Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Desenvolvimento de uma pipeline para interpretação de símbolos em narrativas de ficção baseado em LLMs

Lívia Ribeiro Pessamilio

JUIZ DE FORA
Julho, 2026

Desenvolvimento de uma pipeline para interpretação de símbolos em narrativas de ficção baseado em LLMs

Lívia Ribeiro Pessamilio

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Fabrício Martins Mendonça

JUIZ DE FORA

Julho, 2026

Desenvolvimento de uma pipeline para interpretação de símbolos em narrativas de ficção baseado em LLMs

Lívia Ribeiro Pessamilio

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Fabrício Martins Mendonça
Doutor em Ciência da Informação

Luciana Brugiolo Gonçalves
Doutora em Ciência da Computação

Carolina Alves Magaldi
Doutora em Letras

JUIZ DE FORA

8 DE *Julho*, 2026

Aos meus pais, que sempre me apoiaram de todas as formas possíveis.
Ao meu companheiro, que sempre me lembrava da minha competência quando inevitavelmente eu me sentia nervosa e preocupada.
Aos meus amigos e família, que deram não apenas apoio emocional, mas também apoio concreto ao participarem do meu experimento. Não estaria aqui sem vocês.

Resumo

A análise de simbolismo literário em narrativas de ficção é uma tarefa complexa que envolve interpretação contextual e conhecimento cultural, características historicamente associadas à leitura humana. Com o avanço dos Modelos de Linguagem de Grande Escala (LLMs), surge a questão de em que medida esses modelos são capazes de produzir interpretações simbólicas comparáveis às de leitores humanos. Este trabalho investiga se interpretações simbólicas geradas automaticamente por um LLM podem ser percebidas como humanas e avaliadas como plausíveis por participantes sem conhecimento prévio de sua origem. Para isso, foi desenvolvida uma pipeline que combina técnicas de Processamento de Linguagem Natural (PLN), incluindo tokenização, etiquetagem morfossintática e resolução de correferência, para identificar os objetos mais recorrentes de três contos de domínio público, e um LLM para gerar três interpretações simbólicas por objeto com base nos trechos relevantes do texto. Essas interpretações foram misturadas a anotações humanas e apresentadas a participantes que atribuíram notas de plausibilidade e indicaram a autoria percebida de cada uma. Os resultados mostraram taxa de acerto global de 61,4% na identificação de origem, sendo as anotações humanas mais reconhecidas (71,7%) do que as geradas por LLM (56,3%), e as interpretações automáticas receberam plausibilidade média ligeiramente superior (3,94) à das humanas (3,71) em escala Likert de 1 a 5, indicando que interpretações geradas por LLM podem se aproximar das humanas em qualidade percebida, ainda que apresentem maior tendência a serem confundidas com escrita humana. O escopo do trabalho é limitado: o corpus abrange apenas três contos, o experimento contou com cerca de 22 participantes e cada história recebeu anotações de somente três leitores humanos, o que restringe a generalização dos resultados.

Palavras-chave: Modelos de Linguagem de Grande Escala; Simbolismo Literário; Análise Computacional de Narrativas; Avaliação Humana; Processamento de Linguagem Natural.

Abstract

Analyzing literary symbolism in fictional narratives is a complex task that requires contextual interpretation and cultural knowledge, capabilities historically associated with human reading. With the advancement of Large Language Models (LLMs), the question arises of whether these models can produce symbolic interpretations comparable to those of human readers. This work investigates whether symbolic interpretations automatically generated by an LLM can be perceived as human-written and rated as plausible by participants who are unaware of their origin. To this end, a pipeline was developed that combines Natural Language Processing (NLP) techniques, including tokenization, part-of-speech tagging, and coreference resolution, to identify the most recurrent objects in three public-domain short stories, and an LLM to generate three symbolic interpretations per object based on relevant passages. These interpretations were mixed with human annotations and presented to participants who assigned plausibility scores and indicated the perceived authorship of each interpretation. Results showed an overall accuracy of 61.4% in identifying origin, with human annotations being more reliably recognized (71.7%) than those generated by the LLM (56.3%), and automatic interpretations receiving slightly higher mean plausibility scores (3.94) than human ones (3.71) on a 1–5 Likert scale, suggesting that LLM-generated interpretations can approach human ones in perceived quality, even though they are more often mistaken for human writing. The scope of the work is limited: the corpus covers only three short stories, the experiment involved approximately 22 participants, and each story received annotations from only three human readers, which restricts the generalizability of the findings.

Keywords: Large Language Models; Literary Symbolism; Computational Narrative Analysis; Human Evaluation; Natural Language Processing.

Agradecimentos

Agradeço à minha família, ao meu companheiro e aos meus amigos pelo apoio, incentivo e compreensão ao longo de toda a minha trajetória acadêmica, bem como pela contribuição direta e indireta para a realização deste trabalho.

Ao professor Fabrício, pela orientação, disponibilidade, paciência e pelos conhecimentos compartilhados ao longo deste trabalho.

À professora Luciana, que fez o seu melhor para guiar os primeiros passos deste trabalho, contribuindo com dedicação e apoio durante seu início, quando muitas ideias ainda estavam sendo estruturadas.

Aos professores do Departamento de Ciência da Computação, pelos ensinamentos transmitidos ao longo da graduação, e aos funcionários do curso, que contribuíram para minha formação pessoal e profissional.

Por fim, agradeço a todos os familiares, colegas e demais pessoas que, de alguma forma, contribuíram para que esta etapa fosse concluída.

"A utopia está lá no horizonte. Me aproximo dois passos, ela se afasta dois passos. Caminho dez passos e o horizonte corre dez passos. Por mais que eu caminhe, jamais alcançarei. Para que serve a utopia? Serve para isso: para que eu não deixe de caminhar".

Fernando Birri, citado por Eduardo Galeano in 'Las palabras andantes?' de Eduardo Galeano.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

DCC Departamento de Ciência da Computução

UFJF Universidade Federal de Juiz de Fora

1 Introdução

A interpretação de narrativas literárias envolve a decodificação de camadas de significado que vão além da leitura superficial do texto, incluindo o reconhecimento de metáforas, símbolos e outras formas de linguagem figurativa (CHAKRABARTY; CHOI; SHWARTZ, 2022). Com o avanço dos Modelos de Linguagem de Grande Escala (LLMs), surgiu a oportunidade de investigar essa capacidade no contexto computacional, buscando compreender em que medida sistemas automáticos conseguem replicar o tipo de análise que leitores humanos realizam ao ler um texto (JENNER et al., 2025). Este capítulo apresenta o contexto que motiva essa investigação, a justificativa para a abordagem adotada, a questão de pesquisa que orienta o trabalho e os objetivos propostos.

1.1 Contextualização

Desde mitos antigos até a literatura moderna, as narrativas são um dos pilares do entendimento e da transmissão do conhecimento humano. Podemos definir uma narrativa como um meio para a transmissão de experiências pessoais, sendo um tipo de texto que possui características particulares, não presentes em outras formas de texto, como ponto de vista, personagens e eventos, dispostos em uma determinada ordem. Nesse sentido, narrativas não se limitam à mera sequência de acontecimentos, mas envolvem também a construção de significados que emergem da relação entre elementos textuais, contextuais e simbólicos (ZHU et al., 2023).

Com o avanço da inteligência artificial, do *machine learning* e dos Modelos de Linguagem de Grande Escala (LLMs, do inglês Large Language Models), surgiu a oportunidade de tratar questões narrativas de maneira automática ou semiautomática, incluindo tarefas de interpretação e análise de textos narrativos (JENNER et al., 2025). Visando a essas aplicações, surge o interesse em compreender até que ponto as máquinas são capazes de analisar e representar conteúdos narrativos complexos. Nesse cenário, etapas de pré-processamento do texto têm se mostrado importantes para estruturar e organizar

o conteúdo antes de submetê-lo a modelos de linguagem, que por sua vez ampliam a capacidade de análise semântica e inferência contextual (SANTANA et al., 2023).

Do ponto de vista prático, essa automatização pode ser útil para diferentes perfis de usuários. Para analistas e editores literários, representa uma forma de obter uma leitura preliminar de certos aspectos do texto com maior rapidez e menor custo do que uma revisão humana completa. Para escritores, a questão ganha uma dimensão específica: em gêneros onde a surpresa narrativa é elemento central, como o suspense e o mistério. Um leitor real só pode vivenciar a experiência de ler o texto pela primeira vez uma única vez, o que limita bastante o número de avaliações independentes possíveis. Um modelo de linguagem, por não reter memória entre sessões, pode ser consultado repetidas vezes sem que essa limitação se aplique (JENNER et al., 2025).

Ainda assim, a interpretação automática de narrativas complexas apresenta desafios relevantes. A aplicação direta de LLMs em textos literários enfrenta obstáculos práticos: o processamento de textos extensos eleva o custo computacional e financeiro, e quando expostos a contextos muito longos ou ruidosos, esses modelos tornam-se propensos a respostas imprecisas ou desconexas da obra analisada (HUANG et al., 2025). Um pré-processamento que filtre e delimite o contexto relevante antes de acionar o modelo pode reduzir esses problemas.

Diante disso, este trabalho propõe uma pipeline que parte de um pré-processamento do texto, baseado em técnicas de Processamento de Linguagem Natural (PLN) como tokenização, marcação morfossintática, extração de objetos concretos e resolução de correferência, para selecionar os trechos mais relevantes e submetê-los a um LLM, responsável por identificar e interpretar objetos simbólicos em contos de ficção. A questão central, porém, vai além da combinação técnica em si: busca-se avaliar se as interpretações geradas pela pipeline são comparáveis às produzidas por leitores humanos, verificando se essa abordagem consegue produzir análises simbólicas que se sustentem quando confrontadas com a leitura humana.

1.2 Justificativa e Problema

No formato escrito, uma máquina não é capaz de ler diretamente como nós, tornando necessário o surgimento de uma ponte entre o texto e o computador: o Processamento de Linguagem Natural (PLN). O objetivo central do PLN aplicado à extração de narrativas é traduzir o texto bruto de uma história em informação organizada e legível para sistemas computacionais. Atualmente, o progresso mais significativo da área demonstra-se nas fases iniciais e estruturais do pipeline, notadamente no pré-processamento, no parsing e na extração de componentes narrativos factuais. Avanços consolidados em tarefas como o Reconhecimento de Entidades Nomeadas (NER) e a Marcação de Papéis Semânticos (SRL), discutidas em maior profundidade no próximo capítulo, permitem que as máquinas identifiquem com precisão "quem fez o quê", situando atores em cenários e tempos específicos (KHURANA et al., 2023).

Contudo, as narrativas de ficção envolvem nuances de difícil formalização algorítmica, como o uso de símbolos narrativos, em que um objeto da história atua como substituto de outro conceito abstrato no plano da obra. Esse aspecto é crítico porque a linguagem figurativa é onipresente em textos literários, mas a maioria das pesquisas computacionais ainda se concentra na linguagem literal, negligenciando a natureza não-composicional desse tipo de produção (CHAKRABARTY; CHOI; SHWARTZ, 2022). Embora estudos recentes tenham avançado na decodificação de simbolismo em modelos de linguagem, muitas abordagens ainda tratam o vínculo entre objeto e significado como um mapeamento fixo, em que cada objeto possui um sentido único independente do contexto narrativo, ou se restringem a contextos muito reduzidos. A extração literária permanece incompleta sem a compreensão da rede simbólica que se forma ao longo de um texto (CHAKRABARTY; CHOI; SHWARTZ, 2022).

Para ilustrar essa limitação, considere o clímax do arco do protagonista no filme Up: Altas Aventuras. Em um momento decisivo, ele segura uma corda presa à casa onde viveu com sua falecida esposa, enquanto a criança que o acompanha está prestes a cair. Ele precisa, então, soltar a casa para salvar o menino. Uma extração narrativa puramente literal perceberia apenas a superfície do evento: "o personagem soltou um objeto para alcançar outro"). Contudo, a cena representa a necessidade de abandonar o passado que

o prendia, para que possa viver o presente. Sem reconhecer o que a casa simboliza dentro daquele contexto narrativo específico, a extração computacional falha em captar o ponto essencial da história, resultando em uma representação parcial e limitada da obra.

Diante desses desafios, esta pesquisa apresenta uma pipeline que pré-processa o texto para selecionar e condensar os trechos mais relevantes antes de submetê-los a um LLM. Esse recorte reduz ruídos e tende a minimizar respostas desconexas da obra. O objetivo central, porém, não é medir a eficácia técnica dessa combinação, mas investigar se as interpretações simbólicas geradas por essa abordagem se aproximam das produzidas por leitores humanos.

O problema técnico reside no fato de que a interpretação não literal exige competências típicas do pensamento humano, tais como conhecimento de mundo, raciocínio abstrato e estabelecimento de conexões distantes não literais. Todos esses fatores são difíceis de ensinar passo a passo a um algoritmo, e também não são adquiridos dessa forma por humanos. O pensamento associativo depende de uma rede densa de conexões construídas ao longo da vida. Assim, uma ferramenta que vise emular comportamento análogo ao humano nesse aspecto pode obter melhores resultados ao adotar uma abordagem que se assemelhe à forma como tais competências emergem na mente humana (CHAKRABARTY; CHOI; SHWARTZ, 2022).

Nesse contexto, LLMs introduzem uma abordagem alternativa: treinam-se em grandes conjuntos de dados e aprendem padrões a partir de numerosos exemplos, o que favorece o surgimento de um conhecimento de mundo amplo e capacidades que não emergem em modelos menores (WEI et al., 2022a)(JENNER et al., 2025). Contudo, essas vantagens não são isentas de custos. LLMs enfrentam problemas relevantes para a análise literária, dentre os quais se destacam: o custo elevado (tanto computacional quanto financeiro) associado ao processamento de textos extensos em termos de tokens, a ocorrência de alucinações e a perda de contexto em textos muito longos (HUANG et al., 2025).

Em resumo, a interpretação simbólica exige um tipo de compreensão que vai além do processamento literal, e LLMs representam hoje a abordagem computacional mais próxima dessa capacidade (CHAKRABARTY; CHOI; SHWARTZ, 2022)(JENNER et al., 2025). As limitações práticas dos modelos, no entanto, recomendam que sejam

usados sobre contextos filtrados e delimitados. O problema central deste trabalho está, portanto, na dificuldade de automatizar a interpretação simbólica de objetos em narrativas de ficção, especialmente diante das limitações dos LLMs no tratamento de linguagem figurativa e contextos extensos. A pipeline proposta é apresentada como uma tentativa de contornar parte dessas dificuldades, combinando pré-processamento e LLMs, e tem sua eficácia analisada por meio da comparação com interpretações humanas.

1.2.1 Questão de Pesquisa

A partir do problema apresentado, este trabalho busca responder à seguinte questão de pesquisa principal:

"LLMs são capazes de interpretar objetos simbólicos em narrativas de ficção a partir do contexto em que esses objetos aparecem?"

Dessa questão central, derivam-se as seguintes subquestões:

- As interpretações simbólicas geradas por LLMs são consideradas plausíveis por leitores humanos?
- As interpretações simbólicas geradas por LLMs são distinguíveis daquelas elaboradas por leitores humanos?

1.3 Objetivos

Este projeto se organiza em torno de duas frentes complementares: o desenvolvimento de uma pipeline baseada em LLMs para identificação e interpretação de objetos simbólicos em narrativas de ficção e a avaliação das interpretações por ela produzidas. A avaliação se dá por meio de um experimento em que participantes analisam, sem conhecer a origem, interpretações automáticas e manuais, atribuindo notas de plausibilidade e indicando se consideram cada interpretação de autoria humana ou de IA.

1.3.1 Objetivo Geral

O objetivo geral deste trabalho consiste em:

Desenvolver uma pipeline baseada em LLMs para identificação e interpretação de objetos simbólicos em narrativas de ficção e avaliar, por meio de um experimento com participantes, se as interpretações produzidas são comparáveis, em plausibilidade, às elaboradas por leitores humanos.

1.3.2 Objetivos Específicos

Para o alcance do objetivo geral, definem-se os seguintes objetivos específicos:

- Constituir um corpus de contos de domínio público, selecionados a partir dos critérios de densidade de elementos com potencial simbólico, diversidade de autoria, época e tradição literária, e disponibilidade em versões públicas para leitura pelos anotadores humanos, e coletar anotações humanas de objetos simbólicos e suas interpretações, que servirão como referência para o experimento;
- Implementar uma pipeline modular em Python capaz de pré-processar os textos, extrair objetos candidatos e gerar interpretações simbólicas via LLM;
- Conduzir o experimento de avaliação, apresentando a participantes interpretações humanas e automáticas misturadas, para que atribuam notas de plausibilidade e indiquem a origem percebida de cada uma;
- Analisar os resultados, comparando as notas de plausibilidade e a taxa de acerto na identificação de origem entre interpretações humanas e automáticas.

1.4 Organização do Trabalho

Este trabalho está organizado em sete capítulos. O Capítulo 1 apresentou a introdução, incluindo a contextualização, o problema, a questão de pesquisa e os objetivos. O Capítulo 2 traz a fundamentação teórica, discutindo o conceito de narrativa, o Processamento de Linguagem Natural e os Modelos de Linguagem de Grande Escala. O Capítulo 3 revisa os trabalhos relacionados, situando esta pesquisa diante de estudos sobre análise simbólica, anotação automática e comparação entre LLMs e leitores humanos. O Capítulo 4 detalha a metodologia, descrevendo o tipo de pesquisa adotado e o desenvolvimento da pipeline. O

Capítulo 5 apresenta o experimento conduzido com participantes, e o Capítulo 6 reúne os resultados obtidos e sua análise. Por fim, o Capítulo 7 traz as conclusões e as perspectivas de trabalhos futuros.

2 Fundamentação teórica

Este capítulo apresenta os fundamentos teóricos que embasam o trabalho. Inicialmente, discute-se o conceito de narrativa, seus elementos estruturais e o que a distingue de outros tipos de texto. Em seguida, apresenta-se o Processamento de Linguagem Natural (PLN), com foco nas tarefas relevantes para a análise de textos narrativos. Por fim, abordam-se os Modelos de Linguagem de Grande Escala (LLMs), sua arquitetura e suas aplicações no contexto da análise literária.

2.1 Narrativas

A narrativa, além de transmitir experiências pessoais, é compreendida teoricamente como um tipo de discurso com características próprias que a diferenciam de outros tipos de texto. Seu alcance é amplo e abrange diferentes modalidades: narrativas ficcionais, como novelas, contos, filmes e peças de teatro; narrativas jornalísticas, como reportagens e crônicas; e narrativas históricas, voltadas ao relato de acontecimentos passados, entre outras. Este trabalho concentra-se especificamente nas narrativas ficcionais, por serem o tipo em que os elementos simbólicos aparecem com maior frequência e densidade. Ela é formada por elementos como ponto de vista, personagens e uma sequência organizada de eventos que, combinados, formam um enredo coerente. Esses componentes não atuam isoladamente, mas se relacionam para produzir sentido e orientar a compreensão do leitor ou espectador.

A profundidade de uma narrativa está justamente nas relações entre seus elementos. Não basta acompanhar a sequência dos acontecimentos; é necessário entender as relações de causa e efeito e as conexões entre os personagens. Assim, analisar uma narrativa envolve compreender tanto a ordem dos fatos quanto as ligações que constroem o sentido da história (ZHU et al., 2023).

Narrativas se distinguem de outros tipos de texto pelo foco em eventos e ações de personagens situados em um contexto específico. Enquanto textos expositivos buscam

explicar conceitos e processos, e textos argumentativos organizam ideias e posições em uma estrutura lógica, o texto narrativo centra-se nos acontecimentos e nas ações dos agentes envolvidos. Esse traço é o que diferencia a narrativa dos demais modos discursivos (HENG; PU; LIU, 2023).

Do ponto de vista da narratologia, a análise estrutural de uma narrativa distingue diferentes níveis de organização. Um deles é a fábula, que corresponde à sequência cronológica dos eventos que compõem a história, abstraída de sua forma de apresentação. Outro é a intriga, entendida como a organização concreta desses eventos no texto, com as escolhas de ordem, ritmo e perspectiva que orientam a leitura. A narrativa, nesse enquadramento, é justamente o conjunto formado pela fábula e pela intriga: o que é contado somado ao modo como é contado. Essa separação permite compreender que uma mesma história pode ser contada de formas muito diferentes dependendo das escolhas de tempo, ordem e perspectiva adotadas. Além disso, a narratividade de um texto pode ser entendida como uma propriedade de grau: quanto mais um texto apresenta progressão temporal, encadeamento causal e presença de agentes, mais narrativo ele é (LIU; JOSHI; DAWSON, 2026).

2.2 Processamento de Linguagem Natural

Esta seção discute o Processamento de Linguagem Natural (PLN) e as técnicas empregadas na análise computacional de textos. A subseção de Conceitos Fundamentais define o PLN e apresenta suas principais etapas de análise. A subseção de Uso de PLN em Narrativas aborda como essas técnicas se aplicam especificamente à análise de textos narrativos, contextualizando o pipeline implementado neste trabalho.

2.2.1 Conceitos fundamentais

O Processamento de Linguagem Natural (PLN) é uma subárea da Inteligência Artificial e da Linguística Computacional dedicada ao desenvolvimento de métodos e sistemas capazes de analisar, interpretar e gerar linguagem humana de forma automática, abrangendo tanto textos escritos quanto linguagem falada. O campo pode ser dividido em duas

grandes vertentes: a Compreensão de Linguagem Natural (NLU), voltada para análise e interpretação do texto, e a Geração de Linguagem Natural (NLG), dedicada à produção de texto coerente a partir de representações internas (KHURANA et al., 2023).

Na prática, o processamento de um texto percorre uma sequência de estágios que progressivamente transformam o texto bruto em representações mais ricas. O primeiro desses estágios é a tokenização, que consiste em identificar as unidades mínimas do texto, como palavras e pontuações (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). Na sequência, a etiquetagem morfossintática (POS tagging) atribui a cada token sua classe gramatical com base no contexto em que aparece. As tarefas de normalização lexical, como a radicalização (stemming) e a lematização, reduzem as variantes morfológicas das palavras às suas formas canônicas: enquanto o stemming remove sufixos de forma aproximada, a lematização considera o contexto para retornar a forma base correta (KHURANA et al., 2023). A análise sintática (parsing) agrupa os tokens em sintagmas e estabelece relações de dependência entre eles, revelando a estrutura gramatical da sentença. Em níveis mais avançados, o processamento semântico busca determinar o significado das expressões e o nível discursivo trata de relações que se estendem por mais de uma sentença, como a resolução de anáfora e correferência (KHURANA et al., 2023)(NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). O Reconhecimento de Entidades Nomeadas (NER) é uma tarefa de extração de informação voltada à identificação e categorização de entidades específicas no texto, como pessoas e locais (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011). A Figura 2.1 sintetiza esses estágios.

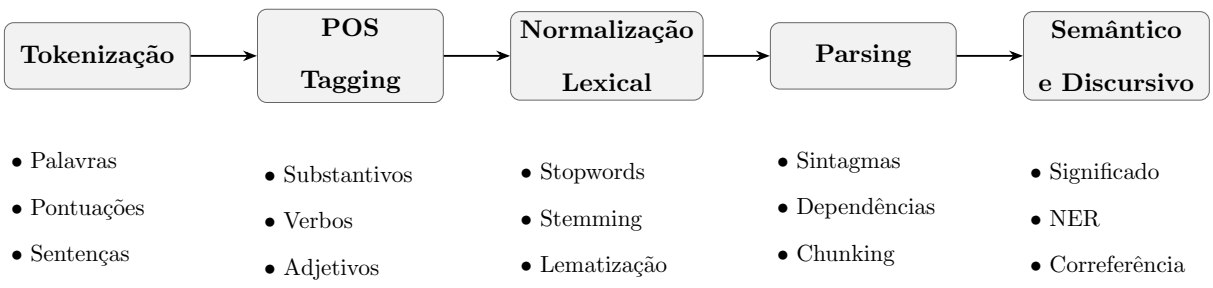

Figura 2.1: Estágios gerais de um pipeline de PLN. Adaptado de Khurana et al. (2023) e Nadkarni, Ohno-Machado e Chapman (2011).

2.2.2 Uso de PLN em Narrativas

Nesse contexto, a aplicação prática do PLN à análise de narrativas pode ser compreendida a partir de um fluxo estruturado de processamento, geralmente organizado em estágios sequenciais. Podemos dividir o processamento em cinco estágios principais, projetados para identificar, extrair e formalizar os elementos que compõem uma história a partir de fontes textuais. O processo inicia-se com o Pré-processamento e Parsing, estágio que utiliza tarefas léxicas e sintáticas, como tokenização e análise de dependências, para padronizar o texto bruto e prepará-lo para análises mais profundas. O segundo estágio é a Identificação e Extração de Componentes, voltado para o reconhecimento dos pilares da narrativa: eventos, participantes (atores), tempo e espaço. No estágio de Conexão de Componentes, o sistema estabelece os vínculos entre as informações extraídas através de raciocínio temporal, anotação de papéis semânticos (SRL) e extração de relações, conferindo uma estrutura global à narrativa. A Representação segue-se como a fase de organização do conteúdo em modelos formais, podendo ser conceitual (através de ontologias e lógica) ou visual (utilizando grafos de conhecimento e linhas do tempo). Finalmente, a Avaliação verifica a eficácia do sistema em capturar a história de forma precisa, utilizando métricas quantitativas e julgamentos humanos qualitativos. (SANTANA et al., 2023)

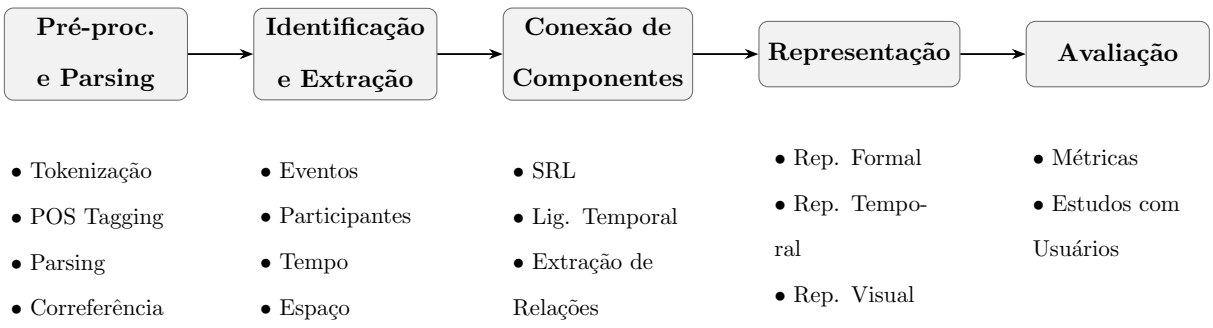

Figura 2.2: Pipeline de extração de narrativas em PLN. Adaptado de Santana et al. (2023).

No presente trabalho, o pipeline implementado abrange os primeiros estágios desse fluxo. O pré-processamento é realizado com a biblioteca spaCy, que executa tokenização, etiquetagem morfossintática e reconhecimento de entidades nomeadas. A

resolução de correferência é conduzida pelo modelo fastcoref, que agrega as diferentes menções de um mesmo objeto ao longo do texto, viabilizando a contagem de frequência necessária para a seleção dos candidatos simbólicos.

Pré-processamento Textual

O primeiro desses estágios corresponde ao pré-processamento textual, que constitui a fase inicial e indispensável em fluxos de PLN, tendo como objetivo principal a padronização e a limpeza dos dados brutos para facilitar as operações computacionais subsequentes. Este processo inicia-se com a tarefa de sentenciação, ou segmentação de sentenças, que visa delimitar as unidades sentenciais do texto, enfrentando desafios consideráveis devido à ambiguidade de certas pontuações que podem indicar tanto o fim de uma frase quanto uma abreviatura ou marcador numérico. Na sequência, realiza-se a tokenização, responsável por fragmentar a sentença em suas unidades linguísticas mínimas, os tokens, utilizando geralmente espaços em branco e sinais de pontuação como critérios de separação. Por fim, a normalização atua na transformação das palavras para formatos padronizados, englobando atividades como a conversão para letras minúsculas, a remoção de termos com baixa carga informativa (stopwords) e a aplicação de métodos de lematização ou radicalização (stemming), que reduzem as variantes morfológicas às suas formas canônicas ou raízes, garantindo a uniformidade do vocabulário e a eficiência das etapas posteriores de análise Por exemplo, a lematização converteria a flexão verbal "corri"em sua forma canônica "correr", ao passo que a radicalização (stemming) reduziria termos como "apresentação"e "apresentando"ao radical comum "apresent". (SANTANA et al., 2023).

Análise Sintática e Morfossintática

Uma vez concluída a padronização lexical do texto, o processamento avança para a análise sintática e morfossintática, frequentemente denominada parsing. Seu propósito é identificar a função gramatical das palavras e a organização estrutural das sentenças. Esse processo fundamenta-se inicialmente na etiquetagem morfossintática (Part-of-Speech tagging), que categoriza cada unidade linguística em classes como substantivos, verbos e adjetivos, considerando tanto sua definição morfológica quanto o contexto em que estão

inseridas. No processamento computacional, a análise sintática é realizada principalmente por duas abordagens: a de constituência, que organiza o texto em unidades hierárquicas chamadas sintagmas, e a de dependência, que estabelece relações de subordinação entre palavras regentes e seus dependentes. Essas estruturas são essenciais para compreender a organização do significado e identificar corretamente os papéis semânticos, permitindo que o sistema interprete as relações internas da oração e sirva de base para tarefas mais complexas de extração de informação. Por exemplo, na frase "O aluno leu o livro", a análise de constituência identifica os sintagmas nominal ("O aluno") e verbal ("leu o livro"), enquanto a análise de dependência mostra que "leu" é o núcleo da oração, do qual dependem "aluno" e "livro". (SANTANA et al., 2023).

Anotação de Papéis Semânticos (SRL)

A partir da estrutura sintática identificada, torna-se possível avançar para níveis mais profundos de interpretação semântica, nos quais se insere a tarefa de Anotação de Papéis Semânticos (SRL). Essa fase trata-se de uma etapa avançada no processamento semântico da linguagem natural, voltada para a identificação da estrutura de predicados e seus respectivos argumentos. Diferentemente da análise sintática, que se concentra na organização gramatical e nas dependências estruturais das sentenças, a SRL atua em um nível de interpretação mais profundo, buscando recuperar a estrutura de significado latente ao responder a questões fundamentais sobre "quem fez o quê a quem", "onde", "quando"e "por que". Esse processo consiste em atribuir papéis semânticos específicos aos constituintes de uma oração, mapeando a relação entre o verbo, que atua como o elo central do evento, e seus participantes, que podem desempenhar funções de agente, paciente, instrumento ou modificadores de tempo e espaço. Especialmente em arquiteturas voltadas para a extração de narrativas, a SRL é considerada essencial para o estabelecimento de conexões globais entre componentes textuais. (SANTANA et al., 2023).

Identificação e Extração de Componentes Narrativos

Com base nas análises sintáticas e semânticas previamente realizadas, o processamento narrativo avança para a identificação e extração dos componentes fundamentais da história:

eventos, participantes, tempo e espaço. Esta fase caracteriza-se por uma transição para a análise semântica, buscando transformar dados não estruturados em informações organizadas e representáveis. A extração de eventos fundamenta-se na detecção de gatilhos linguísticos, como verbos e nomes, que sinalizam ocorrências significativas e suas consequências no mundo narrado. Paralelamente, o reconhecimento de participantes é realizado predominantemente por meio de tarefas de reconhecimento de entidades nomeadas (NER), categorizando atores em classes como pessoas e organizações. Já as informações de tempo e espaço são cruciais para ancorar a cronologia e o cenário da trama, envolvendo a extração e normalização de expressões temporais e referências geográficas. O objetivo global desse processo é fornecer a base necessária para o estabelecimento de conexões lógicas e cronológicas entre os componentes, permitindo uma compreensão automatizada e sistêmica da narrativa. (SANTANA et al., 2023).

Reconhecimento de Entidades Nomeadas (NER)

Entre as tarefas centrais deste estágio, destaca-se o reconhecimento de entidades nomeadas (NER), uma tarefa de extração de informação voltada à localização e categorização de entidades em classes pré-definidas (tais como pessoas, organizações, locais e expressões de tempo). Dentro do contexto de narrativas, esse processo é fundamental para a identificação dos participantes ou atores envolvidos nos eventos, configurando um dos pilares essenciais para a estruturação global da história. Enquanto as abordagens genéricas tratam uma ampla gama de tipos de entidades, no domínio narrativo observa-se uma ênfase na detecção de entidades animadas, que costumam ter maior relevância narrativa. A evolução do reconhecimento de entidades nomeadas tem sido impulsionada por modelos de alta performance que utilizam a concatenação automatizada de embeddings, técnica que permite superar obstáculos como a ambiguidade linguística e a diversidade de gêneros textuais. De maneira simplificada, os embeddings consistem em representações matemáticas de palavras em vetores numéricos, nas quais o significado de cada termo é capturado a partir de sua distribuição e contexto nos textos. Esses avanços aprimoram a identificação de atores e participantes, um passo fundamental no contexto narrativo. (SANTANA et al., 2023).

Resolução de Correferência

Após a identificação das entidades ao longo do texto, torna-se necessário acompanhar suas diferentes menções no discurso, o que é realizado por meio da resolução de correferência. Por exemplo, no trecho "O menino foi na festa. Ele ficou muito feliz.", a máquina não é capaz de saber de imediato a quem o termo "ele"se refere. Essa tarefa é considerada essencial para o entendimento profundo da linguagem, uma vez que o agrupamento desses termos em cadeias de correferência garante a manutenção da coesão e da coerência textual ao longo de uma narrativa. Trata-se de um desafio técnico complexo que exige a integração de múltiplas estratégias diferentes. Historicamente abordada por modelos de regras, a área tem evoluído para o emprego de arquiteturas de redes neurais profundas e modelos de linguagem baseados em Transformers, que possibilitam o reconhecimento de relações semânticas latentes e dependências de longa distância em grandes volumes de dados. (SANTANA et al., 2023).

2.3 LLMs

Esta seção apresenta os Modelos de Linguagem de Grande Escala (LLMs). A Seção de Conceitos Fundamentais cobre a evolução histórica dos modelos de linguagem, a arquitetura Transformer, os tipos de modelos existentes, suas capacidades e limitações estruturais. A Seção de Uso de LLMs em Narrativas discute como essas capacidades se aplicam à análise e interpretação de textos narrativos, fundamentando a abordagem adotada neste trabalho.

2.3.1 Conceitos fundamentais

A Inteligência Artificial (IA) Generativa avançou rapidamente nos últimos anos, permitindo a criação de conteúdos digitais, como textos e imagens fotorrealistas, a partir de descrições fornecidas por usuários humanos. No centro dessa revolução tecnológica estão os Modelos de Linguagem de Grande Escala, conhecidos como Large Language Models (LLMs), que transformaram a maneira como interagimos com a tecnologia e remodelaram diversos setores da indústria (KAMATH et al., 2024).

Tecnicamente, um LLM opera dentro de um framework probabilístico para prever a probabilidade de tokens (palavras ou fragmentos de texto) subsequentes em uma sequência, baseando-se nos padrões aprendidos durante seu treinamento. O que distingue os LLMs de modelos de linguagem menores é, fundamentalmente, a sua escala. Um LLM é caracterizado por possuir de centenas de milhões a trilhões de parâmetros. Três fatores centrais determinam essa escala: o tamanho do corpus de pré-treinamento (volume de dados), o número de parâmetros aprendidos (que determina a complexidade do modelo) e a escala computacional necessária para realizar o treinamento e a inferência. (KAMATH et al., 2024)

Historicamente, a evolução da área partiu de modelos estatísticos simples, como os N-grams, que tentam prever uma palavra com base no histórico de palavras anteriores, evoluindo para o uso de embeddings, que são representações vetoriais densas que capturam o significado semântico dos tokens em um espaço matemático. Um marco nessa transição foi o Word2Vec, que introduziu arquiteturas neurais eficientes para representar palavras como vetores contínuos, capturando relações semânticas e sintáticas a partir de grandes corpora de texto (MIKOLOV et al., 2013). A mudança de paradigma fundamental para o surgimento dos LLMs modernos ocorreu em 2017 com a introdução da arquitetura Transformer. Essa arquitetura baseia-se no mecanismo de atenção (ou self-attention), que permite ao modelo focar seletivamente em diferentes partes da sequência de entrada para entender o contexto global. Tal capacidade é ampliada pelo multi-head attention, que utiliza diversos "cabeçotes"de atenção em paralelo para capturar variados aspectos e dependências dos dados simultaneamente (VASWANI et al., 2017)(KAMATH et al., 2024).

O desenvolvimento desses modelos ocorre em duas etapas principais: o pré-treinamento, onde o modelo aprende habilidades linguísticas fundamentais a partir de volumes massivos de dados não rotulados, e o ajuste fino (fine-tuning), que adapta o modelo pré-treinado para tarefas específicas utilizando dados rotulados. Um fenômeno notável desse processo é o surgimento de habilidades emergentes, que são capacidades de resolver problemas complexos (como raciocínio lógico e aritmética) que surgem apenas quando o modelo atinge escalas massivas, sem que ele tenha sido explicitamente treinado

para tais funções. Entre essas capacidades está o aprendizado zero-shot, pelo qual o modelo realiza tarefas sem que exemplos específicos sejam fornecidos na instrução, ampliando sua aplicabilidade em cenários onde dados rotulados são escassos (WEI et al., 2022a)(KAMATH et al., 2024).

No desenvolvimento desses modelos, a arquitetura pode ser direcionada para diferentes finalidades conforme o objetivo de aprendizado, destacando-se a diferença entre modelos mascarados e autorregressivos. Os modelos de linguagem mascarados (MLM), como o BERT, operam ocultando tokens específicos na sequência de entrada e treinando o modelo para prevê-los, o que permite uma compreensão bidirecional profunda e os torna ideais para tarefas que exigem análise de contexto, como classificação e reconhecimento de entidades (DEVLIN et al., 2019).

Em contrapartida, os modelos autorregressivos, como a série GPT, são projetados para prever o próximo token baseando-se exclusivamente nos tokens precedentes, seguindo uma lógica unidirecional. Essa característica tem implicações diretas na geração interpretativa, pois torna os modelos autorregressivos altamente proficientes em tarefas de geração de texto fluído e assistência na escrita, repetindo o processo de predição e anexação de tokens até que a geração esteja completa (BROWN et al., 2020).

As capacidades atuais dos Modelos de Linguagem de Grande Escala (LLMs) representam um marco na evolução da Inteligência Artificial, permitindo uma interação sem precedentes com a linguagem humana através da compreensão (NLU) e geração (NLG) de texto. Esses modelos demonstram proficiência em tarefas de inferência lógica, raciocínio em múltiplas etapas e habilidades emergentes, como resolução de problemas aritméticos e auxílio na escrita de códigos de programação, para as quais não foram explicitamente treinados. Uma característica central de sua versatilidade é o aprendizado em contexto (in-context learning), que permite ao modelo adaptar-se a novas tarefas durante a execução apenas por meio de exemplos fornecidos no prompt (BROWN et al., 2020).

Atualmente, as aplicações práticas desses sistemas são vastas, abrangendo desde a Inteligência Artificial Conversacional em chatbots e assistentes virtuais até a criação automatizada de conteúdo multimodal (envolvendo áudio, imagem e vídeo) e sistemas de

busca aprimorados via Geração Aumentada de Recuperação (RAG).

O surgimento de habilidades emergentes constitui um fenômeno técnico no qual os LLMs manifestam competências complexas, como aritmética e raciocínio de múltiplas etapas, para as quais não foram explicitamente treinados durante o pré-treinamento. Essas habilidades surgem como uma consequência direta da escala do modelo, manifestando-se frequentemente por meio de aumentos abruptos na precisão de tarefas quando um certo limite de parâmetros é ultrapassado. A ativação dessas capacidades é viabilizada por técnicas de prompting, especificamente o aprendizado em contexto (in-context learning), que permite ao modelo adaptar-se a novos desafios apenas através das instruções fornecidas no momento da inferência, sem a necessidade de atualizar seus parâmetros internos (WEI et al., 2022a).

Uma das metodologias mais eficazes nesse sentido é o Chain-of-Thought (CoT), ou Cadeia de Pensamento, que estrutura os exemplos e comandos de forma a replicar o processo de pensamento sequencial humano, melhorando tanto a acurácia quanto a explicabilidade do resultado final (WEI et al., 2022b). Existem variações como o CoT zero-shot, que utiliza instruções simples como "vamos pensar passo a passo"(let's think step by step), e o CoT manual, que fornece demonstrações explícitas das etapas de raciocínio a serem seguidas. Contudo, a eficácia dessas estratégias está intrinsecamente ligada à escala computacional, uma vez que modelos menores geralmente não exibem os mesmos ganhos de desempenho ao utilizar cadeias de pensamento.

Ainda assim, os LLMs enfrentam limitações estruturais significativas, sendo a alucinação uma das mais críticas, caracterizada pela geração de respostas factualmente incorretas que soam convincentes e gramaticalmente coerentes. Esse fenômeno compromete a confiabilidade dos sistemas baseados em LLMs, especialmente em contextos onde a precisão factual é crítica (HUANG et al., 2025).

Além disso, os modelos herdam vieses estatísticos e preconceitos sociais, culturais e demográficos presentes nos dados massivos coletados da internet, o que pode resultar em tratamentos injustos ou conteúdos tóxicos (BENDER et al., 2021). Outro desafio central é a extrema dependência de prompt, na qual o desempenho do sistema é altamente sensível a pequenas variações na forma como a instrução é escrita, exigindo técnicas

complexas de engenharia para garantir resultados precisos. Por fim, problemas como a sicofancia (tendência de ecoar a perspectiva do usuário em vez de priorizar a verdade) e vulnerabilidades a ataques de jail-breaking reforçam a necessidade de processos contínuos de alinhamento e segurança.

2.3.2 Uso de LLMs em Narrativas

Com o surgimento dos Modelos de Linguagem de Grande Escala (LLMs), abriu-se uma nova oportunidade para tratar questões interpretativas de maneira automática. Esses modelos demonstram proficiência em tarefas de raciocínio de múltiplas etapas e compreensão contextual (NLU), capacidades que emergem com o aumento da escala e não estão presentes em modelos de menor porte (WEI et al., 2022a). Na análise de narrativas, LLMs têm produzido resultados comparáveis aos de analistas humanos (JENNER et al., 2025). No entanto, a aplicação direta de LLMs em obras literárias longas enfrenta obstáculos práticos significativos: o processamento de volumes massivos de texto exige um custo computacional e financeiro elevado. Além disso, esses modelos apresentam tendência a gerar "alucinações", produzindo respostas gramaticalmente coerentes, mas factualmente incorretas ou desconexas do texto analisado (HUANG et al., 2025).

Em conclusão, a interpretação automatizada de conteúdo não literal em narrativas de ficção vai além das tarefas de extração factual já consolidadas. A identificação de simbolismo exige raciocínio contextual e compreensão de linguagem não literal, capacidades que LLMs demonstram ser capazes de aproximar. Para viabilizar essa aplicação de forma prática, um pré-processamento que filtre e condense os trechos relevantes antes de submetê-los ao modelo tende a reduzir ruídos e custos. Com base nesse cenário, este trabalho investiga se uma pipeline com esse design consegue produzir interpretações simbólicas comparáveis às realizadas por leitores humanos.

3 Trabalhos Relacionados

Este capítulo reúne trabalhos relacionados à pesquisa em três frentes: a capacidade de modelos de linguagem em interpretar elementos simbólicos e temáticos em narrativas, a extração de entidades em textos literários, e a comparação entre anotações geradas por LLMs e por humanos. A seleção buscou cobrir tanto estudos próximos ao problema técnico desta pesquisa quanto aqueles que investigam a confiabilidade das saídas de LLMs quando comparadas à análise humana.

3.1 Decoding Symbolism in Language Models

O trabalho de Guo, Hwa e Kovashka (2023), intitulado "Decoding Symbolism in Language Models", investiga a viabilidade de extrair conhecimento de modelos de linguagem (LMs) para a decodificação de simbolismo, definida como a capacidade de reconhecer que um item (significante) atua como substituto de outro conceito (significado),. O objetivo central do estudo é verificar se os modelos de linguagem de última geração encapsulam conhecimento implícito e abstrato suficiente para inferir relações simbólicas, indo além do conhecimento comum superficial.

A pesquisa busca resolver três frentes principais: identificar se os modelos conseguem prever um significado simbólico apropriado para um objeto físico, avaliar o impacto do contexto situacional na interpretação de símbolos que não são fixos por convenção, e investigar como o viés em corpora de pré-treinamento prejudica a identificação de conexões simbólicas mais raras. Para isso, os autores propõem o framework SymbA (Symbolism Analysis), comparando diferentes arquiteturas de modelos e propondo métodos para mitigar vieses estatísticos que favorecem conceitos excessivamente comuns.

Se tratando da metodologia, o trabalho utiliza técnicas de sondagem (probing) para extrair o conhecimento simbólico de modelos baseados em Transformers, comparando arquiteturas mascaradas (BERT e RoBERTa) e autorregressivas (GPT-2 e GPT-J-6B) ao baseline Word2Vec. O estudo aplica a Informação Mútua Pontual (PMI) para medir a

dificuldade dos símbolos. Complementarmente, utiliza-se uma estratégia de desenviesamento que re-ranqueia os candidatos normalizando a probabilidade condicional de cada termo pela sua probabilidade a priori.

Já para a implementação dos experimentos, os autores utilizaram a linguagem Python com o suporte das bibliotecas transformers[1] e gensim[2], realizando o processamento em infraestrutura de hardware NVIDIA Quadro RTX 5000 com suporte a CUDA.

No que diz respeito ao escopo, o estudo concentra-se na interpretação do simbolismo sob uma perspectiva cultural ocidental, empregando conjuntos de dados predominantemente em inglês. Essa delimitação linguística justifica-se por alinhar o domínio de teste aos corpora de pré-treinamento dos modelos de linguagem avaliados. O âmbito de aplicação inclui a análise de textos literários convencionais e a compreensão de narrativas em mídias visuais, em especial anúncios publicitários estáticos. Entre os públicos-alvo e as aplicações práticas estão moderadores de conteúdo em redes sociais, sistemas de tutoria inteligente para escrita e geradores de conteúdo persuasivo.

Em termos de desempenho, os autores observaram que os modelos originais exibiam um viés que priorizava termos frequentes, o que dificultava a identificação de relações simbólicas menos comuns. A introdução da técnica de re-ranqueamento corrigiu essa tendência, resultando em um aumento expressivo na precisão das interpretações, especialmente para símbolos que dependem do contexto. Essa estratégia de re-ranqueamento funciona ajustando a ordem das respostas sugeridas pelo modelo para neutralizar o viés de frequência, que tende a favorecer conceitos muito comuns nos dados de treinamento em detrimento de significados mais raros. Tecnicamente, a abordagem normaliza a probabilidade de uma palavra ser um símbolo dividindo-a pela sua probabilidade geral de ocorrência no idioma (probabilidade a priori), o que permite destacar associações simbólicas genuínas e evitar que termos genéricos dominem o topo da lista de predições.

Graças a esse ajuste, os modelos mais avançados atingiram um patamar de acerto comparável ao humano, superando-o inclusive na decodificação de símbolos literários convencionais. No conjunto de dados voltado para publicidade, o desempenho dos modelos após a correção também se aproximou significativamente dos resultados obtidos por pes-

[1] ⟨https://huggingface.co/docs/transformers⟩
[2] ⟨https://radimrehurek.com/gensim/⟩

soas.

No entanto, os autores identificam limitações relacionadas ao viés linguístico e cultural, uma vez que o estudo adota predominantemente uma perspectiva anglo-saxônica e eurocêntrica. O conjunto de dados apresenta cobertura restrita de símbolos, inclusive no âmbito da tradição literária inglesa, e a análise de simbolismo situado limita-se a anúncios publicitários estáticos, excluindo textos de longa duração e vídeos. Além disso, é desconsiderada a evolução simbólica em narrativas complexas, pois a abordagem emprega pares de palavras tratados como invariantes ao contexto literário.

Em termos de objetivo, ambos os trabalhos investigam se modelos de linguagem encapsulam conhecimento suficiente para lidar com significados não literais e simbólicos. A diferença está na natureza da tarefa: Guo et al. buscam verificar se os LLMs conseguem predizer o significado simbólico correto de um objeto dado um par significante-significado, enquanto o presente trabalho busca gerar interpretações simbólicas livres para objetos em contos completos e verificar se essas interpretações são indistinguíveis das produzidas por humanos. Metodologicamente, ambos utilizam LLMs e recorrem à avaliação com base em referências humanas. O presente trabalho diferencia-se por solicitar ao modelo interpretações textuais abertas fundamentadas em passagens extraídas via correferência, e por conduzir um experimento perceptual com participantes humanos, em vez de medir acurácia contra um gabarito fechado. (GUO; HWA; KOVASHKA, 2023)

3.2 Evaluating Large Language Models for Narrative Topic Labeling

O estudo de Piper e Wu (2025) busca avaliar a capacidade dos modelos de linguagem de grande escala (LLMs) em identificar e rotular tópicos em textos de caráter narrativo. O objetivo central da investigação é verificar se os LLMs podem superar ou igualar o desempenho humano na interpretação de nuances complexas presentes em obras de ficção e notícias, superando as limitações de métodos tradicionais de aprendizado não supervisionado. A pesquisa foca em resolver o desafio da interpretação de contextos implícitos e linguagens figuradas, elementos que dificultam a extração automatizada de temas em

domínios puramente narrativos.

A metodologia fundamenta-se em uma análise comparativa entre diferentes arquiteturas de modelos, contrastando modelos de fronteira, como o GPT-4o, com modelos de pesos abertos, incluindo Llama3 e Gemma2. A rotulação de tópicos narrativos ocorre por meio de uma estratégia de prompt zero-shot (capacidade emergente dos LLMs discutida na Fundamentação Teórica), na qual o modelo realiza a tarefa sem exemplos prévios fornecidos na instrução. O prompt foi estruturado para que os modelos gerassem exatamente três palavras-chave por trecho, organizadas em ordem decrescente de generalidade, visando mitigar problemas de escala e garantir que os tópicos capturassem tanto o tema central quanto nuances mais específicas do texto.

Para validar a qualidade das predições, os autores implementaram um sistema de votação ranqueada através da plataforma Prolific[3], envolvendo a participação de 200 leitores independentes. Nesse processo, cada participante avaliava um trecho textual e ordenava cinco opções de rótulos (provenientes dos quatro modelos de linguagem e de um anotador humano) de acordo com sua preferência, do melhor para o pior. Essa abordagem de validação por múltiplos leitores permitiu verificar se as predições automatizadas eram estatisticamente equivalentes ou superiores às interpretações humanas em diferentes gêneros, como notícias e ficção.

Devido à vasta diversidade de termos gerados pelos modelos de linguagem, os autores aplicaram uma etapa de agregação de tópicos utilizando o modelo GPT-4o1 para consolidar sinônimos e o modelo de embeddings GloVe para o agrupamento semântico de termos em vetores dimensionais. Toda a experimentação foi conduzida sobre um corpus diversificado: aproximadamente 700 romances de domínio público do século XIX disponibilizados pela Chadwyck-Healey, e 6.722 artigos de notícias do Global News Dataset[4], cobrindo veículos como ABC News, Al Jazeera English, BBC News e The Times of India. Todas as anotações produzidas no estudo foram disponibilizadas publicamente pelos autores[5].

Quanto ao contexto e escopo de aplicação, o estudo concentra-se no domínio das

[3] ⟨https://www.prolific.com⟩
[4] ⟨https://www.kaggle.com/datasets/everydaycodings/global-news-dataset⟩
[5] ⟨https://doi.org/10.5683/SP3/MHJRIO⟩

narrativas, explorando tanto a ficção literária quanto o relato factual em notícias. O público-alvo principal da ferramenta são pesquisadores da área de humanidades digitais que necessitam de métodos escaláveis para interpretar grandes repositórios de documentos culturais e históricos. É fundamental destacar que a pesquisa foi conduzida exclusivamente no idioma inglês, utilizando um corpus de romances britânicos do século XIX e uma base global de notícias contemporâneas. Essa especificidade linguística e cultural é um ponto de distinção importante para trabalhos que buscam adaptar tais tecnologias para o contexto do Português do Brasil, dado que a eficácia na captura de nuances pode variar entre diferentes línguas e corpora de treinamento.

Se tratando dos resultados e do desempenho alcançado, a validação foi realizada por meio de um sistema de votação ranqueada, no qual 200 participantes independentes classificaram a qualidade dos rótulos gerados em uma escala de preferência de 1 a 5. Os dados coletados revelaram que os modelos de linguagem de grande escala apresentam um desempenho no mesmo nível de leitores humanos em notícias e superior em textos de ficção. Em termos numéricos, o modelo Gemma2 apresentou a melhor classificação média no gênero de ficção, com um ranking de 2,25, enquanto a média humana para os mesmos textos foi de 3,23. Já no domínio das notícias, o GPT-4o obteve a melhor aceitação pelos usuários, alcançando um ranking médio de 2,40, comparado aos 2,79 obtidos pelos anotadores humanos.

Além das métricas de ranqueamento, a pesquisa observou uma elevada consistência entre as escolhas automatizadas e as humanas, com as anotações feitas por pessoas coincidindo com a saída de pelo menos um modelo em 72% dos casos de ficção e 50% nos de notícias. Esse alto grau de similaridade semântica sugere que os modelos atuais são capazes de capturar a essência temática de histórias complexas sem a necessidade de etapas adicionais de pré-processamento. Por fim, a análise qualitativa indicou que, embora os modelos gerem uma diversidade maior de termos, o uso de técnicas de agregação permite consolidar esses resultados em tópicos inteligíveis que superam significativamente a clareza dos métodos tradicionais de agrupamento estatístico.

Apesar dos avanços apresentados, o estudo de Piper e Wu (2025) possui limitações importantes, como a restrição a apenas dois gêneros narrativos e o uso de um conjunto

reduzido de modelos de linguagem, o que impede a generalização total dos resultados para outros contextos ou arquiteturas. Além disso, os autores destacam que o processo de agregação de tópicos para reduzir a redundância de termos gerados pelos modelos pode introduzir erros interpretativos e ainda precisa ser explorado como um domínio próprio de pesquisa. Outro fator restritivo é o elevado custo computacional e financeiro exigido para o processamento de grandes volumes de texto, o que pode limitar a viabilidade da técnica para projetos com recursos de hardware menos robustos

Em termos de objetivo, ambos os trabalhos investigam se LLMs conseguem realizar análise interpretativa de textos narrativos com qualidade comparável à humana. A diferença está na natureza da tarefa: Piper e Wu avaliam a rotulação de tópicos narrativos em grandes corpora de ficção e notícias, enquanto o presente trabalho foca na geração de interpretações simbólicas de objetos específicos em contos curtos, buscando verificar se essas interpretações são indistinguíveis das produzidas por humanos. Metodologicamente, ambos adotam prompting zero-shot e recorrem a avaliadores humanos para validar os resultados. O presente trabalho diferencia-se por utilizar resolução de correferência para extrair trechos relevantes de forma direcionada (em vez de processar blocos fixos de texto) e por conduzir um experimento perceptual em que participantes julgam a origem das interpretações (humano ou IA) e sua plausibilidade em escala Likert. (PIPER; WU, 2025)

3.3 Are Large Language Models Good Annotators?

Mohta et al. (2023) investigam a viabilidade de empregar grandes modelos de linguagem como substitutos de anotadores humanos em tarefas de processamento de linguagem natural. Publicado na 37ª edição da Conferência sobre Sistemas de Processamento de Informação Neural (NeurIPS 2023), o trabalho avalia o desempenho zero-shot de diferentes LLMs de código aberto para determinar se esses modelos são capazes de gerar rótulos com qualidade comparável à de anotadores humanos, reduzindo os custos e o tempo associados à construção de conjuntos de dados rotulados.

O ponto de partida do estudo é a constatação de que a anotação manual de dados em larga escala é cara, demorada e frequentemente exige conhecimento especializado no

domínio em questão. Com o crescimento dos modelos de linguagem pré-treinados e a demonstração de capacidades zero-shot expressivas por parte de modelos como o GPT-3, surgiu a hipótese de que LLMs poderiam gerar rótulos de qualidade suficiente para treinar modelos menores, tornando a coleta de dados rotulados por humanos dispensável ou ao menos reduzida. O artigo busca testar essa hipótese de forma sistemática e abrangente, incluindo também a dimensão multimodal da anotação.

A avaliação abrange modelos textuais de código aberto (Vicuna-13b, Vicuna-7b, LLaMA-13b, LLaMA-7b e OpenLLaMA-13b) e modelos multimodais baseados na arquitetura InstructBLIP, que utiliza o BLIP-2 com um codificador de imagens ViT-G/14 e um módulo Q-Former para mapear representações visuais ao espaço de embeddings do LLM. Para quantificar o impacto das anotações geradas, os autores treinam um classificador supervisionado MMBT (Multimodal BiTransformers) com labels produzidos tanto por humanos quanto pelos LLMs, e comparam os desempenhos obtidos nos conjuntos de teste.

Durante a inferência, os modelos foram configurados com limite máximo de 1024 tokens, temperatura de 0.8 e nucleus sampling com p = 0.9, com a opção `do_sample` ativada para promover diversidade nas respostas. As saídas dos modelos foram categorizadas por meio de correspondência via expressões regulares. Para os experimentos multimodais, as imagens foram redimensionadas para 224x244 pixels, e o Q-Former extraiu representações de dimensão 32x768.

Os experimentos cobrem quatro conjuntos de dados: MM-IMDB, voltado para classificação de gênero cinematográfico a partir de sinopses e pôsteres; Hateful Memes, para detecção de conteúdo odioso em pares imagem-texto; dois datasets anotados internamente pela equipe (IAD-1 e IAD-2); e o XNLI, utilizado para avaliação multilíngue em francês, holandês e inglês. O foco principal da pesquisa são tarefas de anotação binária ou de classificação com rótulos predefinidos, em inglês.

Entre os modelos avaliados, o Vicuna-13b demonstrou o melhor desempenho geral ao longo das diferentes tarefas de anotação em inglês. Em particular, o Vicuna-7b-v1.5 surpreendeu ao superar o modelo supervisionado em cerca de 2% nos Hateful Memes, embora com taxa de resposta de apenas 60% dos itens durante a inferência. Contudo, ao

substituir os labels humanos pelos labels gerados por LLMs no treinamento do MMBT, observou-se queda consistente de desempenho: de 0,64 para 0,60 nos Hateful Memes e de 0,84 para 0,75 no MM-IMDB. A inclusão da modalidade visual via InstructBLIP não produziu ganho consistente sobre os modelos textuais. Na avaliação multilíngue com XNLI, o desempenho do Vicuna-13b caiu de 0,53 em inglês para 0,41 em francês e 0,42 em holandês, evidenciando a fragilidade dos modelos fora do inglês.

Os autores reconhecem que os LLMs introduzem ruído nos rótulos gerados, o que resulta em perda de desempenho ao treinar modelos supervisionados. A capacidade de resposta dos modelos também foi irregular, com alguns deixando de responder a uma parcela dos itens durante a inferência. O desempenho em línguas diferentes do inglês demonstrou queda expressiva, o que restringe a aplicabilidade da abordagem a contextos majoritariamente anglófonos. Adicionalmente, o trabalho não explora tarefas de geração textual aberta, limitando-se a cenários de classificação com saída predefinida.

Ao contrário de Mohta et al. (2023), que empregam LLMs como anotadores de classificação com rótulos fixos e avaliam o desempenho por acurácia em relação a um gabarito, o presente trabalho utiliza modelos de linguagem para a geração de interpretações textuais abertas no contexto da análise simbólica de contos literários. A tarefa proposta não possui resposta certa ou errada, o que representa um desafio qualitativamente distinto. A avaliação não se baseia em métricas de desempenho de modelos treinados, mas em um experimento de avaliação cega conduzido com participantes humanos, que julgam a plausibilidade das interpretações e tentam identificar sua origem (humana ou gerada por IA). Esse formato permite investigar até que ponto a qualidade percebida das saídas de LLMs se equipara à de anotadores humanos em uma tarefa subjetiva e contextualmente rica, lacuna não abordada pelo artigo de referência. (MOHTA et al., 2023)

3.4 Comparing Large Language Models and Human Annotators in Latent Content Analysis of Sentiment, Political Leaning, Emotional Intensity and Sarcasm

Bojić et al. (2025) investigam em que medida grandes modelos de linguagem conseguem replicar o desempenho de anotadores humanos em tarefas de análise de conteúdo latente em texto, cobrindo quatro dimensões: análise de sentimento, detecção de inclinação política, avaliação de intensidade emocional e detecção de sarcasmo. Publicado na Scientific Reports (Nature Portfolio), o estudo busca quantificar a confiabilidade, a consistência temporal e a qualidade comparativa das anotações geradas por LLMs em relação às produzidas por especialistas humanos.

O ponto de partida do trabalho é a ausência de avaliações abrangentes que comparem LLMs com anotadores humanos de forma simultânea em múltiplas dimensões da análise de conteúdo latente. Questões como confiabilidade entre anotadores, consistência das respostas dos modelos ao longo do tempo e a proximidade qualitativa de suas análises em relação às humanas permanecem pouco exploradas na literatura, especialmente em estudos que considerem diferentes modelos e diferentes tipos de conteúdo ao mesmo tempo.

Para medir a confiabilidade entre anotadores e entre modelos, o estudo adota o alpha de Krippendorff, métrica indicada para dados ordinais com múltiplos avaliadores. A consistência temporal de cada LLM foi mensurada por meio do Coeficiente de Correlação Intraclasse (ICC), calculado a partir de três sessões de anotação distintas sobre o mesmo conjunto de itens. As diferenças de desempenho entre grupos foram analisadas por ANOVAs de uma via e testes-t com correção de Bonferroni para comparações múltiplas.

Os sete LLMs avaliados (GPT-3.5-turbo-16k, GPT-4, GPT-4o, GPT-4o-mini, Gemini 1.5 Pro, Llama-3.1-70B e Mixtral 8x7B) foram acessados via API ou interface web sob condições padronizadas de horário e configuração. Um oitavo modelo, denominado Hard Prompt GPT-4o, utilizou uma instrução adicional que simulava deliberação entre três agentes virtuais, com o objetivo de aproximar o processo do raciocínio humano em

grupo. A análise estatística foi conduzida com Python (pandas, seaborn, matplotlib) e SPSS versão 26.

O corpus foi composto por 100 sentenças em inglês, 25 por dimensão, elaboradas com base em datasets reconhecidos na área, como o Stanford Sentiment Treebank, o EmoBank e o Sarcasm Corpus V2. Do lado humano, participaram 33 anotadores com formação acadêmica avançada, distribuídos por 18 países europeus, todos integrantes da rede COST Action CA21129 (OPINION). O contexto de aplicação é a análise de textos curtos e isolados, em inglês, com foco em dimensões afetivas e discursivas de uso frequente em pesquisa de opinião pública e análise de redes sociais.

Em análise de sentimento, tanto humanos quanto LLMs apresentaram alta confiabilidade interna, com alpha de Krippendorff próximo a 0,95 para ambos os grupos. Na inclinação política, os LLMs (alpha médio $\approx 0{,}80$) superaram os humanos (alpha = 0,55) em consistência, possivelmente por adotarem critérios mais uniformes. Na intensidade emocional, os LLMs mostraram maior consistência entre si, mas os humanos atribuíram escores significativamente mais altos (média 3,44 contra 3,19 dos LLMs), sugerindo que os modelos subestimam a intensidade emocional percebida. Na detecção de sarcasmo, ambos os grupos apresentaram baixa confiabilidade (alpha $\approx 0{,}25$), evidenciando a dificuldade intrínseca da tarefa. Em termos de consistência temporal, os ICCs dos LLMs foram elevados na maioria das dimensões, com exceção notável do Gemini 1.5 Pro na inclinação política (ICC $= -0{,}412$) e no sarcasmo (ICC $= -0{,}065$).

Os autores reconhecem que a amostra de 33 anotadores, embora adequada para análise estatística, pode não capturar a diversidade de interpretações influenciadas por diferenças culturais e individuais. O uso de sentenças curtas e isoladas pode não refletir a complexidade linguística de textos narrativos mais extensos. Além disso, como os LLMs avaliados estão sujeitos a atualizações e ajustes contínuos por parte de seus desenvolvedores, a reprodutibilidade dos resultados em janelas de tempo distintas não é garantida.

Enquanto Bojić et al. (2025) avaliam LLMs em tarefas com dimensões bem definidas, escalas ordinais e sentenças isoladas como unidade de análise, o presente trabalho propõe a comparação em um contexto de geração textual aberta, sem gabarito fixo e com textos narrativos completos como insumo. A tarefa de interpretação simbólica literária

exige raciocínio contextual sobre contos inteiros, o que representa um nível de complexidade distinto das frases avaliadas no artigo de referência. Adicionalmente, a avaliação não se apoia em métricas de confiabilidade entre anotadores, mas em um experimento de avaliação cega no qual participantes julgam a plausibilidade das interpretações e tentam identificar sua origem. Esse formato coloca a questão da qualidade percebida em primeiro plano, aspecto não explorado por Bojić et al., e permite investigar até que ponto interpretações geradas por LLMs são percebidas como equivalentes às humanas em uma tarefa subjetiva e sem resposta única correta. (BOJIĆ et al., 2025)

3.5 Protagonist Tagger

O protagTagger é uma ferramenta de código aberto desenvolvida por Łajewska e Wróblewska (2021), voltada para a anotação semântica de textos literários longos, como romances. O objetivo central da pesquisa é detectar entidades do tipo "pessoa"e atribuir-lhes identidades únicas (Entity Linkage), permitindo diferenciar personagens (especialmente protagonistas) em narrativas complexas onde o mesmo indivíduo pode ser mencionado de diversas formas. O resultado final do processo é o texto anotado, deixando claro qual personagem é referenciado em cada citação. (ŁAJEWSKA; WRÓBLEWSKA, 2021)

A metodologia do sistema compreende dois estágios principais: o Reconhecimento de Entidade Nomeada (NER) e a Desambiguação de Entidade Nomeada (NED). Para o NER, as autoras utilizaram um modelo pré-treinado da biblioteca spaCy[6], que foi refinado (fine-tuned) com sentenças de romances clássicos para melhorar sua recall no domínio literário. No estágio de NED, o sistema utiliza um algoritmo de emparelhamento de texto aproximado (approximate string matching) baseado na distância de Levenshtein para associar os nomes detectados a uma lista de personagens pré-definida, extraída via Wikipedia ou fornecida manualmente.

O trabalho foca exclusivamente no domínio literário, abordando desafios que modelos treinados em notícias ou blogs geralmente não resolvem, como o uso de diminutivos, apelidos e personagens que compartilham o mesmo sobrenome. O escopo inclui o tratamento de títulos pessoais (Sr., Sra.) para identificar gênero e a análise de diminutivos por

[6] ⟨https://spacy.io/⟩

meio de um dicionário externo com mais de 3.300 formas de nomes, visando garantir que variações como "Lizzy"sejam corretamente vinculadas a "Elizabeth Bennet".

O protagTagger demonstrou alto desempenho, alcançando precisão e revocação acima de 83% nos conjuntos de teste. O processo de fine-tuning do modelo NER foi particularmente bem-sucedido, elevando a revocação para mais de 95%, corrigindo falhas de modelos padrão que não reconheciam nomes próprios em contextos literários específicos. Como resultado prático, as autoras disponibilizaram um corpus de 13 romances clássicos anotados, contendo mais de 35.000 menções de personagens.

O protagTagger é relevante para o presente trabalho por evidenciar um desafio central do domínio literário: a mesma entidade pode ser referenciada de formas variadas ao longo do texto — por apelidos, diminutivos ou pronomes. Esse problema se replica no rastreamento de objetos simbólicos recorrentes, onde um mesmo objeto pode aparecer via pronome ou expressão alternativa. Por isso, o pipeline proposto neste trabalho adota resolução de correferência (via fastcoref) para agrupar menções distintas a um mesmo objeto, em vez de depender de listas pré-definidas. A diferença central está no tipo de entidade buscada: enquanto o protagTagger foca em pessoas nomeadas, o presente trabalho lida com substantivos concretos genéricos (como banco, janela, fogo), para os quais o NER convencional não é suficiente. (ŁAJEWSKA; WRÓBLEWSKA, 2021)

4 Metodologia

Este capítulo descreve a metodologia adotada no presente trabalho. A Seção 4.1 apresenta a metodologia de pesquisa, detalhando o design do experimento de avaliação e os critérios de análise dos resultados. A Seção 4.2 descreve a metodologia de desenvolvimento, com foco na arquitetura da pipeline de extração e geração de interpretações simbólicas.

4.1 Metodologia de Pesquisa

A presente pesquisa adota a tipologia metodológica apresentada por Marconi e Lakatos (2017) (MARCONI; LAKATOS, 2017) e classifica-se em quatro dimensões. Quanto à natureza, configura-se como **pesquisa aplicada**, pois envolve o desenvolvimento de uma pipeline computacional que articula técnicas de processamento de linguagem natural e modelos de linguagem de grande escala para gerar interpretações simbólicas a partir de contos literários. Quanto aos procedimentos técnicos, é **pesquisa bibliográfica**, na medida em que reúne e sistematiza a produção já publicada sobre processamento de linguagem natural, LLMs e análise simbólica de narrativas, fornecendo o embasamento teórico que sustenta as decisões da pipeline e do protocolo de avaliação, e também **pesquisa experimental**, pois submete as interpretações geradas a um experimento de avaliação cega no qual a variável manipulada é a origem das interpretações (humana ou automática) e as variáveis observadas são a autoria percebida e a plausibilidade atribuída pelos participantes. Quanto aos objetivos, enquadra-se como **pesquisa exploratório-descritiva**, pois busca descrever de forma detalhada um fenômeno ainda pouco investigado, a saber, a qualidade percebida de interpretações simbólicas geradas por LLMs em comparação às produzidas por leitores humanos, sem o intuito de testar hipóteses causais previamente formuladas. Quanto à abordagem, trata-se de **pesquisa quali-quantitativa**, pois a análise dos dados articula descrições quantitativas, como a taxa de acerto dos participantes na identificação de origem e os escores médios de plausibilidade atribuídos em escala Likert, e descrições qualitativas, voltadas aos padrões nos julgamentos e aos casos em que

a identificação de origem foi sistematicamente equivocada ou em que os escores divergiram de modo expressivo entre as duas condições.

O corpus é composto por três contos curtos de domínio público, redigidos em inglês. A escolha do inglês não é arbitrária: a etapa de resolução de correferência do pipeline exige um modelo treinado nesse idioma. A única solução com desempenho satisfatório testada durante o desenvolvimento (o fastcoref com o modelo biu-nlp/f-coref) utiliza o corpus OntoNotes em inglês para treinamento. Alternativas para o português avaliadas, como o coreferee, mostraram-se incompatíveis com versões atuais do spaCy ou sem suporte ao idioma, inviabilizando o uso de textos em português nesta etapa.

As anotações humanas são coletadas de três leitores independentes por conto. Cada anotador lê o conto na íntegra e escreve entre uma e três interpretações simbólicas para os objetos que identifica como simbolicamente relevantes no texto, cada uma formulada como uma frase curta. O formato é idêntico ao das saídas geradas pelos LLMs, de forma que as respostas não possam ser identificadas como humanas ou automáticas com base em características formais como extensão ou estrutura. Nenhum dos anotadores tem acesso a análises externas sobre os contos: o insumo é apenas o texto em si.

A geração de interpretações automáticas emprega três modelos de linguagem de famílias distintas: ChatGPT (GPT-4o), Claude (Sonnet) e Gemini (2.0 Flash). A opção por múltiplos modelos decorre da evidência de que LLMs distintos produzem resultados diferentes para a mesma tarefa; restringir a análise a um único modelo limitaria o escopo das inferências e tornaria os resultados dependentes de características específicas de um único sistema (MOHTA et al., 2023).

O experimento de avaliação segue um protocolo de avaliação cega. O formulário é dividido em uma seção por conto. Cada seção apresenta o texto completo do conto no início, seguido das interpretações humanas e automáticas intercaladas, cada uma acompanhada do objeto a que se refere, sem qualquer indicação de origem. Participam aproximadamente vinte pessoas, que leem o conto na íntegra antes de iniciar a avaliação. Para cada interpretação, o participante realiza dois julgamentos: indica se percebe a interpretação como de autoria humana ou gerada por IA, e atribui uma nota de plausibilidade em escala Likert de 1 a 5, onde 1 representa “totalmente implausível” e 5 representa “totalmente

plausível". A escala Likert de 5 pontos é o formato de avaliação mais utilizado em estudos de geração de linguagem natural, e os critérios de avaliação devem ser explicitamente definidos e adequados à tarefa em questão (LEE et al., 2019).

A análise dos dados é estruturada em duas frentes. Quantitativamente, compara-se a taxa de acerto na identificação de origem e os escores médios de plausibilidade entre interpretações humanas e automáticas, verificando também se há correlação entre a origem percebida e as notas atribuídas. Para medir o grau de concordância entre os participantes nos seus julgamentos, emprega-se o alpha de Krippendorff, métrica adequada para avaliar confiabilidade entre múltiplos avaliadores em dados categóricos e ordinais (BOJIĆ et al., 2025). Qualitativamente, examinam-se os casos de erro sistemático na identificação de origem e as divergências expressivas de plausibilidade, buscando padrões por conto, por objeto ou por modelo empregado.

A Figura 4.1 sintetiza as três grandes fases da pesquisa: a coleta do corpus e das anotações humanas, o processamento automático pela pipeline computacional, e a avaliação cega conduzida com participantes.

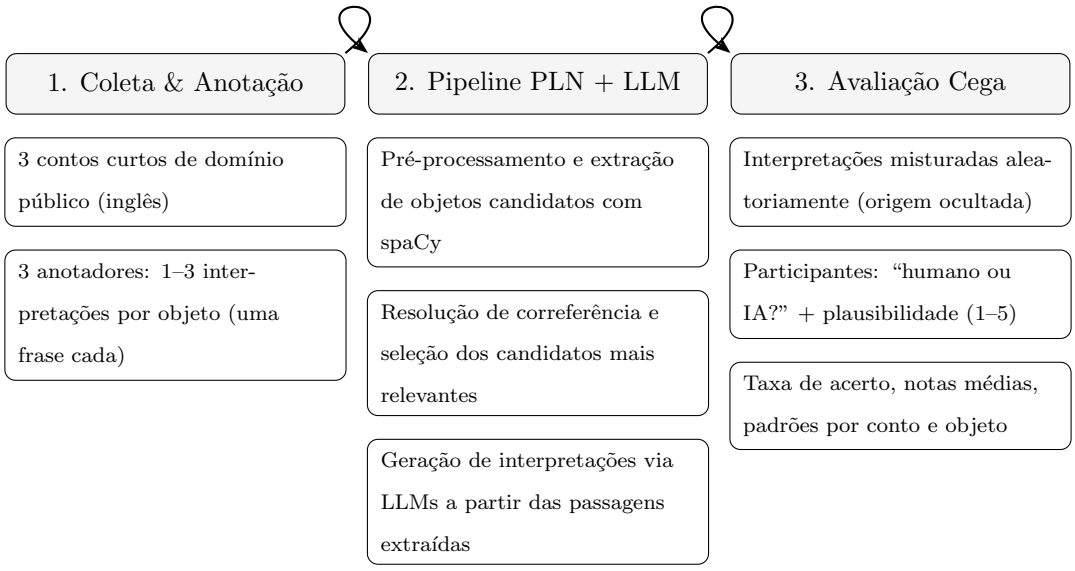

Figura 4.1: Visão geral da metodologia de pesquisa.

4.2 Desenvolvimento da Pipeline

O sistema é implementado como uma pipeline sequencial de sete etapas, cujo fluxo vai do texto bruto do conto até um conjunto de interpretações simbólicas geradas por LLMs. O desenho geral da pipeline foi construído com base nos trabalhos relacionados discutidos no Capítulo 3, especialmente no protagTagger (ŁAJEWSKA; WRÓBLEWSKA, 2021), que evidenciou a importância de tratar as diferentes menções a uma mesma entidade ao longo do texto por meio de agrupamento de referências, princípio aqui estendido do rastreamento de personagens para o de objetos concretos com potencial simbólico, por meio de resolução de correferência. Também foram incorporadas escolhas alinhadas aos demais trabalhos revisados, como o uso de prompting zero-shot com múltiplos modelos e a avaliação por juízes humanos em vez de comparação com um gabarito fechado. Vale registrar que o desenvolvimento da pipeline antecedeu o desenho do experimento de avaliação: apenas em uma etapa posterior, durante a definição do protocolo com participantes, ficou evidente que testar diretamente o impacto do pré-processamento (por exemplo, comparando interpretações geradas com e sem a filtragem prévia) exigiria um volume de leitura por respondente incompatível com o corpus de contos curtos adotado, no qual a vantagem da filtragem sobre o texto integral tende a ser menos pronunciada. Por essa razão, essa comparação não é realizada aqui e permanece indicada como trabalho futuro na Seção 7.1. As etapas e suas principais atividades estão ilustradas na Figura 4.2 e são descritas individualmente ao longo desta seção.

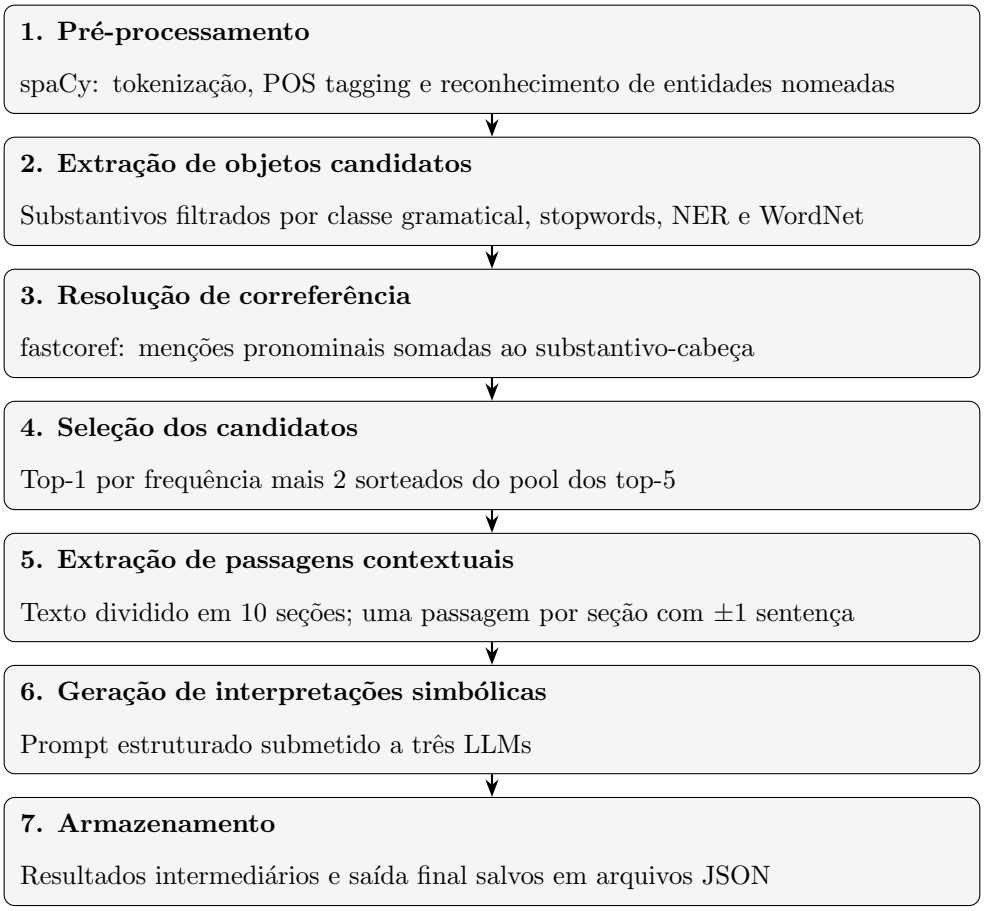

Figura 4.2: Etapas da pipeline de extração de objetos e geração de interpretações simbólicas.

A implementação utiliza Python 3[7] como linguagem principal, escolhida por concentrar as principais bibliotecas maduras de PLN e permitir a integração das diferentes etapas em um único fluxo de execução; o código-fonte completo está disponível em repositório público[8]. Para o processamento linguístico, emprega-se o spaCy[9] com o modelo *en_core_web_sm*[10], responsável pela tokenização, etiquetagem morfossintática (*POS tagging*) e reconhecimento de entidades nomeadas (NER); sua contribuição está em oferecer um pipeline unificado que entrega essas três anotações de forma integrada, o que evita a combinação manual de ferramentas separadas para cada tarefa e ainda expõe a árvore sintática usada na identificação do substantivo-cabeça de cada cluster de correferência. A

[7] ⟨https://www.python.org/⟩
[8] ⟨https://github.com/AngryBeanTrevota/TCCExtracaoSimbolos⟩
[9] ⟨https://spacy.io/⟩
[10] ⟨https://spacy.io/models/en#en_core_web_sm⟩

resolução de correferência é realizada pelo fastcoref[11] com o modelo biu-nlp/f-coref[12], escolhido por ser uma das poucas soluções estáveis para correferência em inglês compatíveis com versões recentes do spaCy e por viabilizar diretamente o agrupamento de menções pronominais aos substantivos, contribuição sem a qual a contagem de frequência de cada objeto seria substancialmente subestimada. A filtragem de substantivos que denotam pessoas faz uso da rede lexical WordNet[13], acessada via NLTK[14], o que permite substituir listas manuais de papéis sociais por uma verificação baseada em hiperônimos e amplia a cobertura da filtragem sem exigir manutenção manual de vocabulário. A versão da biblioteca Transformers[15] é fixada entre 4.11 e abaixo de 5.0 para garantir compatibilidade com o fastcoref, dependência técnica necessária para o funcionamento do modelo de correferência. Na etapa de interpretação simbólica, os modelos empregados são ChatGPT (GPT-4o)[16], Claude (Sonnet)[17] e Gemini (2.0 Flash)[18]. A seleção desses três modelos, todos pertencentes a famílias distintas de LLMs de fronteira, contribui para ampliar a base de comparação da análise e evita que os resultados fiquem restritos ao comportamento de um único sistema, seguindo prática presente em estudos comparativos recentes sobre desempenho de LLMs em tarefas de análise textual. (BOJIĆ et al., 2025)

A **etapa 1, pré-processamento**, aplica o pipeline do spaCy ao texto completo do conto, extraindo *tokens*, lemas, classes gramaticais e entidades nomeadas. Essa anotação fornece os atributos necessários para todas as etapas seguintes, em especial o *POS tagging* utilizado na filtragem de objetos e as entidades nomeadas usadas na exclusão de menções a pessoas.

A **etapa 2, extração de objetos candidatos**, percorre o texto anotado e retém os *tokens* com $pos\_ == \text{“}NOUN\text{”}$, descartando *stopwords* e pontuação e agrupando formas flexionadas pelo lema (ex: *matches* e *match* $\rightarrow$ *match*). A filtragem de pessoas é feita por dois critérios complementares: presença do *token* nos *clusters* de entidades com *label PERSON* detectados pelo NER, e verificação de que o lema do substantivo possui

[11] ⟨https://github.com/shon-otmazgin/fastcoref⟩
[12] ⟨https://huggingface.co/biu-nlp/f-coref⟩
[13] ⟨https://wordnet.princeton.edu/⟩
[14] ⟨https://www.nltk.org/⟩
[15] ⟨https://huggingface.co/docs/transformers⟩
[16] ⟨https://chat.openai.com/⟩
[17] ⟨https://claude.ai/⟩
[18] ⟨https://gemini.google.com/⟩

person.n.01 na cadeia de hiperônimos do primeiro *synset* do WordNet, usando *cache* para evitar consultas repetidas. Essa varredura direta é mantida como etapa autônoma mesmo com a resolução de correferência subsequente, pois o fastcoref não é um contador de frequência, apenas agrupa expressões que referenciam a mesma entidade. Substantivos que ocorrem sem pronomes associados podem não integrar nenhum *cluster*, ficando fora do alcance da correferência. A varredura *token* a *token*, por ser exaustiva, garante a contagem de todas as menções nominais e serve de *baseline* sobre o qual a correferência opera.

A **etapa 3, resolução de correferência**, usa os *clusters* retornados pelo fastcoref, onde cada *cluster* agrupa *spans* de texto que se referem à mesma entidade, como o *cluster* `["the match", "it", "another", "They"]`. Como o fastcoref retorna frases completas e não palavras isoladas, é necessário identificar o substantivo principal de cada *span* para saber qual contador atualizar; essa identificação é feita via `span.root`, atributo do spaCy que aponta para a cabeça sintática da *noun phrase* segundo a *parse tree* já calculada. A seleção do substantivo-cabeça segue dois passos: primeiro, verifica-se se algum *span* do *cluster* tem raiz associada a uma pessoa (por NER ou WordNet); em caso positivo, o *cluster* inteiro é descartado, evitando que pronomes da personagem principal sejam atribuídos a substantivos genéricos como *thing* ou *body*. Segundo, o *root* do primeiro *span* não-pessoa é adotado como cabeça do *cluster*. Por fim, apenas os *spans* pronominais são somados ao contador desse substantivo, pois as menções nominais já foram capturadas pela varredura direta e somar todos os *spans* resultaria em dupla contagem.

A **etapa 4, seleção de candidatos**, garante a inclusão do objeto mais frequente no texto e sorteia mais dois de forma aleatória dentre os quatro seguintes do *ranking* de frequência, formando um conjunto de três candidatos por conto.

A **etapa 5, extração de passagens contextuais**, divide o texto em dez seções de tamanho igual e, para cada seção que contenha ao menos uma menção ao candidato (direta ou coreferente), extrai um único trecho: a sentença da menção mais cedo naquela seção, acrescida de uma sentença de contexto anterior e uma posterior. Janelas estreitas de contexto tendem a capturar bem o significado local de uma palavra sem introduzir ruído de passagens distantes (LISON; KUTUZOV, 2017). Quando o objeto é mencionado

múltiplas vezes dentro de uma mesma seção, apenas a primeira ocorrência é aproveitada e as demais são descartadas, evitando acúmulo de trechos repetitivos da mesma região do texto. O número de passagens geradas é, portanto, igual ao número de seções em que o objeto aparece, com máximo de dez. Essa contenção se justifica pelo fato de que LLMs têm uma janela de contexto efetiva bem inferior ao limite máximo declarado, com degradação expressiva de desempenho quando o volume de *tokens* submetido é alto (PAULSEN, 2025).

A **etapa 6, geração de interpretações simbólicas via LLM**, submete as passagens extraídas aos três modelos selecionados. O *design* do *prompt* segue três padrões catalogados por White et al. (2023): o *Persona Pattern*, que instrui o modelo a atuar como analista literário; o *Template Pattern*, que fornece um modelo concreto de formato de saída com exemplos; e o *Context Manager Pattern*, que instrui o modelo a ignorar conhecimento externo sobre o conto e seu contexto histórico, garantindo que LLM e anotadores humanos partam das mesmas informações (WHITE et al., 2023). A prompt completa utilizada em todos os modelos é apresentada na Figura 4.3.

```
Você é um analista literário. Vou anexar um arquivo JSON com trechos
de um conto.

Cada chave no JSON representa um objeto, símbolo ou elemento
recorrente da história. Use apenas as passagens associadas à chave
correspondente para construir as interpretações. Por favor, ignore
qualquer conhecimento externo que você tenha sobre este conto, seu
autor ou contexto histórico.

Para cada chave, escreva de 1 a 3 interpretações simbólicas. Cada
interpretação deve ter apenas uma frase concisa. Foque no que o
elemento pode representar, sugerir ou evocar simbolicamente dentro do
contexto da história.

Apresente as chaves em ordem aleatória, não necessariamente a do
documento.

Mantenha os nomes das chaves exatamente como aparecem no JSON.

Escreva e formate exatamente no estilo do exemplo abaixo:

Casa:
a. Lugar de pertencimento e memória.
b. Espaço que guarda vínculos afetivos.
c. Desejo de voltar a um ponto de origem.

Botão:
a. Pequena ação capaz de mudar tudo.
b. Controle sobre algo maior do que parece.
c. Gatilho de transformação ou decisão.

Caderno:
a. Registro do que precisa ser lembrado.
b. Tentativa de organizar pensamentos e sentimentos.
c. Espaço onde a identidade vai sendo escrita.
```

Figura 4.3: Prompt utilizada para geração de interpretações simbólicas.

A **etapa 7, armazenamento**, persiste os resultados intermediários, como os

clusters de correferência e as passagens contextuais extraídas, e a saída final das LLMs em arquivos JSON no diretório de saída, permitindo inspeção e depuração a cada etapa da pipeline.

Toda a etapa de PLN é executada localmente, sem necessidade de GPU ou serviços em nuvem. O modelo biu-nlp/f-coref é carregado automaticamente pelo fastcoref a partir do repositório HuggingFace Transformers na primeira execução. O acesso aos LLMs é feito manualmente via interface web: o *script interpretar.py* processa o arquivo de passagens e gera um arquivo de texto com o *prompt* e a instrução de qual arquivo JSON anexar; o usuário cola esse conteúdo em cada um dos três modelos e registra as respostas. Essa abordagem dispensa o uso de chaves de API e permite que o experimento seja reproduzido sem custos adicionais de infraestrutura.

5 Experimento

Este capítulo descreve a execução concreta do protocolo definido no Capítulo 4. A Seção 5.1 detalha o corpus literário utilizado e a sua proveniência. A Seção 5.2 descreve a coleta das anotações humanas, incluindo as instruções repassadas aos anotadores e o formato de resposta esperado. A Seção 5.3 relata como a pipeline foi executada para gerar as interpretações automáticas e como cada conjunto de passagens foi submetido aos três LLMs. A Seção 5.4 explica o processo de montagem do instrumento de avaliação cega, incluindo o esquema de codificação que permitiu rastrear a origem de cada interpretação sem revelá-la aos participantes. Por fim, a Seção 5.5 descreve a aplicação do experimento e apresenta o perfil dos respondentes.

5.1 Corpus literário

O corpus reúne três contos curtos de domínio público, selecionados pela densidade de elementos narrativos com potencial simbólico e pela diversidade de autoria, época e tradição literária. Os textos selecionados foram *O Espelho*, de Machado de Assis (1882)[19], *A Pequena Vendedora de Fósforos*, de Hans Christian Andersen (1848)[20], e *A Máscara da Morte Vermelha*, de Edgar Allan Poe (1842)[21]. As versões em português foram obtidas, respectivamente, no acervo digital da Brigham Young University, no portal da Editora Wish e na Biblioteca Pública Municipal de Torres. Estas versões em português foram apresentadas aos anotadores humanos. As versões em inglês equivalentes, utilizadas como entrada da pipeline computacional, foram obtidas de repositórios também públicos. A divisão entre o idioma de leitura dos anotadores e o idioma de processamento da pipeline decorre da limitação técnica já discutida na Seção 4.1, dado que não há, até o momento, modelo de resolução de correferência estável para o português. Essa diferença de idioma

[19]Disponível em: ⟨https://machado.byu.edu/text/o-espelho/⟩.

[20]Disponível em: ⟨https://www.editorawish.com.br/blogs/contos-de-fadas-originais-completos-e-gratuitos/a-pequena-vendedora-de-fosforos-hans-christian-andersen-1848⟩.

[21]Disponível em: ⟨https://biblioteca.torres.rs.gov.br/wp-content/uploads/2022/06/poe-edgar-allan-a-mascara-da-morte-vermelha.pdf⟩.

entre o texto submetido aos anotadores e o texto processado pela pipeline não influencia o experimento nem a análise dos resultados. Os anotadores humanos escrevem suas interpretações em português, e os LLMs, embora processem o texto em inglês, recebem o prompt em português e devolvem as interpretações também em português, de modo que tanto as anotações humanas quanto as automáticas chegam aos participantes no mesmo idioma e sob o mesmo formato. A avaliação cega compara o conteúdo simbólico das frases, e não o idioma do texto-fonte sobre o qual cada interpretação foi produzida. Diferenças entre as traduções utilizadas (versões em português dos contos originalmente escritos em outros idiomas, como o dinamarquês no caso de Andersen e o inglês no caso de Poe, além do original em português de Machado, e as respectivas versões em inglês empregadas na pipeline computacional) também não afetam o estudo, uma vez que o objeto de análise é a plausibilidade percebida das interpretações e a distinção entre origens (humana ou automática), e não a fidelidade linguística das versões apresentadas.

5.2 Coleta das anotações humanas

A coleta das anotações foi conduzida por meio de formulário eletrônico construído na plataforma Google Forms, no qual os três contos foram disponibilizados em uma única apresentação, na íntegra. Foram recrutados três anotadores independentes, cada um responsável por anotar os três contos. Nenhum dos anotadores teve contato prévio com a saída da pipeline computacional ou com as interpretações geradas pelos LLMs, de modo que as respostas humanas foram produzidas exclusivamente a partir da leitura do texto.

A instrução fornecida aos anotadores no formulário foi a seguinte: cada anotador deveria identificar exatamente três objetos concretos em cada conto que aparentassem possuir valor simbólico na narrativa e escrever, para cada objeto, de uma a três interpretações simbólicas. Cada interpretação deveria ser formulada como uma frase concisa, focada no que o objeto poderia representar, sugerir ou evocar no contexto da história. Foi explicitamente vedado escolher ações ou personagens como candidatos. O formato esperado de resposta foi exemplificado no enunciado por meio de três exemplos ilustrativos, com objetos genéricos (*casa*, *botão* e *caderno*), cada um seguido de três interpretações curtas rotuladas como “a.”, “b.” e “c.”. A escolha por essa forma de exemplificação bus-

cou padronizar a extensão e a estrutura das respostas humanas, evitando que diferenças formais entre interpretações humanas e automáticas servissem de pista de origem na etapa posterior de avaliação cega.

5.3 Geração das interpretações automáticas

A pipeline computacional descrita na Seção 4.2 foi executada sobre os textos em inglês de cada um dos três contos. Como a etapa de seleção de candidatos combina o objeto mais frequente do conto com dois sorteados aleatoriamente entre os quatro seguintes do ranking de frequência, cada execução tende a produzir um conjunto parcialmente diferente de objetos. Para ampliar a variedade dos elementos analisados e reduzir a chance de que um único sorteio determinasse o resultado de todo o experimento, a pipeline foi executada duas vezes por conto, gerando dois conjuntos distintos de passagens contextuais por texto.

Cada conjunto de passagens foi então submetido aos três LLMs definidos na metodologia: ChatGPT (GPT-4o), Claude (Sonnet) e Gemini (2.0 Flash). O envio foi feito manualmente pela interface web de cada modelo, anexando o arquivo JSON de passagens gerado pela etapa 5 da pipeline e colando a prompt apresentada na Figura 4.3. Cada um dos três modelos foi consultado uma vez por execução da pipeline, totalizando seis respostas automáticas por conto, ou seja, duas por modelo. As respostas foram registradas em arquivos de texto separados por conto e por modelo, preservando o formato original devolvido pela LLM.

5.4 Montagem do instrumento de avaliação

A partir das anotações humanas e das interpretações automáticas, foi construído um único formulário de avaliação cega, também na plataforma Google Forms. O formulário foi organizado em três blocos sequenciais, um para cada conto. Cada bloco apresenta o texto integral do conto em sua versão em português logo no início, garantindo que o participante tivesse acesso direto ao material de referência no momento da avaliação. Em seguida ao texto, foram apresentadas todas as interpretações associadas àquele conto, intercaladas em ordem que não denunciava sua origem.

Como o Google Forms não permite, em uma única pergunta, coletar simultaneamente uma classificação categórica e uma nota em escala Likert, optou-se por desdobrar cada interpretação em duas perguntas consecutivas. A primeira pergunta solicita ao participante que indique se considera a interpretação como tendo sido produzida por anotador humano ou por IA. A segunda solicita uma nota de plausibilidade na escala de 1 a 5, em que 1 corresponde a totalmente implausível e 5 a totalmente plausível.

Para preservar o anonimato da origem das interpretações sem perder a capacidade de rastreá-las na etapa de análise, cada interpretação recebeu uma etiqueta de duas letras, aparentemente arbitrária, incluída entre parênteses no enunciado de suas duas perguntas. Internamente, cada etiqueta corresponde de maneira unívoca a um par (conto, origem), onde a origem é uma das nove combinações possíveis: três anotadores humanos, duas execuções da pipeline em ChatGPT, duas em Claude e duas em Gemini. Esse mapeamento foi mantido em arquivo separado e só foi consultado após o encerramento da coleta, na etapa de tabulação dos resultados. Assim, cada conto foi avaliado a partir de nove interpretações distintas, totalizando vinte e sete blocos de duas perguntas no formulário. A Figura 5.1 ilustra o formato de apresentação de um desses blocos no Google Forms.

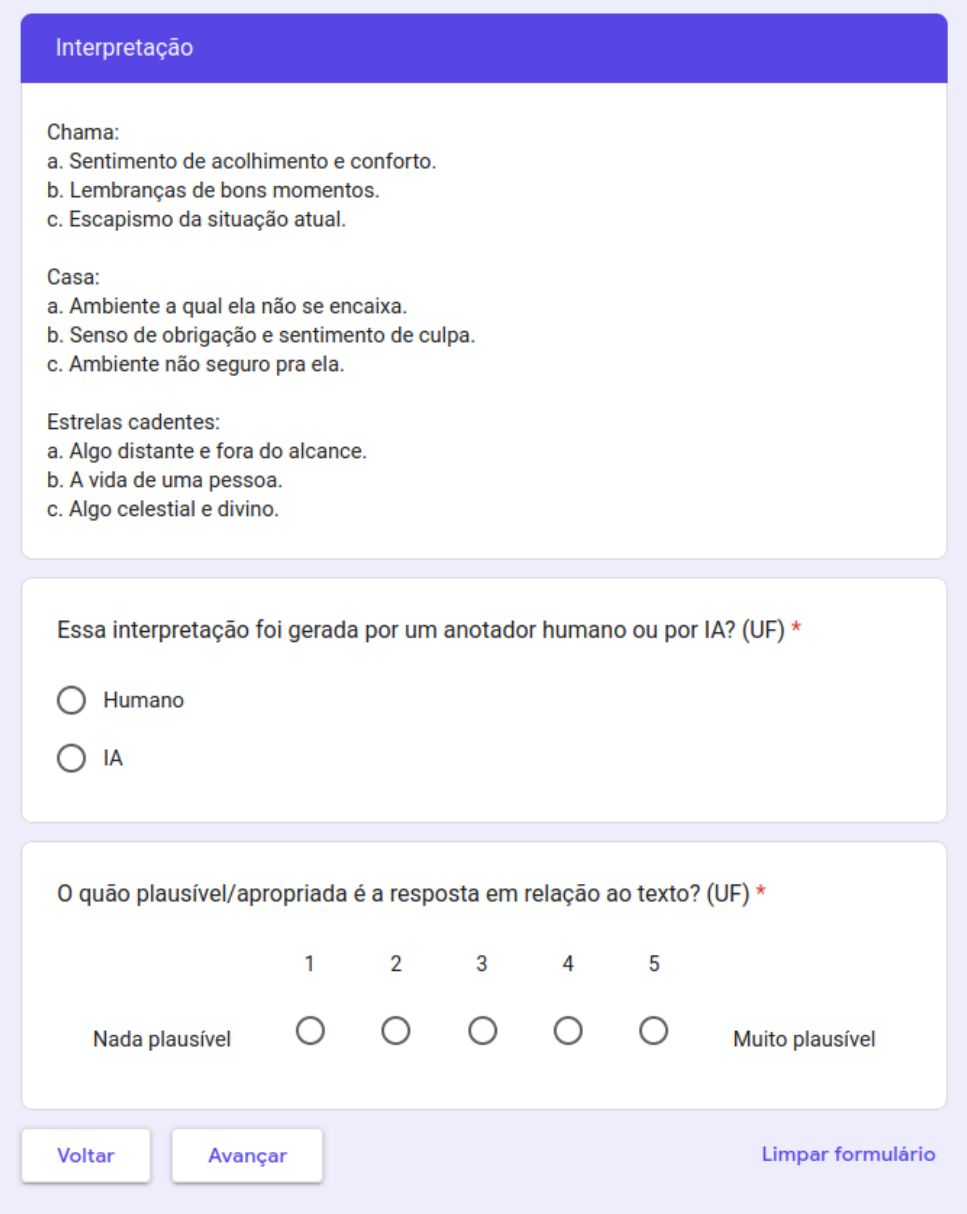

Figura 5.1: Exemplo de um bloco de avaliação do formulário, contendo a interpretação simbólica seguida das duas perguntas associadas (origem percebida e plausibilidade). Ao final de cada título de pergunta, o código está entre parentêses (UF nesse exemplo).

Antes do início das avaliações, o formulário aplicou uma pequena bateria de perguntas demográficas e contextuais: faixa etária, gênero, área de formação, frequência de uso de ferramentas de IA e frequência de leitura de narrativas de ficção. No início de cada um dos três blocos de avaliação, foi adicionada também uma pergunta sobre familiaridade prévia com o conto em questão, com as opções *Sim*, *Não* e *Talvez*, esta última prevista para o caso de o participante reconhecer vagamente a história sem ter certeza de tê-la lido antes.

5.5 Aplicação do experimento e perfil dos participantes

A divulgação do formulário foi feita exclusivamente por contato direto com pessoas conhecidas da autora, sem qualquer divulgação aberta em redes sociais ou canais públicos. Foi estabelecido como critério único de elegibilidade ter idade igual ou superior a dezoito anos. O formulário ficou aberto durante quatorze dias e contabilizou vinte e duas respostas válidas ao final do período de coleta.

O perfil dos respondentes é apresentado na Tabela 5.1. A amostra é predominantemente jovem adulta, com leve maioria feminina, concentração na área de Exatas e baixa familiaridade prévia com os contos avaliados. O contraste de familiaridade entre os três contos foi considerado relevante para a análise posterior, uma vez que o reconhecimento prévio pode influenciar tanto a percepção de plausibilidade quanto a tendência de atribuir uma interpretação a um autor humano.

Tabela 5.1: Perfil dos vinte e dois participantes do experimento.

Dimensão	**Categoria**	**n**
Faixa etária	18–24 anos	20
	25–34 anos	2
Gênero	Feminino	15
	Masculino	7
Área de formação	Exatas	15
	Humanas	6
	Outra	1
Uso de ferramentas de IA	Diário	4
	Semanal	8
	Mensal ou inferior	10
Leitura de ficção	Diariamente	3
	Semanalmente	9
	Mensalmente	3
	Algumas vezes por ano	6
	Raramente ou nunca	1
Familiaridade com *A Pequena Vendedora de Fósforos*	Sim	9
	Não/Talvez	13
Familiaridade com *A Máscara da Morte Vermelha*	Sim	3
	Não/Talvez	19
Familiaridade com *O Espelho*	Sim	0
	Talvez	3
	Não	19

As respostas foram exportadas em formato CSV diretamente da plataforma Google Forms e organizadas para análise sem qualquer edição manual no conteúdo das colunas, preservando os identificadores de duas letras de cada interpretação. Esses dados constituem a base sobre a qual a análise apresentada no Capítulo 6 foi realizada.

6 Resultados e Discussão

Este capítulo apresenta os resultados obtidos a partir das respostas coletadas no formulário descrito no Capítulo 5. A Seção 6.1 examina a tarefa principal do experimento, ou seja, a identificação da origem (humana ou automática) de cada interpretação simbólica. A Seção 6.2 analisa as notas de plausibilidade atribuídas pelos participantes, comparando interpretações humanas e geradas pelos LLMs. A Seção 6.3 investiga se existe relação entre a origem percebida pelo participante e a nota que ele atribuiu. A Seção 6.4 mede a concordância entre os participantes nas duas variáveis por meio do coeficiente alfa de Krippendorff. A Seção 6.5 verifica o efeito da familiaridade prévia com cada conto e das variáveis demográficas. A Seção 6.6 apresenta uma análise qualitativa item a item, por modelo de LLM e por conto. A Seção 6.7 sintetiza os achados.

Antes da análise, as respostas exportadas em CSV foram reorganizadas em formato longo: cada uma das vinte e duas linhas originais, correspondente a um respondente, foi expandida em vinte e sete linhas, correspondentes aos vinte e sete blocos de avaliação do formulário (nove por conto). A cada linha foram associadas a etiqueta de duas letras, a origem real recuperada a partir do mapeamento descrito na Seção 5.4, o veredicto do participante quanto à origem, a nota de plausibilidade na escala de 1 a 5 e as variáveis demográficas e contextuais correspondentes ao respondente. O conjunto resultante reúne quinhentas e noventa e quatro avaliações, número que serve de base para todas as análises subsequentes. A correspondência entre etiquetas e origens foi verificada de modo independente comparando-se o texto de cada bloco no script que gerou o formulário com o conteúdo dos arquivos de respostas dos anotadores e dos LLMs.

6.1 Identificação de origem

A taxa de acerto foi definida como a proporção de avaliações em que o veredicto do participante (humano ou IA) coincide com a origem real do bloco. Sobre o conjunto completo de quinhentas e noventa e quatro avaliações, a taxa de acerto foi de 61,4%. O

valor situa-se acima do desempenho esperado pelo acaso, de 50%, o que indica que os participantes captaram, em média, algum sinal distintivo entre os dois grupos. O ganho sobre o acaso, no entanto, é modesto.

Quando o conjunto é separado pela origem real, a assimetria fica evidente. Entre as 198 avaliações de interpretações humanas, 71,7% foram corretamente identificadas como humanas. Entre as 396 avaliações de interpretações automáticas, apenas 56,3% foram identificadas como tal. Em outras palavras, blocos automáticos passaram por humanos com frequência substancialmente maior do que o oposto, conforme ilustra a Figura 6.1.

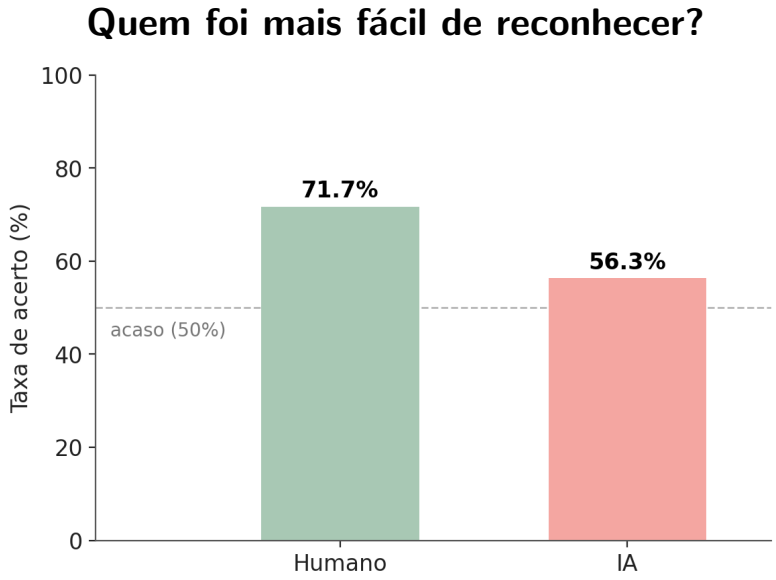

Figura 6.1: Taxa de acerto por origem real do bloco: interpretações humanas foram identificadas como tal com maior frequência que as automáticas. A linha tracejada marca o nível esperado pelo acaso.

A taxa de acerto foi definida como a proporção de avaliações em que o veredicto do participante (humano ou IA) coincide com a origem real do bloco. Sobre o conjunto completo de quinhentas e noventa e quatro avaliações, a taxa de acerto foi de 61,4%, valor sintetizado na Figura 6.2. O resultado situa-se acima do desempenho esperado pelo acaso, de 50%, o que indica que os participantes captaram, em média, algum sinal distintivo entre os dois grupos. O ganho sobre o acaso, no entanto, é modesto.

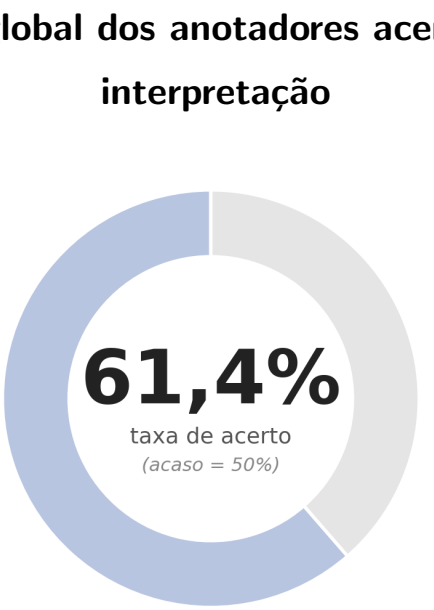

Figura 6.2: Taxa de acerto global na identificação de origem, sobre as 594 avaliações coletadas.

A diferença mais expressiva, no entanto, está entre os três modelos automáticos, como mostra a Figura 6.3. Restringindo o cálculo às interpretações geradas por LLM, a taxa de acerto foi de 37,1% para o ChatGPT, 69,7% para o Claude e 62,1% para o Gemini. Em termos práticos, o ChatGPT foi confundido com produção humana em quase dois terços das avaliações em que apareceu, ficando inclusive abaixo do nível esperado pelo acaso, enquanto o Claude foi corretamente identificado como automático em cerca de 70% das vezes. Entre os três anotadores humanos, as taxas de acerto ficaram entre 68,2% e 77,3%, próximas entre si e sem indicação de que algum anotador tenha soado claramente mais ou menos humano que os demais.

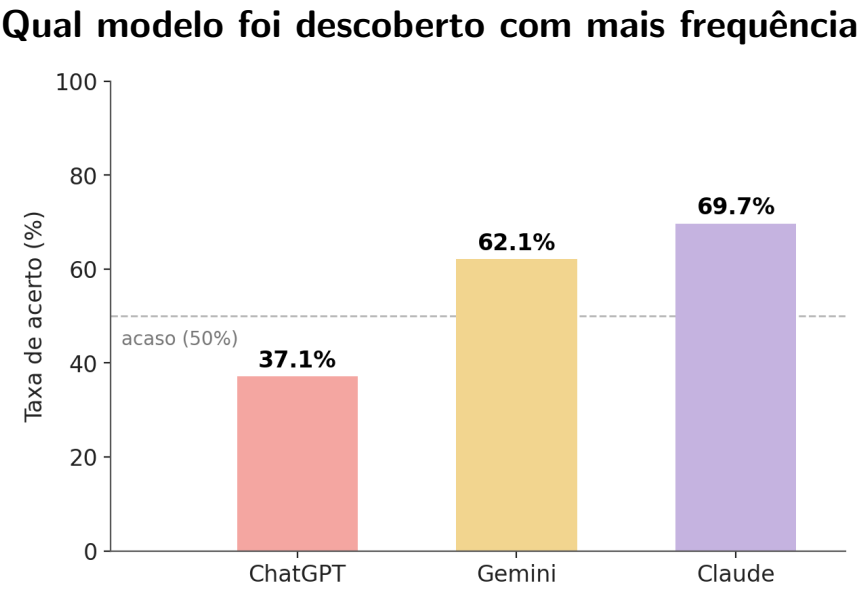

Figura 6.3: Taxa de acerto na identificação de origem por modelo automático: o ChatGPT foi o mais confundido com produção humana, ficando inclusive abaixo do nível esperado pelo acaso.

6.2 Plausibilidade percebida

A plausibilidade média foi calculada como a média aritmética simples das notas atribuídas em cada subconjunto de avaliações, mantendo-se a escala original de 1 a 5. Considerando todas as avaliações, as interpretações automáticas receberam média de 3,94, com desvio padrão de 0,99, ao passo que as interpretações humanas receberam 3,71, com desvio padrão de 1,13. As interpretações geradas pelos LLMs foram, portanto, avaliadas como ligeiramente mais plausíveis que as escritas pelos anotadores humanos, conforme sintetiza a Figura 6.4.

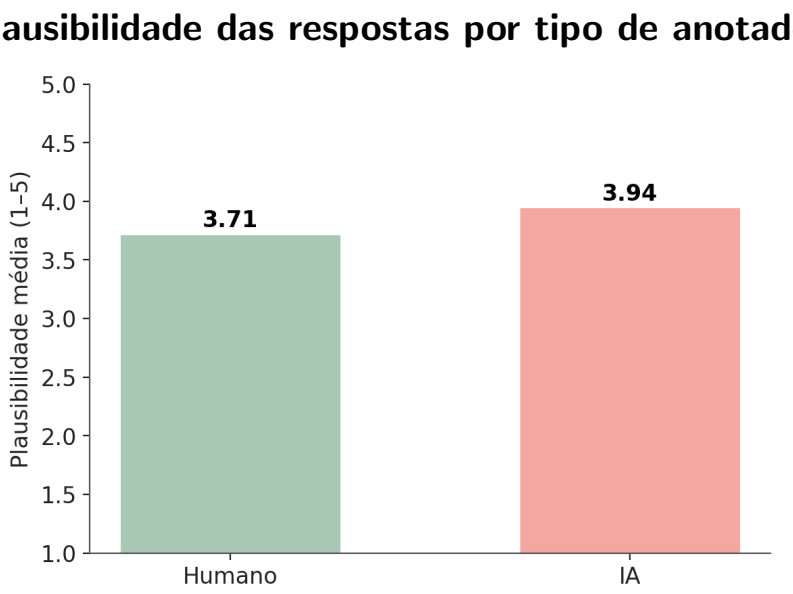

Figura 6.4: Plausibilidade média por origem real do bloco, na escala de 1 a 5: as interpretações automáticas receberam nota ligeiramente superior às escritas pelos anotadores humanos.

Esse padrão, contudo, não é uniforme entre os contos. Em *A Pequena Vendedora de Fósforos*, a média das automáticas foi de 3,87 contra 3,62 das humanas; em *O Espelho*, de 3,95 contra 3,47; em *A Máscara da Morte Vermelha*, as duas origens praticamente se igualaram, com 4,02 para automáticas e 4,05 para humanas. As interpretações associadas a este último conto receberam, no agregado, as notas mais altas, independentemente da origem.

Restringindo a análise às interpretações automáticas e separando-as por modelo, as médias foram de 4,09 para o ChatGPT, 3,88 para o Claude e 3,86 para o Gemini. Entre os anotadores humanos, as médias ficaram entre 3,61 e 3,91, com o segundo anotador apresentando a média mais alta nos três contos.

6.3 Relação entre origem percebida e plausibilidade

Para verificar se existe relação entre o veredicto de origem emitido pelo participante e a nota de plausibilidade que ele atribuiu àquele mesmo bloco, a média de plausibilidade foi recalculada separando as avaliações pelo veredicto, não pela origem real. As avaliações em que o participante considerou a interpretação como sendo de origem humana receberam

média de 4,13. As avaliações em que ele considerou a interpretação como sendo de origem automática receberam média de 3,57. A diferença entre as duas médias é de 0,57 pontos na escala de 1 a 5, conforme ilustra a Figura 6.5.

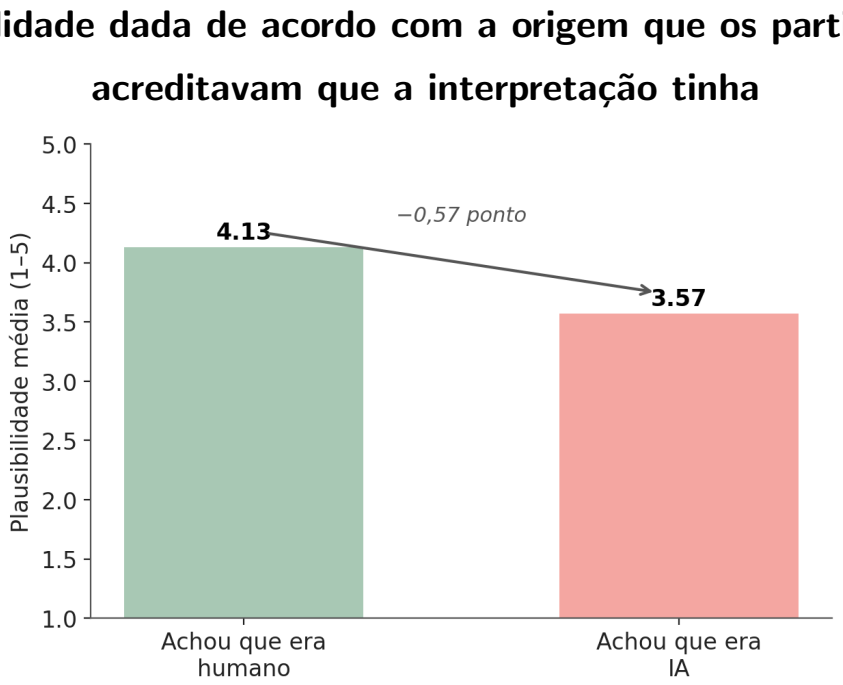

Figura 6.5: Plausibilidade média segundo a origem percebida pelo participante: quando o bloco era classificado como humano, recebia em média 0,57 ponto a mais, independentemente da origem real.

A significância dessa diferença foi avaliada por meio de dois testes complementares. O primeiro foi o teste de Mann-Whitney, que compara se uma das duas amostras tende a apresentar valores sistematicamente mais altos que a outra, sem assumir que as notas sigam uma distribuição normal (FAY; PROSCHAN, 2010). É uma escolha adequada para a escala Likert de 1 a 5, que é ordinal. O teste retornou estatística U igual a 56599,5 e p-valor da ordem de $2{,}2 \times 10^{-10}$, em que o p-valor representa a probabilidade de a diferença observada ter ocorrido por acaso caso as duas distribuições fossem, na realidade, equivalentes. O segundo teste foi o coeficiente de correlação ponto-bisserial entre o veredicto codificado (1 para humano, 0 para IA) e a nota de plausibilidade. O coeficiente varia de -1 a $+1$ e expressa a força da associação entre uma variável binária e uma numérica (HAZRA; GOGTAY, 2016). O valor obtido foi r igual a 0,27, com p-valor da ordem de $1{,}6 \times 10^{-11}$, magnitude usualmente caracterizada como associação moderada.

Os dois resultados, em níveis de significância amplamente abaixo de 0,01, apontam que a percepção sobre a origem está sistematicamente associada à nota atribuída: os participantes tenderam a julgar como mais plausíveis as interpretações que acreditavam ter sido escritas por humanos, independentemente da origem real do bloco.

6.4 Concordância entre participantes

A concordância entre os vinte e dois participantes foi avaliada por meio do coeficiente alfa de Krippendorff (HAYES; KRIPPENDORFF, 2007), uma medida que quantifica o quanto múltiplos avaliadores tendem a dar respostas semelhantes para os mesmos itens. O valor varia de 0, equivalente ao acaso, a 1, equivalente a acordo perfeito, e admite métricas distintas conforme o nível de mensuração da variável. Para o veredicto categórico de origem, foi utilizada a métrica nominal, que considera duas respostas concordantes apenas quando idênticas. Para a nota de plausibilidade, foi utilizada a métrica intervalar, que pondera o grau de discordância pela distância numérica entre as notas, de modo que uma divergência entre 4 e 5 conta menos do que uma divergência entre 1 e 5. A matriz fornecida ao cálculo organizou os participantes como avaliadores nas linhas e as etiquetas como unidades nas colunas. A implementação seguiu a fórmula padrão baseada na matriz de coincidências, com expansão por pares dentro de cada unidade.

O coeficiente para o veredicto de origem, considerando o conjunto completo das vinte e sete etiquetas, foi de 0,14. Separando por conto, o valor foi de 0,19 em *A Pequena Vendedora de Fósforos*, de 0,17 em *O Espelho* e de apenas 0,07 em *A Máscara da Morte Vermelha*. Esses valores são consideravelmente inferiores ao limite de 0,67 usualmente adotado como mínimo para concordância aceitável em estudos de confiabilidade entre avaliadores, e indicam discordância frequente entre os participantes quanto à origem dos blocos.

Para a nota de plausibilidade, o coeficiente foi ainda mais baixo: 0,05 considerando o conjunto completo, e variando entre 0,02 e 0,05 nos três contos isoladamente. Esse valor sugere que cada participante utilizou a escala de modo idiossincrático e que não existe um padrão compartilhado de avaliação de plausibilidade no grupo.

Em conjunto, os dois valores reforçam a leitura de que a tarefa proposta é genui-

namente ambígua. Em condições nas quais a concordância entre humanos é tão baixa, atribuir a um modelo a propriedade de soar humano deixa de ser uma medida absoluta e passa a ser uma medida em comparação a um padrão de juízo que, na prática, varia de pessoa para pessoa.

6.5 Familiaridade prévia e variáveis demográficas

A familiaridade prévia com cada conto foi coletada no início do bloco correspondente, com as opções *Sim*, *Não* e *Talvez*. Cruzando-se a familiaridade declarada com a taxa de acerto na identificação de origem, o desempenho foi semelhante entre as duas respostas predominantes: 63,0% entre os participantes que declararam *Sim*, contra 62,0% entre os que declararam *Não*, com queda para 50,0% no pequeno subconjunto que declarou *Talvez*. Para avaliar formalmente se essa pequena diferença é atribuível ao acaso, foi aplicado o teste qui-quadrado de independência ao subconjunto formado apenas por *Sim* e *Não*, que testa se duas variáveis categóricas estão associadas comparando a tabela cruzada observada com a tabela que seria esperada caso não houvesse associação alguma. O teste retornou p-valor de 0,94, e o teste de Mann-Whitney sobre o acerto retornou p-valor de 0,85. Em ambos, os valores muito acima do limite de 0,05 indicam ausência de efeito da familiaridade sobre a taxa de acerto.

Para a nota de plausibilidade, no entanto, a familiaridade prévia teve efeito marginalmente significativo. Participantes que declararam já conhecer o conto atribuíram nota média de 3,69, enquanto os que declararam não conhecê-lo atribuíram 3,93. O teste de Mann-Whitney aplicado a essas duas distribuições retornou p-valor de 0,031. A direção da diferença sugere que a familiaridade tornou o leitor um pouco mais crítico em relação às interpretações apresentadas, embora a magnitude do efeito seja pequena.

Em relação às variáveis demográficas (uso de IA, frequência de leitura, área de formação, gênero e faixa etária), foram aplicados testes qui-quadrado de independência entre cada variável e a variável binária de acerto. Os p-valores foram de 0,54 para uso de IA, 0,63 para frequência de leitura, 0,25 para área de formação e 0,95 para gênero, todos amplamente acima do limite de 0,05. A única exceção foi faixa etária, com p-valor de 0,0025. Esse resultado, contudo, exige cautela: apenas dois participantes pertencem à

faixa de vinte e cinco a trinta e quatro anos, contra vinte na faixa de dezoito a vinte e quatro, e a estimativa de desempenho do subgrupo mais velho é fundada em um número insuficiente de observações para sustentar conclusão substantiva.

6.6 Análise qualitativa

6.6.1 Itens mais confundidos

A identificação dos blocos mais confundidos com a origem oposta foi feita por meio da taxa de voto humano por etiqueta, calculada como a proporção de participantes que classificaram aquele bloco como humano. Entre as dezoito etiquetas de origem automática, três alcançaram as maiores taxas de voto humano, e todas elas foram geradas pelo ChatGPT. A etiqueta HD, produzida a partir de *A Máscara da Morte Vermelha*, foi classificada como humana por 17 dos 22 participantes (77,3%), com plausibilidade média de 4,23. O texto apresentado aos participantes foi:

> *Relógio de Ébano*: presença constante do tempo que interrompe a ilusão da festa; chamado periódico à reflexão sobre a mortalidade, mesmo entre os despreocupados; força inevitável que acompanha e encerra a vida dos participantes. *Morte Vermelha*: realidade inevitável que atravessa barreiras e desfaz qualquer sensação de segurança. *Salas*: sequência de experiências e estados emocionais que transformam a percepção de quem as percorre; cenário onde fantasia, excesso e inquietação se misturam até perderem seus limites.

As etiquetas BQ, gerada a partir de *A Pequena Vendedora de Fósforos*, e OR, gerada a partir de *O Espelho*, foram ambas classificadas como humanas por 16 dos 22 participantes (72,7%), com plausibilidade média de 4,14 e 4,05, respectivamente. Os textos correspondentes são reproduzidos a seguir.

> **BQ:** *Mão*: sinal de trabalho exausto e de sobrevivência sem amparo; lugar onde a fragilidade da menina aparece como corpo cansado e frio. *Fósforos*: pequena mercadoria que concentra a esperança de ganhar algum sustento;

> faísca breve de consolo, capaz de abrir visões que logo se desfazem; instrumento de fuga simbólica diante de uma realidade intolerável. *Luz*: promessa momentânea de calor, beleza e acolhimento; passagem para uma esfera de consolo que interrompe a dureza do mundo.
>
> **OR:** *Espelho*: superfície que confirma a identidade quando ela ganha forma visível; lugar de encontro entre a pessoa interior e a máscara social. *Alma*: núcleo dividido entre o que sente por dentro e o que busca fora; vínculo com papéis e sinais sociais que podem substituir a pessoa; espaço íntimo que só no sonho consegue agir com liberdade. *Tempo*: força que desgasta, apaga e reordena o que parecia estável.

Entre as nove etiquetas de origem humana, a mais confundida com origem automática foi TQ, do primeiro anotador de *A Máscara da Morte Vermelha*, classificada como humana por apenas 13 dos 22 participantes (59,1%). O texto correspondente foi:

> *Relógio de ébano*: o tempo que leva todos inevitavelmente ao fim; o fim da existência física dos convidados da festa. *Tripés com braseiros*: luz artificial que mantém a ilusão e o clima de sonho; a vida que se apaga quando a morte domina o lugar. *Portões de ferro*: tentativa inútil de se isolar da doença e do sofrimento; falsa sensação de segurança contra o destino inevitável; barreira que transforma o refúgio em uma armadilha mortal.

O texto, apesar de produzido por anotador humano, escolhe objetos pouco usuais do enredo (tripés com braseiros, portões de ferro) e descreve cada um deles com proposições relativamente formais, sem variação estilística marcada, o que pode ter contribuído para que parte significativa dos participantes tenha atribuído origem automática ao bloco.

6.6.2 Padrões por modelo

Comparando-se os três modelos pelas seis execuções da pipeline (duas por conto), observa-se diferença pronunciada no comprimento médio dos blocos e na quantidade média de

interpretações geradas. Os blocos do ChatGPT tiveram, em média, 532 caracteres e 6,5 interpretações. Os do Claude, 735 caracteres e 7,7 interpretações. Os do Gemini, 650 caracteres e 6,3 interpretações. Os blocos dos três anotadores humanos tiveram, em média, 297 caracteres e 6,9 interpretações, comprimento próximo ao do ChatGPT e marcadamente inferior aos dos outros dois modelos. A proporção média de votos como humano nos blocos gerados por cada modelo acompanha esse padrão: 62,9% para o ChatGPT, 37,9% para o Gemini e 30,3% para o Claude. A coincidência entre o comprimento do bloco do ChatGPT e o dos anotadores humanos, somada ao fato de ele ser o mais confundido com humano, sugere relação entre concisão e probabilidade de ser percebido como humano, embora o número de execuções por modelo seja pequeno demais para isolar essa variável de outras características de estilo.

6.6.3 Padrões por conto e objetos selecionados

Ainda no plano qualitativo, observa-se baixa sobreposição entre os objetos escolhidos pelos anotadores humanos e os objetos sorteados pela pipeline. Em cada conto, apenas dois objetos são comuns aos dois grupos. Em *A Pequena Vendedora de Fósforos*, a interseção é formada por *casa* e *fósforo*; em *A Máscara da Morte Vermelha*, por *relógio* e *relógio de ébano*; em *O Espelho*, por *casa* e *espelho*. Os anotadores humanos tenderam a escolher objetos específicos do enredo, como *tripés com braseiros*, *farda de alferes*, *níquel* e *laranja*, enquanto a pipeline tendeu a selecionar objetos centrais e de aparição frequente, como *tempo*, *alma*, *sala* e *janela*. Essa divergência decorre da etapa de seleção de candidatos da pipeline, que utiliza frequência de menção como critério principal, e indica que o conjunto de objetos analisado pelos humanos e o conjunto analisado pelos LLMs nem sempre coincide.

6.7 Síntese dos resultados

Os resultados podem ser organizados em quatro observações principais. Em primeiro lugar, os participantes acertaram a origem em 61,4% das avaliações, modestamente acima do acaso, e a taxa de acerto é assimétrica: interpretações automáticas passam por humanas

com frequência substancialmente maior do que o oposto (56,3% de acerto contra 71,7%). Em segundo lugar, as interpretações automáticas receberam notas de plausibilidade ligeiramente superiores às humanas (3,94 contra 3,71), e a análise da Seção 6.3 identificou um viés cognitivo segundo o qual os participantes atribuíram, em média, 0,57 pontos a mais às interpretações que acreditavam ter origem humana, com p-valor da ordem de 10^{-10} em dois testes complementares. Em terceiro lugar, a concordância entre participantes é muito baixa, com alfa de Krippendorff de 0,14 para o veredicto categórico e 0,05 para a nota de plausibilidade, o que reforça a leitura de que a tarefa é genuinamente ambígua. Por fim, observam-se diferenças marcantes entre os três modelos: o ChatGPT foi o mais frequentemente confundido com humanos e produz blocos cujo comprimento médio coincide com o dos anotadores humanos, ao passo que o Claude e o Gemini produziram blocos mais longos e foram identificados como automáticos com maior frequência.

As limitações do desenho experimental restringem a generalização desses resultados. O número de participantes (vinte e dois), de anotadores por conto (três), de contos (três) e a distribuição desigual nas categorias demográficas reduzem o poder estatístico das comparações por subgrupo. As observações qualitativas extraídas do material, no entanto, permanecem informativas como pontos de partida para investigações subsequentes.

7 Conclusões e Trabalhos Futuros

A interpretação simbólica de objetos em narrativas de ficção foi tratada, neste trabalho, como uma tarefa de difícil automatização, em parte porque envolve linguagem figurativa e dependência de contexto, em parte porque coloca o sistema computacional diante de um tipo de leitura que não se reduz à descrição literal dos eventos. Diante desse problema, foi proposta uma pipeline que combina pré-processamento textual baseado em PLN com LLMs responsáveis pela geração das interpretações, e foi conduzido um experimento de avaliação cega para verificar como leitores humanos reagem ao seu produto.

Os quatro objetivos específicos foram cumpridos. O corpus foi constituído por três contos de domínio público e as anotações humanas correspondentes foram coletadas por meio de formulário aplicado a três anotadores independentes. A pipeline modular em Python foi implementada e executada sobre os textos em inglês, produzindo passagens contextuais que foram submetidas a três LLMs distintos. O experimento de avaliação foi conduzido com vinte e dois participantes, que classificaram a origem e atribuíram notas de plausibilidade às vinte e sete interpretações intercaladas no instrumento. Por fim, as respostas foram analisadas tanto quantitativa quanto qualitativamente, sustentando as observações reunidas no Capítulo 6.

A questão de pesquisa pode então ser respondida de modo afirmativo, com ressalvas. As interpretações geradas pelos LLMs receberam notas de plausibilidade ligeiramente superiores às atribuídas às interpretações humanas, o que responde positivamente à primeira subquestão. Quanto à segunda subquestão, a distinção entre as duas origens mostrou-se apenas parcial: o acerto global ficou modestamente acima do acaso, e as interpretações automáticas passaram por humanas com frequência consideravelmente maior do que o oposto. Esse comportamento, contudo, variou de forma marcante entre os modelos, com o ChatGPT sendo confundido com produção humana na maioria das avaliações em que apareceu, ao passo que Claude e Gemini foram identificados corretamente com frequência substancialmente maior, em proporções comparáveis ou próximas às dos anotadores humanos. No escopo testado, portanto, é razoável afirmar que LLMs

são capazes de interpretar objetos simbólicos a partir do contexto em que esses objetos aparecem, ao menos no sentido de produzirem interpretações que se sustentam diante da leitura humana.

Essa resposta, no entanto, precisa ser lida com três ressalvas que emergiram da própria análise. A primeira é o viés cognitivo identificado na Seção 6.3: os participantes atribuíram notas sistematicamente mais altas às interpretações que acreditavam ter sido escritas por humanos, independentemente de sua origem real, o que indica que a percepção de plausibilidade não é independente da percepção de autoria. A segunda é a baixa concordância entre os participantes, com alfa de Krippendorff distante dos limites usualmente considerados aceitáveis, sugerindo que a tarefa proposta é genuinamente ambígua e que não existe um padrão único e compartilhado de leitura simbólica contra o qual comparar a produção automática. A terceira é a divergência entre os objetos selecionados pelos anotadores humanos e pela pipeline, decorrente do critério de frequência empregado na etapa de seleção de candidatos: a comparação realizada não é, a rigor, entre interpretações para os mesmos objetos, mas entre interpretações produzidas por cada lado dentro do mesmo conto.

Considerando essas ressalvas, o trabalho oferece três contribuições principais. A primeira é a pipeline modular descrita no Capítulo 4, que articula etapas clássicas de PLN, como extração de objetos e resolução de correferência, com um LLM operando como camada interpretativa sobre um contexto previamente filtrado, aplicada a um domínio (interpretação simbólica literária) ainda pouco explorado nesse tipo de combinação. A segunda é a evidência empírica de que LLMs produzem interpretações simbólicas comparáveis às escritas por anotadores humanos no escopo testado: as automáticas passaram por humanas em quase metade das avaliações em que apareceram (43,7%) e receberam, na média, notas de plausibilidade ligeiramente superiores às humanas (3,94 contra 3,71), padrão que respondeu diretamente à questão de pesquisa do trabalho. A terceira é a evidência empírica do viés cognitivo na percepção de autoria, identificado na Seção 6.3 com diferença média de 0,57 ponto na escala de plausibilidade conforme a origem percebida e nível de significância da ordem de 10^{-10} em dois testes independentes, achado cujo alcance extrapola o caso simbólico aqui investigado.

As limitações deste trabalho foram em boa parte já reconhecidas ao longo do texto. O corpus reduzido a três contos, o número de anotadores por conto, o tamanho da amostra de respondentes e a distribuição desigual nas variáveis demográficas restringem a generalização dos resultados. A pipeline opera sobre os textos em inglês enquanto a avaliação humana foi conduzida em português, em decorrência da ausência de modelos estáveis de resolução de correferência para o português. A etapa de seleção de candidatos baseia-se essencialmente em frequência de menção, sem filtro de concretude e sem critério semântico de centralidade narrativa, o que contribui para a divergência entre os objetos analisados por humanos e pela pipeline. Por fim, a avaliação de plausibilidade depende de um julgamento subjetivo cuja variabilidade ficou evidente nos valores de concordância obtidos.

7.1 Trabalhos Futuros

Diversas direções de continuidade decorrem diretamente das limitações apontadas. Uma replicação com corpus mais extenso, abrangendo gêneros e tradições literárias adicionais, permitiria verificar se os padrões aqui observados se sustentam fora do recorte de três contos curtos. Da mesma forma, a ampliação da amostra de respondentes e o equilíbrio entre faixas etárias, áreas de formação e familiaridade com ferramentas de IA possibilitariam análises por subgrupo com maior poder estatístico do que as comparações realizadas neste trabalho.

No plano metodológico, uma extensão natural é forçar o alinhamento de objetos entre humanos e pipeline, solicitando aos LLMs interpretações para exatamente os mesmos objetos escolhidos pelos anotadores. Esse desenho isola a qualidade da interpretação da qualidade da seleção de candidatos e torna a comparação mais direta. Ainda nesse sentido, o viés cognitivo identificado na Seção 6.3 merece investigação dedicada, com experimento em que a mesma interpretação seja apresentada a grupos distintos sob rótulos opostos de origem, de modo a estimar o efeito do rótulo isoladamente.

No plano técnico, a substituição do critério de frequência por uma medida de relevância narrativa ou centralidade simbólica pode aproximar a seleção de candidatos das escolhas feitas por leitores humanos. A incorporação de um filtro entre objetos abs-

tratos e concretos, deixada de fora deste trabalho, permanece como aprimoramento útil para reduzir ruído na etapa de extração. Para superar o descasamento de idioma, vale investigar o uso de modelos multilíngues de correferência ou de estratégias de tradução intermediária com posterior projeção de spans para o texto original em português. Por fim, a repetição do experimento com versões mais recentes dos modelos avaliados permitiria observar como o desempenho na geração de interpretações simbólicas evolui ao longo do ciclo de atualização dos LLMs.

Duas direções adicionais, ambas inviabilizadas por restrições práticas deste trabalho, merecem atenção em estudos futuros. A primeira é a aplicação da pipeline a textos de maior extensão, como novelas e romances, em que a interpretação simbólica costuma se desenvolver ao longo de muitos capítulos e em que a etapa de resolução de correferência tende a impactar o resultado de modo mais significativo do que no recorte de contos curtos. A inclusão desse tipo de obra depende, contudo, da disponibilidade de anotadores e respondentes dispostos a investir tempo de leitura proporcional ao tamanho do material, o que não se mostrou viável no escopo do TCC. A segunda é a realização de um estudo de ablação, técnica adaptada da neurociência por Meyes et al. (2019) e amplamente utilizada em aprendizado de máquina, que consiste em remover ou desativar componentes de um sistema para medir o quanto cada parte contribui para o resultado final. No caso desta pesquisa, o estudo isolaria o efeito do pré-processamento comparando interpretações geradas a partir das passagens filtradas pela pipeline com interpretações geradas a partir do texto integral, sem qualquer filtragem prévia. Esse desenho permitiria quantificar diretamente o quanto a etapa de PLN contribui para a qualidade percebida da saída, hipótese que o presente trabalho assume mas não testa empiricamente.

Bibliografia

BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big?. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. [S.l.: s.n.], 2021. p. 610–623.

BOJIĆ, L. et al. Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, Nature Publishing Group UK London, v. 15, n. 1, p. 11477, 2025.

BROWN, T. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020.

CHAKRABARTY, T.; CHOI, Y.; SHWARTZ, V. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ..., v. 10, p. 589–606, 2022.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. [S.l.: s.n.], 2019. p. 4171–4186.

FAY, M. P.; PROSCHAN, M. A. Wilcoxon-mann-whitney or t-test? on assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys*, v. 4, p. 1, 2010.

GUO, M.; HWA, R.; KOVASHKA, A. Decoding symbolism in language models. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2023. p. 3311–3324.

HAYES, A. F.; KRIPPENDORFF, K. *Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1 (1), 77-89*. 2007.

HAZRA, A.; GOGTAY, N. Biostatistics series module 6: correlation and linear regression. *Indian journal of dermatology*, Medknow, v. 61, n. 6, p. 593–601, 2016.

HENG, R.; PU, L.; LIU, X. The effects of genre on the lexical richness of argumentative and expository writing by chinese efl learners. *Frontiers in Psychology*, Frontiers Media SA, v. 13, p. 1082228, 2023.

HUANG, L. et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, ACM New York, NY, v. 43, n. 2, p. 1–55, 2025.

JENNER, S. et al. Using large language models for narrative analysis: A novel application of generative ai. *Methods in Psychology*, Elsevier, v. 12, p. 100183, 2025.

KAMATH, U. et al. Large language models: A deep dive. *Bridging Theory and Practice, Cham: Springer Nature*, Springer, 2024.

KHURANA, D. et al. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, Springer, v. 82, n. 3, p. 3713–3744, 2023.

ŁAJEWSKA, W.; WRÓBLEWSKA, A. Protagonists' tagger in literary domain–new datasets and a method for person entity linkage. *arXiv preprint arXiv:2110.01349*, 2021.

LEE, C. V. D. et al. Best practices for the human evaluation of automatically generated text. In: *Proceedings of the 12th international conference on natural language generation*. [S.l.: s.n.], 2019. p. 355–368.

LISON, P.; KUTUZOV, A. Redefining context windows for word embedding models: An experimental study. In: *Proceedings of the 21st Nordic Conference on Computational Linguistics*. [S.l.: s.n.], 2017. p. 284–288.

LIU, D. Y.; JOSHI, A.; DAWSON, P. Narrative theory-driven llm methods for automatic story generation and understanding: A survey. *arXiv preprint arXiv:2602.15851*, 2026.

MARCONI, M. d. A.; LAKATOS, E. M. *Metodologia científica*. 8. ed. São Paulo: Atlas, 2017.

MEYES, R. et al. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

MOHTA, J. et al. Are large language models good annotators? In: PMLR. *Proceedings on*. [S.l.], 2023. p. 38–48.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011.

PAULSEN, N. Context is what you need: The maximum effective context window for real world limits of llms. *arXiv preprint arXiv:2509.21361*, 2025.

PIPER, A.; WU, S. Evaluating large language models for narrative topic labeling. In: *Proceedings of the 5th international conference on natural language processing for digital humanities*. [S.l.: s.n.], 2025. p. 281–291.

SANTANA, B. et al. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, Springer, v. 56, n. 8, p. 8393–8435, 2023.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

WEI, J. et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, v. 35, p. 24824–24837, 2022.

WHITE, J. et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

ZHU, L. et al. Are nlp models good at tracing thoughts: An overview of narrative understanding. *arXiv preprint arXiv:2310.18783*, 2023.