



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA
COMPUTAÇÃO

WALCY SANTOS REZENDE RIOS

**Aprimoramento de Dados para SFT em
Português Brasileiro: Um Estudo com
Modelos de Língua e Avaliação
com LLM-as-Judge**

Goiânia
2025



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a [Lei 9.610/98](#), o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

Dissertação Tese Outro*: _____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Walcy Santos Rezende Rios

3. Título do trabalho

Aprimoramento de Dados para SFT em Português Brasileiro: Um Estudo com Modelos de Língua e Avaliação com LLM-as-Judge

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento SIM NÃO¹

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:

- a) consulta ao(à) autor(a) e ao(à) orientador(a);
 - b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
- O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.



Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho, Professor do Magistério Superior**, em 09/07/2025, às 21:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Walcy Santos Rezende Rios, Discente**, em 10/07/2025, às 14:45, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5494246** e o código CRC **99154146**.

WALCY SANTOS REZENDE RIOS

Aprimoramento de Dados para SFT em Português Brasileiro: Um Estudo com Modelos de Língua e Avaliação com LLM-as-Judge

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, da Faculdade do Instituto de Informática, da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Linha de pesquisa: Sistemas Inteligentes e Aplicações.

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho

Goiânia
2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Rios, Walcy Santos Rezende
Aprimoramento de Dados para SFT em Português Brasileiro
[manuscrito] : Um Estudo com Modelos de Língua e Avaliação com LLM
as-Judge / Walcy Santos Rezende Rios. - 2025.
XLII, 42 f.

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto
de Informática (INF), Programa de Pós-Graduação em Ciência da
Computação, Goiânia, 2025.
Bibliografia. Apêndice.
Inclui tabelas, lista de figuras, lista de tabelas.

1. LLM. 2. SFT. 3. dados sintéticos. 4. LLM-as-Judge. I. Galvão
Filho, Arlindo Rodrigues, orient. II. Título.

CDU 004



UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº 14 da sessão de Defesa de Dissertação de **Walcy Santos Rezende Rios**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos dez dias do mês de junho de dois mil e vinte e cinco, a partir das treze horas, via sistema de webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada “**Aprimoramento de Dados para SFT em Português Brasileiro: Um Estudo com Modelos de Língua e Avaliação com LLM-as-Judge**”. Os trabalhos foram instalados pelo Orientador, Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Sávio Salvarino Teles de Oliveira (INF/UFG), membro titular externo; Professora Doutora Telma Woerle de Lima Soares (INF/UFG), membra titular interna. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Arlindo Rodrigues Galvão Filho, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos dez dias do mês de junho de dois mil e vinte e cinco.

TÍTULO SUGERIDO PELA BANCA



Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho, Professor do Magistério Superior**, em 10/06/2025, às 14:08, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Savio Salvarino Teles De Oliveira, Professor do Magistério Superior**, em 10/06/2025, às 14:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Telma Woerle De Lima Soares, Professora do Magistério Superior**, em 10/06/2025, às 14:16, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Walcy Santos Rezende Rios, Discente**, em 10/06/2025, às 15:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fabrizio Alphonsus Alves De Melo Nunes Soares, Coordenador de Pós-Graduação**, em 12/06/2025, às 08:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5424676** e o código CRC **CA40E2B4**.

Referência: Processo nº 23070.025070/2025-17

SEI nº 5424676

Resumo

Rios, Walcy. **Aprimoramento de Dados para SFT em Português Brasileiro: Um Estudo com Modelos de Língua e Avaliação com LLM-as-Judge**. Goiânia, 2025. 42p. Dissertação de Mestrado. Programa de Pós-graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

A carência de recursos de alta qualidade para o português brasileiro (pt-br) dificulta o desenvolvimento de modelos de linguagem eficazes e adaptados às especificidades do idioma. Este trabalho investiga o impacto do aprimoramento sintético de dados conversacionais, utilizando Large Language Models (LLMs), no treinamento supervisionado (SFT) de modelos da família Qwen2.5 (0.5B, 1.5B, 3B). Partindo do dataset SmolTalk, foram geradas duas versões para o pt-br: uma por tradução direta e outra com respostas sinteticamente aprimoradas e reescritas pelo LLM Gemini 2.0 Flash. Os modelos Qwen2.5 foram treinados com ambos os conjuntos de dados e avaliados comparativamente através de benchmarks objetivos padronizados para o português (ENEM, HATEBR, BLUEX, ASSIN2-RTE) e por meio de avaliação qualitativa de geração de texto aberto (Alpaca-Eval-BR), utilizando o Claude 3.5 Haiku como LLM-as-Judge com base em critérios de relevância, precisão, abrangência, utilidade e coerência. Os resultados demonstram uma superioridade significativa dos modelos treinados com dados sintéticos na avaliação qualitativa LLM-as-Judge em todas as métricas. Nesta avaliação, a pontuação média normalizada aumentou expressivamente com dados sintéticos: o modelo 1.5B alcançou 44.45 (vs 14.05 do traduzido, um ganho de 216%) e o modelo 3B atingiu 57.21 (vs 16.79 do traduzido, um ganho de 241%). Em contrapartida, nos benchmarks objetivos, o impacto positivo do aprimoramento sintético foi menos acentuado, sendo mais consistente apenas na versão de 3B parâmetros. Conclui-se que a estratégia de aprimoramento sintético de dados assistida por LLM é eficaz para elevar significativamente a qualidade e o desempenho de modelos de linguagem conversacionais em português brasileiro, representando uma abordagem valiosa para mitigar a escassez de recursos dedicados e avançar no desenvolvimento de tecnologias de NLP mais adaptadas ao contexto nacional.

Palavras-chave

<LLM, SFT, Dados Sintéticos, Pós-treinamento, Avaliação com LLM >

Abstract

Rios, Walcy. <**Data Refinement for SFT in Brazilian Portuguese: A Study with Language Models and Evaluation via LLM-as Judge**>. Goiânia, 2025. 42p. MSc. Dissertation. Programa de Pós-graduação em Ciência da Computação, Instituto de Informática, Universidade Federal de Goiás.

The scarcity of high-quality resources for Brazilian Portuguese (pt-br) hinders the development of effective language models adapted to the language's specificities. This work investigates the impact of synthetic enhancement of conversational data, using Large Language Models (LLMs), on the Supervised Fine-Tuning (SFT) of models from the Qwen2.5 family (0.5B, 1.5B, 3B). Based on the SmolTalk dataset, two versions were generated for pt-br: one by direct translation and another with responses synthetically enhanced and rewritten by the LLM Gemini 2.0 Flash. The Qwen2.5 models were trained on both datasets and comparatively evaluated using standardized objective benchmarks for Portuguese (ENEM, HATEBR, BLUEX, ASSIN2-RTE) and through qualitative evaluation of open-ended text generation (Alpaca-Eval-BR), using Claude 3.5 Haiku as LLM-as-Judge based on relevance, precision, comprehensiveness, usefulness, and coherence criteria. The results demonstrate a significant superiority of the models trained with synthetic data in the qualitative LLM-as-Judge evaluation across all metrics. In this evaluation, the normalized average F1-Score significantly increased with synthetic data: the 1.5B model achieved 44.45 (vs 14.05 for the translated, a 216% gain) and the 3B model reached 57.21 (vs 16.79 for the translated, a 241% gain). In contrast, on the objective benchmarks, the positive impact of synthetic enhancement was less pronounced, being more consistent only in the 3B parameter version. It is concluded that the LLM-assisted synthetic data enhancement strategy is effective in significantly raising the quality and performance of conversational language models for Brazilian Portuguese, representing a valuable approach to mitigate the scarcity of dedicated resources and advance the development of NLP technologies better adapted to the national context.

Keywords

<LLM, SFT, LLM-as-Judge, Qwen2.5, Synthetic Data, Post-Training>

Sumário

Lista de Figuras	6
Lista de Tabelas	7
1 Introdução	8
1.0.1 Problema de Pesquisa	9
1.0.2 Objetivos	10
1.0.3 Contribuições	10
1.0.4 Organização do Trabalho	11
2 Fundamentação Teórica	12
2.1 Modelos Generativos de Texto Fundacionais	12
2.2 Avanços Recentes e Estado da Arte em Modelos Generativos	15
2.3 Modelos de Linguagem para o Português Brasileiro	17
2.4 LLMs Privados do Estado da Arte	19
2.5 LLM como Avaliador (LLM-as-a-Judge)	19
3 Materiais e Métodos	21
3.1 Materiais	21
3.1.1 Dataset SmolTalk	21
3.1.2 Proposta de Tradução e Aprimoramento	21
3.1.3 Arquitetura Qwen2.5	22
3.1.4 Treinamento Supervisionado com Interação Multi-turno	23
3.1.5 Configuração do Treinamento	24
3.2 Métodos de Avaliação	24
3.2.1 Avaliação com benchmarks	24
3.2.2 Avaliação de Geração de texto aberto com Alpaca-Eval-BR	25
4 Resultados	27
4.0.1 Análise do Dataset	27
4.0.2 Resultados do Treinamento	27
4.0.3 Resultados dos Benchmarks	28
4.0.4 Resultados da Geração de Texto Aberto no Alpaca-Eval	29
5 Conclusão	32
Referências Bibliográficas	34
A Apêndice	40

Lista de Figuras

2.1	Curva de perda no treinamento do GPT-3 com variação de parâmetros treináveis em escala logaritmica. Eixo x: Tokens Processados em Bilhões. Eixo y: Valor do erro da entropia cruzada. Fonte: [Brown et al. 2020]	14
2.2	Expressões regulares de pré-tokenização do GPT-2 e GPT-4, decompõem o texto em subpartes funcionais através de regex e outra versão chamada Punct, introduzimos pelo autor [Dagan, Synnaeve e Roziere 2024] como estudo ablativo simplificando para a língua inglesa. Fonte: [Dagan, Synnaeve e Roziere 2024]	16
2.3	Disposição da camada de atenção: Multi-Head (MHA) - performance, Grouped-Query - balanceamento de qualidade e desempenho e MQA - velocidade. Fonte: [Ainslie et al. 2023]	16
4.1	Gráfico comparativo da função de custo ao longo do treinamento para diferentes configurações e tamanhos de modelos	28
4.2	F1-Score entre benchmarks HATEBR, ASSIN2RT2, ENEM e BLUEX. Qwen2.5 0.5b, 1.5b e 3b, cenário traduzido e sintético.	30
4.3	Métricas F1-Score normalizado nas categorias Relevância, Precisão, Abrangência, Utilidade e Coerência, modelo Qwen2.5 1.5b e 3b, no cenário traduzido e sintético. Avaliado pelo Claude 3.5 Haiku <i>LLM-as-Judge</i>	30
4.4	Preferência do <i>LLM as Judge</i> (Claude 3.5 Haiku) na resposta gerada e resposta referência no Alpaca-Eval-Pt. Vitória como verde, e Empate em amarelo.	31

Lista de Tabelas

3.1	Critérios de Avaliação de Respostas Abertas	26
4.1	Análise estatística comparativa entre dados traduzidos e sintéticos	27
4.2	Tabela de comparação de frases e modificações	31

Introdução

O surgimento de modelos robustos de linguagem (LLMs, do inglês *Large Language Models*) representou um avanço significativo em toda a área de Processamento de Linguagem Natural (NLP, em inglês *Natural Language Processing*), trazendo melhorias consideráveis em uma vasta gama de tarefas, com mínima necessidade de supervisão [Thirunavukarasu et al. 2023, Lai et al. 2024, Meyer et al. 2023, Gan et al. 2023, Li et al. 2023]. Nos últimos anos, a rápida evolução desses modelos foi impulsionada por esforços globais em aumentar progressivamente o tamanho dos modelos e conjuntos de dados empregados. Como exemplo dessa trajetória, modelos passaram rapidamente de poucos bilhões de parâmetros há cinco anos — GPT-2 com 1,5 bilhão de parâmetros — para modelos com centenas de bilhões de parâmetros na atualidade, tal como DeepSeek-v3 com 671B [Liu et al. 2024], Llama-3.1 com 405B [Grattafiori et al. 2024], Llama 4 com 400B [Meta AI 2024] e Megatron Turing com 530 bilhões de parâmetros [Smith et al. 2022].

Entretanto, a abordagem centrada em modelos generalistas extremamente grandes, embora capaz de atingir competências multilíngues impressionantes, enfrenta desafios significativos em relação ao custo computacional e acessibilidade para pesquisadores com recursos limitados. Diante dessa limitação, uma alternativa promissora tem sido o desenvolvimento de modelos especializados, que se destacam por oferecer maior eficiência e desempenho em tarefas específicas, sem que seja necessário ampliar drasticamente o número de parâmetros [Ling et al. 2023, Fu et al. 2023, Saadi et al. 2025].

Esse direcionamento para modelos especializados alinha-se com as necessidades observadas no contexto da língua portuguesa, em especial o português brasileiro. Além das considerações gerais sobre eficiência computacional e desempenho focado, a especialização é motivada pela escassez significativa de modelos robustos já disponíveis e pela limitada disponibilidade de recursos e materiais adequados, sobretudo conjuntos de dados específicos. Adicionalmente, constata-se que a maioria dos modelos multilíngues existentes apresenta desempenho desproporcional para o português, geralmente enviesado para línguas de maior recurso como o inglês, o que reforça a busca por soluções dedicadas [Garcia et al. 2024, Saadi et al. 2025]. A dificuldade é comple-

mentada pela notória carência de conjuntos de dados extensos e de alta qualidade para a fase de pós-treino em português brasileiro, um fator que dificulta o desenvolvimento eficaz de modelos bem adaptados. Em resposta a esses desafios, iniciativas recentes como a coleção do Sabiá [Pires et al. 2023, Almeida et al. 2024, Abonizio et al. 2025], Cabrita [Larcher et al. 2023], TeenyTinyLlama [Corrêa et al. 2024] e Tucano [Corrêa et al. 2024] têm buscado justamente superar essas limitações, adotando estratégias específicas para aumentar a qualidade e o desempenho na geração textual em português.

Diante desse cenário de escassez, uma estratégia que tem ganhado destaque é a geração de dados sintéticos, que se referem a dados gerados artificialmente visando mimetizar as características e padrões de dados reais. Os próprios modelos de língua emergiram como ferramentas centrais nesse processo, possibilitando a criação unificada de conjuntos de dados sintéticos em larga escala [Huang et al. 2024]. A hipótese, explorada em trabalhos como [Gunasekar et al. 2023] e evidenciada pelo desenvolvimento de modelos como Phi-3 [Abdin et al. 2024], é que dados sintéticos cuidadosamente curados, mesmo que não em volumes massivos, podem ser extremamente eficazes para treinar modelos de alto desempenho. O próprio ato de gerar datasets sintéticos também oferece uma perspectiva valiosa para a compreensão das capacidades e limitações dos LLMs [Dekoninck et al. 2024].

No entanto, reconhecer que o uso de LLMs como geradores de dados, embora possa aumentar a diversidade, também carrega o risco de introduzir ou amplificar vieses indesejados, exigindo atenção cuidadosa durante o processo [Yu et al. 2023]. Quando aplicada criteriosamente, a geração sintética permite não apenas suprir a falta de dados reais, mas também controlar características do dataset para promover melhor equilíbrio e generalização, contribuindo para a eficiência e robustez de modelos especializados, como investigado por diversas pesquisas recentes [Liu et al. 2024, Guo e Chen 2024, Long et al. 2024].

1.0.1 Problema de Pesquisa

A presente pesquisa aborda a lacuna existente no desenvolvimento de recursos linguísticos computacionais de alta qualidade para o português brasileiro. Apesar da expressiva base de falantes e da significativa presença digital do idioma, há uma carência de conjuntos de dados e modelos de linguagem abertos e adaptados às suas especificidades [Garcia et al. 2024, Lopes, Magalhães e Semedo 2024], resultando em uma dependência de modelos multilíngues que frequentemente falham em capturar as nuances linguísticas e culturais do pt-br.

Diante disso, o problema central investigado é: **qual o impacto do aprimoramento sintético de um conjunto de dados conversacional existente, utilizando um**

Large Language Model (LLM) de grande porte para reescrita e detalhamento, no desempenho de modelos de linguagem abertos de diferentes tamanhos (Qwen2.5 0.5B, 1.5B, 3B) treinados especificamente para o português brasileiro, quando comparado ao treinamento com dados apenas traduzidos?

1.0.2 Objetivos

O objetivo geral deste trabalho é investigar e avaliar a eficácia do aprimoramento sintético de dados conversacionais, via LLM, como estratégia para melhorar a qualidade e o desempenho de modelos de linguagem abertos ajustados para o português brasileiro. Para alcançar este objetivo geral, foram definidos os seguintes objetivos específicos:

- Adaptar o conjunto de dados conversacional *SmolTalk* para o português brasileiro através de tradução turno a turno com interação com Gemini 2.0 Flash;
- Gerar uma versão sinteticamente aprimorada do *SmolTalk* em português, utilizando o modelo Gemini 2.0 Flash para reescrever e enriquecer as respostas do assistente;
- Realizar o Treinamento Supervisionado (SFT) dos modelos Qwen2.5 nas versões de 0.5B, 1.5B e 3B parâmetros, utilizando separadamente os conjuntos de dados traduzido e sintético;
- Avaliar comparativamente o desempenho dos modelos treinados em benchmarks objetivos padronizados para o português brasileiro (ENEM, HATEBR, BLUEX, ASSIN2-RTE) utilizando a métrica F1-Score;
- Avaliar comparativamente a qualidade da geração de texto aberto dos modelos treinados (1.5B e 3B) utilizando a metodologia LLM-as-Judge com o modelo Claude 3.5 Haiku, com base nos critérios de relevância, precisão, abrangência, utilidade e coerência;
- Analisar e validar a hipótese de que o conjunto de dados sinteticamente aprimorado resulta em modelos com maior qualidade nas respostas generativas em comparação com o conjunto apenas traduzido.

1.0.3 Contribuições

Este trabalho apresenta as seguintes contribuições principais para a área de Processamento de Linguagem Natural (NLP) para o português brasileiro:

- A criação e disponibilização de duas novas versões do dataset conversacional *SmolTalk* adaptadas para o português brasileiro: uma versão traduzida e uma versão sinteticamente aprimorada com o auxílio do Gemini 2.0 Flash;

- O treinamento e a análise comparativa de modelos da família Qwen2.5 (0.5B, 1.5B, 3B) ajustados especificamente para tarefas conversacionais em pt-br, utilizando dados de diferentes processos de criação (tradução vs. aprimoramento sintético);
- A demonstração empírica e quantificação da eficácia do aprimoramento sintético de dados, mediado por LLM, como uma estratégia viável e vantajosa para melhorar o desempenho de modelos de linguagem abertos em pt-br, especialmente em tarefas de geração de texto aberto e qualitativas;
- Uma análise do impacto diferencial do tamanho do modelo (0.5B, 1.5B vs. 3B) na capacidade de se beneficiar do aprimoramento dos dados de treinamento.

1.0.4 Organização do Trabalho

Este documento está estruturado da seguinte forma: O Capítulo 2 aborda o referencial teórico, discutindo conceitos fundamentais sobre modelos de linguagem, avanços de importantes nos LLM, treinamento supervisionado, arquiteturas relevantes (Qwen2.5), LLMs utilizados (Gemini, Claude), avaliação de modelos LLM-as-Judge e o cenário do NLP para o português. O Capítulo 3 detalha a metodologia empregada, incluindo a adaptação do dataset, o processo de geração sintética, a configuração do treinamento dos modelos e os protocolos de avaliação. O Capítulo 4 apresenta e analisa os resultados obtidos, incluindo a análise do dataset, as curvas de treinamento e o desempenho dos modelos nos benchmarks objetivos e na avaliação de geração de texto aberto. Finalmente, o Capítulo 5 sumariza os principais achados do estudo, discute suas implicações, limitações e aponta direções para trabalhos futuros.

Fundamentação Teórica

2.1 Modelos Generativos de Texto Fundacionais

Os modelos de linguagem grandes (Large Language Models - LLMs) representaram um avanço na capacidade de gerar texto com coerência e fluidez [Brown et al. 2020]. O GPT-2 (Generative Pre-trained Transformer 2) [Radford et al. 2019] e seu sucessor, GPT-3 [Brown et al. 2020], demonstraram a eficácia de arquiteturas baseadas em Transformers [Vaswani et al. 2017] pré-treinadas em vastos volumes de dados textuais. O paradigma de treinamento fundamental para esses modelos é autorregressivo. Dado uma sequência de tokens $X = (x_1, x_2, \dots, x_n)$, o modelo aprende a prever o próximo token x_i com base nos tokens anteriores $x_{<i} = (x_1, \dots, x_{i-1})$. A função objetivo é tipicamente a maximização da verossimilhança do conjunto de dados de treinamento, que é fatorada como o produto das probabilidades condicionais:

$$P(X; \theta) = \prod_{i=1}^n P(x_i | x_1, \dots, x_{i-1}; \theta) \quad (2-1)$$

onde θ representa os parâmetros do modelo. Na prática, isso corresponde à minimização da verossimilhança logarítmica negativa (negative log-likelihood) sobre o corpus de treinamento.

Para processar o texto bruto, esses modelos empregam técnicas de tokenização como o Byte Pair Encoding (BPE) [Sennrich, Haddow e Birch 2015]. O BPE inicia com um vocabulário base de caracteres individuais e, iterativamente, mescla os pares de tokens adjacentes mais frequentes no corpus de treinamento para formar novas unidades de subpalavra (subword units). Esse método permite representar um vocabulário extenso de forma eficiente, balanceando a granularidade entre caracteres e palavras inteiras, e lidando com palavras raras ou fora do vocabulário ao decompô-las em subunidades conhecidas.

A arquitetura central a esses modelos é o Transformer [Vaswani et al. 2017], que superou as redes recorrentes e convolucionais tradicionais em favor de mecanismos de auto-atenção *self-attention*. A auto-atenção permite ao modelo ponderar a importância

de diferentes tokens na sequência de entrada ao computar a representação de um token específico. A operação central é a atenção de produto escalar escalonado (scaled dot-product attention), definida como:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2-2)$$

onde Q (Query), K (Key) e V (Value) são matrizes obtidas por projeções lineares das representações dos tokens de entrada, d_k é a dimensão dos vetores de chave (e query), e o fator de escala $\frac{1}{\sqrt{d_k}}$ previne que os produtos escalares se tornem excessivamente grandes, estabilizando os gradientes durante o treinamento. O mecanismo de Multi-Head Attention aprimora isso, permitindo que o modelo atenda conjuntamente a informações de diferentes subespaços de representação em diferentes posições, projetando Q , K e V múltiplas vezes (h vezes, ou "cabeças") com diferentes projeções lineares aprendidas, computando a atenção em paralelo para cada cabeça e, subsequentemente, concatenando e projetando linearmente os resultados. A arquitetura Transformer completa consiste em múltiplas camadas desses blocos de atenção, intercalados com redes feed-forward, conexões residuais e normalização de camada *layer normalization*, além de codificações posicionais *positional encodings* adicionadas às entradas para adicionar informação sobre a ordem dos tokens, já que o mecanismo de atenção por si só é invariante a permutações.

O lançamento do GPT-3 [Brown et al. 2020] marcou um ponto de inflexão, não apenas pela escalada em magnitude — saltando de 1.5 bilhão de parâmetros no maior modelo GPT-2 para 175 bilhões — mas principalmente pela demonstração de capacidades emergentes associadas a essa escala. Uma característica notável observada foi a sua habilidade como um "few-shot learner". Diferentemente do paradigma predominante que requeria fine-tuning específico para cada tarefa subsequente, o GPT-3 demonstrou proficiência em realizar uma ampla gama de tarefas de processamento de linguagem natural recebendo apenas uma descrição da tarefa e/ou alguns poucos exemplos diretamente em seu prompt de entrada (input prompt), um fenômeno denominado aprendizado no contexto *in-context learning*. Essa capacidade de adaptação rápida, manifestando-se como zero-shot, one-shot e few-shot learning, dispensou a necessidade de atualizações de gradiente ou ajustes finos nos parâmetros do modelo para muitas aplicações. Além disso, o GPT-3 exibiu um nível de generalização sem precedentes, performando de maneira competente em tarefas para as quais não foi explicitamente treinado, incluindo tradução, resposta a perguntas, sumarização, geração de código e raciocínio, muitas vezes com desempenho comparável ou superior a modelos estado-da-arte que haviam sido especificamente ajustados para aquelas tarefas. Essas habilidades sugeriram que o aumento massivo na quantidade de parâmetros e no volume e diversidade dos dados de pré-treinamento poderia induzir a aquisição de meta-aprendizagem e uma compreensão mais abstrata da linguagem e do

raciocínio subjacente às tarefas.

A Figura 2.1 do estudo [Brown et al. 2020] ilustra quantitativamente o impacto da escala nos modelos da família GPT-3. As curvas de treinamento e validação demonstram a relação entre o tamanho do modelo (variando de centenas de milhões até 175 bilhões de parâmetros, conforme indicado pela escala de cores), a quantidade de dados processados (até 300 bilhões de tokens) e a performance, medida pela perda de entropia cruzada (*cross-entropy loss*). Observa-se que modelos com maior número de parâmetros consistentemente atingem níveis de perda mais baixos, tanto no conjunto de treinamento quanto no de validação, para uma mesma quantidade de tokens processados. A perda de validação continua a diminuir mesmo com centenas de bilhões de tokens, indicando que os modelos maiores continuam a generalizar melhor com mais dados, sem saturação aparente no regime de dados apresentado. Adicionalmente, conforme apontado por [Brown et al. 2020], a diferença entre a perda de treinamento e a perda de validação cresce de forma apenas marginal com o aumento do tamanho do modelo e do tempo de treinamento. Isso sugere que, apesar da vasta capacidade dos modelos maiores, o overfitting não aumentou drasticamente com a escala, e a melhoria na performance de validação reflete um aprendizado das complexidades da distribuição de linguagem natural presente nos dados.

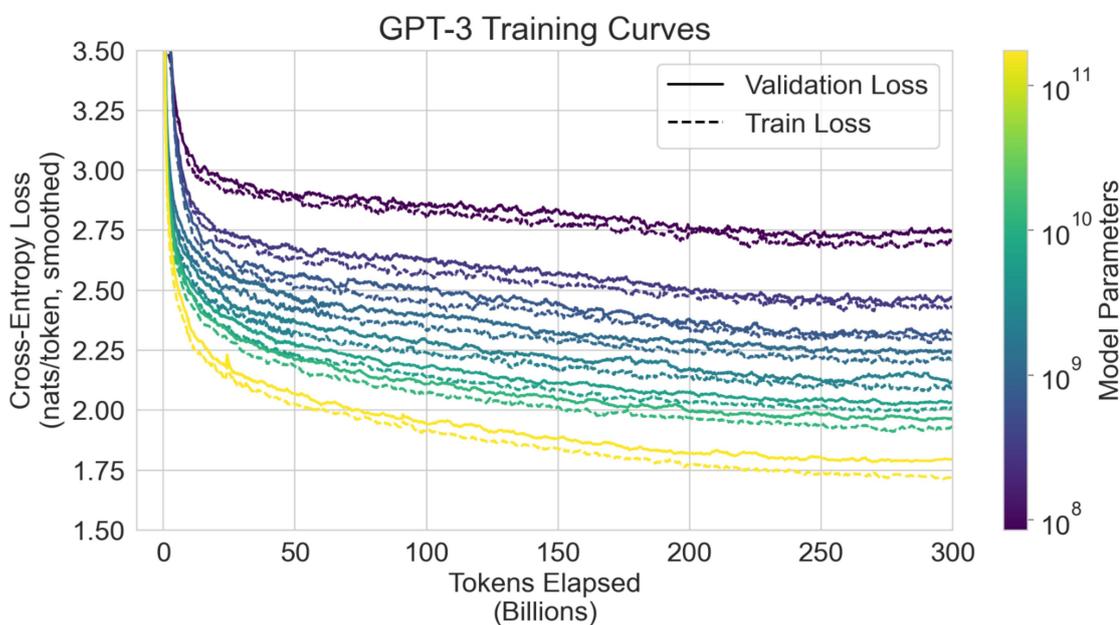


Figura 2.1: Curva de perda no treinamento do GPT-3 com variação de parâmetros treináveis em escala logarítmica. Eixo x: Tokens Processados em Bilhões. Eixo y: Valor do erro da entropia cruzada.

Fonte: [Brown et al. 2020]

2.2 Avanços Recentes e Estado da Arte em Modelos Generativos

Após os modelos fundacionais como o GPT-3, a pesquisa em modelos generativos de texto continuou a avançar em múltiplas frentes, refinando as metodologias de treinamento e arquitetura. A exploração das leis de escala (*scaling laws*) [Kaplan et al. 2020] forneceu um entendimento mais preditivo sobre a relação entre o tamanho do modelo, a quantidade de dados e o desempenho computacional, orientando o desenvolvimento de modelos subsequentes. Modelos recentes, como os da família Llama 3 [Grattafiori et al. 2024] e Qwen 2.5 [Qwen et al. 2025], exemplificam a continuidade dessa tendência, sendo pré-treinados em volumes de dados ainda maiores, frequentemente na escala de múltiplos trilhões de tokens, e por períodos de tempo mais extensos, visando extrair o máximo potencial das arquiteturas.

Paralelamente à escala, uma ênfase crescente tem sido colocada na qualidade e diversidade dos dados de pré-treinamento. Em vez de apenas aumentar a quantidade bruta, a curadoria rigorosa dos datasets tornou-se um fator crítico. Isso envolve filtragem extensiva para remover conteúdo de baixa qualidade ou indesejado, técnicas de deduplicação para aumentar a eficiência da aprendizagem e um esforço consciente para diversificar as fontes de dados, cobrindo uma gama mais ampla de domínios, estilos e idiomas [Allal et al. 2024, Penedo et al. 2024]. A importância da qualidade dos dados foi particularmente destacada por trabalhos como a série de modelos Phi [Gunasekar et al. 2023], que demonstraram que modelos com contagens de parâmetros relativamente menores podem alcançar desempenho notável em tarefas de raciocínio e codificação quando treinados em dados de alta qualidade, comparáveis a "textbooks", mesmo que em volumes menores que os gigantes da indústria.

Avanços também ocorreram na tokenização. Embora o BPE e suas variantes permaneçam comuns, houve esforços para otimizar a fase de pré-tokenização com regex, ilustrado na Figura 2.2, até a otimização dos vocabulários, buscando um melhor balanceamento para cenários multilíngues e maior eficiência na representação de linguagens de programação ou outros domínios específicos [Dagan, Synnaeve e Roziere 2024]. Modelos como Qwen [Qwen et al. 2025] e Llama 3 [Grattafiori et al. 2024] frequentemente empregam vocabulários maiores e mais adaptados às suas massivas bases de dados multilíngues [Tao et al. 2024].

No nível arquitetônico, embora a base do Transformer permaneça, otimizações para melhorar a eficiência computacional, especialmente durante a inferência, ganharam tração. Um exemplo notável é a Atenção de Consulta Agrupada (*Grouped-Query Attention* - GQA) [Ainslie et al. 2023], que oferece um meio-termo entre a *Multi-Head Attention* (MHA) padrão e a *Multi-Query Attention* (MQA) [Shazeer 2019], reduzindo os

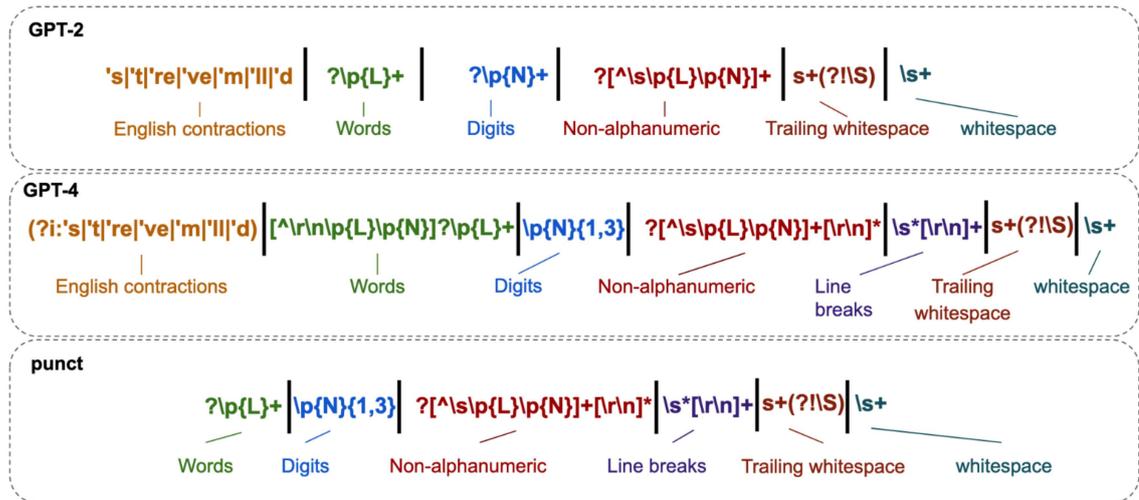


Figura 2.2: Expressões regulares de pré-tokenização do GPT-2 e GPT-4, decompõem o texto em subpartes funcionais através de regex e outra versão chamada Punct, introduzimos pelo autor [Dagan, Synnaeve e Roziere 2024] como estudo ablativo simplificando para a língua inglesa.

Fonte: [Dagan, Synnaeve e Roziere 2024]

requisitos de memória e acelerando a geração de tokens sem sacrificar significativamente a qualidade do modelo. A Figura 2.3 ilustra o balanceamento da velocidade entre MQA, com proposta de apenas um vetor de escrita, com a performance da tradicional da MHA.

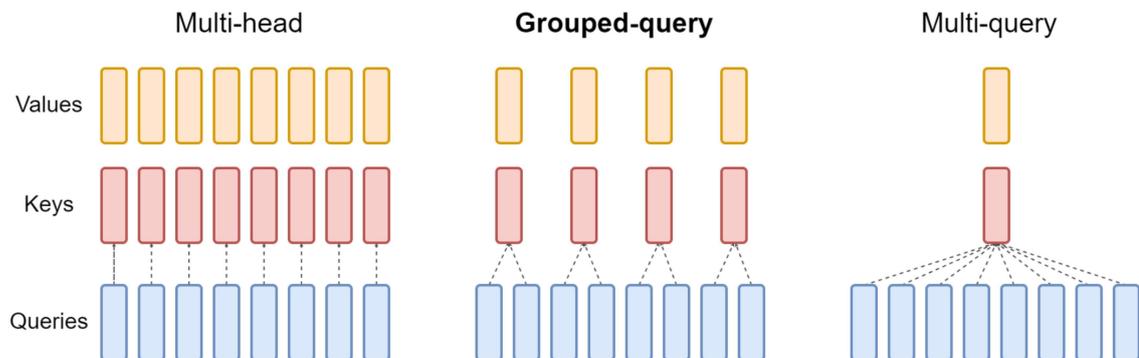


Figura 2.3: Disposição da camada de atenção: Multi-Head (MHA) - performance, Grouped-Query - balanceamento de qualidade e desempenho e MQA - velocidade.

Fonte: [Ainslie et al. 2023]

Além dos avanços em arquitetura e dados, o desenvolvimento de técnicas de otimização de treinamento foi indispensável para escalar os LLMs contemporâneos. Duas inovações de destaque abordaram gargalos críticos de memória e computação. Primeiramente, o Otimizador de Redundância Zero (ZeRO), introduzido com o DeepSpeed [Rasley et al. 2020], combate a limitação de memória da paralelização de dados padrão

(DP) ao eliminar a replicação de estados do modelo. Através do particionamento dos estados do otimizador, gradientes e até mesmo dos parâmetros (em seus estágios progressivos, culminando no ZeRO-1/2/3) entre os dispositivos de computação (multi-GPU/multi-nó), o ZeRO reduz drasticamente a pegada de memória *memory footprint* por dispositivo, permitindo o treinamento eficiente de modelos cujos requisitos excedem a capacidade de uma única GPU. Em segundo lugar, o FlashAttention (1 e 2) [Dao et al. 2022, Dao 2023] reformulou o cálculo da auto-atenção. Evitando a materialização da matriz de atenção completa ($N \times N$) na memória HBM da GPU e utilizando algoritmos I/O-aware com tiling e recomputação na SRAM, o FlashAttention reduz a complexidade de memória da atenção em relação ao comprimento da sequência (N) de $O(N^2)$ para $O(N)$, acelera significativamente a operação e habilita o uso de janelas de contexto muito mais longas.

Finalmente, a fase de pós-treinamento tornou-se um componente ainda mais central e sofisticado no desenvolvimento de LLMs. O Ajuste Fino Supervisionado (Supervised Fine-Tuning - SFT) agora utiliza conjuntos de dados de instruções de alta qualidade, muitas vezes com curadoria humana detalhada, para ensinar o modelo a seguir instruções complexas e a interagir de maneira útil. Técnicas de alinhamento subsequentes, como o Aprendizado por Reforço com Feedback Humano (RLHF) ou alternativas como o Direct Preference Optimization (DPO), são aplicadas para refinar ainda mais o comportamento do modelo, alinhando-o com as preferências humanas em termos de utilidade, honestidade e segurança. Modelos como Llama 3 [Grattafiori et al. 2024] e Qwen 2.5 [Qwen et al. 2025] destacam o uso intensivo dessas técnicas com grandes volumes de dados de feedback e preferência para alcançar seu desempenho no estado da arte. Esses avanços combinados resultaram em modelos que não apenas possuem forte capacidade generativa bruta, mas também são mais controláveis, alinhados e eficazes em seguir instruções complexas.

2.3 Modelos de Linguagem para o Português Brasileiro

O desenvolvimento de modelos de linguagem grandes (LLMs) de alta capacidade para o Português Brasileiro (PT-BR) enfrenta desafios particulares, a disparidade em recursos computacionais e datasets larga escala quando comparado ao ecossistema da língua inglesa. Contudo, diversas iniciativas têm buscado superar essas barreiras, gerando modelos progressivamente mais proficientes no idioma e contexto cultural brasileiro.

O trabalho com o Sabiá-1 [Pires et al. 2023] foi um marco inicial importante. Ao realizar um pré-treinamento continuado monolíngue em português sobre modelos como LLaMA, demonstrou-se que mesmo com um orçamento computacional modesto, era possível obter melhorias substanciais de desempenho em tarefas específicas do idioma (avaliadas no benchmark Poeta), com o Sabiá-65B atingindo paridade com o GPT-3.5-

turbo da época. Este estudo demonstrou os benefícios da especialização monolíngue para capturar conhecimento de domínio e cultural, e uma versão menor (Sabiá-7B) foi disponibilizada abertamente para a comunidade.

Em contraste com a adaptação de modelos existentes, o projeto Tucano [Corrêa et al. 2024] focou na criação de recursos fundamentais abertos. Foi desenvolvido o corpus GigaVerbo (aprox. 200 bilhões de tokens em português) e, a partir dele, treinada a família de modelos Tucano. A disponibilização aberta tanto do corpus quanto dos modelos representou uma contribuição valiosa para fomentar a pesquisa e desenvolvimento independentes na área. Além disso, o estudo criticou a adequação de benchmarks existentes para PT-BR, apontando para a necessidade de métricas mais refinadas.

Posteriormente, a iniciativa privada Maritaca AI apresentou as famílias Sabiá-2 [Almeida et al. 2024] e, mais recentemente, Sabiá-3 [Abonizio et al. 2025]. O Sabiá-2 destacou-se pela sua avaliação em exames complexos do contexto brasileiro, onde o modelo Sabiá-2 Medium demonstrou competitividade notável frente ao GPT-4 em diversas áreas, reforçando os ganhos da especialização. Seguindo essa linha, o Sabiá-3 foi apresentado como o novo modelo de ponta da empresa, treinado em um corpus ainda maior e focado no Brasil [Abonizio et al. 2025]. Relata-se que o Sabiá-3 oferece melhorias significativas sobre o Sabiá-2, especialmente em tarefas de raciocínio, codificação e manipulação de textos longos, atingindo um desempenho médio comparável aos modelos de fronteira globais em tarefas relevantes para o português, mantendo uma vantagem de custo por token devido à especialização. Contudo, os próprios desenvolvedores apontam para limitações em tarefas *multi-step* e no seguimento de instruções complexas como áreas para desenvolvimento futuro.

É crucial notar, entretanto, que diferentemente do Sabiá-1 (7B) e do projeto Tucano, os modelos Sabiá-2 e Sabiá-3, assim como seus datasets de treinamento e metodologias detalhadas, são proprietários e não foram disponibilizados abertamente. Embora representem avanços técnicos e de performance para o PT-BR, alcançando o estado da arte em diversas métricas, essa natureza fechada impõe uma limitação significativa ao avanço mais amplo do campo no Brasil. A falta de acesso aberto a modelos de ponta, aos dados massivos e curados utilizados em seu treinamento, e aos detalhes de implementação dificulta a pesquisa acadêmica independente, a inovação comunitária e a replicação de resultados. Enquanto iniciativas como o Tucano fornecem infraestrutura aberta, a ausência de modelos abertos que acompanhem o desempenho dos melhores sistemas proprietários focados em PT-BR, somada às restrições gerais de recursos computacionais, deixa o ecossistema de pesquisa e desenvolvimento em LLMs para o português brasileiro, em certa medida, "encilhado", dependente de esforços comerciais fechados para definir o estado da arte.

2.4 LLMs Privados do Estado da Arte

Dois expoentes recentes no campo dos Grandes Modelos de Linguagem relevantes para este trabalho são as famílias Gemini, desenvolvida pelo Google DeepMind, e Claude 3.5, da Anthropic. A família Gemini foi introduzida com o Gemini 1.0, destacando-se por sua arquitetura multimodal nativa, capaz de processar e raciocinar sobre texto, imagem, áudio e vídeo, e oferecida em diferentes níveis de capacidade (Ultra, Pro e Nano) para diversas aplicações [Team et al. 2023]. Atualizações subsequentes, como a introdução da família Gemini 1.5, trouxeram avanços significativos, notavelmente com o Gemini 1.5 Pro apresentando uma janela de contexto estendida de até 1 milhão de tokens (ou mais em contextos de pesquisa) e uma arquitetura Mixture-of-Experts (MoE) para maior eficiência computacional [Google DeepMind 2025]. Dentro desta linha, variantes como o Gemini 1.5 Flash foram otimizadas para oferecer um balanço entre alta performance, velocidade e custo-efetividade, tornando-se adequadas para tarefas de geração em larga escala, como o aprimoramento sintético de dados realizado nesta pesquisa. Paralelamente, a Anthropic lançou a família Claude 3.5, cujo modelo de entrada, Claude 3.5 Haiku, foi projetado para ser o mais rápido e acessível da sua classe [Anthropic 2025]. O Haiku visa fornecer respostas quase instantâneas e lidar com alto volume de interações, sendo ideal para aplicações como chatbots de suporte, moderação de conteúdo e, crucialmente para este estudo, como um avaliador automatizado (LLM-as-Judge) eficiente e capaz, papel que desempenhou na avaliação qualitativa das gerações dos modelos Qwen2.5 treinados. Por tratar-se de modelos de língua privados, não é disponibilizado informações técnicas entre as versões, como por exemplo Gemini 1.0 para o 2.0, apenas melhoria nos conjuntos de avaliações publicados.

2.5 LLM como Avaliador (LLM-as-a-Judge)

A capacidade dos Modelos de Linguagem Grandes de processar informações, emular o raciocínio humano e avaliar entradas com base em regras predefinidas impulsionou o paradigma conhecido como "LLM-as-a-Judge" (LLM como Avaliador) [Gu et al. 2024]. Este conceito refere-se à utilização de LLMs para realizar tarefas de avaliação que tradicionalmente eram executadas por humanos. A adoção dessa abordagem tem crescido rapidamente devido a fatores como escalabilidade, adaptabilidade e custo-efetividade, permitindo que LLMs gerenciem um volume crescente de tarefas avaliativas em diversos cenários e com diferentes objetivos. O avanço de técnicas de treinamento como o Aprendizado por Reforço com Feedback Humano (RLHF) foi fundamental nesse processo, ao alinhar os LLMs mais estreitamente com os valores e processos de raciocínio humanos, facilitando sua transição de tarefas puramente generativas para funções

avaliativas [Li et al. 2024].

O termo "LLM-as-a-Judge" abrange um espectro amplo de papéis avaliativos. Conforme descrito por [Gu et al. 2024], isso inclui LLMs atuando como Classificadores, Avaliadores, Críticos, Verificadores, Examinadores e Modelos de Recompensa ou Ranking (Reward/Ranking Models), entre outros. Em essência, trata-se de usar um LLM para julgar objetos, ações ou decisões com base em critérios, regras ou preferências estabelecidas.

Uma das abordagens mais comuns para implementar o LLM-as-a-Judge é através do Aprendizado no Contexto (In-Context Learning - ICL). Nesse método, a tarefa de avaliação é especificada diretamente no prompt fornecido ao LLM, que pode incluir instruções detalhadas e exemplos ou demonstrações para guiar o raciocínio e julgamento do modelo.

Dentro do design do prompt para ICL, diversas estratégias podem ser adotadas para extrair a avaliação [Gu et al. 2024]. Uma abordagem intuitiva é solicitar a geração de scores, que podem ser discretos (e.g., escalas de 1-5 ou 1-10) ou contínuos (e.g., 0-1 ou 0-100), com critérios de pontuação que variam do simples ao complexo, podendo envolver múltiplas dimensões avaliativas e escalas. Outras estratégias incluem formular a avaliação como uma questão binária (Sim/Não), realizar comparações par-a-par (solicitando ao LLM que escolha a melhor entre duas opções), ou apresentar a tarefa como uma seleção de múltipla escolha. A escolha da estratégia depende da natureza da tarefa avaliativa e do tipo de saída desejada.

Materiais e Métodos

3.1 Materiais

3.1.1 Dataset SmolTalk

O dataset *SmolTalk* [Allal et al. 2025] é um conjunto sintético desenvolvido especificamente para o ajuste supervisionado (*Supervised Fine-tuning - SFT*) de modelos de linguagem. Este conjunto possui cerca de um milhão de amostras em inglês, abrangendo diversas tarefas como edição de texto, reescrita, sumarização e raciocínio lógico. O SmolTalk foi construído com uma composição diversificada de datasets públicos e sintéticos, destacando-se: OpenHermes 2.5 [Teknium 2023], MetaMath [Yu et al. 2023], NuminaMath [LI et al. 2024], LongAlign [Bai et al. 2024], Everyday Conversations [Face 2024], APIGen [Liu et al. 2024] e Explore-Instruct [Fanqi et al. 2023]. Para a criação deste dataset, foram utilizados LLM recentes de geração de dados, incluindo técnicas como o uso de prompts específicos e filtragens para garantir a qualidade dos dados. O desempenho superior do SmolTalk foi evidenciado em benchmarks como IFEval e MT-Bench, quando comparado a outros datasets públicos, como OpenHermes e Magpie-Pro.

3.1.2 Proposta de Tradução e Aprimoramento

Para a adaptação do dataset SmolTalk ao contexto de pesquisa, foram realizadas duas etapas utilizando o modelo Gemini 2.0 Flash [Pichai 2024]:

Tradução Turno a Turno: Aplicação de tradução específica para cada turno das conversas originais do inglês para o português brasileiro, seguindo um prompt estruturado para preservar códigos e cálculos matemáticos disposto completamente [A](#):

Você está responsável por traduzir um trecho de uma conversa do Inglês para o Português brasileiro. Traduza tudo, com exceção de código e cálculos matemáticos. ### Trecho da conversa:

Aprimoramento das Respostas do Assistente: Após a tradução inicial, especificamente os turnos do assistente foram reescritos utilizando um prompt direcionado para elevar a qualidade das respostas, disposto integralmente em [A](#), com a hipótese de que **respostas mais detalhadas e claras impactam positivamente o desempenho do modelo final**:

"Você é um assistente de IA especializado em aprimorar as respostas de outros assistentes de IA. Sua tarefa é refinar e melhorar o turno do 'assistant' para a língua Portuguesa em uma conversa, mantendo a integridade e o contexto da discussão. Siga as seguintes diretrizes:

- Mantenha o objetivo geral da conversa original de forma concisa.
- Se alguma resposta original for vaga ou incompleta, adicione detalhes e informações relevantes.
- Se alguma resposta original contiver erros ou imprecisões, corrija-os.
- Adicione exemplos ou explicações para tornar as respostas mais claras.
- Adote um tom profissional, educado e prestativo.
- Não se desvie do assunto original da conversa.
- Formate as respostas de forma clara e organizada, usando parágrafos, listas ou outros elementos para facilitar a leitura.

Estas duas etapas são alinhadas à literatura, indicando que a qualidade elevada dos dados impacta diretamente na eficácia dos modelos treinados [[Guo e Chen 2024](#), [Long et al. 2024](#)].

3.1.3 Arquitetura Qwen2.5

A escolha do modelo Qwen2.5 [[Qwen et al. 2025](#)] baseou-se principalmente nas melhorias introduzidas durante a fase de pré-treino em comparação com seu antecessor, o Qwen2. Entre as melhorias destacam-se: (1) uma filtragem aprimorada dos dados de treinamento, onde modelos Qwen2-Instruct são utilizados para realizar análises abrangentes e multidimensionais, permitindo uma retenção mais efetiva de dados de alta qualidade e exclusão de exemplos de menor relevância; (2) a inclusão de dados especializados em matemática e codificação provenientes dos modelos Qwen2.5-Math e Qwen2.5-Coder, o que proporciona ao modelo capacidade aprimorada em raciocínio matemático e geração de código; (3) utilização ampliada de dados sintéticos gerados por Qwen2-72B-Instruct e Qwen2-Math-72B-Instruct, refinados posteriormente com modelos específicos de recompensa; e (4) uma mistura de dados otimizada, onde ocorre um balanceamento estratégico dos conteúdos para corrigir a superrepresentação de domínios como comércio eletrônico

e mídias sociais e aumentar a representação de domínios como tecnologia e pesquisa acadêmica.

Essas estratégias resultaram em uma expansão significativa do conjunto de dados pré-treino, passando de 7 trilhões para 18 trilhões de tokens. Além disso, técnicas específicas foram adotadas para ajustar os hiperparâmetros ideais durante o treinamento, como taxa de aprendizado e tamanho do lote, usando leis de escalonamento com base em diversos tamanhos de modelos e quantidades de dados.

Adicionalmente, o Qwen2.5 realiza um pré-treino de contexto longo em duas fases distintas. Inicialmente, com comprimento de contexto fixo em 4.096 tokens, seguido por uma expansão para 32.768 tokens na fase final, utilizando técnicas como RoPE (Rotary Positional Embedding) com frequência base aumentada. Especificamente, o Qwen2.5-Turbo adota uma expansão progressiva ainda mais extensa do contexto durante o treinamento, atingindo contextos de até 262.144 tokens. Essa abordagem progressiva permite melhor adaptação e generalização do modelo a diferentes comprimentos de sequências. Técnicas adicionais como YARN e Dual Chunk Attention (DCA) também foram implementadas para aumentar ainda mais a capacidade do modelo em lidar com longas sequências durante a inferência, reduzindo a perplexidade e garantindo desempenho consistente em diferentes tamanhos de sequência.

Essas inovações no pré-treinamento proporcionaram ao modelo Qwen2.5 desempenho superior em benchmarks comparativos, destacando-se frente a modelos concorrentes como o Gemma e Llama, especialmente em tarefas de compreensão e raciocínio lógico.

3.1.4 Treinamento Supervisionado com Interação Multi-turno

O treinamento supervisionado fino (SFT) realizado utilizou diálogos estruturados no formato multi-turno, aplicando o template ChatML, proposta pela OpenAI para formatar as interações [OpenAI 2023]. Neste treinamento, adotou-se a prática de atualizar os pesos do modelo apenas para as respostas geradas pela função `assistant`, mantendo inalteradas as entradas geradas pela função `user`. Essa abordagem objetiva a capacidade generativa em turnos, e não apenas uma geração de uma única iteração.

Para acomodar as necessidades específicas do treinamento multi-turno, foram feitas alterações no tokenizador, introduzindo tokens especiais adaptados ao formato ChatML. Notavelmente, o token que marca o fim das respostas passou a ser `<im_end>`, facilitando uma segmentação precisa das interações. Um exemplo do formato utilizado está descrito a seguir:

```
<im_start>user
```

```
Texto da pergunta do usuário aqui.<im_end>
```

<im_start>assistant

Texto da resposta do assistente aqui.<im_end>

3.1.5 Configuração do Treinamento

O treinamento SFT foi realizado diretamente a partir das versões base do modelo Qwen2.5 após a etapa inicial de pré-treino. Foram utilizadas as versões 0.5B, 1.5B e 3B, permitindo a análise do impacto do tamanho do modelo no desempenho pós-treinamento. Os conjuntos de dados empregados incluíram dados traduzidos de outras línguas para o português brasileiro e dados sintéticos gerados especificamente para aumentar a diversidade e qualidade das interações.

As configurações do treinamento adotaram parâmetros fixos para todas as versões: epoch= 1, learning rate= 2×10^{-5} , batch size = 64, learning rate scheduler= cosine, optimizer = AdamW8bit

Durante o treinamento, foi utilizada a técnica *Cut Cross Entropy* (CCE), proposta por [Wijmans et al. 2025], que visa resolver o problema do alto consumo de memória causado pela função de perda em modelos com grandes vocabulários. Ao contrário da entropia cruzada tradicional, que materializa toda a matriz de logits em memória global, a CCE calcula apenas o logit do token correto e realiza a operação de log-sum-exp dinamicamente, evitando a necessidade de armazenar grandes matrizes. Esta abordagem reduz o uso de memória global, possibilitando o treinamento integral dos parâmetros dos modelos sem comprometimento na velocidade de treinamento ou convergência, principalmente em arquiteturas com tamanhos menores, cujo principal gargalo é na última camada com mapeamento para o tamanho do vocabulário [Wijmans et al. 2025]. Todos os experimentos foram executados em uma GPU NVIDIA A100 com 80 GB de memória disponível.

3.2 Métodos de Avaliação

3.2.1 Avaliação com benchmarks

A avaliação dos modelos foi conduzida utilizando quatro benchmarks em português brasileiro: HATEBR [Vargas e Carvalho 2024], ASSIN2RTE [Real, Fonseca e Oliveira 2020], ENEM [Nunes et al. 2023] e BLUEX [Almeida et al. 2023]. Cada benchmark apresenta características específicas conforme descrito a seguir:

- **ENEM**: composto por questões de múltipla escolha de duas edições do Exame Nacional do Ensino Médio, exame padronizado para ingresso nas universidades brasileiras.
- **HATEBR**: contendo comentários do Instagram classificados como linguagem abusiva ou não, resultando em uma classificação binária.
- **BLUEX**: constituído por questões extraídas das provas de ingresso das universidades Unicamp (Convest) e USP (Fuvest), ambas com questões de múltipla escolha.
- **ASSIN2RTE**: formado por pares de sentenças simples, com classificação binária de similaridade semântica.

A avaliação dos modelos nestes benchmarks foi realizada sob uma configuração few-shot, onde foram utilizados 5 exemplos de perguntas e respostas como contexto-guia. Esses exemplos foram selecionados aleatoriamente para garantir variabilidade estatística. Foram executadas três rodadas independentes de inferências para cada benchmark, alterando-se os exemplos de contexto entre cada execução. A métrica empregada para avaliação foi o *F1 score*, por se tratar de datasets binários ou de múltipla escolha, semelhante aos estudos [Garcia et al. 2024, Larcher et al. 2023].

3.2.2 Avaliação de Geração de texto aberto com Alpaca-Eval-BR

Para avaliação aberta, foi utilizado o dataset Alpaca-Eval-BR [Li et al. 2023], que contém 805 instruções em português brasileiro. Cada instrução foi submetida ao modelo com geração aberta em configuração de amostragem ($top_p = 1$, *repetition penalty* = 1), com três gerações distintas para cada instrução, resultando em múltiplas amostras por instrução.

Essas amostras foram posteriormente avaliadas utilizando a metodologia *LLM-as-Judge*. Nesta etapa, optou-se pelo modelo Claude-3.5-Haiku, da Anthropic, visando reduzir possíveis vieses provenientes de modelos previamente utilizados. Cada resposta gerada foi comparada com a resposta referência do dataset segundo os critérios de relevância, precisão, abrangência, utilidade/informatividade, clareza e coerência, além de originalidade e criatividade. Prompt completo no Apêndice A

A avaliação foi estruturada em uma escala de 0 a 5, onde 0 indica completa inadequação e 5 indica excelência ou superação da resposta referência, seguindo o seguinte prompt padronizado para garantir uniformidade nas avaliações:

Não foram empregadas métricas tradicionais de sobreposição textual como BLEU ou Rouge, uma vez que o objetivo principal não é a geração espelhada da resposta esperada, mas sim a qualidade intrínseca e a comparação analítica das respostas geradas com as respostas referências por meio de um avaliador externo baseado em LLM.

Critério	Descrição
Relevância	Se a resposta gerada responde diretamente à pergunta e fornece informações pertinentes ao contexto.
Precisão	Se as informações na resposta gerada são factualmente corretas e não contradizem a resposta de referência ou conhecimento geral.
Abrangência	Se a resposta gerada cobre os pontos principais da pergunta, sem omitir informações importantes da referência e possivelmente oferecendo uma cobertura mais completa.
Utilidade/Informatividade	Se a resposta gerada é útil para o usuário, fornecendo insights, explicações relevantes e agregando valor à resposta de referência.
Clareza e Coerência	Se a resposta gerada é fácil de entender, com linguagem clara, concisa e organização lógica.

Tabela 3.1: Critérios de Avaliação de Respostas Abertas

Resultados

4.0.1 Análise do Dataset

O dataset utilizado neste estudo, após tradução, possui um total de 1.098.583 amostras, divididas em dois conjuntos principais: 1.043.649 amostras para treinamento e 54.934 para teste.

Tipo	Média de Tokens	Desvio Padrão	Amostras com Contexto > 3072
Traduzido	764	669	1%
Sintético	1208	859	12%

Tabela 4.1: Análise estatística comparativa entre dados traduzidos e sintéticos

Foi realizada uma análise exploratória para identificar características importantes do dataset, particularmente relacionadas à complexidade e extensão dos diálogos, é inerente maior verbosidade na geração de LLM's de grande porte, ou seja, a resposta tem um teor menos objetivo e mais completo e amigável. Na tabela 4.1 está disposto o tamanho médio das sentenças, 58% de aumento médio do tamanho da sentença do traduzido para o sintético e a quantidade de amostras maiores do que o tamanho de sentença máximo escolhido, 3072 tokens, um aumento de 1% para 12%. Observou-se que 95% das amostras apresentam diálogos com até 6 turnos de interação, indicando que a maior parte das conversas possui interações relativamente curtas. No entanto, foi registrado um limite superior de até 20 turnos para algumas amostras, demonstrando a existência de conversas extensas dentro do dataset.

4.0.2 Resultados do Treinamento

Durante a fase de treinamento, foi analisada a curva de perda obtida através da ferramenta de observabilidade, Weights and Biases (wandb). A Figura 4.1 apresenta o comportamento da função de perda ao longo do treinamento, permitindo comparações entre diferentes configurações e tamanhos de modelos.

Analisando o gráfico observa-se uma redução consistente da perda ao longo das atualizações, o que indica um aprendizado contínuo e eficiente dos modelos avaliados. Destacam-se duas principais observações:

- Os modelos maiores apresentam uma redução mais rápida e estável da função de custo, indicando uma capacidade superior em assimilar informações complexas durante o treinamento. É notado também alguns picos de instabilidade no treinamento do modelo de 3 bilhões de parâmetros, sendo sugerido uma otimização dos hiperparâmetros utilizados, sendo o principal a taxa de aprendizado.
- Não pode-se inferir qualidade comparativa entre as variantes de treinamento sintético e traduzido, apenas pela curva do treinamento. Convergência e estabilidade são fatores a serem analisados na queda do erro intrínseco ao mesmo experimento.



Figura 4.1: Gráfico comparativo da função de custo ao longo do treinamento para diferentes configurações e tamanhos de modelos

4.0.3 Resultados dos Benchmarks

A Figura 4.2 dispõe o desempenho dos modelos avaliados em benchmarks específicos, múltipla escolha, considerando o F1-Macro como métrica de avaliação. Observa-se um padrão: o treinamento no dataset sintético melhora consistentemente a pontuação apenas na versão Qwen2.5 3B. Especificamente, o modelo 3B treinado com dados sintéticos (barras verde-escuras) superou sua contraparte treinada com dados traduzidos (barras azul-escuras) em todos os benchmarks avaliados, alcançando F1-Scores de 0.57 (+0.17 pontos ou +42.5% de melhoria) no HATEBR, 0.79 (+0.16 pontos ou +25.4%) no ASSIN2RTE, 0.46 (+0.13 pontos ou +39.4%) no ENEM, e 0.37 (+0.07 pontos ou +23.3%) no BLUEX, em comparação aos scores de 0.40, 0.63, 0.33 e 0.30, respectivamente, do modelo 3B traduzido.

Em contraste, para os modelos menores (0.5B e 1.5B), o treinamento com dados sintéticos frequentemente resultou em desempenho inferior ou marginalmente diferente ao obtido com dados traduzidos. Por exemplo, no ASSIN2RTE, o modelo 1.5B caiu de 0.66 (traduzido) para 0.62 (sintético), e o 0.5B caiu de 0.53 para 0.47. No HATEBR, o 1.5B foi o único modelo menor a mostrar melhora significativa (de 0.38 para 0.51), enquanto o 0.5B piorou (0.42 para 0.40). Essa divergência pode ser atribuída à natureza mais direta e objetiva das tarefas de benchmark, onde o padrão de resposta mais elaborado e complexo gerado pelo LLM para os dados sintéticos (com exemplificações, markdown, etc.) pode ter desestabilizado o ajuste fino dos modelos com menor capacidade (0.5B e 1.5B), enquanto o modelo maior (3B) conseguiu capitalizar sobre a maior riqueza semântica do dado.

4.0.4 Resultados da Geração de Texto Aberto no Alpaca-Eval

Os resultados obtidos a partir da inferência no Alpaca-Eval e com avaliação do modelo Claude 3.5 Haiku as *LLM-as-Judge* são apresentados na Figura 4.3. Nesta análise, as respostas foram avaliadas quanto à relevância, precisão, abrangência, utilidade e coerência. Visto que trata-se de tarefas de geração de texto aberto, e não mais de múltipla escolha.

Analisando as médias de F1-Score normalizado para cada cenário experimental, a diferença de desempenho é notável, especialmente entre os tipos de dados de treinamento. Os modelos treinados com dados sintéticos apresentaram um rendimento substancialmente superior: o modelo 1.5B sintético (média 44.45) superou sua contraparte traduzida (média 14.05) em mais de 30 pontos, enquanto o modelo 3B sintético (média 57.21) excedeu o 3B traduzido (média 16.79) em mais de 40 pontos. Comparando o tamanho dos modelos, o aumento de 1.5B para 3B parâmetros resultou em um ganho modesto de aproximadamente 2.7 pontos com dados traduzidos (16.79 vs 14.05), mas um ganho bem mais expressivo de cerca de 12.7 pontos com dados sintéticos (57.21 vs 44.45). Fica evidente que o treinamento com dados sintéticos proporcionou um salto de qualidade muito grande, com o modelo Qwen2.5 3B sintético atingindo a maior média geral (57.21), contrastando fortemente com a menor média do 1.5B traduzido (14.05). Destaca-se a categoria de coerência, que obteve as variações de pontuações mais altas entre as variações analisadas, demonstrando a capacidade superior em gerar respostas logicamente organizadas e claras para os usuários a partir de dados melhorados.

A Figura 4.4 detalha a preferência do avaliador externo Claude 3.5 Haiku em comparação com as respostas referência do conjunto Alpaca-Eval-Pt, destacando a predileção clara para as variantes com dados sintéticos, que demonstraram maior eficácia em geração de respostas abertas, seguida por uma expressiva taxa de empate, sendo

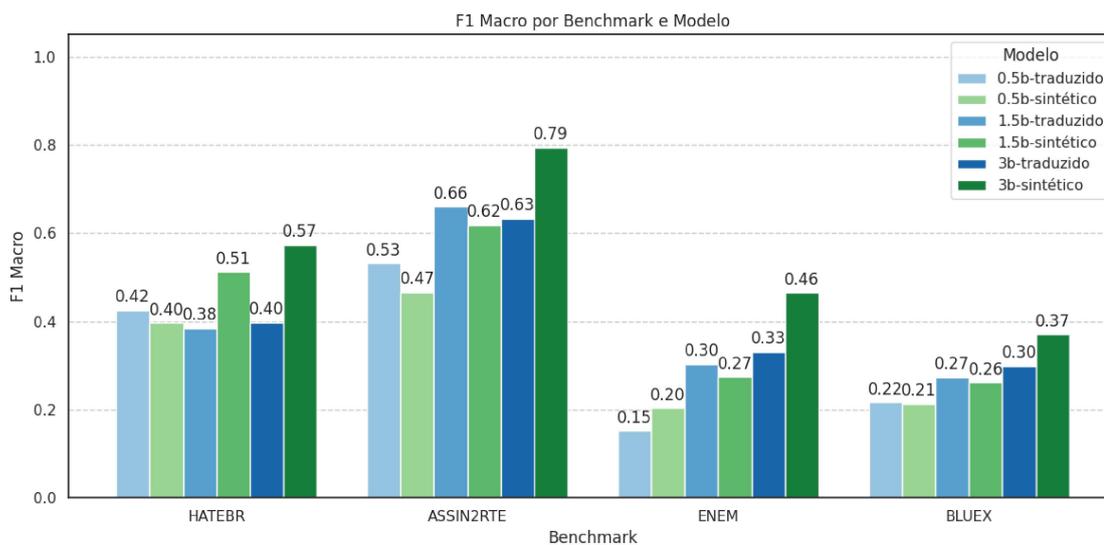


Figura 4.2: F1-Score entre benchmarks HATEBR, ASSIN2RT2, ENEM e BLUEX. Qwen2.5 0.5b, 1.5b e 3b, cenário traduzido e sintético.

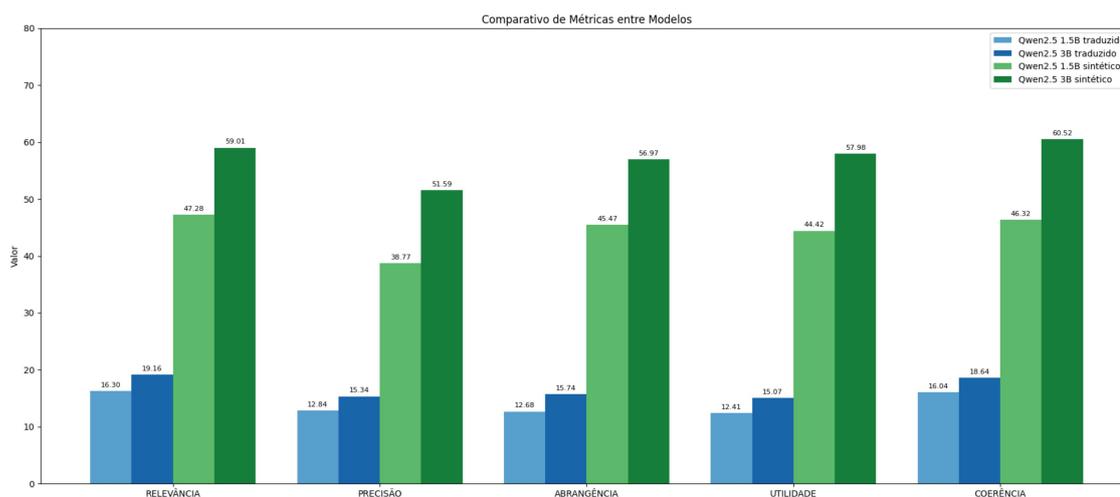


Figura 4.3: Métricas F1-Score normalizado nas categorias Relevância, Precisão, Abrangência, Utilidade e Coerência, modelo Qwen2.5 1.5b e 3b, no cenário traduzido e sintético. Avaliado pelo Claude 3.5 Haiku *LLM-as-Judge*

considerado algo positivo empatar com a resposta esperada.

Na tabela 4.2 está disposto duas amostras consideradas pelo avaliador que as respostas dos modelos de 3B são consideradas melhores que a referência. É observado que o comportamento do modelo com conjunto de dados com resposta aprimorada, complementa mais do que se solicita e já em formato markdown com intenção de facilitar a leitura, comportamento usual dos modelos de língua privados.

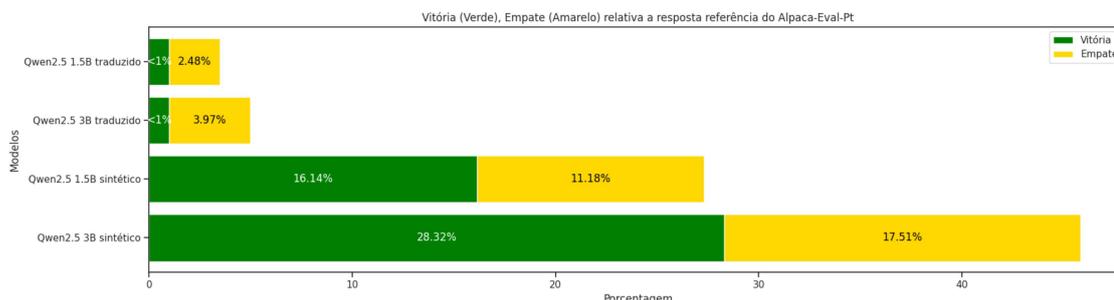


Figura 4.4: Preferência do *LLM as Judge* (Claude 3.5 Haiku) na resposta gerada e resposta referência no Alpaca-Eval-Pt. Vitória como verde, e Empate em amarelo.

Categoria	Descrição
Instrução	Edite a seguinte frase para torná-la mais concisa: "Estaremos fazendo mudanças para nos ajustarmos a esses tempos incertos"
Resposta Referência	"Vamos nos ajustar a esses tempos incertos."
Qwen2.5-3B traduzido	"Nos ajustamos aos tempos incertos."
Qwen2.5-3B sintético	<p>Em resposta às mudanças inesperadas, ajustaremos nossa postura.</p> <p>**Explicação das alterações:**</p> <p>* **"Em resposta às mudanças inesperadas"*: Essa frase substitui a expressão "Adapting to these unknown times" por uma construção mais direta e facilmente compreendida. Ela indica que a ação de "ajustar" é motivada por "mudanças inesperadas". * **"Estaremos ajustando nossa postura"*: A parte "will be adjusting our approach" foi reescrita para "ajustaremos nossa postura". A concisão e clareza são aprimoradas ao usar a forma verbal "ajustar" na primeira pessoa do plural, tornando a frase mais direta e ativa. A expressão "nossa postura" substitui "our approach", mantendo o sentido original, mas com uma linguagem mais enxuta.</p> <p>A frase revisada mantém o significado original da frase original, mas de forma mais concisa e clara.</p>

Tabela 4.2: Tabela de comparação de frases e modificações

Conclusão

Considerando que este estudo integra um esforço mais amplo para suprir a lacuna no desenvolvimento de recursos linguísticos de alta qualidade para o português — idioma com expressiva base de falantes e presença digital significativa — a criação de conjuntos de dados dedicados, como o proposto neste trabalho, representa um esforço significativo para o contexto tecnológico brasileiro. Diante disso, essa iniciativa contribui para reduzir a dependência de modelos multilingues de código aberto, que têm se mostrado limitados na captura das especificidades linguísticas do português.

Diante disso, foram investigados a adaptação e o aprimoramento do conjunto de dados de conversação *SmolTalk*, para o português brasileiro, com o objetivo de avaliar seu impacto no treinamento supervisionado (SFT) de modelos da arquitetura Qwen2.5, nas versões de 0.5B, 1.5B e 3B parâmetros. O processo envolveu a tradução turno a turno e a reescrita das respostas do assistente, realizadas com o modelo Gemini 2.0 Flash, visando aumentar a qualidade, com mais detalhes e correção linguística. Tal investigação foi baseada na hipótese central de que o aprimoramento de dados de forma sintética resulta em um padrão de maior qualidade nas respostas generativas.

A partir disso, de forma a referenciar a qualidade das respostas, foram conduzidas duas avaliações: (i) benchmarks padronizados em português (ENEM, HATEBR, BLUEX e ASSIN2-RTE), utilizando a métrica F1-Score em configuração few-shot; e (ii) avaliação de geração de texto aberto por meio do Alpaca-Eval-BR, com análise automática conduzida pelo modelo Claude 3.5 Haiku, conforme os critérios de relevância, precisão, abrangência, utilidade e coerência. Os resultados das avaliações indicaram um impacto positivo do aprimoramento sintético dos dados, com efeitos distintos entre as metodologias de avaliação.

Na primeira avaliação, utilizando benchmarks objetivos (F1-Macro), apenas o modelo Qwen2.5 3B apresentou melhorias consistentes ao ser treinado com dados sintéticos. Exemplos notáveis incluem os ganhos percentuais de +42.5% no HATEBR (F1-Score de 0.57 vs 0.40) e +25.4% no ASSIN2RTE (0.79 vs 0.63) em comparação com o treinamento usando dados traduzidos.

Já a segunda avaliação, realizada via LLM-as-Judge utilizando o Claude 3.5

Haiku, mensurou os aspectos de relevância, precisão, abrangência, utilidade e coerência da resposta gerada. Confirmando a hipótese da superioridade dos dados sintéticos, a média geral de notas normalizadas para os modelos sintéticos foi de 44.45 (1.5B) e 57.21 (3B), valores significativamente maiores que os 14.05 (1.5B traduzido) e 16.79 (3B traduzido). Isso representa um desempenho aproximadamente 3.2 vezes superior para o modelo 1.5B sintético e 3.4 vezes superior para o 3B sintético, em comparação com suas contrapartes treinadas com dados apenas traduzidos. Observou-se também que o modelo 3B superou o 1.5B em ambos os cenários, mas a diferença de desempenho foi maior entre os modelos sintéticos (+12.7 pontos) do que entre os traduzidos (+2.7 pontos). Estes resultados quantitativos evidenciam o forte impacto positivo do aprimoramento sintético especialmente em tarefas que demandam maior qualidade e complexidade na geração textual.

Dessa forma, fica demonstrado que o processo investigado neste trabalho – envolvendo a adaptação do conjunto de dados SmolTalk, que resultou na criação e disponibilização de duas novas versões para o português brasileiro, uma versão traduzida e outra sinteticamente aprimorada via reescrita turno a turno com o LLM Gemini 2.0 Flash – constitui uma estratégia eficaz para elevar a qualidade de modelos de linguagem treinados para o português brasileiro. A superioridade consistente dos modelos Qwen2.5 treinados com esses dados refinados, especialmente na avaliação qualitativa de geração aberta conduzida pelo Claude 3.5 Haiku, valida a hipótese central e sublinha o potencial dessa abordagem. Conclui-se, portanto, que a geração e o refinamento de dados de treinamento assistidos por LLMs de grande porte são ferramentas valiosas para desenvolver modelos com capacidades conversacionais avançadas, particularmente relevantes para idiomas como o português, onde a disponibilidade de recursos linguísticos de alta qualidade ainda é um desafio.

De forma a aprofundar na pesquisa, propõe-se, como próximos passos, a otimização de hiperparâmetros, a extensão dos experimentos para modelos de maior porte ou arquiteturas alternativas (como LLaMA e Gemma), e a incorporação de capacidades de uso de ferramentas (*tool use*) nos modelos gerados, com vistas a ampliar sua aplicabilidade prática.

Referências Bibliográficas

- [Abdin et al. 2024]ABDIN, M. et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [Abonizio et al. 2025]ABONIZIO, H. et al. *Sabiá-3 Technical Report*. 2025. Disponível em: <<https://arxiv.org/abs/2410.12049>>.
- [Ainslie et al. 2023]AINSLIE, J. et al. *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints*. 2023. Disponível em: <<https://arxiv.org/abs/2305.13245>>.
- [Allal et al. 2025]ALLAL, L. B. et al. *SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model*. 2025. Disponível em: <<https://arxiv.org/abs/2502.02737>>.
- [Allal et al. 2024]ALLAL, L. B. et al. *Cosmopedia*. 2024. Disponível em: <<https://huggingface.co/datasets/HuggingFaceTB/cosmopedia>>.
- [Almeida et al. 2024]ALMEIDA, T. S. et al. *Sabiá-2: A New Generation of Portuguese Large Language Models*. 2024. Disponível em: <<https://arxiv.org/abs/2403.09887>>.
- [Almeida et al. 2023]ALMEIDA, T. S. et al. *BLUEX: A benchmark based on Brazilian Leading Universities Entrance eXams*. 2023.
- [Anthropic 2025]Anthropic. *Claude Haiku*. fev. 2025. <https://www.anthropic.com/claude/haiku>. Accessed: 2025-04-13. Disponível em: <<https://www.anthropic.com/claude/haiku>>.
- [Bai et al. 2024]BAI, Y. et al. *LongAlign: A Recipe for Long Context Alignment of Large Language Models*. 2024. Disponível em: <<https://arxiv.org/abs/2401.18058>>.
- [Brown et al. 2020]BROWN, T. B. et al. *Language Models are Few-Shot Learners*. 2020. Disponível em: <<https://arxiv.org/abs/2005.14165>>.

- [Corrêa et al. 2024]CORRÊA, N. K. et al. Teenytinyllama: Open-source tiny language models trained in brazilian portuguese. *Machine Learning with Applications*, Elsevier BV, v. 16, p. 100558, jun. 2024. ISSN 2666-8270. Disponível em: <<http://dx.doi.org/10.1016/j.mlwa.2024.100558>>.
- [Corrêa et al. 2024]CORRÊA, N. K. et al. *Tucano: Advancing Neural Text Generation for Portuguese*. 2024. Disponível em: <<https://arxiv.org/abs/2411.07854>>.
- [Dagan, Synnaeve e Roziere 2024]DAGAN, G.; SYNNAEVE, G.; ROZIERE, B. Getting the most out of your tokenizer for pre-training and domain adaptation. *arXiv preprint arXiv:2402.01035*, 2024.
- [Dao 2023]DAO, T. *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning*. 2023. Disponível em: <<https://arxiv.org/abs/2307.08691>>.
- [Dao et al. 2022]DAO, T. et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. Disponível em: <<https://arxiv.org/abs/2205.14135>>.
- [Dekoninck et al. 2024]DEKONINCK, J. et al. *Understanding Large Language Models Through the Lens of Dataset Generation*. 2024. Disponível em: <<https://openreview.net/forum?id=miGplhquyB>>.
- [Face 2024]FACE, H. *Everyday Conversations for LLMs*. 2024. <https://huggingface.co/datasets/HuggingFaceTB/everyday-conversations-llama3.1-2k>.
- [Fanqi et al. 2023]FANQI, W. et al. *Explore-Instruct: Enhancing Domain-Specific Instruction Coverage through Active Exploration*. 2023.
- [Fu et al. 2023]FU, Y. et al. Specializing smaller language models towards multi-step reasoning. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2023. p. 10421–10430.
- [Gan et al. 2023]GAN, W. et al. Large language models in education: Vision and opportunities. In: IEEE. *2023 IEEE international conference on big data (BigData)*. [S.l.], 2023. p. 4776–4785.
- [Garcia et al. 2024]GARCIA, G. L. et al. *Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task*. 2024. Disponível em: <<https://arxiv.org/abs/2401.02909>>.
- [Google DeepMind 2025]Google DeepMind. *Gemini model updates February 2025*. fev. 2025. <https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/>. Accessed: 2025-04-13. Disponível

em: <<https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/>>.

[Grattafiori et al. 2024]GRATTAFIORI, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Grattafiori et al. 2024]GRATTAFIORI, A. et al. *The Llama 3 Herd of Models*. 2024. Disponível em: <<https://arxiv.org/abs/2407.21783>>.

[Gu et al. 2024]GU, J. et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[Gunasekar et al. 2023]GUNASEKAR, S. et al. *Textbooks Are All You Need*. 2023. Disponível em: <<https://arxiv.org/abs/2306.11644>>.

[Guo e Chen 2024]GUO, X.; CHEN, Y. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*, 2024.

[Huang et al. 2024]HUANG, Y. et al. Datagen: Unified synthetic dataset generation via large language models. In: *The Thirteenth International Conference on Learning Representations*. [S.l.: s.n.], 2024.

[Kaplan et al. 2020]KAPLAN, J. et al. *Scaling Laws for Neural Language Models*. 2020. Disponível em: <<https://arxiv.org/abs/2001.08361>>.

[Lai et al. 2024]LAI, J. et al. Large language models in law: A survey. *AI Open*, Elsevier, 2024.

[Larcher et al. 2023]LARCHER, C. et al. Cabrita: closing the gap for foreign languages. *arXiv preprint arXiv:2308.11878*, 2023.

[Li et al. 2024]LI, D. et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.

[LI et al. 2024]LI, J. et al. *NuminaMath*. [S.l.]: Numina, 2024. [<https://huggingface.co/AI-MO/NuminaMath-CoT>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).

[Li et al. 2023]LI, X. et al. *AlpacaEval: An Automatic Evaluator of Instruction-following Models*. [S.l.]: GitHub, 5 2023. https://github.com/tatsu-lab/alpaca_eval.

[Li et al. 2023]LI, Y. et al. Large language models in finance: A survey. In: *Proceedings of the fourth ACM international conference on AI in finance*. [S.l.: s.n.], 2023. p. 374–382.

- [Ling et al. 2023]LING, C. et al. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*, 2023.
- [Liu et al. 2024]LIU, A. et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [Liu et al. 2024]LIU, R. et al. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*, 2024.
- [Liu et al. 2024]LIU, Z. et al. Apigen: Automated pipeline for generating verifiable and diverse function-calling datasets. *arXiv preprint arXiv:2406.18518*, 2024.
- [Long et al. 2024]LONG, L. et al. On llms-driven synthetic data generation, curation, and evaluation: A survey, 2024. URL <https://arxiv.org/abs/2406.15126>, 2024.
- [Lopes, Magalhães e Semedo 2024]LOPES, R.; MAGALHÃES, J.; SEMEDO, D. *Glória – A Generative and Open Large Language Model for Portuguese*. 2024. Disponível em: <<https://arxiv.org/abs/2402.12969>>.
- [Meta AI 2024]Meta AI. *Llama 4: Advancing multimodal intelligence*. abr. 2024. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>. Accessed: 2025-04-13. Disponível em: <<https://ai.meta.com/blog/llama-4-multimodal-intelligence/>>.
- [Meyer et al. 2023]MEYER, J. G. et al. Chatgpt and large language models in academia: opportunities and challenges. *BioData mining*, Springer, v. 16, n. 1, p. 20, 2023.
- [Nunes et al. 2023]NUNES, D. et al. *Evaluating GPT-3.5 and GPT-4 Models on Brazilian University Admission Exams*. 2023.
- [OpenAI 2023]OpenAI. *Introducing ChatGPT and Whisper APIs*. 2023. <https://openai.com/index/introducing-chatgpt-and-whisper-apis/>. Accessed: 2025-04-12.
- [Penedo et al. 2024]PENEDO, G. et al. The fineweb datasets: Decanting the web for the finest text data at scale. In: *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. [s.n.], 2024. Disponível em: <<https://openreview.net/forum?id=n6SCkn2QaG>>.
- [Pichai 2024]PICHAI, S. *Introducing Gemini 2.0: our new AI model for the agentic era*. dez. 2024. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>. Accessed: 2025-4-7.
- [Pires et al. 2023]PIRES, R. et al. Sabiá: Portuguese large language models. In: _____. *Intelligent Systems*. Springer Nature Switzerland, 2023. p. 226–240. ISBN 9783031453922. Disponível em: <http://dx.doi.org/10.1007/978-3-031-45392-2_15>.

- [Qwen et al. 2025]QWEN et al. *Qwen2.5 Technical Report*. 2025. Disponível em: <<https://arxiv.org/abs/2412.15115>>.
- [Radford et al. 2019]RADFORD, A. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019.
- [Rasley et al. 2020]RASLEY, J. et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. [S.l.: s.n.], 2020. p. 3505–3506.
- [Real, Fonseca e Oliveira 2020]REAL, L.; FONSECA, E.; OLIVEIRA, H. G. The assin 2 shared task: a quick overview. In: SPRINGER. *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*. [S.l.], 2020. p. 406–412.
- [Saadi et al. 2025]SAADI, N. et al. Bridging language barriers in healthcare: A study on arabic llms. *arXiv preprint arXiv:2501.09825*, 2025.
- [Sennrich, Haddow e Birch 2015]SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- [Shazeer 2019]SHAZEER, N. *Fast Transformer Decoding: One Write-Head is All You Need*. 2019. Disponível em: <<https://arxiv.org/abs/1911.02150>>.
- [Smith et al. 2022]SMITH, S. et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- [Tao et al. 2024]TAO, C. et al. *Scaling Laws with Vocabulary: Larger Models Deserve Larger Vocabularies*. 2024. Disponível em: <<https://arxiv.org/abs/2407.13623>>.
- [Team et al. 2023]TEAM, G. et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [Teknium 2023]TEKNIUM. *OpenHermes 2.5: An Open Dataset of Synthetic Data for Generalist LLM Assistants*. HuggingFace, 2023. Disponível em: <<https://huggingface.co/datasets/teknium/OpenHermes-2.5>>.
- [Thirunavukarasu et al. 2023]THIRUNAVUKARASU, A. J. et al. Large language models in medicine. *Nature medicine*, Nature Publishing Group US New York, v. 29, n. 8, p. 1930–1940, 2023.

- [Vargas e Carvalho 2024]VARGAS, F.; CARVALHO, I. *franciellevargas/HateBR*. Zenodo, mar. 2024. Disponível em: <<https://doi.org/10.5281/zenodo.10794024>>.
- [Vaswani et al. 2017]VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.
- [Wijmans et al. 2025]WIJMANS, E. et al. *Cut Your Losses in Large-Vocabulary Language Models*. 2025. Disponível em: <<https://arxiv.org/abs/2411.09009>>.
- [Yu et al. 2023]YU, L. et al. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- [Yu et al. 2023]YU, Y. et al. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, v. 36, p. 55734–55784, 2023.

Apêndice

Prompts

Prompt de Tradução Inglês-Português

Você está responsável por traduzir um trecho de uma conversa do inglês para o português brasileiro. O objetivo é garantir uma tradução fiel ao conteúdo original, mantendo a naturalidade e clareza em português.

Diretrizes para a tradução:

1. Traduza todo o conteúdo textual da conversa, exceto trechos de código ou cálculos matemáticos explícitos.
2. Preserve o sentido original das mensagens, respeitando o contexto da conversa.
3. Adote uma linguagem natural em português, compatível com o uso em interações com assistentes de IA.
4. Não inclua comentários, explicações adicionais ou metadados. Traduza apenas o conteúdo textual.

Entrada:

- **Trecho da Conversa:** {conversation}

Saída Esperada:

O conteúdo traduzido para o português brasileiro, com preservação dos trechos técnicos (código ou fórmulas) no idioma original.

Prompt de Aprimoramento do turno Assistant

Você é um assistente de IA especializado em aprimorar respostas de outros assistentes. Sua tarefa é reescrever os turnos do papel *assistant* em uma conversa, refinando o conteúdo em língua portuguesa, preservando o contexto e a intenção original da interação.

Siga as seguintes diretrizes ao realizar o aprimoramento:

1. Mantenha o objetivo geral da conversa original de forma concisa.
2. Se a resposta original for vaga ou incompleta, complemente com detalhes e informações relevantes.
3. Se houver erros factuais ou imprecisões, corrija-os adequadamente.
4. Adicione exemplos ou explicações que melhorem a clareza da resposta.
5. Adote um tom profissional, educado e prestativo.
6. Não desvie do assunto original da conversa.
7. Organize o conteúdo de forma clara, utilizando parágrafos, listas ou outros recursos estruturais.
8. Retorne apenas o conteúdo aprimorado do turno do *assistant*, em português. Não inclua informações de papel ou marcações adicionais.

Entrada:

• **Conversa:** {conversation}

Saída Esperada:

Apenas o conteúdo reescrito do turno do *assistant*, em português, com melhorias conforme as diretrizes acima.

Prompt de Avaliação LLM-as-Judge

Você é um avaliador crítico de respostas geradas por um modelo de linguagem. Sua tarefa é comparar uma resposta gerada com uma resposta de referência para a mesma pergunta e fornecer uma avaliação detalhada com uma pontuação estruturada.

Considere os seguintes aspectos na sua avaliação:

1. **Relevância:** A resposta gerada aborda diretamente a pergunta? Ela fornece informações pertinentes ao contexto?
2. **Precisão:** As informações apresentadas na resposta gerada são factualmente corretas? Elas contradizem a resposta referência ou informações amplamente aceitas?
3. **Abrangência:** A resposta gerada cobre os principais pontos da pergunta? Ela omite informações importantes presentes na resposta referência? Ou, inversamente, ela oferece uma cobertura mais completa?
4. **Utilidade/Informatividade:** A resposta gerada é útil para o usuário? Ela fornece insights, explicações ou informações práticas relevantes? Ela agrega valor em relação à resposta referência?
5. **Clareza e Coerência:** A resposta gerada é fácil de entender? A linguagem é clara, concisa e organizada logicamente?

Entradas:

- Pergunta:** {instruction}
- Resposta Referência:** {output}
- Resposta Gerada:** {answer}

Avaliação:

Forneça uma análise textual detalhada comparando a Resposta Gerada com a Resposta Referência em relação aos aspectos listados acima. Destaque os pontos fortes e fracos da Resposta Gerada em comparação com a Resposta Referência.

Pontuação Estruturada:

Atribua uma pontuação de 0 a 5 para cada critério, onde:

- 0:** Completamente inadequado/incorreto/irrelevante.
- 1:** Muito ruim/impreciso/pouco relevante.
- 2:** Razoável/parcialmente correto/alguma relevância.
- 3:** Bom/preciso/relevante.
- 4:** Muito bom/altamente preciso/muito relevante.
- 5:** Excelente/excepcionalmente preciso/extremamente relevante e/ou superior à resposta Referência.

Esquema de Pontuação:

```
{{
  "avaliacao_detalhada": "[SUA ANÁLISE TEXTUAL AQUI]",
  "pontuacao": {
    "relevancia": [0-5],
    "precisao": [0-5],
    "abrangencia": [0-5],
    "utilidade": [0-5],
    "clareza_coerencia": [0-5]
  },
  "melhor_resposta": [referencia, gerada, empate]
  "justificativa_melhor_que_referencia": "[EXPLICAÇÃO SE A RESPOSTA GERADA FOR CO
}}}
```