shot* indica que a inclusão de exemplos no *prompt* foi decisiva para mitigar a tendência de supergeneralização nas anotações. No contexto *zero-shot*, observou-se atribuição indevida (mais frequente) de etiquetas de EN a *tokens* "não-entidade", sendo que, no cenário *few-shot*, o modelo exibiu uma filtragem mais precisa. Isso pode ser ilustrado com o exemplo da Figura 15. Nela, observa-se que os *tokens* "Ele", "sanções", "políticos" e "plebiscito", os quais não foram anotados como EN na anotação de referência, receberam erroneamente etiquetas de EN no cenário *zero-shot,* mas não no cenário *few-shot.*

Figura 15. Redução da superanotação de “não-entidades” no cenário *few-shot*.

Ele, porém, não mencionou que sanções serão aplicadas aos políticos que impulsionam o plebiscito.			
Tokens	**Anotação referência**	**Anotação Zero-shot**	**Anotação Few-shot**
Ele	_	S-Pessoa	_
,	_	_	_
porém	_	_	_
,	_	_	_
não	_	_	_
mencionou	_	_	_
que	_	_	_
sanções	_	S-Organização	_
serão	_	_	_
aplicadas	_	_	_
aos	_	_	_
a	_	_	_
os	_	_	_
políticos	_	S-Pessoa	_
que	_	_	_
impulsionam	_	_	_
o	_	_	_
plebiscito	_	S-Local	_
.	_	_	_

Fonte: A autora (2025)

Essa calibração mostrou-se particularmente eficaz para as categorias PESSOA e LOCAL, que apresentaram redução substancial de ruído e melhora consistente no desempenho. Ainda assim, a categoria ORGANIZAÇÃO permanece como a mais problemática, refletindo sua maior variabilidade lexical e fronteiras menos estáveis (cf. Figura 16).

Figura 16. Superanotação de “não-entidades” como ORG nos dois cenários.

A minha família inteira, que é bastante humilde, sempre trabalhou na indústria.			
Tokens	**Anotação referência**	**Anotação Zero-shot**	**Anotação Few-shot**
A	_	S-Pessoa	_
minha	_	I-Pessoa	_
família	_	B-Organização	B-Organização
inteira	_	I-Organização	E-Organização
,	_	_	_
que	_	_	_
é	_	_	_
bastante	_	_	_
humilde	_	_	_
,	_	_	_
sempre	_	_	_
trabalhou	_	_	_
na	_	_	_
em	_	_	_
a	_	_	_
indústria	_	B-Organização	_
.	_	_	_

Fonte: A autora (2025)

A Figura 16 ilustra bem esse comportamento. Embora no cenário *few-shot* o modelo tenha corrigido a superanotação observada no *zero-shot*, reconhecendo corretamente que "A minha" não constitui uma entidade da categoria PESSOA, a expressão "família inteira" continua sendo rotulada indevidamente como ORGANIZAÇÃO em ambos os cenários. Isso evidencia que, mesmo com exemplos adicionais, o modelo ainda tende a superestender essa etiqueta para sintagmas nominais genéricos, o que reforça a dificuldade associada à categoria ORGANIZAÇÃO.

No que se refere à dificuldade de distinção entre LOCAL e ORGANIZAÇÃO, nota-se, na Figura 17, que, embora o modelo tenha acertado a classificação de "Catedral de St. Patrick" no cenário *few-shot*, a anotação incorreta de "Rockefeller Center" como ORGANIZAÇÃO se manteve, evidenciando a dificuldade do modelo em distinguir o espaço físico de suas instituições.

Figura 17. Confusão entre LOCAL e ORGANIZAÇÃO no cenário *few-shot*.

Na vizinhança, ficam também a Catedral de St. Patrick e o Rockefeller Center.			
Tokens	**Anotação referência**	**Anotação Zero-shot**	**Anotação Few-shot**
Na	_	_	_
Em	_	_	_
a	_	_	_
vizinhança	_	_	_
,	_	_	_
ficam	_	_	_
também	_	_	_
a	_	B-Pessoa	_
Catedral	B-Local	B-Local	B-Local
de	I-Local	_	I-Local
St.	I-Local	S-Pessoa	I-Local
Patrick	E-Local	E-Pessoa	E-Local
e	_	_	_
o	_	_	_
Rockefeller	B-Local	B-Organização	B-Organização
Center	E-Local	E-Organização	E-Organização
.	_	_	_

Fonte: A autora (2025)

Já na Figura 18, observa-se a dificuldade do modelo em delimitar ou segmentar as ENs, principalmente da categoria ORGANIZAÇÃO. Diz-se isso porque, apesar do modelo *few-shot* ter corrigido o erro inicial do *zero-shot* ao classificar o sintagma "União Americana" como ORGANIZAÇÃO (e não como PESSOA), ele o considerou como uma EN única e não como

parte integrante de “União Americana pelas Liberdades Civis”. Em outras palavras, o LLM segmentou erroneamente “União Americana pelas Liberdades Civis” em dois trechos (“União Americana” e “Liberdades Civis”), excluindo a contração “pelas” (“por”+”as”) que une os dois.

Figura 18. Desafio de delimitação de ORGANIZAÇÃO nos dois cenários.

O laço representa o apoio à União Americana pelas Liberdades Civis (Aclu, na sigla em inglês).			
Tokens	**Anotação referência**	**Anotação Zero-shot**	**Anotação Few-shot**
O	_	_	_
laço	_	_	_
representa	_	_	_
o	_	_	_
apoio	_	_	_
à	_	_	_
a	_	_	_
a	_	_	_
União	B-Organização	B-Pessoa	B-Organização
Americana	I-Organização	I-Pessoa	E-Organização
pelas	_	_	_
por	I-Organização	_	_
as	I-Organização	_	_
Liberdades	I-Organização	B-Organização	B-Organização
Civis	E-Organização	I-Organização	E-Organização
(	_	_	_
Aclu	S-Organização	B-Organização	S-Organização
,	_	_	_
na	_	_	_
em	_	_	_
a	_	_	_
sigla	_	_	_
em	_	_	_
inglês	_	_	_
)	_	_	_
.	_	_	_

Fonte: A autora (2025)

Por fim, destaca-se um efeito colateral relevante da adição de exemplos ao *prompt* na categoria PESSOA. Como pode ser observado na Figura 19, entidades como “Alemanha”, “Espanha” e “Inglaterra”, anotadas na referência como PESSOA devido ao uso metonímico para designar seleções nacionais, foram corretamente identificadas no cenário *zero-shot*, mas passaram a ser rotuladas como LOCAL no cenário *few-shot*.

Esse comportamento sugere que, ao receber exemplos, o modelo tornou-se mais rígido e literal na interpretação dos tokens, priorizando a leitura geográfica canônica de nomes de países e reduzindo sua sensibilidade a usos metonímicos. Assim, embora o *few-shot prompting* reduza ruído e promova maior consistência em diversas categorias, ele também pode limitar a flexibilidade interpretativa do modelo em contextos ambíguos, fazendo-o perder casos que o *zero-shot* conseguia capturar.

Figura 19. Ilustração do viés literal introduzido pelo *few-shot prompting*.

Alemanha, Espanha e Inglaterra são três campeãs com a ida assegurada para a Copa de 2018.			
Tokens	**Anotação referência**	**Anotação Zero-shot**	**Anotação Few-shot**
Alemanha	S-Pessoa	S-Pessoa	S-Local
,	_	_	_
Espanha	S-Pessoa	S-Pessoa	S-Local
e	_	_	_
Inglaterra	S-Pessoa	S-Pessoa	S-Local
são	_	_	_
três	_	_	_
campeãs	_	_	_
com	_	_	_
a	_	_	_
ida	_	_	_
assegurada	_	_	_
para	_	_	_
a	_	_	_
Copa	_	_	_
de	_	_	_
2018	_	_	_
.	_	_	_

Fonte: A autora (2025)

6. Considerações finais

Os resultados obtidos neste trabalho mostram que, no cenário *zero-shot*, o modelo Llama 3.1 alcançou 37,7% de medida-F1, enquanto no cenário *few-shot*, com *prompting* instruído e estruturado, a medida-F1 subiu para 66%. Esses valores permanecem abaixo do estado-da-arte treinado (BERTimbau, 78,6%), o que era esperado devido a fatores como a ausência de *fine-tuning*, a não utilização de milhões de parâmetros adicionais para *embeddings* ou *self-verification*, além da própria complexidade da tarefa de REN.

Embora ainda distante dos resultados alcançados por modelos ajustados (como o mencionado BERTimbau) ou métodos com infraestrutura adicional (como PromptNER (83,48%) e GPT-Ner (90,91%)), a abordagem baseada exclusivamente em *prompting* (especialmente no cenário *few-shot*) alcança desempenho competitivo (66% medida-F1) e oferece uma alternativa leve e promissora para o REN em português, evidenciando que estratégias como *persona prompting*, *rule-based prompting* e *structured output formatting* podem mitigar supergeneralizações e melhorar significativamente a qualidade das anotações. Por "leve", entende-se de baixo custo computacional, implementação rápida, menor dependência de dados anotados (massivos) e agilidade.

No entanto, é importante destacar que este estudo utilizou um *corpus* relativamente pequeno, composto por 471 sentenças e 9.470 *tokens*, manualmente anotados com entidades PLO, totalizando 559 ENs. Essa limitação amostral recomenda cautela na interpretação dos resultados. Isso quer dizer que, embora a análise crítica indique que o desempenho de 66% medida-F1 seja competitivo para um método sem treinamento, tais conclusões devem ser relativizadas, dado o tamanho restrito do *corpus*. Adicionalmente, a anotação manual é uma tarefa intrinsecamente subjetiva e, mesmo quando guiada por um manual de diretrizes, pode apresentar inconsistências ou divergências interpretativas, impactando a confiabilidade e a generalização dos resultados.

Mesmo assim, acredita-se que este trabalho contribuiu para as pesquisas do PLN, principalmente no que diz respeito à tarefa de REN em língua portuguesa, explorando a Engenharia de *Prompt* e técnicas de *In-Context Learning* até então não testadas na tarefa.

Referências

AMARAL, D.; VIEIRA, R. NERPCRF: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. Linguamática, v. 6, n. 1, p. 41–49, 2014.

ASHOK, D.; LIPTON, Z. C. PromptNER: Prompting For Named Entity Recognition. 2023. arXiv preprint arXiv:2305.15444. Disponível em: https://arxiv.org/abs/2305.15444

BICK, E. Functional aspects in Portuguese NER. In: INTERNATIONAL WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE (PROPOR). Proceedings [...]. [S.l.: s.n.], 2006. P. 80–89.

BROWN, T. B. *et al.* Language models are few-shot learners. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS 2020), 33., 2020, Vancouver. Proceedings [...]. Red Hook: Curran Associates, 2020. p. 1877–1901.

CASELI, H. M.; NUNES, M. G. V. (Org.). **Processamento de Linguagem Natural**: conceitos, técnicas e aplicações em português. 1. ed. São Carlos: Brasileiras em PLN, 2023. Disponível em: https://brasileiraspln.com/livro-pln/1a-edicao/.

CASTRO, P. *et al.* Portuguese named entity recognition using LSTM-CRF. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE (PROPOR), 13., 2018, Canela. Proceedings [...]. Cham: Springer, 2018. p. 83-92. (Lecture Notes in Computer Science, v. 11122). Disponível em: https://doi.org/10.1007/978-3-319-99722-3_9.

CHENG, K. *et al*. Structure guided prompt: instructing large language model in multi-step reasoning by exploring graph structure of the text. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), 2024, Miami. Proceedings [...]. Miami: Association for Computational Linguistics, 2024. p. 9407-9430. Disponível em: https://aclanthology.org/2024.emnlp-main.528/.

COELHO, G. *et al.* Information Extraction in the Legal Domain: Traditional Supervised Learning vs. ChatGPT. INSTICC; SciTePress, 2024. Disponível em: https://www.scitepress.org/Link.aspx?doi=10.5220/0012499800003690

CUI, L. *et al*. Template-based named entity recognition using BART. 2021. arXiv preprint arXiv:2106.01760. Disponível em: https://arxiv.org/abs/2106.01760.

DEVLIN, J. *et al.* BERT: Pre-training of deep bidirectional transformers for language understanding. 2018. Disponível em: https://arxiv.org/abs/1810.04805

DODDINGTON, G. *et al.* The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 4. Proceedings [...]. Lisbon, Portugal: European Language Resources Association (ELRA), mai. 2004. Disponível em: http://www.lrec-conf.org/proceedings/lrec2004/pdf /5.pdf

DONG, Q. *et al.* A Survey on In-context Learning. *In*: PROCEEDINGS OF THE 2024 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2024, Miami. Anais [...]. Miami: Association for Computational Linguistics, 2024. p. 1107–1128. Disponível em: https://aclanthology.org/2024.emnlp-main.64. Acesso em: 19 out. 2025.

DURAN, M. S. *et al*. The dawn of the Porttinari multigenre treebank: introducing its journalistic portion. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14., 2023, Belo Horizonte. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 115-124. Disponível em: https://doi.org/10.5753/stil.2023.233975.

DURAN, M.; PARDO, T. Anotação de corpus, um lugar privilegiado de observação linguística: o estudo das aposições do português brasileiro segundo o modelo Universal Dependencies. In: ENCONTRO DE LINGUÍSTICA DE CORPUS (ELC), 16. Anais [...]. Porto Alegre: [s.n.], 2024.

FREITAS, C. Linguística Computacional. [S.l.]: Parábola Editorial, 2022.

FREITAS, C.; SOUZA, E. *et al*. Recursos linguísticos para o PLN específico de domínio: o Petrolês. Linguamática, v. 15, n. 2, p. 51–68, dez. 2023. DOI: 10.21814/lm.15.2.412. Disponível em: https://linguamatica.com/index.php/linguamatica/article/view/412.

GRATTAFIORI, A. *et al*. The Llama 3 Herd of Models. [S.L.: s.n]: arXiv, 2024.

GRISHMAN, R.; SUNDHEIM, B. M. Message understanding conference-6: A brief history. In: International Conference on Computational Linguistics (COLING), 1996.

HAGÈGE, C.; BAPTISTA, J.; MAMEDE, N. Identificação, classificação e normalização de expressões temporais do português: a experiência do Segundo HAREM e o futuro. *In*: MOTA, C.; SANTOS, D. (org.). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. [S.l.]: Linguateca, 2008. cap. 2, p. 33-54.

JÚNIOR, C. M. *et al*. Paramopama: a Brazilian Portuguese corpus for named entity recognition. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC). Proceedings [...]. [S.l.]: SBC, 2015.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 3rd (draft). [S.l.: s.n.], 2025. Disponível em: https://web.stanford.edu/~jurafsky/slp3/ . Acesso em: 11 out. 2025.

LEE, D. *et al*. Good examples make a faster learner: simple demonstration-based learning for low-resource NER. 2021. arXiv preprint arXiv:2110.08454. Disponível em: https://arxiv.org/abs/2110.08454.

LOPES, F. *et. al.* Named entity recognition in portuguese neurology text using CRF. In: EPIA CONFERENCE ON ARTIFICIAL INTELLIGENCE. Proceedings [...]. [S.l.: s.n.], 2019. P. 336–348.

MADUREIRA, B. Avaliação de tecnologias de linguagem. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 3. ed. [S.l.]: BPLN, 2024. cap. 14. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-dados-avaliacao/cap-avaliacao/capavaliacao.html

MARRERO, M. *et al*. Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces, v. 35, n. 5, p. 482–489, 2013. ISSN 0920-5489. DOI: https://doi.org/10.1016/j.csi.2012.09.004. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920548912001080.

MCENERY, T. *et al*. Corpus-based language studies: an advanced resource book. London: Routledge, 2006.

MOTA, C., SANTOS, D. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008. Disponível em: https://www.linguateca.pt/LivroSegundoHAREM/ . Acesso em: 16 out. 2025.

NIVRE, J. MARNEFFE, M,; GINTER, F.; HAIJIE, J.; MANNING, C.; PYYSALO, S.; SCHUSTER, S.; TYERS, R.; ZEMAN, D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC). Marseille, France: 2020. p. 4034-4043. Disponível em: https://aclanthology.org/2020.lrec-1.497.pdf. Acesso em: 10 out. 2025.

NUNES, R. O. *et al*. Out of Sesame Street: A Study of Portuguese Legal Named Entity Recognition Through In-Context Learning. INSTICC; SciTePress, a2024. Disponível em: https://www.scitepress.org/Link.aspx?doi=10.5220/0012624700003690.

OLIVEIRA, V. *et al*. Combining prompt-based language models and weak supervision for labeling named entity recognition on legal documents. Artificial Intelligence and Law, p. 1–21, fev. 2024. Disponível em: https://link.springer.com/article/10.1007/s10506-023-09388-1

OpenAI. 2023. Gpt-4 technical report.

PIAI, L.. Anotação de corpus: caracterização de entidades nomeadas em tweets do mercado financeiro. 2025. Dissertação (Mestrado em Linguística) – Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, São Carlos, 2025.

PIRES, R. *et al*. Sabiá: Portuguese Large Language Models. (M. C. Naldi, R. A. C. Bianchi, Eds.)Intelligent Systems. Anais...Cham: Springer Nature Switzerland, 2023.

PIROVANI, J.; OLIVEIRA, E. Portuguese named entity recognition using conditional random fields and local grammars. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 11., 2018, Miyazaki. Proceedings [...]. Paris: ELRA, 2018. p. 4452-4456. Disponível em: https://aclanthology.org/L18-1704.pdf.

RAMSHAW, L. A.; MARCUS, M. P. Text chunking using transformation-based learning. *In*: Natural language processing using very large corpora. Dordrecht: Springer, 1999. p. 157-176. Disponível em: https://doi.org/10.1007/978-94-017-2390-9_10.

RATINOV, L.; ROTH, D. Design Challenges and Misconceptions in Named Entity Recognition. In: CONFERENCE on Computational Natural Language Learning (CoNLL), 13. Boulder, Colorado: Association for Computational Linguistics, jun. 2009. P. 147–155. Disponível em: https://aclanthology.org/W09-1119/

ROCHA, C. et al. PAMPO: using pattern matching and pos-tagging for effective named entities recognition in portuguese. Conference on Computational Processing of the Portuguese Language (PROPOR), 2016.

SANG, E. F. T. K.; MEULDER, F. D Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Em 7 th Conference on Natural Language Learning (CoNLL), 142-147. 2003.

SANTOS, C. N. dos; GUIMARÃES, V. Boosting Named Entity Recognition with Neural Character Embeddings. In: NAMED ENTITY WORKSHOP, 5. Proceedings [...]. [S.l.: s.n.], 2015. P. 25–33.

SANTOS, D. *et al*. HAREM: an advanced NER evaluation contest for Portuguese. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 5., 2006, Gênova. Proceedings [...]. Paris: ELRA, 2006. p. 1986-1991. Disponível em: http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.

SANTOS, D.; CARDOSO, N. Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. [S.l.]: Linguateca, 2007. DOI: http://hdl.handle.net/10400.26/380. Acesso em: 14 out. 2025.

SANTOS, D.; FREITAS, C. Avaliação conjunta em português. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português. 3. ed. [S.l.]: BPLN, 2024. cap. 15. ISBN 978-65-01-20581-6. Disponível em: https://brasileiraspln.com/livro-pln/3a-edicao/parte-dados-avaliacao/capavaliacao-conjunta/cap-avaliacao-conjunta.html.

SANTOS, J. *et al*. Assessing the impact of contextual embeddings for Portuguese named entity recognition. *In*: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS), 8., 2019, Salvador. Proceedings [...]. [S.l.]: IEEE, 2019. p. 437-442. Disponível em: https://doi.org/10.1109/BRACIS.2019.00083.

SARCINELLI, J. L. L. L. Grandes Modelos de Linguagem Reduzidos para Reconhecimento de Entidades Nomeadas em Português. Dissertação (Mestrado em Ciências) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

SARMENTO, L.; PINTO, A. S.; CABRAL, L. REPENTINO – A Wide-Scope Gazetteer for Entity Recognition in Portuguese. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE. Proceedings [...]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. P. 31–40. ISBN 978-3-540-34046-1.

SHANANHAN, M.; MCDONELL, K.; REYNOLDS, L. Role play with large language models. Nature, London, v. 623, n. 7987, p. 493-498, nov. 2023. Disponível em: https://www.nature.com/articles/s41586-023-06647-8.

SINCLAIR, J. Corpus and Text - Basic Principles. In: WYNNE, M. (Ed.). Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books, 2005. Acesso em: 30/10/2006. P. 1–16. Disponível em: http://ahds.ac.uk/linguistic-corpora/

SOLORIO, T. MALINCHE: a NER system for portuguese that reuses knowledge from Spanish. In: RECONHECIMENTO DE ENTIDADES MENCIONADAS EM PORTUGUÊS: DOCUMENTAÇÃO E ACTAS DO HAREM A PRIMEIRA AVALIAÇÃO CONJUNTA NA ÁREA. Proceedings [...]. [S.l.]: Linguateca, 2007. P. 123–136.

SOUZA, F. *et al*. Portuguese named entity recognition using BERT-CRF. 2019. arXiv preprint arXiv:1909.10649. Disponível em: https://arxiv.org/abs/1909.10649

VIEIRA E SILVA, A. Uma revisão para o Reconhecimento de Entidades Nomeadas aplicado à língua portuguesa. Linguamática, v. 15, n. 2, p. 69-85, 30 Dez. 2023. Disponível em: https://linguamatica.com/index.php/linguamatica/article/view/396. Acesso em: 14 out. 2025.

WAGNER F. *et al*. The brWaC corpus: a new open resource for Brazilian Portuguese. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 11., 2018, Miyazaki. Proceedings [...]. Paris: ELRA, 2018. p. 4339-4344. Disponível em: https://aclanthology.org/L18-1686.pdf.

WANG, Shuhe et al. GPT-NER: named entity recognition via large language models. 2023. arXiv preprint arXiv:2304.10428. Disponível em: https://arxiv.org/abs/2304.10428 .

WEI, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS 2022), 35., 2022, New Orleans. Proceedings [...]. Red Hook: Curran Associates, 2022. p. 24824–24835.

WEI, J. *et al*. Finetuned language models are zero-shot learners. 2021. arXiv preprint arXiv:2109.01652. Disponível em: https://arxiv.org/abs/2109.01652.

WEISCHEDEL, R. *et al*. OntoNotes Release 5.0. Philadelphia: Linguistic Data Consortium, 2013. (LDC2013T19). Disponível em: https://catalog.ldc.upenn.edu/LDC2013T19.

ZHANG, Y. *et al.* SynLLM: A Comparative Analysis of Large Language Models for Medical Tabular Synthetic Data Generation via Prompt Engineering. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2508.08529. Acesso em: 19 out. 2025.

ZHAO, W. X. *et al*. A Survey of Large Language Models. 2023

APÊNDICE A

Diretrizes de anotação de entidades nomeadas em uma parcela do Portinari-base

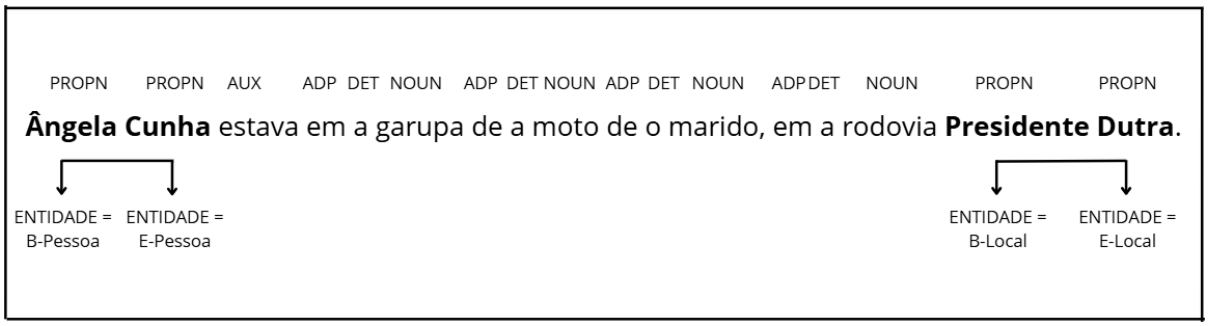
1. Formato de anotação

O padrão de anotação adotado neste trabalho foi o BIOES (Jurafsky; Martin, 2025). As *tags* desse esquema ou padrão funcionam da seguinte forma: para entidades que correspondem a apenas um *token,* aplica-se o prefixo "S-" (*Single*). Já as entidades que correspondem a multipalavras, ou seja, compostas por dois ou mais *tokens*, usa-se a marcação "B-" (*Begin*) para indicar o início da entidade, a marcação "I-" (*Inside*) para os intermediários e "E-" (*End*) delimitar o final. Para indicar *tokens* não anotados, que não correspondem a entidades, usa-se de modo implícito o prefixo "O-" (*Outside*). Considere o exemplo (1) e, em seguida, a Figura 1, que apresenta graficamente como funciona o processo de identificação e classificação de ENs multipalavras seguindo o formato BIOES.

(1) Ângela Cunha estava na garupa da moto do marido, na rodovia Presidente Dutra.

Na Figura 1 veremos que as contrações presentes na sentença foram desfeitas devido ao padrão de representação UD.

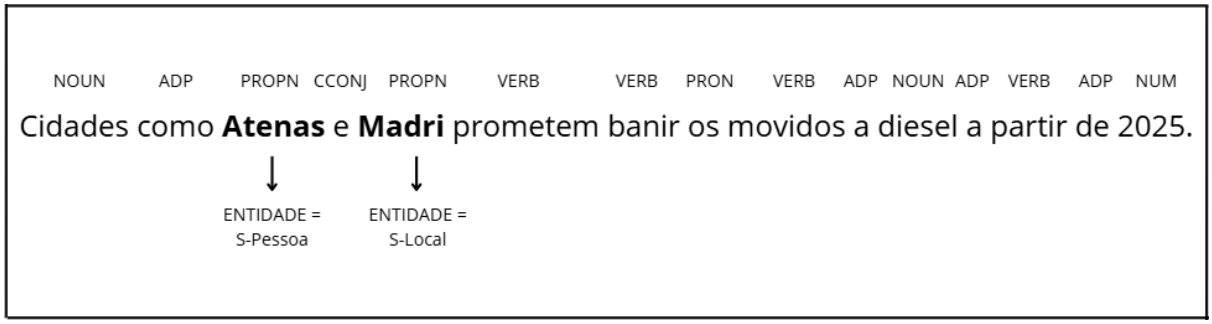
Figura 1: Exemplo de identificação e delimitação de ENs multipalavras no modelo BIOES.



Fonte: A autora (2025).

Na Figura 2 veremos graficamente como a identificação e classificação de ENs simples ocorrem no formato de anotação BIOES.

Figura 2: Exemplo de identificação e delimitação de ENs simples no modelo BIOES.



Fonte: A autora (2025).

O processo de anotação de entidades nomeadas no *corpus* Porttinari-base foi feito sobre os arquivos no formato CONLL-U, resultantes da anotação do *corpus* segundo o modelo gramatical *Universal Dependencies* (UD) (Nivre *et al.* 2020). Esses arquivos armazenam as informações morfológicas e sintáticas de uma sentença em 10 colunas (Figura 3).

Figura 3: Arquivo no formato CONLL-U referente ao modelo UD.

# sent_id = FOLHA_DOC003104_SENT003									
# text = Em comunicado, o Hamas, que governa a Faixa de Gaza, afirmou que aceitou condições exigidas pelo Fatah para que um acordo entre os grupos se concretize.									
1	Em	em	ADP	_	_	2	mark	_	_
2	comunicadc	comunicadc	NOUN	_	Gender=Masc|Number=Sing	14	obl	_	SpaceAfter=No
3	,	,	PUNCT	_	_	2	punct	_	_
4	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	5	det	_	_
5	Hamas	Hamas	PROPN	S-Pessoa	_	14	nsubj	_	SpaceAfter=No
6	,	,	PUNCT	_	_	8	punct	_	_
7	que	que	PRON	_	PronType=Rel	8	nsubj	_	_
8	governa	governar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fir	5	acl:relcl	_	_
9	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	10	det	_	_
10	Faixa	Faixa	PROPN	B-Local	_	8	obj	_	_
11	de	de	ADP	I-Local	_	12	case	_	Proper=Yes
12	Gaza	Gaza	PROPN	E-Local	_	10	flat:name	_	SpaceAfter=No
13	,	,	PUNCT	_	_	8	punct	_	_
14	afirmou	afirmar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fir	0	root	_	_
15	que	que	SCONJ	_	_	16	mark	_	_
16	aceitou	aceitar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fir	14	ccomp	_	_
17	condições	condição	NOUN	_	Gender=Fem|Number=Plur	16	obj	_	_
18	exigidas	exigir	VERB	_	Gender=Fem|Number=Plur|VerbForm=Part|Voice=Pass	17	acl	_	_
19-20	pelo	_	_	_	_	_	_	_	_
19	por	por	ADP	_	_	21	case	_	_
20	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	21	det	_	_
21	Fatah	Fatah	PROPN	S-Organizaçã	_	18	obl:agent	_	_
22	para	para	ADP	_	_	30	mark	_	_
23	que	que	SCONJ	_	_	22	fixed	_	_
24	um	um	DET	_	Definite=Ind|Gender=Masc|Number=Sing|PronType=Art	25	det	_	_
25	acordo	acordo	NOUN	_	Gender=Masc|Number=Sing	30	nsubj	_	_
26	entre	entre	ADP	_	_	28	case	_	_
27	os	o	DET	_	Definite=Def|Gender=Masc|Number=Plur|PronType=Art	28	det	_	_
28	grupos	grupo	NOUN	_	Gender=Masc|Number=Plur	25	nmod	_	_
29	se	se	PRON	_	Case=Nom|Person=3|PronType=Prs	30	expl	_	_
30	concretize	concretizar	VERB	_	Mood=Sub|Number=Sing|Person=3|Tense=Pres|VerbForm=Fi	16	advcl	_	SpaceAfter=No
31	.	.	PUNCT	_	_	14	punct	_	SpaceAfter=No

Fonte: A autora (2025).

Cada coluna de um arquivo CONLL-U é responsável por uma informação linguística:

1. ID: identificador da posição do token em uma sentença.
2. Form: *token* na forma que aparece na sentença

3. Lemma: lema ou radical do *token*
4. Upos tag: etiqueta que categoriza a classe de palavra (*part-of-speech* (PoS) tag)
5. Xpos tag: etiqueta PoS específica da língua
6. Feats: traços morfológicos do *token*
7. Head: ID do head do token dependente em questão
8. DepRel: relação de dependência entre o *token* ao seu *head*
9. Deps: delação de dependência *enhanced* (ou enriquecida) do *token*
10. Misc: informações adicionais do *token*

Para este trabalho, selecionou-se, como se observa na Figura 3, a quinta coluna desse formato/arquivo para a inserção das anotações de ENs, uma vez que a informação originalmente prevista para esta coluna (XPoS, isto é, etiquetas morfossintáticas específicas de uma língua) não está sendo usado pelo projeto POeTISa, responsável pela anotação UD do *corpus.*, um procedimento que pode ser visualizado em detalhe na Figura 3.

2. Diretrizes

Aqui apresentam-se as diretrizes de delimitação e classificação gerais e específicas por categorias das ENs do Portinari-base. Estas orientações foram baseadas na taxonomia do HAREM, levando em consideração suas duas publicações: o Primeiro (Cardoso; Santos, 2007) e o Segundo HAREM (Mota; Santos, 2008). Todos os exemplos utilizados foram retirados do Portinari-base.

2.1 Diretrizes de delimitação e classificação gerais:

1. Anotar apenas as ENs das categorias: Pessoa, Local e Organização (sem tipificação).
2. Anotar apenas uma categoria por EN, isto é, caso uma entidade possa pertencer a duas categorias diferentes é preciso escolher a que melhor se encaixa no contexto. Esse é um mecanismo para evitar problemas de ambiguidade.
3. Anotar apenas EN que tenha no mínimo uma letra maiúscula.
4. Identificar as ENs utilizando a etiqueta PoS PROPN (nomes próprios) da anotação UD do *corpus* como pista linguística
5. Anotar apenas os elementos constitutivos de contrações (do = de + o) que compõem ENs (por exemplo: Estado do (de+o) Missouri); isso significa que a contração em si

("do") não será anotada. Isso se deve ao fato de que as contrações foram decompostas segundo as diretrizes de *tokenização* do modelo UD.

6. Para cada entidade nomeada, identifique e anote apenas a sequência mais abrangente e coesa de *tokens* que expressa unicamente a entidade no contexto, descartando entidades aninhadas (*nested entities*). Assim, em "União Americana pelas Liberdades Civis", "União Americana" não será anotada.

2.2 Diretrizes de classificação específicas

2.2.1 Pessoa

Usa-se essa etiqueta para classificar:

(i) indivíduo, incluindo o título/função que está ligado a essa pessoa.
EXEMPLO: Inclusive nas obras de **Romero Brito**, pra usar um artista que o prefeito conhece.

(ii) grupo de indivíduos que não possui nome estabelecido, como um coletivo governamental.
(iii) cargos e posições ocupadas por um indivíduo, que pode ser ocupado por outras pessoas no futuro.
EXEMPLO: De todo modo, as negociações dentro da União Europeia costumam caber à própria **Chancelaria** e permaneceriam nas mãos de Merkel.

(iv) conjunto de indivíduos através de um cargo.
EXEMPLO: Na última terça, a **PF** encontrou em um apartamento em Salvador um montante de R$ 51 milhões.

(v) indivíduos que são caracterizados a partir da organização que representam.
(vi) conjunto de indivíduos participantes de uma organização ou conceito semelhante, como seitas e grupos terroristas.
EXEMPLO: Ainda não está claro se o **Hamas** pretende colocar as suas forças de segurança sob o controle de Abbas ou como a proposta permitirá uma aproximação entre ambos os lados.
(vii) situações em que um lugar ou coletivo de pessoas é personificado para descrever ações coletivas.

EXEMPLO: Já o **Palmeiras** caiu de 89% para 81%.

2.2.2 Local

Usa-se essa etiqueta para classificar:

(i) lugares geográficos físicos, sendo eles nomeados pelo Homem ou não, como rios e ilhas.
EXEMPLO: A **Chapada dos Guimarães**, em MT, perdeu 4.300 hectares entre agosto e setembro.

(ii) lugares físicos criados e/ou delimitados pelo Homem, como construções e regiões.
EXEMPLO: Contudo, a **Pompeia**, nova casa, só lhe dava o teto.

(iii) lugares abstratos ou virtuais, como meios de comunicação sociais e espaços eletrônicos.
EXEMPLO: O trio somaria cerca de 52,6% dos votos, segundo a projeção do **ZDF**.

2.2.3 Organização

Usa-se essa etiqueta para classificar:

(i) organizações administrativas, que são responsáveis pela governança de territórios em todas as instâncias, nacionais, internacionais e supranacionais.
EXEMPLO: **Justiça** solta Garotinho no dia de São Cosme e Damião

(ii) organizações que visam fins lucrativos, como empresas e sociedades.
EXEMPLO: A **Disney** divulgou na noite desta segunda-feira (9) o novo trailer de "Star Wars: Os Últimos Jedi", oitavo episódio da saga.

(iii) organizações que não buscam fins lucrativos, mas que não possuem papel direto de governança, como universidades e partidos políticos.
EXEMPLO: Que foi herança do **PT**, que nos deixou esse rombo, disse Doria.

APÊNDICE B

Versão traduzida da Figura 2. Exemplo de *prompt* no modelo PromptNER, feita em tradução livre.

Input

Defn: Uma entidade é uma pessoa(pessoa), universidade(universidade), cientista(cientista), ... evento(evento), prêmio(prêmio) ou teoria(teoria). Conceitos cinetíficos abstratos podem ser entidades se tiverem um nome associado a eles. Se uma entidade não se encaixa acima ela é (misc) . Datas, horas, adjetivos e verbos não são entidades.

Q: Dado o parágrafo abaixo, identifique uma lista de possíveis entidades e para cada entrada explique o porquê isso é ou não é uma entidade.

Parágrafo: Ele atendeu o Instituto de Tecnologia das Forças Aéreas dos EUA por um ano, ganhando um diploma de bacharel em aero mecânica e recebeu seu teste de treinamento de pilo na Base de Força Aérea de Edwards antes do seu exame de piloto na Base de Força Aérea de Wright-Patterson em Ohio.

Resposta:
1.Instituto de Tecnologia de Força Aérea dos EUA | Verdadeiro |como ele atendeu a esse instituto é provável que seja uma universidade (universidade)
2. diploma de bacharel | Falso | como não é uma universidade, prêmio ou nenhum outro tipo de entidade
3. aeromecânica | Verdadeiro | por ser uma disciplina científica (disciplina)
4. Base de Força Aérea de Edwards | Verdadeiro | como uma base de força aérea é uma unidade organizada (organização)
5. California | Verdadeiro |nesse caso California se refere ao próprio estado da California (localização)
6. Base de Força Aérea de Wright-Patterson | Verdadeiro| como uma base de força aérea é uma unidade organizada (organização)
7. Ohio | Verdadeiro | por ser um estado (localização)