

*UNIVERSIDADE DO ESTADO DO AMAZONAS
ESCOLA SUPERIOR DE TECNOLOGIA*

DANILO BRUNO DA SILVA

*ENGENHARIA DE PROMPTS PARA RESPOSTAS DE RESTAURANTES EM
PORTUGUÊS: AVALIAÇÃO DE TOM, EMPATIA, FLUÊNCIA E CONSISTÊNCIA
DE MARCA COM LLMs*

Manaus
2026

DANILO BRUNO DA SILVA

*ENGENHARIA DE PROMPTS PARA RESPOSTAS DE RESTAURANTES EM
PORTUGUÊS: AVALIAÇÃO DE TOM, EMPATIA, FLUÊNCIA E CONSISTÊNCIA
DE MARCA COM LLMs*

Orientador: Tiago Eugenio de Melo, Dr.

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).
Sistema Integrado de Bibliotecas da Universidade do Estado do Amazonas.

D111e	da Silva, Danilo Bruno ENGENHARIA DE PROMPTS PARA RESPOSTAS DE RESTAURANTES EM PORTUGUÊS : AVALIAÇÃO DE TOM, EMPATIA, FLUÊNCIA E CONSISTÊNCIA DE MARCA COM LLMs / Danilo Bruno da Silva. Manaus : [s.n], 2026. 93 f.: color.; 21.0 cm. TCC - Graduação em Engenharia de Computação- Universidade do Estado do Amazonas, Manaus, 2026. Inclui Apêndice. Orientador: Tiago Eugenio de Melo. 1. Engenharia de Prompt. 2. Modelos de Linguagem em Larga Escala (LLMs),. 3. Processamento de Linguagem Natural (PLN). 4. Avaliação Multidimensional. 5. Automação. I. Tiago Eugenio de Melo (Orient.) II. Universidade do Estado do Amazonas. III. Título CDU(1997)62:004
-------	--



FOLHA DE APROVAÇÃO

Engenharia de Prompts para Respostas de Restaurantes em Português: Avaliação de Tom, Empatia, Fluência e Consistência de Marca com LLMs

Danilo Bruno Da Silva

Trabalho de Conclusão de Curso defendido e aprovado pela banca avaliadora constituída pelos professores:

Tiago Eugenio de Melo
Presidente

Danielle Pompeu Noronha Pontes
Avaliador

Carlos Mauricio Serodio Figueiredo
Avaliador

Manaus, 02 de junho de 2026.

AGRADECIMENTOS

Aos meus pais e irmãos, que me incentivaram nos momentos difíceis desta caminhada e sempre compreenderam a verdadeira importância da educação como parte imprescindível da minha formação moral e intelectual.

À minha amada namorada que ofereceu apoio incondicional ao longo de todo o percurso acadêmico. Obrigado por compreender minhas ausências e por ser meu porto seguro.

Aos amigos que estiveram ao meu lado, pela amizade e pelo apoio demonstrado ao longo de todo o período de tempo em que me dediquei a este trabalho.

Ao professor Tiago de Melo, por ter sido meu orientador e ter desempenhado tal função com dedicação e clareza. Aos professores, pelas correções e ensinamentos que me permitiram apresentar um melhor desempenho no meu processo de formação profissional ao longo do curso.

A todos que participaram, direta ou indiretamente do desenvolvimento deste trabalho de pesquisa, enriquecendo o meu processo de aprendizado.

Às pessoas que convivi ao longo desses anos de curso, que me incentivaram e que certamente tiveram impacto na minha formação acadêmica.

RESUMO

A gestão automatizada de avaliações online no setor gastronômico exige respostas que conciliem tom, empatia, fluência e consistência de marca, dimensões ainda pouco exploradas em modelos de linguagem de larga escala (*LLMs*) aplicados ao português brasileiro. Este trabalho avalia a viabilidade e a maturidade de três *LLMs* (*GPT-5.5*, *Gemini 3.1* e *Llama 4*) para essa tarefa. A partir de um corpus de 200 comentários reais, segmentado em dois períodos (pré e pós-popularização das *LLMs*), conduziu-se um experimento fatorial cruzando modelos, personas de marca (formal e informal) e estratégias de engenharia de *prompt* (*zero-shot* e *few-shot*). A avaliação seguiu metodologia mista, combinando o julgamento de 15 avaliadores humanos em escala *Likert* com métricas automáticas multidimensionais: perplexidade (fluência), *ToneCal* (adequação de tom), similaridade estilométrica (consistência de marca) e modelo *WASSA* (empatia percebida). Os resultados indicam que nenhum modelo é universalmente superior: o *Llama 4* lidera em empatia e adequação de tom, enquanto o *GPT-5.5* prevalece em fluência e consistência de marca, com diferenças estatisticamente significativas entre modelos. A estratégia *zero-shot*, apoiada em descrições robustas de persona, tende a superar a *few-shot* nas dimensões pragmáticas. A inversão direcional do efeito de persona, informal preferida pelos avaliadores humanos, porém formal favorecida pelas métricas automáticas, evidencia um viés sistemático dos *proxies* computacionais para o registro formal. A principal contribuição é a proposta de um *framework* de avaliação multidimensional para geração de texto em português brasileiro, cujos resultados sugerem que a triangulação entre avaliação humana e automática é desejável para conclusões mais robustas no cenário avaliado.

Palavras-chave: Engenharia de *Prompt*, Modelos de Linguagem em Larga Escala (*LLMs*), Processamento de Linguagem Natural (PLN), Avaliação Multidimensional, Automação.

ABSTRACT

Automating responses to online restaurant reviews requires balancing tone, empathy, fluency, and brand consistency, dimensions still underexplored for large language models (LLMs) applied to Brazilian Portuguese. This work evaluates the viability and maturity of three LLMs (GPT-5.5, Gemini 3.1, and Llama 4) for this task. Drawing on a corpus of 200 real reviews, partitioned into two temporal periods (pre- and post-LLM popularization), a factorial experiment was conducted crossing models, brand personas (formal and informal), and prompt engineering strategies (zero-shot and few-shot). The evaluation followed a mixed-methods design, combining Likert-scale judgments from 15 independent human raters with multidimensional automatic metrics: perplexity (fluency), ToneCal (tone adequacy), stylometric similarity (brand consistency), and the WASSA model (perceived empathy). Results show that no model is universally superior: Llama 4 leads in empathy and tone adequacy, while GPT-5.5 prevails in fluency and brand consistency, with statistically significant differences across models. The zero-shot strategy, supported by well-defined personas, tends to outperform few-shot in pragmatic dimensions. A directional inversion of the persona effect, informal preferred by human raters, yet formal favored by automatic metrics, reveals a systematic bias of computational proxies toward formal register. The main contribution is a multidimensional evaluation framework for text generation in Brazilian Portuguese, with results suggesting that triangulation between human and automatic evaluation is advisable for more robust conclusions in the evaluated scenario.

Keywords: Prompt Engineering, Large Language Models (LLMs), Natural Language Processing (NLP), Multidimensional Evaluation, Automation.

Lista de Figuras

1	Comentário negativo real coletado do <i>Instagram</i> , utilizado como cenário ilustrativo na Introdução.	15
2	Comentário sobre data comemorativa coletado do <i>Instagram</i> , ilustrando a perda de contexto emocional em respostas sem ajuste de persona.	21
3	Resposta genérica a comentário retirado do <i>Instagram</i>	22
4	Resposta inconsistente de restaurante "requintado" a comentário retirado do Instagram	23
5	Desenho experimental para avaliação de LLMs.	28
6	<i>Coleta de 200 comentários, engenharia de prompts com duas personas e duas estratégias, geração de 2.400 respostas por três LLMs, e avaliação multidimensional com análise comparativa.</i>	41
7	72 pares comentário-resposta selecionados por amostragem estratificada, distribuídos individualmente com ordem aleatorizada e cegamento ao modelo. Os 15 avaliadores avaliaram quatro dimensões em escala Likert 1–5, com concordância inter-avaliadores de $\kappa = 0,76$ – $0,86$ (Fleiss).	50
8	<i>Template do prompt Zero-Shot</i>	57
9	<i>Template do prompt Few-Shot</i>	58
10	Mapa de calor das médias Likert por modelo e dimensão na avaliação humana.	61
11	Distribuição dos scores Likert por modelo e dimensão (média dos 15 avaliadores por item, $n = 72$ itens totais, ≈ 24 por modelo). <i>Gemini 3.1</i> apresenta maior dispersão e medianas inferiores em empatia, consistência e tom nas condições experimentais avaliadas (interface web, sem controle de parâmetros de <i>decoding</i> — cf. Seção 7.4). Fluência apresenta efeito teto para todos os modelos.	64
12	Comparação entre os três modelos nas avaliações humana (esquerda, escala Likert 1–5) e automática (direita, escalas individuais por métrica).	65
13	Perfil multidimensional dos modelos combinando métricas humanas e automáticas (valores normalizados 0–1).	66
14	Dispersão entre avaliações humana (eixo X) e automática (eixo Y) por modelo e persona. Quadrados representam a persona formal (P1), círculos a informal (P2).	73

Lista de Tabelas

1	Caracterização Quantitativa do <i>Corpus de Origem</i>	42
2	Comparativo de Adequação de Persona e Tempo de Processamento por LLM	46
3	Resumo da Estrutura do Conjunto de Dados Gerado	46
4	Configuração dos Modelos e Parâmetros de Reprodutibilidade	53
5	Concordância Inter-Avaliadores por Dimensão (15 avaliadores, 72 itens) . .	60
6	Avaliação Humana Global por Modelo (escala Likert 1–5)	61
7	Teste de Kruskal-Wallis — Efeito de Modelo na Avaliação Humana	62
8	Comparações Pareadas — Teste de Mann-Whitney U com correções para comparações múltiplas (Holm-Bonferroni e Benjamini-Hochberg). Valores em negrito: $p \leq 0,05$ após a correção indicada.	63
9	Métricas Automáticas Globais por Modelo	64
10	Avaliação Humana por Modelo e Persona (escala Likert 1–5)	66
11	Avaliação Humana por Modelo e Estratégia de <i>Prompt</i> (escala Likert 1–5)	67
12	Avaliação Humana por Modelo e Período (escala Likert 1–5)	68
13	Top 5 e Bottom 5 Cenários por Empatia na Avaliação Humana	69
14	Convergência de Rankings: Avaliação Humana vs. Métricas Automáticas .	72

LISTA DE ABREVIATURAS E SIGLAS

API	<i>Application Programming Interface</i> (Interface de Programação de Aplicações)
BERT	<i>Bidirectional Encoder Representations from Transformers</i>
BX	<i>Brand Experience</i> (Experiência da Marca)
CX	<i>Customer Experience</i> (Experiência do Cliente)
GPT	<i>Generative Pre-trained Transformer</i>
GPU	<i>Graphics Processing Unit</i>
ICC	<i>Intraclass Correlation Coefficient</i> (Coeficiente de Correlação Intraclass)
ICL	<i>In-Context Learning</i> (Aprendizado em Contexto)
LLM	<i>Large Language Model</i> (Modelo de Linguagem em Larga Escala)
LSTM	<i>Long Short-Term Memory</i>
NLP	<i>Natural Language Processing</i> (Processamento de Linguagem Natural)
PLN	Processamento de Linguagem Natural
RLHF	<i>Reinforcement Learning from Human Feedback</i>
RNN	Redes Neurais Recorrentes
TQS	<i>Tone Quality Score</i> (Pontuação de Qualidade de Tom)
UX	<i>User Experience</i> (Experiência do Usuário)
WASSA	<i>Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis</i>

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Justificativa	16
1.2	Objetivo Geral	17
1.3	Objetivos específicos	17
1.4	Organização do Trabalho	18
1.5	Contribuições do Trabalho	19
2	FUNDAMENTAÇÃO TEÓRICA	20
2.1	O problema da geração automática de comentários	20
2.2	Processamento de Linguagem Natural (<i>PLN</i>)	23
2.3	<i>Large Language Models</i>	24
2.4	Engenharia de prompts	26
2.5	Desenho Experimental para Avaliação de <i>LLMs</i>	26
2.6	Métricas de Avaliação	27
2.6.1	Avaliação Humana com Escala <i>Likert</i>	29
2.6.2	Perplexidade como Métrica de Fluência	29
2.6.3	ToneCal: Adequação de Tom	30
2.6.4	<i>Writeprints</i> : Consistência Estilométrica de Marca	30
2.6.5	WASSA: Empatia Percebida	30
3	TRABALHOS RELACIONADOS	31
3.1	Resposta Gerencial a Avaliações Online	31
3.2	<i>LLMs</i> e Geração de Respostas no Domínio de Serviço	33
3.3	Modelos de Linguagem e <i>PLN</i> para o Português Brasileiro	34
3.4	Eficácia das Estratégias de <i>Prompting</i> (<i>Few-Shot vs. Zero-Shot</i>)	35
3.5	Personas e Controle de Estilo em <i>LLMs</i>	36
3.6	Metodologias de Avaliação de Texto Gerado	37
3.7	Empatia Computacional e Avaliação de Tom	38
4	METODOLOGIA	40
4.1	Arquitetura e Fluxo de Dados	40
4.1.1	<i>Pipeline</i> Experimental	40
4.2	Conjunto de Dados	41
4.2.1	Caracterização Geral do Corpus de Origem	42
4.2.2	Crítérios de Seleção, Inclusão e Exclusão	42
4.2.3	Procedimento de Anonimização	43

4.2.4	Marco Temporal e Divisão do Corpus	44
4.2.5	Uso do corpus como referência humana nas métricas automáticas	44
4.2.6	Corpus Resultante	45
4.2.7	Considerações Éticas e Tratamento dos Dados	46
4.2.8	Uso Consciente de Inteligência Artificial Generativa	47
4.3	Métricas de Avaliação	48
4.3.1	Avaliação Humana (Qualitativa)	49
4.3.2	Avaliação Automática (Quantitativa)	51
4.4	<i>Large Language Models</i>	52
4.5	Configuração dos Modelos	53
4.6	Parâmetros e Procedimento de Geração	54
4.7	<i>Prompts</i> Empregados	55
4.7.1	Engenharia de <i>Prompt</i> : Princípios Adotados	55
4.7.2	Estrutura Base do <i>Prompt</i> (Componente Fixo)	55
4.7.3	Estratégia Experimental 1: <i>Prompt Zero-Shot</i>	56
4.7.4	Estratégia Experimental 2: <i>Prompt Few-Shot</i>	57
5	EXPERIMENTOS E RESULTADOS	60
5.1	Concordância Inter-Avaliadores	60
5.2	Visão Geral dos Resultados	60
5.3	Avaliação Humana	61
5.4	Escolha dos Testes Estatísticos	62
5.5	Métricas Automáticas	64
5.6	Análise Estratificada	66
5.6.1	Interação Modelo $\times$ Persona	66
5.6.2	Interação Modelo $\times$ Estratégia de <i>Prompt</i>	67
5.6.3	Interação Modelo $\times$ Período Temporal	68
5.6.4	Melhores e Piores Cenários	68
5.7	Análise Qualitativa: Exemplos Ilustrativos	69
5.8	Convergências e Divergências entre Avaliação Humana e Automática	72
5.8.1	Convergências	73
5.8.2	Divergências	73
5.8.3	Viés de Registro Linguístico	74
5.8.4	Natureza Aproximativa das Métricas Automáticas	74
6	CONCLUSÕES	75
6.1	Síntese dos Resultados	75
6.2	Contribuições do Trabalho	76
6.3	Implicações Práticas	78
6.4	Trabalhos Futuros	78

7	LIMITAÇÕES DO ESTUDO	79
7.1	Corpus, Amostra e Avaliação Humana	79
7.2	Limitações das Métricas Automáticas	80
7.3	Análise Estatística e Estrutura de Dependência	80
7.4	Condições Experimentais e Configuração dos Modelos	80
	REFERÊNCIAS	82
A	REPOSITÓRIO DO PROJETO	89
B	Formulário de Avaliação Humana	89

1 INTRODUÇÃO

Com o advento das redes sociais, a escolha de um estabelecimento gastronômico passou a ser profundamente impactada por manifestações públicas de comentários, experiências e avaliações publicadas em plataformas como *Instagram*, *TripAdvisor*, *Google Reviews* etc. (SILVA; BIANCOLINO; SPERS, 2019). Assim, consolidando uma dinâmica de influência informacional, onde o usuário sente-se persuadido a pautar-se no conteúdo dessas mídias.

Um exemplo para ajudar a ilustrar esse cenário: em um relato no *Instagram* de um restaurante, um cliente fez um comentário negativo em relação à demora no atendimento e serviço, avaliando-o como decepcionante: *“Estávamos hospedados em um hotel de alto padrão, mas o restaurante foi decepcionante. O atendimento demorou, o prato infantil veio errado e o meu risoto tinha quase só beterraba, com quase nada de cogumelo. Em um restaurante desse nível, isso é inadmissível. Não voltaria ao restaurante.”*. Embora breve, um comentário desse tipo pode influenciar enfaticamente a percepção de futuros clientes. Nesse caso, a resposta do estabelecimento foi: *“Lamentamos muito pela sua experiência, em especial pela troca do prato e pela demora no atendimento. Estamos revisando nossos processos para evitar que situações assim se repitam. Esperamos ter a oportunidade de recebê-lo novamente e oferecer uma experiência condizente com o que valorizamos.”*. Visto isso, uma possível resposta de *LLMs*, sem ajuste, pode soar genérica e antipática: *“Obrigado pelo seu comentário. Sentimos muito pela experiência e valorizamos seu feedback para melhorarmos nossos serviços.”*. Essa comparação evidencia o tamanho do desafio e a complexidade na geração de respostas automáticas que soem naturais. A Figura 1 apresenta o comentário original utilizado neste cenário.

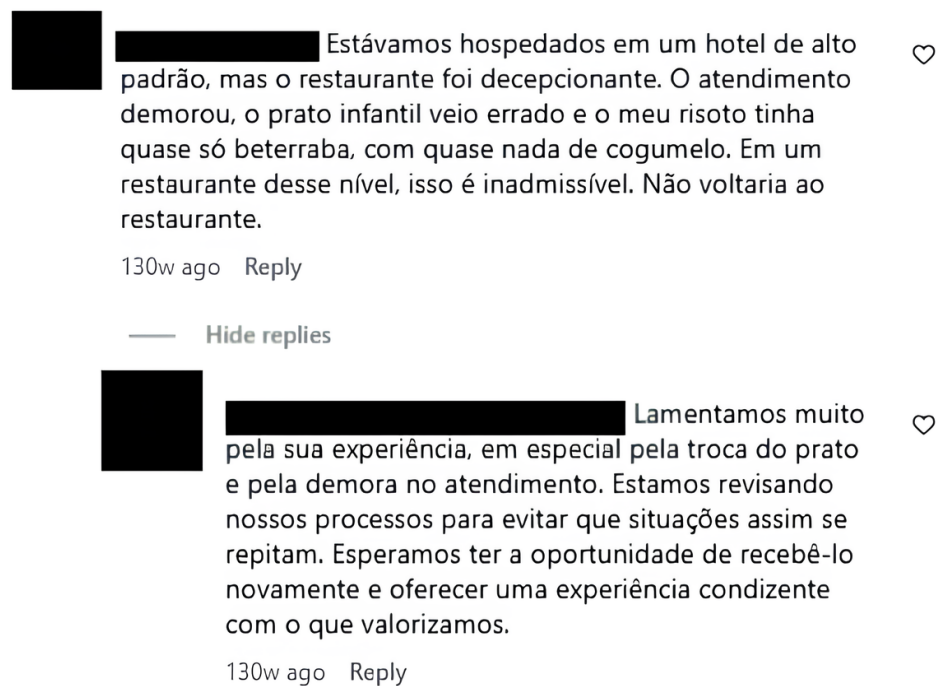


Figura 1 – Comentário negativo real coletado do *Instagram*, utilizado como cenário ilustrativo na Introdução.

Dada a repercussão que essas avaliações podem causar sobre a reputação de um estabelecimento, o tamanho desse desafio é expressivo. Plataformas como o *TripAdvisor* acumulam bilhões de avaliações oriundas de diferentes plataformas (Tripadvisor, Inc., 2024). Gerenciar manualmente esse volume, mantendo a qualidade e a agilidade na comunicação, torna-se uma tarefa desafiadora. É nesse cenário que as *LLMs* apresentam-se como uma ferramenta promissora para a automação dessa tarefa (GAMBOA-CRUZADO et al., 2023). Por um lado, sua capacidade de processar e gerar texto em grande escala oferece agilidade, permitindo que nenhum comentário fique sem resposta. Contudo, em contraponto a essa eficiência, a comunicação humana eficaz demanda sutilezas que vão além de adequação e correção gramatical, exige-se uma percepção de tom para diferenciar uma crítica construtiva de um desabafo, empatia para validar a experiência do cliente e uma compreensão contextual que preserve a identidade do estabelecimento (CROLIC et al., 2023).

No entanto, o desafio não reside apenas na automação da resposta, mas na capacidade desses sistemas em gerar interações que, de fato, pareçam humanas e inspirem cuidado e confiança. Pesquisas recentes indicam que a empatia percebida e o tom emocional de agentes de *IA* impactam diretamente a satisfação do consumidor e a intenção de recompra (ROHDEN; ESPARTEL, 2024), reforçando a criticidade dessa lacuna.

Em suma, este trabalho avalia a qualidade, viabilidade e maturidade das *LLMs* disponíveis para a geração de respostas a comentários no setor gastronômico, com foco exclusivo

na língua portuguesa, para assim capturar melhor as nuances linguísticas e culturais inerentes ao português (ARAUJO; MELO; FIGUEIREDO, 2024). O estudo compreende um levantamento dos modelos mais destacados, assim como uma análise aprofundada das estratégias de *prompt*, para adequar variáveis como tom, empatia, fluência e consistência de marca. Alinhado à recomendação da literatura por avaliações holísticas e multidimensionais (LIANG et al., 2023; CELIKYILMAZ; CLARK; GAO, 2021), a avaliação é conduzida por meio de um *framework* que combina métricas automáticas especializadas, fluência via perplexidade (PISTOTTI et al., 2025), adequação de tom (DANESCU-NICULESCU-MIZIL et al., 2013), similaridade estilométrica (ABBASI; CHEN, 2008) e empatia percebida (BARRIERE et al., 2023), com a avaliação humana de 15 avaliadores independentes, aplicada sobre um corpus expandido de 2.400 respostas geradas. Por fim, busca-se oferecer um panorama crítico e prático sobre os limites e possibilidades dessa tecnologia, contribuindo para um diálogo sobre a automação consciente e eficaz no sensível campo da interação com o cliente.

1.1 JUSTIFICATIVA

Estudos indicam que, no setor gastronômico, as avaliações online impactam diretamente a escolha do consumidor. Esse cenário sujeita os estabelecimentos a uma dura realidade: a necessidade de engajamento constante com o público e o desafio de fazer isso de forma rápida, incessante e, acima de tudo, humana (AURELIANO-SILVA; LEUNG; SPERS, 2021).

Nesse cenário, as *LLMs* surgem como uma alternativa escalável e promissora na automação dessas interações, permitindo que empresas gerenciem grandes volumes de dados sem perder a capacidade de resposta operacional. Estudos indicam que fatores como gênero e percepção de lotação podem influenciar significativamente a forma como os consumidores interpretam avaliações e respostas, reforçando a necessidade de respostas contextuais e empáticas (ALI; KIM; RYU, 2021). A sua capacidade de compreender perguntas e cenários é o que lhe confere o poder de formular respostas compatíveis, tanto em contexto, quanto em tom, empatia e assertividade. No entanto, essa poderosa ferramenta traz consigo um desafio igualmente relevante: a qualidade da comunicação. Respostas mecânicas, repetitivas ou impessoais podem agravar percepções negativas, comprometendo a imagem do estabelecimento em vez de protegê-la, visto que a falta de calor humano é frequentemente penalizada pelos consumidores (SANDS et al., 2020).

É justamente nesse ponto que este estudo se faz necessário. Atualmente, observa-se uma lacuna de conhecimento sobre a real qualidade e maturidade das *LLMs* em português para o setor gastronômico. Embora existam avanços em modelos para o português brasileiro, como o *BERTimbau* (SOUZA; NOGUEIRA; LOTUFO, 2020), a aplicação generativa focada em nuances culturais de serviço ainda carece de exploração aprofundada.

Além disso, a literatura apresenta uma carência de *frameworks* de avaliação multidimensional que mensurem, de forma integrada e com métricas especializadas, aspectos como fluência, tom, consistência de marca e empatia em respostas geradas automaticamente em português, uma lacuna corroborada por Liang et al. (2023), que advogam por avaliações holísticas que vão além de métricas agregadas, e por Howcroft e Rieser (2021), que indicam que avaliações humanas em *PLN* são frequentemente subdimensionadas. No campo da empatia computacional, embora Sharma et al. (2020) tenham proposto *frameworks* para mensurar empatia em texto e as *shared tasks* da WASSA (BARRIERE et al., 2022, 2023; GIORGI et al., 2024) tenham avançado na predição de empatia em reações a textos, a transferência dessas abordagens para o domínio de respostas comerciais em português permanece inexplorada. De forma análoga, a análise estilométrica de consistência de marca, fundamentada em técnicas de *Writeprints* (ABBASI; CHEN, 2008), e a avaliação automática de adequação de tom, investigada computacionalmente desde os trabalhos seminais de Danescu-Niculescu-Mizil et al. (2013) e sistematizada por Priya, Firdaus e Ekbal (2024), não foram aplicadas ao contexto de respostas automáticas de restaurantes. Assim, este trabalho propõe-se a ir além da simples automação, conduzindo uma análise comparativa capaz de investigar não apenas a correção linguística, mas também a habilidade dos modelos em modular variáveis como tom, empatia, fluência e preservação da identidade da marca. Dessa forma, a pesquisa se justifica por sua relevância prática, ao oferecer diretrizes seguras para o uso consciente dessa tecnologia, contribuindo com um esquema de avaliação multidimensional voltado à automatização de interações em contextos sensíveis.

1.2 OBJETIVO GERAL

O objetivo geral deste trabalho é avaliar a viabilidade do uso de *LLMs* na tarefa de geração automática de respostas para comentários do domínio de restaurantes, com foco exclusivo na língua portuguesa.

1.3 OBJETIVOS ESPECÍFICOS

Os objetivos específicos deste estudo são:

- Coletar, organizar e sanitizar um conjunto de comentários reais de avaliações sobre restaurantes feitos em plataformas digitais como *Google Maps*, *TripAdvisor*, *Instagram* etc.;
- Desenvolver um conjunto de *prompts* para guiar a geração de respostas das *LLMs*, explorando variáveis como tom, empatia, fluência e alinhamento com a identidade do estabelecimento;
- Estudar e desenvolver estratégias de interação com as *LLMs*, visando maximizar a acurácia e a adequação contextual das respostas, buscando o alinhamento com a

identidade do estabelecimento;

- Submeter o conjunto de comentários e os *prompts* desenvolvidos nas principais *LLMs* selecionadas para o estudo (*ex: GPT-5.5, Gemini, Llama 4* etc.) para gerar um conjunto de respostas automáticas;
- Analisar e comparar o desempenho das diferentes *LLMs*, utilizando uma metodologia mista que avalie diversas métricas pertinentes às avaliações qualitativas, quantitativas e humanas.

1.4 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está estruturado em sete seções principais, organizadas para guiar o leitor desde a fundamentação teórica até a análise dos resultados experimentais.

A *Seção 1 (Introdução)* contextualiza o problema da gestão de avaliações *online* no setor gastronômico, apresenta a justificativa para o uso de *LLMs* na automação de respostas e define os objetivos gerais e específicos da pesquisa.

A *Seção 2 (Fundamentação Teórica)* estabelece as bases conceituais necessárias para o estudo, abordando o estado da arte das *LLMs*, os princípios da Engenharia de *Prompts* (incluindo estratégias *Zero-Shot* e *Few-Shot*), as métricas de avaliação de texto gerado e os fundamentos do desenho experimental adotado para a comparação multidimensional de modelos.

A *Seção 3 (Trabalhos Relacionados)* situa o estudo no contexto da literatura existente, discutindo trabalhos sobre *LLMs* no domínio de serviço, engenharia de *prompts*, metodologias de avaliação de texto gerado e abordagens computacionais para empatia e adequação de tom.

A *Seção 4 (Metodologia)* detalha a arquitetura experimental adotada, descrevendo a coleta e composição do *Corpus de Origem* (200 comentários reais) e do *Corpus Resultante* (2.400 respostas), as considerações éticas no tratamento dos dados, a justificativa para a seleção dos modelos avaliados (*GPT-5.5, Gemini 3.1* e *Llama 4*), os parâmetros e o procedimento de geração das respostas, a estrutura dos *prompts* empregados nas duas estratégias experimentais, e o conjunto de métricas que sustentam a avaliação multidimensional, fluência via perplexidade, adequação de tom via *ToneCal*, consistência de marca via similaridade estilométrica (*Writeprints*) e empatia percebida via modelo *WASSA*, complementado pela avaliação humana conduzida com 15 avaliadores independentes.

A *Seção 5 (Experimentos e Resultados)* apresenta e discute os dados obtidos. A análise contempla a concordância inter-avaliadores, os resultados da avaliação humana e das métricas automáticas, a análise estratificada por modelo, persona, estratégia de *prompt* e período temporal, exemplos qualitativos ilustrativos extraídos do *Corpus Resultante*, e a

triangulação entre avaliação humana e métricas automáticas, com discussão das convergências, divergências e da natureza aproximativa dos *proxies* computacionais adotados.

A Seção 6 (Conclusões) sintetiza as descobertas do estudo, discute as contribuições metodológicas e as implicações práticas dos achados. Por fim, a Seção 7 (Limitações do Estudo) reconhece as restrições do desenho experimental, discute ameaças à validade e aponta direções para investigações futuras.

1.5 CONTRIBUIÇÕES DO TRABALHO

As principais contribuições deste trabalho organizam-se em três eixos complementares:

1. *Proposta de um protocolo de avaliação multidimensional*: estrutura-se um *framework* reproduzível, articulando quatro dimensões de avaliação humana (empatia, consistência de marca, adequação de tom e fluência) em escala *Likert* com quatro *proxies* computacionais especializados (perplexidade GPT-2PT-BR, *ToneCal*, similaridade estilométrica *Writeprints* e modelo *WASSA*), aplicável a outras tarefas de geração de texto em português brasileiro.
2. *Comparação sistemática entre modelos, personas e estratégias de prompt*: por meio de um desenho fatorial completo ($3 \times 2 \times 2 \times 2 = 24$ cenários, 2.400 respostas), o estudo oferece evidências comparativas controladas entre três *LLMs* de fronteira (*GPT-5.5*, *Gemini 3.1* e *Llama 4 Maverick*), duas personas de marca (formal e informal) e duas estratégias de engenharia de *prompt* (*zero-shot* e *few-shot*) no domínio gastronômico em português brasileiro.
3. *Análise da convergência e divergência entre avaliação humana e métricas automáticas*: a triangulação sistemática entre as duas modalidades, sustentada por concordância inter-avaliadores robusta ($\kappa = 0,76\text{--}0,86$), permite identificar vieses sistemáticos dos *proxies* computacionais, em particular a preferência das métricas automáticas pelo registro formal, em oposição à preferência humana pelo registro acolhedor, evidenciando os limites legítimos do uso isolado dessas métricas para tomadas de decisão sobre qualidade comunicacional.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta as bases teóricas do trabalho, essenciais para o estudo realizado. Assim, é necessária a compreensão desses conceitos de forma mais clara, definindo também os termos técnicos que serão utilizados.

2.1 O PROBLEMA DA GERAÇÃO AUTOMÁTICA DE COMENTÁRIOS

A comunicação digital entre empresas e consumidores transformou-se em um fenômeno de alta velocidade e visibilidade pública. Em especial, no setor gastronômico, avaliações e comentários publicados em plataformas como *Google Reviews*, *TripAdvisor* e redes sociais desempenham papel determinante na formação da reputação de um estabelecimento (SILVA; BIANCOLINO; SPERS, 2019). O que antes se limitava à experiência individual do cliente passou a configurar um conteúdo de influência coletiva, no qual cada opinião expressa pode impactar significativamente o comportamento de futuros consumidores.

Esse novo cenário de engajamento constante, em que o volume de interações cresce de forma exponencial, impõe uma demanda operacional significativa aos estabelecimentos. Gerenciar milhares de avaliações com consistência, rapidez e sensibilidade humana torna-se uma tarefa logisticamente desafiadora (LEE; LEE; CHOI, 2018). É nesse contexto que surgem as abordagens baseadas em *LLMs*, cuja capacidade de compreender e gerar linguagem natural oferece uma alternativa escalável à redação manual de respostas (GAMBOA-CRUZADO et al., 2023; ARAUJO; MELO; FIGUEIREDO, 2024).

Contudo, a adoção de sistemas automáticos de resposta envolve um desafio de natureza qualitativa: a preservação da empatia e do alinhamento comunicacional da marca. Diferentemente de tarefas puramente informativas, a resposta a avaliações negativas requer compreensão contextual, reconhecimento emocional e adaptação de tom (CROLIC et al., 2023). Respostas genéricas, embora sintaticamente corretas, falham em estabelecer conexão com o cliente e podem, inclusive, intensificar percepções negativas sobre o atendimento.

Um exemplo para ajudar a ilustrar esse cenário: em um relato no *Instagram*, um cliente expressou frustração ao ter uma data comemorativa arruinada pela desorganização da casa: *“Fiz a reserva para comemorar nosso aniversário de 10 anos de casados e pedi uma mesa tranquila. Nos colocaram ao lado da porta da cozinha, com barulho de pratos quebrando e garçons esbarrando na gente o tempo todo. A comida estava boa, mas o clima romântico foi zero. Uma pena.”*. Visto isso, uma possível resposta de *LLMs*, sem ajuste de persona, tende a ignorar o peso emocional da ocasião: *“Olá! Agradecemos o seu feedback sobre a comida. Sentimos muito pelo desconforto com a mesa e o barulho. Vamos passar suas observações para a gerência para melhorarmos nosso atendimento. Esperamos vê-los*

em breve!”. Já a resposta humana do estabelecimento demonstrou acolhimento e leitura do contexto: “*Ficamos verdadeiramente tristes em saber que não conseguimos proporcionar a atmosfera mágica que os seus 10 anos de união mereciam. A falha na alocação da mesa foi nossa e entendemos perfeitamente sua frustração. Gostaríamos de uma chance para reverter essa memória e oferecer a experiência exclusiva que vocês buscaram.*”. Essa comparação evidencia o tamanho do desafio e a complexidade na geração de respostas automáticas que soem naturais e empáticas. A Figura 2 ilustra o comentário original deste cenário.

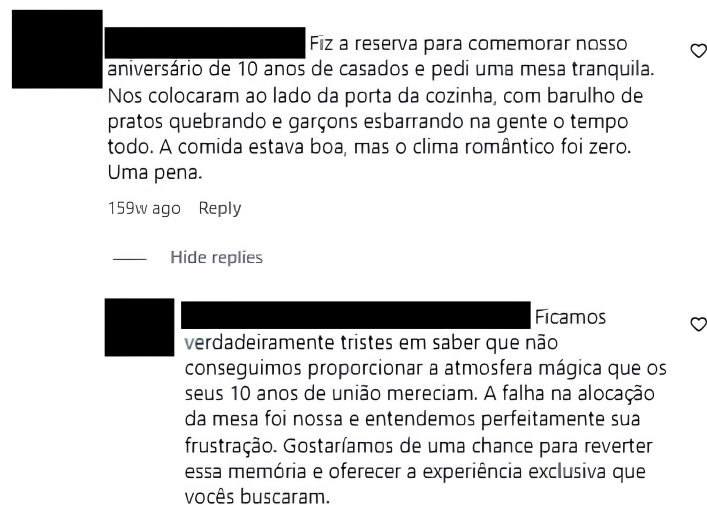


Figura 2 – Comentário sobre data comemorativa coletado do *Instagram*, ilustrando a perda de contexto emocional em respostas sem ajuste de persona.

A literatura recente classifica esse problema em duas dimensões principais:

- a) *Respostas genéricas*: aquelas que, embora corretas e polidas, ignoram o contexto específico do comentário e, portanto, carecem de empatia situacional (LI et al., 2016). A Figura 3 exemplifica esse padrão.

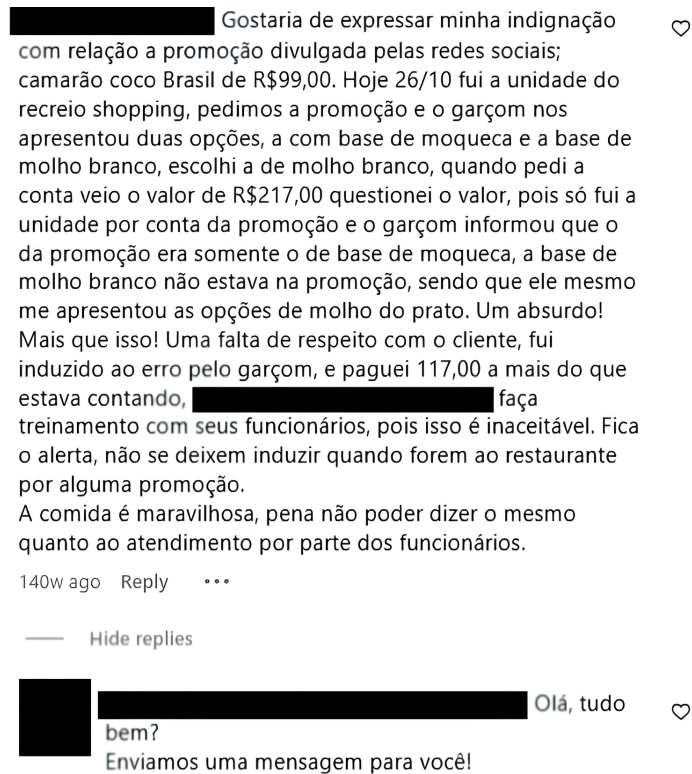


Figura 3 – Resposta genérica a comentário retirado do *Instagram*

- b) *Inconsistência de marca*: ocorre quando o tom da resposta não reflete o posicionamento comunicacional do estabelecimento, gerando dissonância entre a linguagem institucional e a interação pública (WANG et al., 2025). A Figura 4 ilustra esse tipo de ocorrência.

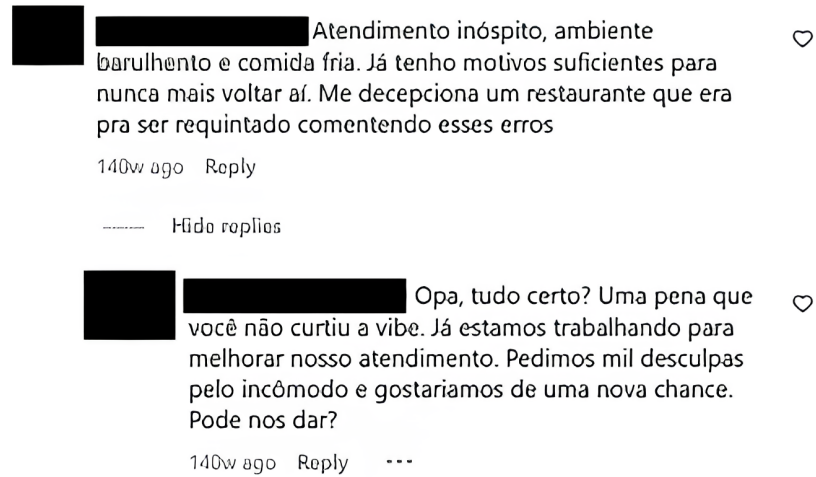


Figura 4 – Resposta inconsistente de restaurante "requintado" a comentário retirado do Instagram

Abordagens tradicionais de automação, como *ChatBots* baseados em regras, indicam limitações expressivas ao lidar com a subjetividade e a ambiguidade típicas do discurso humano (ORTIZ-GARCES et al., 2024). Por essa razão, o estudo e a aplicação de técnicas de *prompt engineering* tornam-se fundamentais para orientar as *LLMs* na geração de respostas mais empáticas, contextualizadas e coerentes com a identidade de marca. Essa investigação representa um passo essencial rumo à automação comunicacional sensível e alinhada à experiência humana.

2.2 PROCESSAMENTO DE LINGUAGEM NATURAL (*PLN*)

O Processamento de Linguagem Natural (*PLN*), ou *Natural Language Processing* (*NLP*), é uma subárea interdisciplinar da Ciência da Computação e da Inteligência Artificial que se preocupa com as interações entre computadores e linguagens humanas (naturais). O objetivo fundamental do *PLN* é ler, decifrar, compreender e dar sentido à linguagem humana de uma maneira que seja valiosa (CHOWDHARY, 2020). Diferente de linguagens de programação, que são precisas e livres de ambiguidade, a linguagem natural é inerentemente complexa, polissêmica e dependente de contexto, tornando seu processamento computacional um dos desafios mais persistentes da *IA* (JURAFSKY;

MARTIN, 2024).

Historicamente, a evolução do *PLN* pode ser categorizada em três eras distintas: a simbólica, a estatística e a neural. As abordagens iniciais (simbólicas), predominantes entre as décadas de 1950 e 1980, baseavam-se em conjuntos complexos de regras manuais e lógica formal para codificar a gramática e o conhecimento. Embora eficazes em domínios restritos, esses sistemas falhavam em capturar a variabilidade e as nuances da fala cotidiana (HIRSCHBERG; MANNING, 2015).

A partir da década de 1990, houve uma mudança de paradigma em direção a métodos estatísticos. Com o aumento da disponibilidade de grandes *corpora* digitais, algoritmos de aprendizado de máquina, como Modelos Ocultos de *Markov* (*HMMs*), passaram a ser utilizados para aprender padrões probabilísticos da linguagem, melhorando significativamente o desempenho em tarefas como tradução automática e reconhecimento de fala (MANNING; SCHÜTZE, 1999).

No entanto, a revolução mais recente e impactante ocorreu com o advento das abordagens neurais e do Aprendizado Profundo (*Deep Learning*). Um marco fundamental foi o desenvolvimento de *Word Embeddings* (vetores de palavras), como o *Word2Vec* proposto por Mikolov et al. (MIKOLOV et al., 2013). Essa técnica permitiu representar palavras como vetores densos em um espaço multidimensional contínuo, onde a proximidade geométrica reflete a similaridade semântica. Por exemplo, o vetor da palavra "rei" menos o vetor de "homem" mais o vetor de "mulher" resulta em um vetor muito próximo de "rainha".

Essa capacidade de representação vetorial é a base das métricas semânticas modernas. Ela permite que o computador compreenda que, em um contexto de restaurante, "demora" e "espera longa" são conceitos semântica e matematicamente próximos, mesmo sendo lexicamente distintos. Essa evolução preparou o terreno para o surgimento dos Modelos de Linguagem em Larga Escala (*LLMs*), que expandiram essas representações para sequências inteiras e contextos complexos.

2.3 LARGE LANGUAGE MODELS

As *LLMs* representam uma mudança de paradigma no campo do Processamento de Linguagem Natural (*PLN*), transitando de abordagens baseadas em regras e modelos estatísticos simples para redes neurais profundas treinadas em *datasets* de escala web. A base arquitetural que viabilizou essa revolução é o *Transformer*, introduzido por Vaswani et al. (VASWANI et al., 2017). Diferentemente das Redes Neurais Recorrentes (*RNNs*) e *Long Short-Term Memory* (*LSTMs*), que processavam dados sequencialmente, limitando a paralelização e o aprendizado de dependências longas, os *Transformers* utilizam o mecanismo de *Self-Attention* (Autoatenção). Esse mecanismo permite que o modelo pondere a relevância de cada palavra em relação a todas as outras em uma frase simultaneamente,

capturando contextos complexos e nuances sintáticas independentemente da distância entre os termos (VASWANI et al., 2017).

O treinamento desses modelos ocorre, primariamente, através de uma etapa de pré-treinamento auto-supervisionado (*Self-Supervised Learning*). Modelos como o *BERT* (*Bi-directional Encoder Representations from Transformers*) e a família *GPT* (*Generative Pre-trained Transformer*) são expostos a terabytes de texto, onde o objetivo é prever a próxima palavra (*token*) em uma sequência ou preencher lacunas mascaradas (DEVLIN et al., 2018; BROWN et al., 2020). Segundo Zhao et al. (ZHAO et al., 2023), esse processo permite que a rede aprenda uma distribuição de probabilidade complexa sobre sequências de palavras, internalizando não apenas gramática e vocabulário, mas também fatos sobre o mundo e capacidades de raciocínio implícito.

Uma característica distintiva das *LLMs* atuais é a observância das "Leis de Escala" (*Scaling Laws*). Kaplan, McCandlish et al. (2020) demonstraram empiricamente que o desempenho do modelo tem uma relação de lei de potência com o aumento do tamanho do modelo (número de parâmetros), o tamanho do *dataset* e o poder computacional empregado. A partir de um certo limiar de escala (geralmente acima de dezenas de bilhões de parâmetros), observam-se as chamadas "Habilidades Emergentes" (*Emergent Abilities*), que são capacidades que não estavam presentes em modelos menores e que não foram explicitamente treinadas, como aritmética, tradução entre línguas com poucos recursos e raciocínio lógico em cadeia (WEI et al., 2022a).

Contudo, apenas o pré-treinamento não é suficiente para tornar um modelo útil e seguro para interação humana, como no caso de respostas a clientes. Um modelo "cru" pode gerar texto tóxico (GEHMAN et al., 2020; WEIDINGER et al., 2021) ou não seguir instruções. Para mitigar isso, utiliza-se o *Reinforcement Learning from Human Feedback* (*RLHF*) ou Aprendizado por Reforço com *Feedback* Humano. Ouyang et al. (OUYANG; WU et al., 2022) detalham que essa técnica alinha o modelo às intenções humanas, utilizando um modelo de recompensa treinado com preferências humanas para punir ou premiar as saídas da *LLM*. É o *RLHF* que permite que o *GPT-5.5* ou o *Llama 4* ajam como assistentes prestativos, modulando o tom para ser "empático" ou "formal" conforme instruído no *prompt*.

Apesar de seu poder, as *LLMs* enfrentam o desafio das "alucinações", a geração de conteúdo factualmente incorreto ou sem sentido, mas com alta confiança aparente (JI; LEE et al., 2023). No contexto de restaurantes, isso poderia resultar na invenção de pratos que não existem ou políticas de cancelamento falsas. Portanto, a aplicação dessas ferramentas exige estratégias robustas de *Prompt Engineering* e validação, especialmente em línguas com morfologia rica como o português, onde a tokenização e a representação semântica podem diferir dos padrões da língua inglesa predominantes nos dados de treino (ARAÚJO; MELO; FIGUEIREDO, 2024).

2.4 ENGENHARIA DE PROMPTS

A interação com as *LLMs* se dá, fundamentalmente, através de "*prompts*", os quais são as entradas fornecidas ao modelo para guiar sua resposta. Um *prompt* pode variar de uma simples instrução textual a combinações mais complexas envolvendo exemplos, contexto adicional ou até imagens. No entanto, obter resultados consistentes e de alta qualidade de uma *LLM*, especialmente em tarefas que exigem nuances como empatia, tom específico e aderência à identidade de marca, raramente é uma questão de apenas fornecer uma instrução direta. Frequentemente, a formulação inicial de um *prompt* resulta em respostas genéricas, desalinhadas ou que não capturam a intenção do usuário, tornando o processo, para muitos, uma questão de tentativa e erro (DANG et al., 2022).

É nesse contexto que surge a *Engenharia de Prompts (Prompt Engineering)*. Trata-se do processo iterativo de projetar, refinar e otimizar os *prompts* fornecidos a uma *LLM* para extrair dela a resposta desejada de forma mais eficaz e confiável. Mais do que uma simples formulação de perguntas ou diretivas, a engenharia de *prompts* é uma prática que explora a maneira como as *LLMs* interpretam instruções e exemplos para maximizar seu desempenho em tarefas específicas (WEI et al., 2022b). A forma como o *prompt* é estruturado pode impactar significativamente a capacidade do modelo de realizar raciocínios complexos (SCHULHOFF et al., 2025).

Uma das abordagens fundamentais na engenharia de *prompts* é o aprendizado *In-Context Learning (ICL)*. Nessa modalidade, em vez de exigir re-treinamento ou ajuste fino do modelo, o próprio *prompt* é utilizado para "ensinar" a tarefa desejada à *LLM* no momento da entrada. Isso pode ser feito de diferentes maneiras (DIAO et al., 2024):

a) *Zero-Shot Learning*: O *prompt* contém apenas a descrição da tarefa ou a instrução direta, sem nenhum exemplo. Por exemplo: "Responda ao seguinte comentário de cliente de forma empática: [comentário]".

b) *Few-Shot Learning*: O *prompt* inclui, além da instrução, um ou mais exemplos (chamados *exemplars* ou *shots*) de como a tarefa deve ser realizada. Por exemplo, fornecer alguns pares de "comentário do cliente" e "resposta ideal" antes de apresentar o novo comentário a ser respondido. A inclusão de poucos exemplos pode melhorar drasticamente o desempenho do modelo em comparação com a abordagem zero-shot (DANG et al., 2022).

2.5 DESENHO EXPERIMENTAL PARA AVALIAÇÃO DE *LLMs*

A avaliação rigorosa de *LLMs* demanda um desenho experimental estruturado que permita isolar o efeito de variáveis independentes sobre a qualidade das respostas geradas. A literatura recente consolidou um paradigma de avaliação baseado em *pipelines* modulares, nos quais as etapas de coleta de dados, geração de texto e avaliação são segmentadas

em fases independentes e reproduzíveis.

O *framework* HELM (*Holistic Evaluation of Language Models*), proposto por Liang et al. (2023), estabeleceu a referência metodológica para a avaliação multidimensional de modelos de linguagem. O HELM organiza a avaliação em um *pipeline* sequencial: definição de cenários (tarefas), construção padronizada de *prompts*, execução da inferência em múltiplos modelos e aplicação de métricas por dimensão. Esse desenho modular permite a comparação controlada entre modelos sob condições experimentais idênticas.

De forma complementar, Fabbri et al. (2021) demonstraram, no *framework* SummEval, a viabilidade de triangular avaliação humana e métricas automáticas em um mesmo *pipeline*. Os autores avaliaram 16 sistemas de sumarização em quatro dimensões (coerência, consistência, fluência e relevância) utilizando tanto anotadores humanos quanto métricas automáticas, revelando que a correlação entre as duas modalidades varia significativamente conforme a dimensão avaliada, e que nenhuma métrica automática isolada substitui adequadamente o julgamento humano.

No âmbito metodológico mais amplo, o paradigma de pesquisa de métodos mistos (*mixed methods*), formalizado por Creswell e Clark (2018), fundamenta a integração de dados qualitativos (avaliação humana) e quantitativos (métricas automáticas) em um único estudo. O desenho convergente paralelo (*convergent parallel design*), no qual ambas as modalidades são coletadas simultaneamente e comparadas na fase de interpretação, é particularmente adequado para a avaliação de sistemas de geração de linguagem natural, pois permite identificar convergências e divergências entre a percepção humana e os *proxies* computacionais.

A combinação de desenho fatorial com avaliação mista foi validada empiricamente em estudos recentes de avaliação de *LLMs*. Nori et al. (2023) compararam múltiplos modelos sob condições controladas de *prompt* (*zero-shot* vs. *few-shot*) em domínios especializados, demonstrando que o cruzamento sistemático de fatores experimentais é essencial para evitar conclusões enviesadas por uma única configuração. De forma análoga, Schulhoff et al. (2025) sistematizaram as técnicas de engenharia de *prompt* e advogam pela comparação fatorial entre estratégias como requisito metodológico para resultados generalizáveis.

A arquitetura experimental adotada neste trabalho fundamenta-se nesses precedentes, organizando a avaliação em um *pipeline* modular com cruzamento fatorial completo (3 modelos $\times$ 2 personas $\times$ 2 estratégias $\times$ 2 períodos) e avaliação convergente paralela (humana e automática), conforme detalhado na Seção 4.

2.6 MÉTRICAS DE AVALIAÇÃO

A avaliação da qualidade das respostas geradas por *LLMs* é uma etapa crucial, porém complexa (CROLIC et al., 2023), especialmente em domínios como o atendimento ao



Figura 5 – Desenho experimental para avaliação de LLMs.

cliente no setor gastronômico. Nesse setor, a experiência do usuário (*UX*), a experiência do cliente (*CX*) e a experiência de marca (*BX*) estão profundamente interligadas, onde fatores como o afeto e o valor percebido pelo usuário impactam diretamente o valor da marca (LEE; LEE; CHOI, 2018). As nuances interpessoais como tom, empatia, fluência e alinhamento com a marca são, portanto, essenciais (CELIKYILMAZ; CLARK; GAO, 2021), e o desafio reside em preencher a "lacuna humano-IA" na experiência afetiva e social do cliente (LIU-THOMPkins; OKAZAKI; LI, 2022). A natureza aberta da geração de linguagem natural, que permite múltiplas respostas válidas para uma mesma entrada, dificulta a aplicação de métricas simples e diretas.

Para fundamentar a análise, é essencial definir estes conceitos chave:

- *Tom*: O tom de um texto escrito refere-se ao sentimento do escritor em relação ao assunto, que deve ser ajustado com base no conteúdo, na situação e nos valores percebidos do leitor para alcançar um objetivo específico (VÁSQUEZ, 2013). Ele define o "*clima*" do texto e pode ser formal, informal, alegre, sério etc.
- *Empatia*: A empatia, por sua vez, é a capacidade de "sentir, entender e compartilhar os pensamentos e sentimentos de outra pessoa". Sua função não é meramente rotular emoções mas "reconhecer como é experienciar diretamente algo" (HALPERN, 2003), envolvendo tanto ser movido pela experiência do outro (HALPERN, 2003) quanto a tomada de perspectiva cognitiva (tentar compreender como o outro entende a situação).

- *Fluência*: A fluência linguística refere-se à capacidade de um texto ser lido de forma natural, coesa e gramaticalmente correta, sem rupturas que comprometam a compreensão (CELIKYILMAZ; CLARK; GAO, 2021). Em textos gerados por *LLMs*, a fluência envolve não apenas a correção gramatical, mas também a naturalidade lexical e a coerência do fluxo textual, sendo frequentemente operacionalizada por meio de métricas de perplexidade (LIANG et al., 2023).
- *Alinhamento de marca*: Por fim, o alinhamento de marca é o equilíbrio que garante a sincronia entre a "identidade da marca" (o que ela diz ser) (BALMER, 2012), as "ações" (como ela gere seus problemas) e a identidade real da organização (o que de fato ela é, sendo assim, a junção de "identidade e ação" que são percebidas no ambiente) (BALMER, 2012), assegurando a credibilidade.

Visto tais conceitos, este estudo empregará uma metodologia mista, combinando avaliação humana detalhada com métricas automáticas quantitativas, conforme detalhado a seguir.

2.6.1 Avaliação Humana com Escala *Likert*

A avaliação humana é considerada o *gold-standard* para aferição da qualidade de textos gerados por sistemas de geração de linguagem natural (CELIKYILMAZ; CLARK; GAO, 2021). Conforme argumentam Amidei, Piwek e Willis (2019), a avaliação automática isolada é insuficiente para capturar dimensões pragmáticas e subjetivas, como adequação contextual, naturalidade e empatia, sendo indispensável a participação de juízes humanos para validar os resultados obtidos por métricas computacionais.

A escala *Likert* ordinal é amplamente adotada na avaliação de sistemas de geração de linguagem natural (HOWCROFT; RIESER, 2021), permitindo que avaliadores expressem graus de concordância em escalas discretas (tipicamente de 1 a 5). Para assegurar a reprodutibilidade dos julgamentos, a literatura recomenda a mensuração da concordância inter-avaliadores por meio de coeficientes complementares: *Fleiss' Kappa* (κ), para concordância categórica entre múltiplos avaliadores; *Krippendorff's Alpha* (α), por sua robustez a dados faltantes e escalas ordinais; e o *Intraclass Correlation Coefficient* (*ICC*), para consistência absoluta em variáveis contínuas ou ordinais. A Seção 4 detalha a composição do painel avaliador e os procedimentos adotados neste estudo.

2.6.2 Perplexidade como Métrica de Fluência

A perplexidade (*perplexity*, PPL) é uma métrica intrínseca de modelos de linguagem que quantifica o grau de "surpresa" do modelo diante de uma sequência de *tokens* (SERRANO; BRUMBAUGH; SMITH, 2023). Valores mais baixos indicam que o texto é mais previsível e, conseqüentemente, mais fluente segundo o modelo avaliador. Originalmente formulada no contexto da teoria da informação, a perplexidade consolidou-se como padrão

para avaliação de fluência em geração de linguagem natural, sendo particularmente eficaz quando calculada por modelos de linguagem treinados no idioma-alvo (PISTOTTI et al., 2025). A definição formal e os parâmetros de implementação adotados neste estudo são apresentados na Seção 4 (Equação 4.1).

2.6.3 ToneCal: Adequação de Tom

A métrica ToneCal (*Tone Quality Score*, TQS) avalia a adequação de tom de textos, produzindo pontuações na escala $[0, 1]$. Sua formulação fundamenta-se no *framework* computacional de polidez linguística proposto por Danescu-Niculescu-Mizil et al. (2013), que operacionaliza estratégias de polidez positiva e negativa a partir de marcadores textuais, como saudações, marcadores de deferência, escolha de modo verbal e expressões de agradecimento, extraídos automaticamente. O *ToneCal* estende esse *framework* ao incorporar indicadores de formalidade e cortesia (BALSHETWAR; TUGANAYAT; REGULWAR, 2019), atribuindo pontuações proporcionais à presença e à adequação desses marcadores ao contexto comunicativo.

2.6.4 Writeprints: Consistência Estilométrica de Marca

A similaridade estilométrica, operacionalizada pelo método *Writeprints* (ABBASI; CHEN, 2008), avalia a consistência autoral por meio da extração de perfis multidimensionais. O método baseia-se na premissa de que cada autor, ou persona de marca, possui um “DNA estilístico” identificável, composto por centenas de características agrupadas em categorias lexicais (riqueza vocabular, distribuição de frequência), sintáticas (comprimento e complexidade das sentenças), estruturais (uso de parágrafos, pontuação) e específicas de conteúdo (expressões idiomáticas, registro). A similaridade do cosseno entre vetores *Writeprints* permite quantificar a aderência de um texto gerado ao perfil estilístico de referência, produzindo um índice no intervalo $[0, 1]$.

2.6.5 WASSA: Empatia Percebida

A avaliação automática de empatia percebida baseia-se nas *Shared Tasks WASSA* (*Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*), que propõem a predição de pontuações contínuas de empatia na escala $[1, 5]$ a partir de anotações humanas (BARRIERE et al., 2022, 2023; GIORGI et al., 2024). O arcabouço teórico subjacente é o modelo multidimensional de empatia de Sharma et al. (2020), que distingue entre reações emocionais (*emotional reactions*), explorações (*explorations*) e interpretações (*interpretations*). Embora promissora, essa abordagem apresenta limitações conhecidas de transferência *cross-lingual* e de domínio quando aplicada a idiomas e contextos distintos dos dados de treinamento originais.

3 TRABALHOS RELACIONADOS

3.1 RESPOSTA GERENCIAL A AVALIAÇÕES ONLINE

O comportamento de consumidores no setor de hospitalidade tem sido reconfigurado pela proliferação de avaliações em plataformas digitais. Ye et al. (2011) demonstraram empiricamente que avaliações online influenciam significativamente decisões de reserva em hotéis, estabelecendo o *eWOM* (*electronic word-of-mouth*) como variável crítica para o desempenho de estabelecimentos de hospitalidade. Mauri e Minazzi (2013) reforçam essa evidência ao mostrar que a leitura de comentários online molda expectativas pré-consumo, especialmente quando há respostas gerenciais associadas. No contexto gastronômico, a dependência de plataformas como *TripAdvisor* e *Google Maps* tornou a presença e a qualidade das respostas gerenciais um componente estratégico da reputação digital.

A literatura em *hospitality management* sustenta que a resposta gerencial não apenas atenua os efeitos de avaliações negativas, como pode converter avaliadores insatisfeitos em defensores da marca. Proserpio e Zervas (2017) estimaram, com base em um experimento natural em larga escala envolvendo hotéis, que estabelecimentos que adotaram respostas sistemáticas a avaliações observaram melhoria mensurável em suas notas agregadas, resultado robusto a controles geográficos e temporais. Liu, Schuckert e Law (2018) ampliaram essa análise no contexto asiático, evidenciando que a velocidade, a personalização e o tom da resposta interagem entre si: respostas tempestivas, porém genéricas, produzem efeito inferior a respostas mais lentas, mas individualizadas.

A questão de *como* responder é tão central quanto a de *se* responder. Sparks e Bradley (2017) propuseram uma tipologia tríplice de respostas a avaliações negativas, sintetizada em três dimensões (*accommodative*, *assertive* e *aggressive*), demonstrando que respostas acomodativas, aquelas que reconhecem a queixa, expressam empatia e oferecem solução, maximizam a percepção de credibilidade gerencial. Em consonância, Min, Lim e Magnini (2015) identificaram três fatores determinantes para a eficácia de respostas a avaliações negativas em hotéis: empatia (reconhecimento explícito do sentimento do cliente), paráfrase (demonstração de que o conteúdo da queixa foi compreendido) e velocidade de resposta. Esses três fatores aparecem como variáveis recorrentes na literatura empírica posterior, evidenciando que dimensões linguístico-discursivas como tom, empatia e registro são determinantes para o sucesso da estratégia.

A consolidação dos *Large Language Models* no setor de hospitalidade desencadeou, a partir de 2024, uma nova vertente de pesquisa orientada à aplicação prática de IA generativa nesse domínio. Dwivedi et al. (2024), em um dos trabalhos seminais sobre o tema, propõem uma agenda de pesquisa estruturada para a adoção do *ChatGPT* e de tecnologias correlatas em turismo e hospitalidade, destacando impactos transformacionais sobre operações de *front-office* e *back-office*, bem como desafios associados à alucinação, ao viés

e à percepção de confiança pelos consumidores. Tuyen et al. (2025), em revisão sistemática que sintetiza a produção científica do biênio 2023–2025, mapeiam os principais eixos de aplicação do *ChatGPT* no setor, identificando a comunicação automatizada com clientes, a geração de conteúdo e a recomendação personalizada como linhas de investigação predominantes. No contexto específico da gastronomia, Ali et al. (2025) apresentam evidência empírica de que a adoção de IA generativa em restaurantes afeta significativamente a percepção de *autenticidade da marca*, a imagem institucional e a intenção comportamental dos consumidores, demonstrando que o impacto da automação não se restringe à eficiência operacional, mas atinge diretamente dimensões intangíveis da reputação.

A percepção do consumidor diante de conteúdo gerado por IA constitui outro eixo investigativo emergente. Choi et al. (2024) comparam experimentalmente descrições produzidas por *ChatGPT* a avaliações orgânicas no *Airbnb* e indicam que textos gerados por *LLMs* podem ser tão ou mais persuasivos que conteúdo humano sob determinadas condições, com efeitos mensuráveis sobre confiança e intenção de reserva. Jia, Chi e Chi (2025) estendem essa análise aos sumários de avaliações: sumários gerados por IA e por humanos produzem efeitos distintos sobre a intenção de reserva, modulados pela autenticidade percebida e pela qualidade argumentativa do texto. Complementarmente, Baek, Kim e Kim (2025) sugerem que o *estilo de linguagem* (figurativo *vs.* literal) e a humanização aparente do agente moderam a aceitação de recomendações geradas por IA em hospitalidade, fornecendo suporte direto à hipótese de que dimensões estilísticas como tom, registro e voz da marca são variáveis críticas para a recepção de textos automatizados pelo consumidor final.

A despeito do volume crescente da literatura internacional, há uma lacuna sistemática quando o foco recai sobre o domínio gastronômico em português brasileiro. Os estudos predominantes utilizam corpus de hotéis em língua inglesa (PROSERPIO; ZERVAS, 2017; LIU; SCHUCKERT; LAW, 2018; SPARKS; BRADLEY, 2017), e os trabalhos lusófonos restringem-se majoritariamente a análises de conteúdo descritivas. Mais ainda, embora a literatura recente em hospitalidade já investigue *efeitos* do uso de IA generativa sobre a percepção do consumidor (CHOI et al., 2024; JIA; CHI; CHI, 2025; ALI et al., 2025), são escassos os estudos que avaliam, com rigor metodológico, a *qualidade comunicacional* de respostas gerenciais produzidas por *LLMs* em português brasileiro, especialmente no domínio gastronômico. Este trabalho posiciona-se nessa interseção, articulando os achados consolidados da pesquisa em *hospitality* sobre os fatores que tornam uma resposta gerencial eficaz com a engenharia de *prompts* para *LLMs* em português, investigando se a automação pode preservar as dimensões qualitativas (empatia, tom e consistência com a identidade da marca) valorizadas por consumidores reais.

3.2 LLMs E GERAÇÃO DE RESPOSTAS NO DOMÍNIO DE SERVIÇO

A aplicação de modelos generativos para responder a usuários não é uma novidade, mas a abordagem mudou drasticamente com a chegada dos *Transformers*. Estudos anteriores, como o de (FU; FENG, 2020), exploraram o uso de redes *Seq2Seq* para gerar respostas em sistemas de suporte ao cliente. Embora tecnicamente viáveis, os autores notaram que esses modelos tendiam a gerar respostas curtas e genéricas (“Obrigado pelo contato”), falhando em abordar os pontos específicos levantados pelo cliente. Posteriormente, (HU; CHEN et al., 2023) realizaram um levantamento sobre o uso de *LLMs* modernas em tarefas de diálogo orientado a tarefas, concluindo que, embora a fluência tenha atingido níveis humanos, a alucinação de fatos e a inconsistência de tom permanecem desafios abertos. Os autores destacam que, sem controle explícito via *prompts* ou *fine-tuning*, o modelo tende a adotar um tom neutro padrão, inadequado para situações que exigem alta empatia, como reclamações de clientes. No contexto específico de pesquisas de produtos e serviços, (LI et al., 2016) já apontavam a necessidade de diversidade nas respostas geradas para evitar a percepção robótica do interlocutor.

Estudos publicados a partir de 2024 sugerem que, apesar do salto qualitativo proporcionado pelos *LLMs* de fronteira, o desempenho desses modelos em atendimento ao cliente do mundo real ainda apresenta lacunas substantivas. Gao et al. (2025) introduzem o *OlaBench*, um *benchmark* de serviço industrial multidimensional, e indicam que modelos de ponta como *GPT-5.2* e *Gemini 3 Pro* atingem apenas 70,58 e 70,84 pontos, respectivamente, na avaliação consolidada de qualidade de serviço, evidenciando que métricas tradicionais de sucesso de tarefa subestimam a complexidade real do atendimento. Os autores propõem o *OlaMind*, modelo adaptado por destilação de padrões dialógicos de especialistas e por aprendizado por reforço baseado em rubricas, que supera os modelos comerciais no *benchmark* e alcança redução de 6,6% na taxa de transferência para atendentes humanos em testes A/B online. Em paralelo, Hong et al. (2025) apresentam um *framework* *LLM-in-the-loop* para a estruturação de intenções em diálogos de atendimento ao cliente em larga escala, demonstrando concordância superior a 95% entre o julgamento do *LLM* e o de avaliadores humanos quanto à coerência semântica de *clusters* de intenção. Esses achados consolidam o uso de *LLMs* não apenas como geradores, mas também como avaliadores confiáveis de qualidade no domínio de serviço, embora restritos, predominantemente, a corpus em inglês e chinês.

O controle explícito de dimensões pragmáticas, como empatia e voz institucional, tem mobilizado uma linha específica de pesquisa em geração com *LLMs*. Cao et al. (2024) propõem o *TOOL-ED*, arquitetura que integra bases de conhecimento de senso comum (*COMET*) ao processo de decodificação via *tool-calling*, demonstrando ganhos sistemáticos em métricas de empatia textual sobre o *benchmark* *EmpatheticDialogues*. Tang, Yang e Abbasi (2025), por sua vez, propõem o *PersonaFuse*, *framework* de pós-treinamento

ancorado na *Trait Activation Theory* e no modelo dos cinco grandes fatores de personalidade, que habilita *LLMs* a expressarem dinamicamente diferentes traços conforme o contexto conversacional. Em avaliações específicas de atendimento ao cliente baseado em *reviews*, o *PersonaFuse* alcança desempenho competitivo com modelos substancialmente maiores como *GPT-4o*, fornecendo evidência empírica direta de que a ativação controlada de persona, eixo central do presente trabalho, é um mecanismo eficaz para adaptar tom e estilo a contextos comunicacionais específicos.

Não obstante o avanço da literatura, persiste uma lacuna metodológica relevante: os estudos recentes concentram-se predominantemente em inglês e chinês, valem-se de *benchmarks* sintéticos ou de domínios distintos (saúde mental, suporte técnico, varejo) e raramente comparam *LLMs* comerciais distintos sob o mesmo protocolo experimental no contexto de atendimento gastronômico em língua portuguesa. O presente trabalho avança sobre essas questões ao investigar, em corpus real de comentários de restaurantes em português brasileiro e com avaliação humana estruturada, se *LLMs* comerciais atuais conseguem superar a barreira da resposta genérica por meio de *personas* bem definidas, articulando os achados internacionais sobre controle de persona e qualidade de serviço a um contexto linguístico e cultural ainda pouco explorado.

3.3 MODELOS DE LINGUAGEM E *PLN* PARA O PORTUGUÊS BRASILEIRO

A maior parte das contribuições em *PLN* originadas a partir do paradigma *Transformer* concentra-se em corpus de língua inglesa, refletindo um viés de recursos que historicamente desfavoreceu línguas não majoritárias, incluindo o português brasileiro. Iniciativas dedicadas ao desenvolvimento de recursos lexicais e modelos pré-treinados para o português têm buscado mitigar essa lacuna. Filho et al. (2018) introduziram o *brWaC*, corpus web de aproximadamente 2,7 bilhões de *tokens* em português brasileiro, que se tornou recurso de referência para o pré-treinamento de modelos voltados ao idioma. Esse corpus serviu de base para iniciativas posteriores, evidenciando que a qualidade e o volume dos dados de pré-treinamento permanecem fatores determinantes para o desempenho final dos modelos.

No âmbito dos modelos pré-treinados, Souza, Nogueira e Lotufo (2020) desenvolveram o *BERTimbau*, variante do *BERT* (DEVLIN et al., 2018) pré-treinada em *brWaC*, que apresentou ganhos substanciais em tarefas como reconhecimento de entidades nomeadas, similaridade textual e inferência de linguagem natural quando comparado a modelos multilíngues. Em paralelo, Carmo et al. (2020) estenderam o paradigma *Encoder-Decoder* ao português ao apresentar o *PTT5*, variante do *T5* (RAFFEL et al., 2020) treinada em corpus lusófono. Ambos os modelos demonstraram que o pré-treinamento monolíngue, mesmo em volume inferior ao de modelos em inglês, supera consistentemente as variantes multilíngues em tarefas que dependem de nuances semânticas e morfossintáticas do

português.

A geração de *benchmarks* acompanhou o desenvolvimento dos modelos. Real, Fonseca e Oliveira (2020) consolidaram o conjunto *ASSIN 2*, voltado para inferência textual e similaridade semântica em português, estabelecendo um marco comparativo para avaliação de modelos. Mais recentemente, esforços como o *HateBR* (VARGAS et al., 2022) ampliaram o escopo para detecção de discurso ofensivo em *Instagram*, demonstrando a importância de corpus anotados especificamente para o português brasileiro e suas variedades dialetais.

Com a popularização das *LLMs*, observa-se uma segunda onda de modelos especializados em português. Almeida et al. (2024) apresentaram a família *Sabiá-2*, modelos com arquitetura *decoder-only* treinados em corpus majoritariamente lusófono, demonstrando que modelos de tamanho moderado especializados em uma língua podem competir, em tarefas específicas, com modelos generalistas de proporções substancialmente maiores. Esses esforços, somados à oferta crescente de modelos comerciais multilíngues com bom desempenho em português, como *GPT-4*, *Gemini* e *Llama 3* (ACHIAM et al., 2023; TEAM et al., 2024; DUBEY et al., 2024), configuram um cenário em que múltiplas alternativas estão disponíveis para tarefas comerciais aplicadas. Apesar desses avanços, há ainda escassez de trabalhos comparativos que avaliem o desempenho desses modelos em tarefas gerativas situadas, como a resposta a avaliações de consumidores em português brasileiro, lacuna que o presente trabalho contribui para preencher.

3.4 EFICÁCIA DAS ESTRATÉGIAS DE *PROMPTING* (*FEW-SHOT VS. ZERO-SHOT*)

A Engenharia de *Prompts* consolidou-se como uma metodologia científica para controlar o comportamento de *LLMs* sem a necessidade de re-treinamento custoso.

O trabalho seminal de (BROWN et al., 2020) introduziu formalmente o conceito de *In-Context Learning*, demonstrando que o desempenho do modelo *GPT-3* em tarefas de tradução e sumarização aumentava logarithmicamente com a adição de exemplos no contexto (*Few-Shot*). Eles observaram que fornecer exemplos de "entrada e saída ideal" permitia ao modelo "aprender" o formato e o estilo desejados instantaneamente.

Investigando essa estratégia em domínios específicos, (MELLO et al., 2024) comparou 128 combinações de *prompts* para a correção automática de respostas em português. Os resultados indicaram que a estratégia *Few-Shot* não apenas melhorou a acurácia, mas foi determinante para estabilizar a saída do modelo. De forma similar, (WEI et al., 2022b) demonstraram que exemplos no *prompt* funcionam como âncoras cognitivas, guiando o raciocínio do modelo.

Entretanto, há uma lacuna na literatura sobre o custo-benefício dessas estratégias para respostas de restaurantes. Embora a literatura aponte vantagens recorrentes do *Few-Shot* em tarefas lógicas, este trabalho investiga se essa tendência se mantém em

tarefas criativas e subjetivas, como a expressão de empatia em português.

3.5 PERSONAS E CONTROLE DE ESTILO EM *LLMs*

O conceito de persona, originário das ciências da comunicação e do marketing, foi recentemente apropriado pela engenharia de *prompts* como um mecanismo de controle de estilo, registro e voz em *LLMs*. Shanahan, McDonell e Reynolds (2023) discutem, em ensaio teórico publicado na revista *Nature*, que *LLMs* contemporâneos não devem ser entendidos como agentes com identidade fixa, mas como *simuladores* capazes de assumir múltiplos papéis a partir de instruções textuais. Segundo os autores, a atribuição de uma persona ao modelo equivale a condicionar sua distribuição de probabilidade sobre *tokens* em direção a um subconjunto estilístico específico do espaço de geração. Essa formulação fornece o arcabouço conceitual para o uso sistemático de personas em aplicações comerciais.

Empiricamente, a eficácia das personas como mecanismo de controle vem sendo documentada por estudos recentes. Kong et al. (2024) demonstraram que a estratégia *Role-Play Prompting*, na qual o modelo é instruído a assumir um papel definido antes de responder, melhora significativamente o desempenho em tarefas *zero-shot* de raciocínio, sugerindo que o efeito da persona extrapola questões puramente estilísticas. Em paralelo, Tseng et al. (2024), em pesquisa de levantamento sistemático, distinguem duas vertentes complementares no uso de personas: *role-playing* (em que o modelo emula um agente fictício com características explicitamente declaradas) e *personalization* (em que o modelo adapta seu comportamento ao perfil do usuário). Para aplicações comerciais como atendimento ao cliente, a primeira vertente é mais relevante, pois define o tom institucional independentemente da identidade do interlocutor.

A relação entre persona e estilo conecta-se diretamente à literatura de *style transfer*. Reif et al. (2022) propuseram um *pipeline* para transferência de estilo arbitrária baseado exclusivamente em *prompting* de *LLMs*, demonstrando que descrições verbais de estilo (“reescreva no estilo formal”) podem substituir, em muitos cenários, o treinamento supervisionado em pares paralelos. Esse achado é particularmente relevante para o presente trabalho, pois fundamenta a viabilidade de definir personas operacionais (formal versus informal) por instrução direta, sem necessidade de *fine-tuning*. Em complemento, Madaan et al. (2020) já haviam demonstrado, em um contexto pré-*LLMs* modernas, que a polidez é um atributo controlável por transferência de estilo, abrindo caminho para a manipulação fina de tom em respostas geradas.

Apesar da maturidade crescente da literatura, o uso de personas em domínios comerciais aplicados ao português brasileiro permanece pouco investigado. Os estudos disponíveis concentram-se predominantemente em inglês e em contextos genéricos como assistentes conversacionais, suporte técnico e raciocínio matemático, deixando em aberto questões so-

bre a eficácia de personas em tarefas situadas culturalmente, como a resposta a avaliações de restaurantes. O presente trabalho contribui ao operacionalizar duas personas comerciais distintas, formal e informal, e ao avaliar empiricamente seus efeitos sobre dimensões qualitativas (tom, empatia, consistência de marca) por meio de avaliação humana estruturada e métricas automáticas, fornecendo evidência aplicada à interseção entre engenharia de *prompts* e linguagem comercial em português.

3.6 METODOLOGIAS DE AVALIAÇÃO DE TEXTO GERADO

A avaliação de texto gerado por Inteligência Artificial constitui um desafio metodológico central em PLN, dada a natureza multidimensional da qualidade textual (CELIKYLMAZ; CLARK; GAO, 2021). Novikova et al. (2017) demonstraram que métricas baseadas em sobreposição lexical apresentam correlação fraca ou inexistente com o julgamento humano em tarefas de geração de diálogo, evidenciando que uma resposta pode ser perfeitamente adequada sem compartilhar *tokens* com a referência. Essa constatação impulsionou uma mudança paradigmática na área: a adoção de avaliação humana estruturada complementada por métricas automáticas especializadas por dimensão (HOWCROFT et al., 2020).

No âmbito da avaliação humana, Amidei, Piwek e Willis (2019) conduziram uma revisão sistemática sobre o uso de escalas *Likert* em avaliação de *NLG*, recomendando protocolos com definições operacionais explícitas para cada nível da escala a fim de reduzir a variabilidade inter-avaliador. Complementarmente, Howcroft e Rieser (2021) alertaram que tratar dados ordinais como intervalares, prática recorrente na literatura, pode inflacionar o poder estatístico dos testes, resultando em conclusões incorretas. Estudos aplicados, como o de (SEDOC et al., 2019), operacionalizam protocolos de avaliação humana para respostas de *chatbots*, demonstrando a viabilidade de capturar distinções qualitativas de tom e empatia por meio de avaliadores treinados.

No polo das métricas automáticas, a literatura recente converge para *frameworks* holísticos e modulares. Liang et al. (2023) propuseram o *HELM* (*Holistic Evaluation of Language Models*), que avalia modelos de linguagem em dezenas de dimensões, incluindo fluência, robustez e viés, consolidando o paradigma de avaliação multidimensional. Na mesma direção, Lee et al. (2025) introduziram o *CheckEval*, um *framework LLM-as-a-Judge* baseado em *checklists* binárias, demonstrando que a decomposição da avaliação em critérios atômicos aumenta significativamente a concordância entre avaliadores automáticos. Para diálogos especificamente, Zhou, Farag e Vlachos (2024) propuseram um *framework* baseado em *features* extraídas por *LLMs* para avaliar a construtividade de interações, reforçando a tendência de métricas especializadas por dimensão comunicativa.

Com base nessa literatura, este trabalho adota avaliação humana estruturada, escalas *Likert* de 5 pontos com definições operacionais por dimensão complementada por métricas

automáticas especializadas, conforme detalhado na Seção 4.

3.7 EMPATIA COMPUTACIONAL E AVALIAÇÃO DE TOM

A modelagem computacional de empatia e tom em texto constitui um campo interdisciplinar em expansão, situado na intersecção entre *PLN*, psicologia e ciência da computação (YALÇIN, 2019). Diferentemente de tarefas tradicionais de análise de sentimento, que classificam a polaridade emocional de um texto, a empatia computacional busca capturar a capacidade do locutor de reconhecer, compreender e responder adequadamente ao estado emocional do interlocutor (SHARMA et al., 2020).

A predição automática de empatia ganhou impulso com o trabalho de Buechel et al. (2019), que introduziram a modelagem de empatia e *distress* em reações a notícias, estabelecendo a base empírica para os *benchmarks WASSA*. Em paralelo, Sharma et al. (2020) propuseram um *framework* multidimensional que operacionaliza a empatia textual em três mecanismos comunicativos distintos: *emotional reactions* (expressões de emoção compartilhada), *interpretations* (reformulações que indicam compreensão cognitiva) e *explorations* (perguntas que aprofundam a compreensão do estado emocional do outro). Essas duas linhas de pesquisa convergiram na consolidação das *Shared Tasks WASSA*: Barriere et al. (2022) estenderam a tarefa para contextos conversacionais; Barriere et al. (2023) incorporaram a predição conjunta de empatia e personalidade; e Giorgi et al. (2024) ampliaram o escopo para interações dialógicas completas, evidenciando a crescente sofisticação dos *benchmarks* disponíveis.

A detecção computacional de empatia tem sido investigada sob diferentes perspectivas metodológicas. Montiel-Vázquez, Uresti e Loyola-González (2022) aplicaram uma abordagem de IA explicável (*Explainable AI*) à detecção de empatia em comunicação textual, demonstrando que *features* como escolha lexical, tom e estrutura discursiva são preditores significativos, e que a transparência do modelo favorece a interpretabilidade em contextos aplicados. No domínio de diálogos, Rashkin et al. (2019) construíram o *benchmark EmpatheticDialogues*, evidenciando que modelos treinados para reconhecer e expressar empatia geram respostas percebidas como mais naturais e engajantes por avaliadores humanos.

No que concerne à avaliação de tom, a polidez linguística computacional oferece um *framework* complementar. O trabalho seminal de Danescu-Niculescu-Mizil et al. (2013) formalizou a polidez como um fenômeno computável, extraindo marcadores textuais de estratégias de polidez positiva (saudações, agradecimentos, expressões de solidariedade) e negativa (deferência, atenuação, desculpas) com base na teoria de Brown e Levinson. Priya, Firdaus e Ekbal (2024), em uma revisão abrangente sobre polidez computacional, mapearam a evolução do campo desde abordagens baseadas em regras até modelos neurais, identificando que a integração de polidez com análise de tom (BALSHETWAR;

TUGANAYAT; REGULWAR, 2019) e com transferência de estilo (MADAAN et al., 2020) abre perspectivas para o controle fino do registro em respostas geradas automaticamente.

A aplicação dessas dimensões no contexto de serviço ao cliente tem recebido atenção crescente. Rohden e Espartel (2024) investigaram o impacto da empatia e do tom emocional de *chatbots* sobre a satisfação do consumidor, demonstrando que a combinação de reconhecimento emocional com tom adequado ao contexto, empático para reclamações, entusiástico para elogios, eleva significativamente o *word-of-mouth* positivo. Contudo, os estudos existentes concentram-se predominantemente em inglês e em domínios como saúde mental ou suporte técnico, deixando uma lacuna na avaliação integrada de empatia e tom para respostas comerciais em português brasileiro, lacuna que este trabalho busca preencher ao articular métricas automáticas especializadas com avaliação humana estruturada, conforme detalhado na Seção 4.

4 METODOLOGIA

4.1 ARQUITETURA E FLUXO DE DADOS

A metodologia deste trabalho foi estruturada sob a forma de um *pipeline* modular de *PLN*, desenhado para garantir a rastreabilidade dos dados e a reprodutibilidade dos experimentos. A arquitetura experimental, ilustrada na Figura 6, foca na validação de hipóteses e comparação de modelos, integrando coleta de dados, engenharia de *prompts*, geração de respostas e avaliação dual (humana e automática).

4.1.1 *Pipeline* Experimental

A Figura 6 ilustra o fluxo de dados do *pipeline* experimental. Este *pipeline* foi projetado seguindo as diretrizes de reprodutibilidade em Inteligência Artificial generativa propostas por Kapoor e Narayanan (2023), estabelecendo controles rígidos sobre as variáveis de entrada. O processo é segmentado em quatro módulos sequenciais:

1. *Coleta e Pré-processamento (ETL)*: A etapa inicial consistiu na extração e tratamento dos dados brutos (*Raw Data*). Para garantir a conformidade ética e a integridade da avaliação, aplicou-se um processo de sanitização para remoção de Informações Pessoalmente Identificáveis (PII - *Personally Identifiable Information*), como nomes de funcionários ou dados de contato, prevenindo vieses nos modelos e protegendo a privacidade dos envolvidos (BROWN et al., 2020). A heterogeneidade foi assegurada pela coleta cross-platform *Google Reviews*, *TripAdvisor*, *Instagram*.
2. *Design Sistemático de Prompts*: O módulo de engenharia de *prompt* não foi tratado como uma entrada estática, mas como uma variável experimental controlada. Baseado na taxonomia de Wei et al. (2022b), foram desenvolvidos *templates* estruturados que variam em duas dimensões ortogonais:
 - *Personalidade (Persona)*: Definição explícita do tom de voz (Sofisticada/Formal vs. Acolhedora/Informal).
 - *Contexto (Shot-Strategy)*: Variação da densidade de informação contextual, comparando a capacidade de generalização *Zero-Shot* com a capacidade de aprendizado em contexto *Few-Shot*.
3. *Motor de Inferência (LLMs)*: A execução dos experimentos utilizou três arquiteturas de estado da arte: *GPT-5.5* (proprietário, acesso via API), *Gemini 3.1* (proprietário, acesso via interface web) e *Llama 4 Maverick* (pesos abertos, execução local). Para preservar variabilidade lexical sem comprometer a coerência, adotou-se temperatura $\tau = 0,7$ e *top-p* = 0,9 sempre que esses parâmetros eram acessíveis ao usuário, com os demais hiperparâmetros mantidos em valores neutros, conforme detalhado na Seção 4.5 e na Tabela 4 (ZHAO et al., 2023).

4. *Framework de Avaliação Híbrida*: A validação dos resultados adotou uma abordagem *Human-in-the-Loop* (HITL), combinando a escalabilidade das métricas automáticas com a precisão do julgamento humano.

- *Qualitativa*: Avaliação cega realizada por 15 avaliadores humanos independentes, focada em quatro dimensões subjetivas e pragmáticas (empatia, consistência de marca, adequação de tom e fluência) via escala *Likert* de 5 pontos.
- *Quantitativa*: Aplicação de métricas automáticas especializadas para medir dimensões complementares da qualidade textual: fluência via perplexidade (GPT-2 PT-BR), adequação de tom (ToneCal), consistência de marca por similaridade estilométrica (*Writeprints*) e empatia percebida (modelo WASSA) (PIS-TOTTI et al., 2025; DANESCU-NICULESCU-MIZIL et al., 2013; ABBASI; CHEN, 2008; BARRIERE et al., 2023).

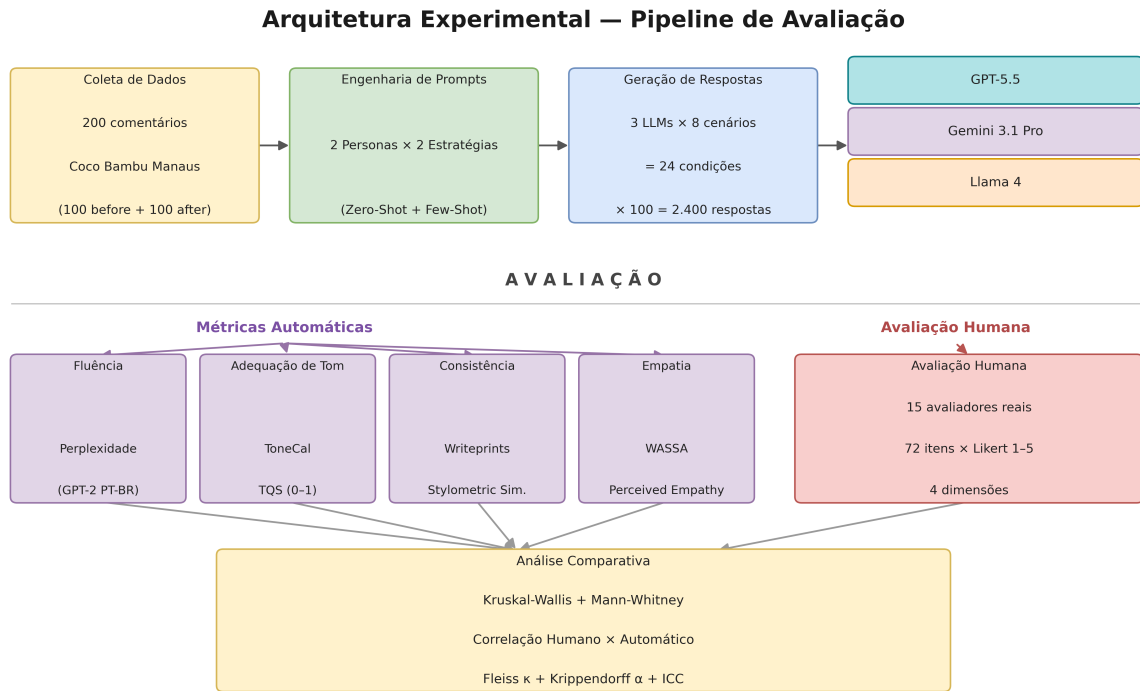


Figura 6 – Coleta de 200 comentários, engenharia de prompts com duas personas e duas estratégias, geração de 2.400 respostas por três LLMs, e avaliação multidimensional com análise comparativa.

4.2 CONJUNTO DE DADOS

O conjunto de dados utilizado nesta pesquisa é composto por duas partes principais: o *Corpus de Origem*, que reúne os comentários reais coletados das plataformas digitais, e o *Corpus Resultante*, que compreende as 2.400 respostas automáticas produzidas pelos LLMs ao longo do *pipeline* experimental. Esta seção descreve a origem, os critérios

de seleção, a anonimização e as características gerais do *Corpus de Origem*; o *Corpus Resultante* é detalhado na Seção 4.2.5.

4.2.1 Caracterização Geral do Corpus de Origem

O *Corpus de Origem* é composto por 200 comentários textuais em português brasileiro, coletados manualmente no mês de março de 2026, a partir de três plataformas públicas de avaliação de estabelecimentos gastronômicos: *Google Maps*, *TripAdvisor* e *Instagram*. Todos os comentários referem-se a um único estabelecimento (*Coco Bambu Manaus*), escolhido por concentrar volume expressivo de avaliações textuais públicas em ambos os períodos temporais analisados (pré e pós-popularização das *LLMs*) e por apresentar perfil heterogêneo de clientela.

A Tabela 1 sintetiza as principais características quantitativas do corpus.

Tabela 1 – Caracterização Quantitativa do *Corpus de Origem*

Característica	Valor
Número total de comentários	200
Comentários no período <i>before</i> (até 30/11/2022)	100
Comentários no período <i>after</i> (após 30/11/2022)	100
Comprimento médio (palavras por comentário)	18,6
Comprimento mínimo (palavras)	10
Plataformas de origem	<i>Google Maps, TripAdvisor, Instagram</i>
Distribuição de polaridade (positivos / neutros / negativos)	≈45% / 15% / 40% ^a
Idioma	Português brasileiro
Período de coleta dos comentários publicados	2019–2025
Período de extração e anonimização	Março de 2026

A polaridade foi atribuída manualmente pelo autor a partir da leitura integral de cada comentário, considerando o conteúdo textual independentemente da nota numérica eventualmente associada na plataforma.

4.2.2 Critérios de Seleção, Inclusão e Exclusão

A seleção dos 200 comentários seguiu critérios objetivos de inclusão e exclusão, aplicados sistematicamente sobre o universo de avaliações publicamente disponíveis no estabelecimento alvo:

Critérios de inclusão:

- Presença de conteúdo opinativo, descritivo ou emocional explícito sobre a experiência no estabelecimento;

- Comprimento mínimo de 10 palavras, garantindo material textual suficiente para a geração de uma resposta contextualizada;
- Clareza semântica adequada para interpretação humana, sem dependência de imagens, vídeos ou contexto externo;
- Diversidade de polaridade (positivos, negativos e neutros), preservando a heterogeneidade das situações reais de atendimento;
- Disponibilidade pública da resposta gerencial associada (utilizada como referência humana para a métrica estilométrica, cf. Seção 4.2.5).

Crítérios de exclusão:

- Comentários extremamente curtos sem conteúdo opinativo substantivo (e.g., "bom", "ótimo", "recomendo");
- Comentários duplicados ou semanticamente redundantes oriundos do mesmo autor;
- Mensagens classificáveis como *spam*, propaganda ou conteúdo não relacionado ao estabelecimento;
- Comentários compostos majoritariamente por emojis, símbolos ou abreviações que comprometessem a leitura;
- Comentários em idiomas distintos do português brasileiro;
- Comentários cujo conteúdo expusesse, mesmo após anonimização, terceiros não associados ao estabelecimento.

4.2.3 Procedimento de Anonimização

Todos os comentários selecionados foram submetidos a um procedimento sistemático de anonimização, conduzido manualmente pelo autor antes de sua incorporação ao *corpus*. Foram removidos ou ocultados:

- Nomes de usuários e identificadores de conta nas plataformas de origem;
- Fotografias de perfil e quaisquer elementos visuais identificáveis;
- *URLs* associadas a perfis pessoais ou conteúdos externos;
- Menções a terceiros (familiares, acompanhantes) que pudessem permitir identificação indireta;
- Quaisquer informações sensíveis de saúde, financeiras ou de localização precisa.

Nomes próprios mencionados no corpo dos comentários e referentes a funcionários do estabelecimento, comumente citados em elogios ou críticas pontuais, foram *preservados* na forma original quando integrantes do registro público de avaliações. Esta decisão foi tomada por se tratar de informação amplamente acessível na própria plataforma, sem

expor dado adicional não disponibilizado pelo titular. O tratamento adotado observa os princípios de finalidade, necessidade e transparência previstos na Lei Geral de Proteção de Dados (*LGPD*, Lei n. 13.709/2018) e está descrito em detalhe na Seção 4.2.7.

4.2.4 Marco Temporal e Divisão do Corpus

Uma decisão metodológica central para a integridade comparativa do estudo foi a definição de um marco temporal rígido para a partição do *corpus*, estabelecido em 30 de novembro de 2022, data oficial de lançamento do *ChatGPT* pela *OpenAI*. A divisão em dois subconjuntos balanceados (100 comentários em cada período) busca preservar uma *baseline* de resposta humana minimamente contaminada por ferramentas de IA generativa amplamente acessíveis:

- *Período Before (até 30/11/2022)*: 100 pares comentário–resposta publicados antes do lançamento do *ChatGPT*, considerados representativos de respostas redigidas sem auxílio de *LLMs* comercialmente disponíveis. Este subconjunto opera como referência aproximada de redação humana espontânea.
- *Período After (após 30/11/2022)*: 100 pares comentário–resposta publicados após a popularização das *LLMs*, refletindo o contexto atual de interação digital.

Cabe ressaltar que, embora o marco temporal mitigue a contaminação por respostas geradas com auxílio de *LLMs*, não é possível garantir, para o período *after*, que todas as respostas humanas tenham sido redigidas sem qualquer assistência de IA generativa. Esta limitação é discutida na Seção 7.4.

4.2.5 Uso do corpus como referência humana nas métricas automáticas

A separação entre os períodos pré e pós-popularização das *LLMs* sustenta a análise comparativa descrita acima. No entanto, para a métrica de Similaridade Estilométrica (*Writeprints*), que demanda um perfil agregado da *voz* da marca, optou-se por compor a referência humana a partir do conjunto completo das 200 respostas originais do restaurante, tanto do período pré quanto do pós-popularização das *LLMs*. Esta decisão metodológica baseia-se em duas justificativas:

- A métrica estilométrica captura padrões agregados de estilo (distribuição lexical, complexidade sintática, padrões de pontuação e *n-gramas*) que caracterizam a identidade comunicacional da marca como um todo, independentemente da autoria individual de cada resposta;
- A ampliação da base de referência aumenta a estabilidade estatística do perfil estilométrico extraído, reduzindo o impacto de variações características presentes em amostras reduzidas.

Reconhece-se, contudo, que respostas do período pós-popularização das *LLMs* podem ter sido elaboradas com auxílio de ferramentas de IA generativa, o que constitui uma limitação inerente ao uso de dados públicos contemporâneos. Para as demais métricas (Avaliação Humana, Perplexidade, *ToneCal* e *WASSA*), que avaliam cada resposta gerada de forma independente, esta consideração não se aplica, pois não há comparação direta com uma referência humana específica.

4.2.6 Corpus Resultante

O corpus final é composto por 200 comentários reais em português brasileiro, com comprimento variável, refletindo a linguagem natural utilizada em plataformas digitais. Os comentários foram divididos igualmente entre dois períodos temporais (pré e pós 30/11/2022), garantindo consistência na comparação experimental.

O *Corpus Resultante* é o principal resultado do *Pipeline* Experimental da arquitetura. A partir dos 200 comentários do *Corpus de Origem*, foi gerado um conjunto de 2.400 respostas automáticas.

Este conjunto de dados foi estruturado em torno de quatro eixos experimentais para permitir uma análise comparativa robusta:

- *Modelos de LLM*: Para cada comentário, foram geradas respostas de três modelos distintos: *Gemini 3.1*, *GPT-5.5* e *Llama 4*.
- *Estratégia de Prompt*: Duas técnicas de engenharia de *prompt* foram aplicadas:

Zero-Shot: A *LLM* recebeu uma instrução direta sem exemplos prévios.

Few-Shot: A *LLM* recebeu a instrução acompanhada de alguns exemplos de respostas ideais.

- *Persona da Marca*: Duas *personas* distintas foram simuladas para avaliar a consistência de marca, conforme informado:

Persona 1: Sofisticada e formal.

Persona 2: Acolhedora e informal.

- *Período temporal*: O experimento foi aplicado a ambos os subconjuntos do *Corpus de Origem* (*Antes das LLMs* e *Depois das LLMs*).

Essa estrutura resulta em 8 arquivos de dados gerados:

- *After-Llm-Release-Comments-Generated-Responses-Persona-1-Few-Shot*
- *After-Llm-Release-Comments-Generated-Responses-Persona-1-Zero-Shot*
- *After-Llm-Release-Comments-Generated-Responses-Persona-2-Few-Shot*

- *After-Llm-Release-Comments-Generated-Responses-Persona-2-Zero-Shot*
- *Before-Llm-Comments-Generated-Responses-Persona-1-Few-Shot*
- *Before-Llm-Comments-Generated-Responses-Persona-1-Zero-Shot*
- *Before-Llm-Comments-Generated-Responses-Persona-2-Few-Shot*
- *Before-Llm-Comments-Generated-Responses-Persona-2-Zero-Shot*

Cada um desses 8 arquivos contém 100 linhas que apresentam o comentário gerado pela devida LLM (*Gemini*, *GPT*, *Llama*) para aquela condição específica (ex: Período "*After*", *Persona* "*Sofisticada*", Estratégia "*Few-shot*").

A Tabela 2 sugere a estrutura padronizada dentro dos arquivos com as respostas geradas.

<i>GPT-5.5 'Persona'</i>	<i>Gpt Processing Time</i>	<i>Gemini 3.1 'Persona'</i>	<i>Gemini Processing Time</i>	<i>Llama 4 'Persona'</i>	<i>Llama Processing Time</i>
"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"
"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"
"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"
"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"	"Comentário gerado"	"X segundos"

Tabela 2 – Comparativo de Adequação de Persona e Tempo de Processamento por LLM

A Tabela 3 resume a estrutura do conjunto de dados gerado.

<i>Eixo Experimental</i>	<i>Variáveis</i>
Comentários-Fonte	200 (100 antes, 100 depois)
Modelos LLM	3 (<i>Gemini</i> , <i>GPT</i> , <i>Llama</i>)
Personas	2 (Sofisticada e Formal, Acolhedora e Informal)
Estratégias de <i>Prompt</i>	2 (<i>Zero-Shot</i> , <i>Few-Shot</i>)
<i>Total de Respostas Geradas</i>	<i>2.400</i> ($200 \times 3 \times 2 \times 2$)

Tabela 3 – Resumo da Estrutura do Conjunto de Dados Gerado

O cruzamento completo dos quatro eixos experimentais, 3 modelos $\times$ 2 personas $\times$ 2 estratégias $\times$ 2 períodos, define 24 cenários experimentais únicos. Cada cenário corresponde a uma combinação específica (e.g., *GPT-5.5* + *Persona Sofisticada* + *Zero-Shot* + Período *After*) e contém 100 respostas geradas (uma por comentário do corpus de origem). Essa estrutura fatorial completa permite isolar o efeito de cada variável independente sobre a qualidade das respostas.

4.2.7 Considerações Éticas e Tratamento dos Dados

A presente pesquisa foi conduzida com observância dos princípios éticos aplicáveis ao tratamento de dados textuais coletados em plataformas digitais públicas. Todos os comentários que compõem o *Corpus de Origem* foram extraídos de *reviews* disponibilizados

publicamente pelos próprios autores em plataformas como *Google Maps*, *TripAdvisor* e *Instagram*, nas quais os usuários, ao publicarem suas avaliações, consentem com a visibilidade pública do conteúdo. Os dados foram tratados de forma anonimizada, com remoção ou ocultação de nomes de usuários, fotografias de perfil, identificadores de conta, *URLs* associadas e quaisquer outros elementos suscetíveis à identificação direta dos autores. As capturas de tela e exemplos textuais apresentados ao longo deste trabalho foram utilizados exclusivamente para fins acadêmicos de ilustração metodológica, tendo informações sensíveis sistematicamente mascaradas. Eventuais nomes próprios mencionados no corpo dos comentários (e.g., nomes de funcionários do estabelecimento citados em elogios ou críticas) foram preservados na forma original quando integrantes do registro público de avaliações, sem qualquer juízo de valor ou exposição adicional por parte da pesquisa.

Os dados não foram empregados para fins comerciais, perfilamento individual ou qualquer aplicação que extrapole o escopo acadêmico-científico aqui descrito. O tratamento adotado observa os princípios de finalidade, necessidade e transparência previstos na Lei Geral de Proteção de Dados (*LGPD*, Lei n. 13.709/2018), tratando os comentários como dados pessoais de natureza pública divulgados pelo próprio titular, com uso restrito ao interesse legítimo de pesquisa científica. O *Corpus de Origem* e o *Corpus Resultante* disponibilizados no repositório do projeto (Apêndice A) seguem o mesmo procedimento de anonimização adotado neste documento, assegurando que sua reutilização em trabalhos futuros preserve os princípios éticos aqui descritos.

Anonimização e Tratamento de Dados Todos os dados passaram por um processo de anonimização, com remoção de nomes de usuários, identificadores de conta, links/*URLs* e informações potencialmente identificáveis.

4.2.8 Uso Consciente de Inteligência Artificial Generativa

Considerando que o objeto de estudo desta pesquisa são precisamente os *Large Language Models*, registra-se de forma explícita e transparente que ferramentas de Inteligência Artificial generativa foram utilizadas como apoio em etapas auxiliares da elaboração deste trabalho. Esse uso seguiu princípios de transparência, supervisão crítica e responsabilidade autoral. A autoria intelectual, conceitual e científica do trabalho permanece exclusivamente do autor, com orientação acadêmica institucional. Nenhuma ferramenta de *IA* é creditada como coautora, em alinhamento com as diretrizes editoriais correntes em periódicos científicos sobre o uso responsável de *IA* na pesquisa.

Especificamente, recursos de *IA* generativa foram empregados como assistentes nas seguintes atividades auxiliares:

- Revisão ortográfica, padronização de acentuação e uniformização de termos técnicos ao longo do documento, com validação final realizada pelo autor;

- Geração de *scripts* de visualização de dados em *Python* (*boxplots*, gráficos de dispersão, gráfico polar e barras comparativas), todos posteriormente revisados, executados e validados quanto à correção dos dados representados;
- Apoio à pesquisa bibliográfica, com sugestões iniciais de literatura potencialmente relevante, sendo cada referência verificada manualmente quanto à existência, à fidelidade da citação e à pertinência ao argumento desenvolvido;
- Apoio à concepção visual de figuras e diagramas conceituais, incluindo sugestões de paleta de cores, *layout* e organização de elementos, sob direção editorial integral do autor.

De maneira igualmente importante, registra-se que ferramentas de *IA* generativa *não* foram utilizadas nas seguintes atividades centrais à pesquisa:

- Coleta de dados, realizada manualmente nas plataformas *Google Maps*, *TripAdvisor* e *Instagram*;
- Geração das respostas avaliadas, que constituem o próprio objeto de estudo e não foram revisadas ou pós-processadas por *IA* antes da avaliação;
- Análise estatística, incluindo os testes de *Kruskal-Wallis*, *Mann-Whitney U*, *Fleiss κ* , *Krippendorff α* e *ICC*, conduzidos pelo autor sem assistência de *IA*;
- Decisões metodológicas, formulação de hipóteses de pesquisa e interpretação dos resultados, integralmente de autoria humana.

Toda saída produzida com apoio de *IA* passou por revisão crítica do autor, que assume responsabilidade integral pelo conteúdo final apresentado neste documento.

4.3 MÉTRICAS DE AVALIAÇÃO

A avaliação da qualidade de textos gerados por *LLMs* é uma tarefa intrinsecamente complexa, especialmente ao tentar mensurar variáveis subjetivas como tom, empatia, fluência e consistência de marca. A literatura acadêmica aponta que não há uma métrica automática singular capaz de capturar a totalidade da adequação direta e da percepção humana em tarefas de geração de linguagem natural (HOWCROFT et al., 2020).

Por essa razão, este trabalho emprega uma metodologia mista, alinhada com os objetivos da pesquisa. Esta abordagem combina uma rigorosa avaliação humana qualitativa, considerada o "*padrão-ouro*" (*gold standard*) para este tipo de tarefa (CELIKYLMAZ; CLARK; GAO, 2021), com métricas quantitativas automáticas para objetividade e escalabilidade.

4.3.1 Avaliação Humana (Qualitativa)

A principal forma de avaliação adotada neste estudo é qualitativa, fundamentada em julgamento humano. Reconhecendo que a clientela de restaurantes é intrinsecamente heterogênea, a composição da banca de avaliadores foi desenhada estrategicamente para evitar vieses de bolhas tecnológicas ou geracionais.

Foram selecionados 15 avaliadores independentes, com idades entre 21 e 71 anos (média de 40 anos), distribuídos em quatro faixas etárias para assegurar representação intergeracional: quatro avaliadores entre 21 e 29 anos, cinco entre 30 e 39 anos, quatro entre 40 e 59 anos e dois com 60 anos ou mais. Quanto à atuação profissional, o painel reuniu perfis heterogêneos: quatro profissionais da área da saúde (dois médicos, um dentista e um profissional de educação física), quatro engenheiros (dois engenheiros de *software*, um engenheiro florestal e um engenheiro elétrico), três estudantes universitários, uma advogada, um corretor de imóveis e dois avaliadores em situação ocupacional distinta (uma autônoma e um aposentado). Essa composição contempla diferentes níveis de letramento digital e familiaridade com tecnologias de *IA*, bem como diferentes contextos de consumo gastronômico, mitigando vieses de percepção associados a bolhas tecnológicas, ocupacionais ou geracionais e aproximando o painel da heterogeneidade da clientela real de restaurantes (HOWCROFT; RIESER, 2021).

Cada avaliador analisou um conjunto cego de 72 itens (respostas geradas), selecionados por amostragem estratificada proporcional: 3 itens por cenário experimental (3×24 cenários = 72), garantindo representação de todas as combinações de modelo, persona, estratégia e período. A seleção dentro de cada cenário foi realizada por subamostragem aleatória dentre as 100 respostas disponíveis naquele estrato, assegurando cobertura sistemática do espaço experimental. Os itens foram apresentados sem identificação do modelo de origem, utilizando um *Scorecard* (Apêndice B) com escala *Likert* de 5 pontos para quatro dimensões: Empatia, Consistência de Marca, Adequação de Tom e Fluência.

Os avaliadores receberam instruções padronizadas, entregues por escrito junto ao formulário, contendo: (i) uma breve descrição do contexto da pesquisa (resposta a avaliações de restaurantes); (ii) definições operacionais explícitas para cada uma das quatro dimensões avaliadas (empatia, consistência de marca, adequação de tom e fluência), com exemplos comentados de pontuações 1, 3 e 5; (iii) orientações sobre o uso da escala *Likert* de 5 pontos; e (iv) recomendação de leitura integral do comentário antes da resposta. Não houve treinamento prévio nem rodada de calibração entre avaliadores, de modo a preservar a heterogeneidade de julgamentos típica da clientela real do estabelecimento; a consistência foi aferida *a posteriori* pelos coeficientes apresentados na Tabela 5.

O instrumento de coleta consistiu em um formulário estruturado no *Google Forms* (Apêndice B), com uma página por item avaliado. Cada item apresentava: o comentário

original anonimizado, a resposta gerada (sem identificação do modelo, persona ou estratégia) e quatro escalas *Likert* de 1 a 5 (uma por dimensão), todas obrigatórias. A ordem dos 72 itens foi *aleatorizada de forma independente para cada avaliador*, evitando efeitos de ordem sistemáticos, e a identificação do modelo, da persona e da estratégia de *prompt* permaneceu cega durante toda a avaliação. Os avaliadores não tiveram contato entre si durante o preenchimento e receberam o prazo de uma semana para conclusão, permitindo pausas voluntárias e mitigando fadiga cognitiva. Para reduzir ainda mais a fadiga, o número de itens por participante foi limitado a 72, distribuídos de forma equilibrada entre os 24 cenários experimentais (3 itens por cenário).

A concordância inter-avaliadores foi aferida por três coeficientes complementares (cf. Seção 2.6): *Fleiss' Kappa* (κ), que mede a concordância nominal além do acaso; *Krippendorff's Alpha* (α) com ponderação ordinal, que leva em conta a magnitude das discordâncias; e o *Intraclass Correlation Coefficient* ($ICC_{2,1}$), que avalia a consistência absoluta entre avaliadores tratados como efeito aleatório. Os valores obtidos, κ entre 0,76 e 0,86 (substantial a quase perfeita), α entre 0,82 e 0,94, e ICC entre 0,88 e 0,94, atestam a robustez e a reprodutibilidade dos julgamentos humanos coletados.

A Figura 7 ilustra o processo completo de avaliação humana adotado neste estudo.

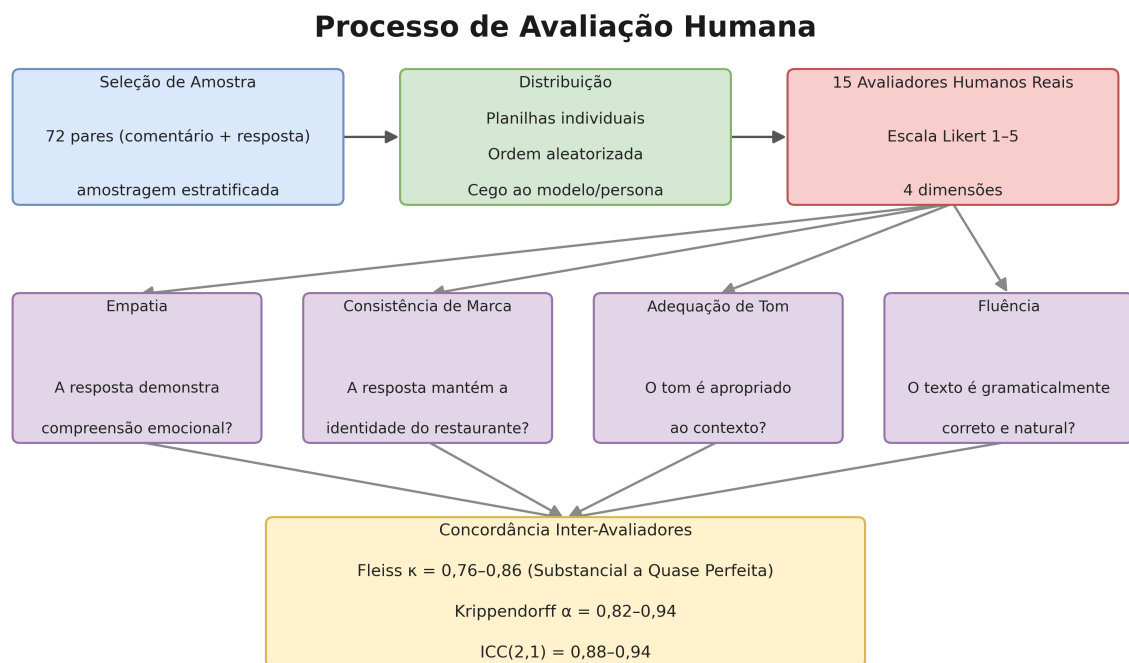


Figura 7 – 72 pares comentário-resposta selecionados por amostragem estratificada, distribuídos individualmente com ordem aleatorizada e cegamento ao modelo. Os 15 avaliadores avaliaram quatro dimensões em escala Likert 1–5, com concordância inter-avaliadores de $\kappa = 0,76\text{--}0,86$ (Fleiss).

4.3.2 Avaliação Automática (Quantitativa)

Para complementar a avaliação humana com objetividade e escalabilidade, foi empregado um *framework* de quatro métricas automáticas, cada uma capturando uma dimensão distinta da qualidade textual. Esta abordagem multimétrica alinha-se à recomendação de Liang et al. (2023) por avaliações holísticas que transcendam uma métrica agregada.

Fluência Textual — Perplexidade (GPT-2 PT-BR). A fluência das respostas geradas foi aferida por meio da perplexidade (*PPL*), calculada com o modelo *GPT-2* pré-treinado para o português brasileiro (*pierreguillou/gpt2-small-portuguese*). Formalmente, para uma sequência de *tokens* $w_1, w_2, \dots, w_N$:

$$PPL(w_1, \dots, w_N) = \exp \left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i \mid w_1, \dots, w_{i-1}) \right) \quad (4.1)$$

A perplexidade mede a fluência *probabilística* do texto, quantificando a previsibilidade do próximo *token* dado o contexto precedente. Seus valores devem ser interpretados como *proxy* de naturalidade linguística, considerando que textos formais e mais longos tendem a obter valores mais baixos independentemente da qualidade pragmática.

Adequação de Tom — ToneCal (TQS). A adequação de tom foi avaliada pela métrica *ToneCal* (*Tone Quality Score*), na escala $[0, 1]$, implementada com base no *framework* de polidez de Danescu-Niculescu-Mizil et al. (2013) e na sistematização de Priya, Firdaus e Ekbal (2024). A métrica opera sobre marcadores lexicais de superfície, saudações, expressões de deferência, escolha de modo verbal e indicadores de cortesia, sendo mais sensível a indicadores explícitos de polidez do que a nuances pragmáticas de tom implícito. Neste estudo, o ToneCal analisa marcadores de cortesia, deferência e assertividade contextualizada, calibrados para as duas personas de atendimento definidas (Tabela 3).

Consistência de Marca — Similaridade Estilométrica (*Writeprints*). A consistência de marca foi mensurada por similaridade estilométrica via *Writeprints* (ABBASI; CHEN, 2008). A implementação extrai perfis multidimensionais (distribuição lexical, complexidade sintática, frequência de *n-gramas* de caracteres e padrões de pontuação) do corpus de referência humano (200 respostas originais do restaurante) e de cada resposta gerada, computando a similaridade na escala $[0, 1]$. Conforme detalhado na Seção 4.2.5, o perfil estilométrico de referência foi extraído do conjunto completo das 200 respostas humanas do restaurante.

Empatia Percebida — Modelo WASSA. A empatia percebida foi aferida pelo modelo treinado na *WASSA Shared Task* (BARRIERE et al., 2023), produzindo pontuações

na escala [1, 5]. Embora o modelo original tenha sido desenvolvido em inglês para o domínio jornalístico, sua arquitetura baseada em *transformers* multilíngues permite a captura de padrões empáticos transferíveis entre línguas e domínios (BARRIERE et al., 2023). Neste estudo, a métrica é empregada como indicador comparativo entre modelos e estratégias, permitindo identificar diferenças relativas nos níveis de empatia das respostas geradas.

É importante destacar que as quatro métricas automáticas adotadas neste estudo devem ser interpretadas como *proxies* computacionais aproximativos de qualidade textual, e não como equivalentes diretos da percepção humana. A perplexidade, por exemplo, mede a previsibilidade do texto segundo um modelo de linguagem específico (*GPT-2 PT-BR*), não a fluência efetivamente percebida por leitores humanos. De forma análoga, a similaridade estilométrica (*Writeprints*) captura padrões agregados de léxico, sintaxe e pontuação, mas não a percepção subjetiva de identidade de marca; o *ToneCal* é sensível a marcadores explícitos de polidez de superfície, podendo não capturar plenamente nuances de tom implícito; e o modelo *WASSA*, treinado em corpus jornalístico em inglês, fornece estimativas comparativas de empatia, mais informativas para diferenciar modelos do que para valores absolutos no contexto português brasileiro.

Por essas razões, os resultados obtidos por meio dessas métricas são analisados em conjunto com a avaliação humana ao longo de toda a Seção 5, e as divergências entre as duas modalidades, quando ocorrem, são interpretadas com a cautela exigida pela natureza aproximativa dos *proxies* computacionais adotados (cf. Seção 5.8).

4.4 LARGE LANGUAGE MODELS

A seleção dos modelos para este estudo (Objetivo Específico 4) foi pautada na necessidade de avaliar o desempenho do atual estado da arte em geração de linguagem, abrangendo os principais exemplares de *LLMs* disponíveis (ZHAO et al., 2023). O objetivo é avaliar a viabilidade e maturidade dessas ferramentas para o contexto de restaurantes.

Para realizar uma análise comparativa robusta, foram selecionados três modelos de ponta que representam as duas principais vertentes do ecossistema de *LLMs*: os modelos proprietários (fechados) e os modelos de pesos abertos (*open-weight*).

A escolha específica dos modelos *GPT-5.5*, *Gemini 3.1* e *Llama 4* permite a este estudo cobrir o que há de mais avançado em cada uma dessas vertentes:

- *GPT-5.5 (OpenAI)*: Selecionado por ser amplamente considerado o *benchmark* da indústria em termos de capacidade de raciocínio, geração de texto e, crucialmente, seguimento de instruções complexas (como as de persona e tom) (MINAEE et al., 2025). Sua performance é frequentemente utilizada como a linha de base (*baseline*) para o que é considerado o estado da arte, sendo essencial para estabelecer o teto

de performance esperado para a tarefa.

- *Gemini 3.1 (Google)*: Incluído como o principal competidor do *GPT-5.5*. Os modelos da família *Gemini* têm demonstrado capacidades de ponta, nativamente multi-modais e com forte compreensão de contexto (Google, 2025). Sua inclusão permite uma comparação direta entre as duas *LLMs* de grande escala mais proeminentes no mercado, avaliando qual delas oferece melhor adequação para a geração de respostas empáticas.
- *Llama 4 (Meta)*: Representa o estado da arte dos modelos de pesos abertos (*open-weight*). O *Llama 4*, especificamente, foi treinado em um conjunto de dados massivo e diverso, com melhorias notáveis em seu *tokenizador* para abranger de forma mais eficiente múltiplas línguas, incluindo o português (Meta AI, 2025). Sua inclusão é vital para responder uma questão chave da pesquisa: se soluções *open-source*, que oferecem maior controle e possibilidade de *fine-tuning*, já atingem um nível de maturidade comparável ao dos modelos proprietários em tarefas de alta nuance, como as definidas neste trabalho.

4.5 CONFIGURAÇÃO DOS MODELOS

Os três modelos avaliados foram selecionados por representarem distintas arquiteturas, regimes de disponibilidade (proprietário *vs.* pesos abertos) e modos de acesso. A Tabela 4 consolida as informações de reprodutibilidade adotadas neste estudo.

Tabela 4 – Configuração dos Modelos e Parâmetros de Reprodutibilidade

Parâmetro	GPT-5.5	Gemini 3.1	Llama 4 Maverick
Provedor	OpenAI	Google DeepMind	Meta AI
Identificador do modelo	<code>gpt-5.5</code>	<i>Gemini 3.1 (Pro)</i>	<code>llama4:maverick</code> (17B/128E)
Tipo	Proprietário	Proprietário	Pesos abertos
Modo de acesso	API oficial (<i>Responses API</i>)	Interface web oficial	Execução local via <i>Ollama</i>
Endpoint	<code>api.openai.com/v1/responses</code>	<code>gemini.google.com</code>	<code>localhost:11434</code> (<i>Ollama</i>)
Temperatura	0,7	Padrão da interface ^a	0,7
<i>Top-p</i>	0,9	Padrão da interface ^a	0,9
Máx. <i>tokens</i> de saída	512	512 (truncado em pós-processamento) ^a	512
Penalidade de freq./pres.	0,0 (neutra)	N/D ^a	0,0 (neutra)
Amostras por entrada	1	1	1
<i>Fine-tuning</i> aplicado	Não	Não	Não
Período de coleta	Março de 2026	Março de 2026	Março de 2026

A interface web pública do *Gemini 3.1* não expõe ao usuário final os parâmetros de *decoding* (temperatura, *top-p*, penalidades), de modo que as respostas foram coletadas com as configurações *default* adotadas pelo provedor. Esta diferença de modo de acesso constitui uma limitação assumida do estudo e é discutida na Seção 7.4.

A escolha da temperatura $\tau = 0,7$ e do *top-p* = 0,9, valores adotados de forma uniforme onde acessíveis, segue convenções consolidadas em tarefas criativas de geração

textual em linguagem natural (HOLTZMAN et al., 2020; PEEPERKORN et al., 2024; RENZE; GUVEN, 2024), equilibrando expressividade lexical e coerência. Para o *GPT-5.5*, as chamadas foram realizadas via *SDK* oficial em *Python* (método `client.responses.create`); para o *Llama 4 Maverick*, a inferência foi executada localmente por meio do *runtime Ollama*, garantindo controle integral sobre os parâmetros de *decoding*; para o *Gemini 3.1*, em razão de a *API* programática estável não estar disponível no período de execução, optou-se pela coleta manual via interface web oficial, com *prompts* idênticos aos demais modelos. Não foi aplicado nenhum procedimento de ajuste fino (*fine-tuning*) nem de otimização específica por modelo. Cada entrada foi processada uma única vez, em conformidade com a opção metodológica de evitar regeneração seletiva (cf. Seção 4.7).

Os *prompts* seguiram um *template* padronizado entre os três modelos, com variações controladas apenas no componente de persona e na inclusão/ausência de exemplos demonstrativos, conforme descrito na Seção 4.7.

4.6 PARÂMETROS E PROCEDIMENTO DE GERAÇÃO

A execução experimental foi conduzida de forma programática, em ambiente *Python*, com chamadas sequenciais aos três modelos avaliados. Para cada um dos 200 comentários do *Corpus de Origem*, o *prompt* estruturado (descrito na Seção 4.7) foi submetido a cada modelo (*GPT-5.5*, *Gemini 3.1* e *Llama 4*), combinando-se as duas personas e as duas estratégias de *prompting*, totalizando os 24 cenários experimentais e as 2.400 respostas que compõem o *Corpus Resultante*.

Os modelos proprietários (*GPT-5.5* e *Gemini 3.1*) foram acessados por meio de suas *APIs* oficiais, enquanto o modelo de pesos abertos *Llama 4* foi executado em ambiente local controlado, garantindo a reprodutibilidade integral do procedimento.

A fim de promover variabilidade lexical suficiente sem comprometer a coerência das respostas, adotou-se o parâmetro de temperatura igual a 0,7 em todos os modelos, valor intermediário entre a geração determinística ($T = 0$) e a geração altamente estocástica ($T = 1$), comumente empregado em tarefas criativas de geração textual em linguagem natural (HOLTZMAN et al., 2020; PEEPERKORN et al., 2024; RENZE; GUVEN, 2024).

Os demais parâmetros de *decoding* foram explicitamente controlados para garantir consistência experimental. O *top-p* (*nucleus sampling*) foi fixado em 0,9, enquanto o limite máximo de *tokens* de saída foi definido como 512 para todos os modelos avaliados.

Os parâmetros de penalização de frequência e presença foram mantidos em seus valores neutros (0,0), evitando interferência adicional na geração textual.

Cada chamada gerou uma única resposta, registrada juntamente com seu respectivo tempo de processamento. Não foram aplicados mecanismos de regeneração automática nem critérios de descarte seletivo: todas as 2.400 saídas produzidas foram preservadas e

submetidas integralmente às etapas posteriores de avaliação. Esta escolha metodológica evita vieses de seleção *a posteriori* e assegura que os resultados reflitam o comportamento real de cada modelo na tarefa proposta. O código-fonte completo, os *prompts* empregados e os dados utilizados encontram-se disponíveis no repositório indicado no Apêndice A.

4.7 PROMPTS EMPREGADOS

O desenvolvimento dos *prompts* (Objetivo Específico 2) é um pilar central desta metodologia. Os *prompts* não são apenas a entrada, mas o principal instrumento de controle experimental usado para guiar as *LLMs* na geração de respostas que se alinhem às variáveis de análise (tom, empatia, fluência e consistência de marca) (MELLO et al., 2024).

Conforme a estrutura do *Corpus Resultante* (Seção 4.2.5), o experimento foi desenhado para comparar duas estratégias de *prompt* sobre duas personas distintas. As próximas subseções apresentam, nesta ordem: (i) os princípios de engenharia de *prompt* adotados, (ii) a estrutura base comum aos dois experimentos, e (iii) as duas estratégias experimentais comparadas — *Zero-Shot* e *Few-Shot*.

4.7.1 Engenharia de *Prompt*: Princípios Adotados

A engenharia de *prompt* é a disciplina de projetar instruções textuais de entrada que modulam o comportamento de uma *LLM* sem alterar seus pesos, controlando registro, formato, contexto e estilo da saída (MELLO et al., 2024; MESKÓ, 2023). Conforme apresentado na Seção 2.4 (Fundamentação Teórica), não há *prompt* único correto: cada decisão de projeto encerra um *trade-off* entre especificidade, generalidade e custo cognitivo do modelo.

Neste estudo, a engenharia de *prompt* é tratada como variável experimental controlada, e não como entrada estática. Os *prompts* foram desenhados segundo quatro princípios: (i) atribuição explícita de papel (persona), (ii) definição clara da tarefa e do tom esperado, (iii) diretrizes condicionais para sentimentos distintos (positivo, negativo, neutro) e (iv) restrições de saída para evitar metatexto. Esses princípios compõem a **estrutura base do *prompt***, descrita a seguir, e são aplicados consistentemente em ambas as estratégias experimentais comparadas (*Zero-Shot* e *Few-Shot*).

4.7.2 Estrutura Base do *Prompt* (Componente Fixo)

Todos os *prompts* utilizados, independentemente da estratégia (*Zero-Shot* ou *Few-Shot*), compartilharam uma estrutura base arquitetada para padronizar o comportamento do modelo antes da inserção de quaisquer exemplos. Baseada em técnicas de engenharia de *prompt* (MESKÓ, 2023), esta estrutura fixa continha os seguintes componentes:

- **Atribuição de Papel:** A instrução inicial ancora a identidade da *LLM* (ex: "Você é o gerente do *Instagram*...") e define a *persona* específica a ser adotada (Acolhedora ou Sofisticada), estabelecendo o viés comportamental da resposta.
- **Definição da Tarefa e Tom:** Uma descrição concisa do objetivo ("responder ao Comentário do Cliente") combinada com uma diretriz de estilo explícita ("tom compreensivo e empático"), garantindo que a modulação emocional seja priorizada.
- **Diretrizes Condicionais:** Um bloco de regras lógicas que orienta a tomada de decisão do modelo com base na análise de sentimento do comentário. O *prompt* instrui explicitamente como reagir a cenários distintos: oferecer desculpas e soluções em casos negativos, ou demonstrar alegria e exaltar a experiência em casos positivos/neutros.
- **Restrições de Saída:** Uma diretiva de formatação ("Não precisa demonstrar seu pensamento. Retorne apenas a resposta...") utilizada para impedir que o modelo gere metatexto, explicações ou cadeias de pensamento visíveis, assegurando que o *output* seja estritamente a resposta final ao cliente.

4.7.3 Estratégia Experimental 1: *Prompt Zero-Shot*

Na condição *Zero-Shot*, o modelo recebe apenas a estrutura base do *prompt* e o novo comentário do cliente. Nenhum exemplo de resposta é fornecido.

Esta abordagem testa a capacidade de "raciocínio em-contexto" (*zero-shot reasoning*) da *LLM* (KOJIMA et al., 2022). O objetivo é avaliar a maturidade de cada modelo (*GPT-5.5*, *Gemini*, *Llama 4*) para seguir instruções complexas e subjetivas (como "ser empático") sem qualquer treinamento prévio para a tarefa específica, baseando-se apenas em seu conhecimento pré-treinado (ZHAO et al., 2023). Esta condição serve como a (*baseline*) do desempenho. A Figura 8 apresenta o *template* utilizado.

Estrutura Prática (Template):

Template do Prompt Zero-Shot
Contexto e Papel Você é o gerente de mídias sociais de um restaurante em Manaus e responde comentários públicos de clientes em português do Brasil. Persona de marca: {PERSONA} → sofisticada e formal acolhedora e informal
Objetivo Escrever uma resposta curta, contextualizada ao comentário, preservando: • adequação de tom • percepção de empatia • consistência de marca • fluência em português
Instruções de Comportamento • Soe formal/refinada (ou próxima/calorosa, conforme persona) • Demonstre consideração genuína pela experiência do cliente • Evite soar fria, genérica ou mecanicamente corporativa • Leia o comentário com atenção e responda aos pontos mencionados • Não copie a resposta de referência humana • Não invente fatos, políticas, compensações, nomes ou promessas • Não use emojis, hashtags ou aspas
Regras de Adaptação • Comentário positivo → agradeça, valorize elogios, convide a retornar • Comentário negativo → reconheça insatisfação, expresse pesar, acolha feedback • Comentário misto → reconheça elogios e pontos de insatisfação • Ocasão especial/frustração → sensibilidade e sobriedade • Menção a prato/ambiente/demora → responda esses elementos
Restrições de Saída • Apenas uma resposta final ao cliente • Entre 2 e 4 frases • Específica para o comentário recebido • Evite estruturas repetitivas
Entrada [Comentário do cliente: {customer_comment}]

Figura 8 – Template do prompt Zero-Shot

4.7.4 Estratégia Experimental 2: Prompt Few-Shot

Na condição *Few-Shot*, o *prompt* é aumentado com a inclusão de exemplos de alta qualidade antes de apresentar o comentário a ser respondido.

Justificativa Metodológica: Esta abordagem explora o *In-Context Learning (ICL)*, onde o modelo aprende a tarefa a partir dos exemplos fornecidos no próprio *prompt* (DANG et al., 2022). O objetivo é medir o ganho de performance em relação à abordagem *Zero-Shot*. Espera-se que, ao fornecer exemplos de como a persona e a empatia devem ser aplicadas na prática, a *LLM* consiga gerar respostas mais alinhadas e consistentes, reduzindo a generalidade ou a "robotização" das respostas. A Figura 9 apresenta o *template* utilizado.

Template do Prompt Few-Shot

Contexto, Papel e Objetivo

Você é o gerente de mídias sociais de um restaurante em Manaus.
Persona de marca: {PERSONA}
→ sofisticada e formal | acolhedora e informal

Preservar: adequação de tom | empatia | consistência de marca | fluência

Exemplos Demonstrativos (In-Context Learning)

EXEMPLO 1

Comentário: "Amei a experiência e o atendimento foi incrível."
Resposta: "Ficamos honrados com seu retorno tão positivo."
É uma satisfação saber que sua experiência foi marcante e que o atendimento atendeu às suas expectativas. Esperamos recebê-lo novamente."

EXEMPLO 2

Comentário: "A comida estava boa, mas o atendimento demorou bastante."
Resposta: "Agradecemos por compartilhar sua experiência conosco."
Ficamos satisfeitos que apreciou a refeição, mas lamentamos pela demora. Esperamos proporcionar uma experiência mais harmoniosa."

Instrução Final

Responda ao comentário abaixo sem copiar os exemplos.

Regras

- Não invente fatos, compensações, contatos ou promessas
- Não use emojis, hashtags ou aspas
- Responda em 2 a 4 frases
- Seja específica para o comentário

Entrada

[Comentário do cliente: {customer_comment}]

Figura 9 – Template do prompt Few-Shot

A aplicação dessas duas estratégias (*Zero-Shot* e *Few-Shot*) em todas as combinações

de modelos e personas permitiu a criação sistemática das 2.400 respostas que compõem o *Corpus Resultante*, possibilitando uma análise estatística e qualitativa robusta sobre qual abordagem é mais eficaz para a tarefa.

5 EXPERIMENTOS E RESULTADOS

Esta seção apresenta a análise dos dados obtidos a partir das 2.400 respostas geradas, estratificadas pelos 24 cenários experimentais (3 modelos $\times$ 2 personas $\times$ 2 estratégias de *prompt* $\times$ 2 períodos temporais). A avaliação integra os resultados da avaliação humana (15 avaliadores, 72 itens, escala *Likert* 1–5) com as quatro métricas automáticas (Perplexidade, ToneCal, Similaridade Estilométrica e Empatia WASSA), visando uma análise multidimensional da qualidade das respostas.

5.1 CONCORDÂNCIA INTER-AVALIADORES

Antes de analisar os resultados da avaliação humana, é fundamental validar a confiabilidade dos julgamentos coletados. A Tabela 5 apresenta as métricas de concordância inter-avaliadores para cada dimensão avaliada.

Tabela 5 – Concordância Inter-Avaliadores por Dimensão (15 avaliadores, 72 itens)

Dimensão	Fleiss' κ	Krippendorff's α	ICC(2,1)
Empatia	0,818	0,910	0,928
Consistência de marca	0,760	0,866	0,875
Adequação de tom	0,860	0,938	0,943
Fluência	0,794	0,815	0,876

Os valores de κ entre 0,76 e 0,86 classificam-se como concordância *substancial* a *quase perfeita* segundo a escala de Landis e Koch, enquanto os valores de α entre 0,82 e 0,94 indicam boa confiabilidade segundo Krippendorff. Esses resultados sustentam a robustez dos julgamentos humanos e apoiam sua utilização como referência principal do estudo no cenário avaliado.

5.2 VISÃO GERAL DOS RESULTADOS

A apresentação dos resultados está organizada em quatro etapas complementares. Primeiro, a Seção 5.3 reporta os resultados da avaliação humana, incluindo os rankings por modelo e os testes de significância estatística. Em seguida, a Seção 5.5 apresenta os resultados das quatro métricas automáticas utilizadas como *proxies* computacionais. A Seção 5.6 detalha a análise estratificada pelos fatores experimentais (persona, estratégia de *prompt* e período temporal). Por fim, a Seção 5.8 examina as convergências e divergências entre as avaliações humana e automática, incluindo a identificação de vieses sistemáticos nos *proxies* computacionais.

5.3 AVALIAÇÃO HUMANA

A Tabela 6 apresenta as médias globais da avaliação humana por modelo, agregando as quatro dimensões avaliadas em todos os cenários experimentais.

Tabela 6 – Avaliação Humana Global por Modelo (escala Likert 1–5)

Modelo	Empatia	Consist. Marca	Adeq. Tom	Fluência
<i>GPT-5.5</i>	3,983	3,203	3,975	4,997
<i>Llama 4</i>	4,247	3,103	4,275	4,714
<i>Gemini 3.1</i>	3,444	2,639	3,453	4,792

Os resultados revelam que nenhum modelo domina todas as dimensões simultaneamente. O cenário do *Llama 4* apresenta as maiores pontuações em empatia (4,247) e adequação de tom (4,275), enquanto o *GPT-5.5* apresenta superioridade em fluência (4,997) e consistência de marca (3,203). O *Gemini 3.1* **apresenta as menores pontuações** em três das quatro dimensões avaliadas (empatia, consistência de marca e adequação de tom), sendo segundo em fluência (4,792) — **resultado que, conforme detalhado nas Seções 4.5 e 7.4, deve ser interpretado considerando a coleta via interface web sem controle de parâmetros de *decoding*.** A Figura 10 apresenta esses resultados em formato de mapa de calor para visualização imediata dos perfis de competência de cada modelo.

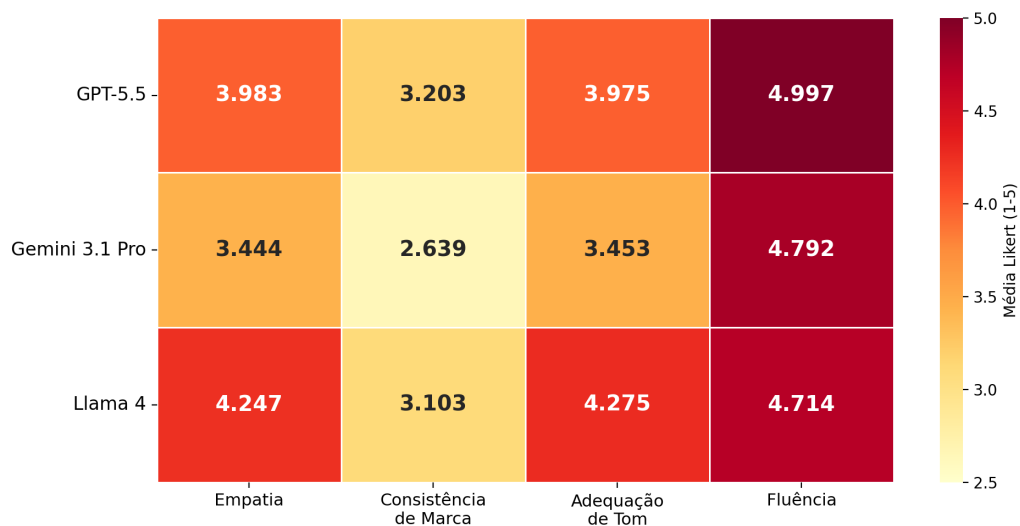


Figura 10 – Mapa de calor das médias Likert por modelo e dimensão na avaliação humana.

A composição do painel de 15 avaliadores, os procedimentos de cegamento, randomização e calibração, bem como o instrumento utilizado para coleta dos julgamentos em escala *Likert* de 1–5 nas quatro dimensões avaliadas (empatia, consistência de marca,

adequação de tom e fluência), encontram-se descritos integralmente na Seção 4.3.1. As métricas de concordância inter-avaliadores que sustentam os achados subsequentes foram apresentadas na Tabela 5.

5.4 ESCOLHA DOS TESTES ESTATÍSTICOS

As variáveis-resposta deste estudo são pontuações em escala *Likert* de 1 a 5 (variável ordinal discreta), provenientes de avaliações humanas e métricas automáticas que produzem distribuições assimétricas e tipicamente afastadas da normalidade. Em virtude da natureza ordinal dos dados e do não atendimento aos pressupostos paramétricos clássicos (normalidade e homogeneidade de variâncias), optou-se por testes não paramétricos baseados em *ranks*. O teste de *Kruskal-Wallis* foi adotado para verificar a existência de efeito global de modelo em cada dimensão, por se tratar da extensão não paramétrica da ANOVA de um fator para mais de dois grupos independentes. Identificada significância, o teste de *Mann-Whitney U* foi aplicado em comparações pareadas, por ser o equivalente não paramétrico do teste *t* para amostras independentes. O nível de significância adotado foi $\alpha = 0,05$, com resultados de $p < 0,01$ considerados robustos e valores próximos ao limiar interpretados com cautela.

Para verificar se as diferenças entre modelos são estatisticamente significativas, aplicou-se o teste de *Kruskal-Wallis* a cada dimensão. A Tabela 7 apresenta os resultados.

Tabela 7 – Teste de Kruskal-Wallis — Efeito de Modelo na Avaliação Humana

Dimensão	<i>H</i>	<i>p</i> -valor	Significativo ($\alpha = 0,05$)
Empatia	6,44	0,040	Sim
Consistência de marca	8,14	0,017	Sim
Adequação de tom	10,66	0,005	Sim ($p < 0,01$)
Fluência	8,46	0,015	Sim

O efeito global de modelo é estatisticamente significativo em todas as quatro dimensões ($p < 0,05$). Para identificar quais pares de modelos diferem entre si, realizaram-se 12 comparações pareadas com o teste de *Mann-Whitney U* (4 dimensões $\times$ 3 pares). Dado o número de testes simultâneos, e considerando o risco de inflação do erro tipo I, os *p*-valores brutos foram corrigidos por dois procedimentos complementares: **Holm-Bonferroni**, que controla a taxa de erro por família (*family-wise error rate*, *FWER*) de forma sequencialmente conservadora (HOLM, 1979), e **Benjamini-Hochberg (BH)**, que controla a taxa de descobertas falsas (*false discovery rate*, *FDR*) e oferece maior poder estatístico em estudos exploratórios com múltiplas comparações (BENJAMINI; HOCHBERG, 1995). A Tabela 8 reporta os *p*-valores brutos e ambos os ajustes.

Tabela 8 – Comparações Pareadas — Teste de Mann-Whitney U com correções para comparações múltiplas (Holm-Bonferroni e Benjamini-Hochberg). Valores em negrito: $p \leq 0,05$ após a correção indicada.

Dimensão	Comparação	p bruto	p (Holm)	p (BH/FDR)
Empatia	<i>GPT-5.5</i> vs. <i>Llama 4</i>	0,391	1,000	0,469
Empatia	<i>GPT-5.5</i> vs. <i>Gemini</i>	0,097	0,388	0,129
Empatia	<i>Llama 4</i> vs. <i>Gemini</i>	0,015	0,135	0,045
Consist. de marca	<i>GPT-5.5</i> vs. <i>Llama 4</i>	0,696	1,000	0,696
Consist. de marca	<i>GPT-5.5</i> vs. <i>Gemini</i>	0,009	0,090	0,036
Consist. de marca	<i>Llama 4</i> vs. <i>Gemini</i>	0,026	0,182	0,052
Adequação de tom	<i>GPT-5.5</i> vs. <i>Llama 4</i>	0,075	0,375	0,113
Adequação de tom	<i>GPT-5.5</i> vs. <i>Gemini</i>	0,046	0,276	0,079
Adequação de tom	<i>Llama 4</i> vs. <i>Gemini</i>	0,003	0,036	0,018
Fluência	<i>GPT-5.5</i> vs. <i>Llama 4</i>	0,004	0,044	0,018
Fluência	<i>GPT-5.5</i> vs. <i>Gemini</i>	0,018	0,144	0,043
Fluência	<i>Llama 4</i> vs. <i>Gemini</i>	0,523	1,000	0,571

Os resultados após correção indicam que duas diferenças permanecem significativas mesmo sob o critério mais conservador (Holm-Bonferroni, $p \leq 0,05$): a superioridade do *Llama 4* sobre o *Gemini* em adequação de tom ($p_{\text{Holm}} = 0,036$) e a superioridade do *GPT-5.5* sobre o *Llama 4* em fluência ($p_{\text{Holm}} = 0,044$). Sob o critério de *FDR* (Benjamini-Hochberg), que é o mais adequado para estudos exploratórios com múltiplas comparações, cinco diferenças permanecem significativas: *Llama 4* > *Gemini* em adequação de tom (0,018) e em empatia (0,045); *GPT-5.5* > *Gemini* em consistência de marca (0,036); e *GPT-5.5* > *Llama 4* e *GPT-5.5* > *Gemini* em fluência (0,018 e 0,043). Os pares *GPT-5.5* vs. *Llama 4* em empatia e em consistência de marca não diferem significativamente em nenhuma das correções, corroborando a interpretação de *trade-off* entre estes dois modelos.

Em síntese, mesmo após a correção para múltiplas comparações, o padrão central dos achados é preservado: (i) **nas condições de teste deste estudo** (cf. Seção 4.5), o *Gemini 3.1* apresenta as menores pontuações nas dimensões pragmáticas (empatia, consistência de marca, adequação de tom), e (ii) o *GPT-5.5* é superior aos demais em fluência. **Cabe ressaltar que o acesso ao *Gemini 3.1* foi realizado via interface web, sem controle sobre os parâmetros de *decoding*, diferentemente dos demais modelos avaliados via API ou execução local; essa assimetria é discutida na Seção 7.4 e deve ser considerada na interpretação relativa dos resultados.** As comparações marginais e não significativas após correção devem ser interpretadas com

cautela, em conformidade com a recomendação de tratar resultados de testes múltiplos como exploratórios quando não atingem significância sob o ajuste apropriado.

A Figura 11 apresenta a distribuição completa dos scores Likert por modelo e dimensão, permitindo a visualização da dispersão, medianas e *outliers*.

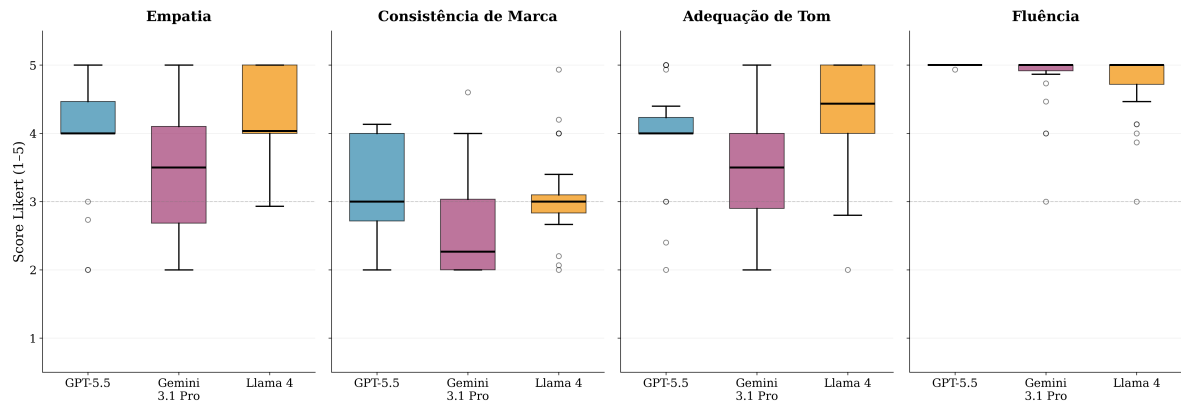


Figura 11 – Distribuição dos scores Likert por modelo e dimensão (média dos 15 avaliadores por item, $n = 72$ itens totais, ≈ 24 por modelo). *Gemini 3.1* apresenta maior dispersão e medianas inferiores em empatia, consistência e tom **nas condições experimentais avaliadas** (interface web, sem controle de parâmetros de *decoding* — cf. Seção 7.4). Fluência apresenta efeito teto para todos os modelos.

5.5 MÉTRICAS AUTOMÁTICAS

A Tabela 9 apresenta os resultados das quatro métricas automáticas utilizadas como *proxies* computacionais para cada dimensão de qualidade.

Tabela 9 – Métricas Automáticas Globais por Modelo

Modelo	Empatia WASSA (1–5 $\uparrow$)	ToneCal TQS (0–1 $\uparrow$)	Fluência PPL ($\downarrow$)	Estilometria (0–1 $\uparrow$)
<i>GPT-5.5</i>	3,360	0,544	118,3	0,061
<i>Llama 4</i>	3,280	0,582	229,4	0,052
<i>Gemini 3.1</i>	3,113	0,513	179,5	0,070

Na dimensão de fluência (perplexidade GPT-2), o *GPT-5.5* apresenta a menor perplexidade média (118,3), indicando maior fluência textual, seguido pelo *Gemini* (179,5) e *Llama 4* (229,4). No ToneCal (adequação de tom), o *Llama 4* lidera com TQS médio de 0,582, seguido pelo *GPT-5.5* (0,544) e *Gemini* (0,513). Na empatia percebida (WASSA), o *GPT-5.5* obtém o maior escore (3,360), seguido pelo *Llama 4* (3,280) e *Gemini* (3,113), nota-se que, diferentemente da avaliação humana, a métrica WASSA posiciona o *GPT-5.5* em primeiro lugar. Na similaridade estilométrica (*Writeprints*), o *Gemini* apresenta o maior valor (0,070), seguido pelo *GPT-5.5* (0,061) e *Llama 4* (0,052). Os valores absolutos de similaridade estilométrica são uniformemente baixos ($< 0,08$), refletindo a diferença

inerente entre texto gerado por *LLMs* e texto humano; a utilidade principal da métrica reside na comparação relativa entre modelos.

Esses resultados corroboram parcialmente a avaliação humana: os rankings de fluência e adequação de tom convergem entre as duas modalidades, enquanto os de empatia e consistência de marca divergem. As divergências são analisadas em detalhe na Seção 5.8.

A Figura 12 apresenta uma comparação visual lado a lado entre avaliação humana e métricas automáticas.

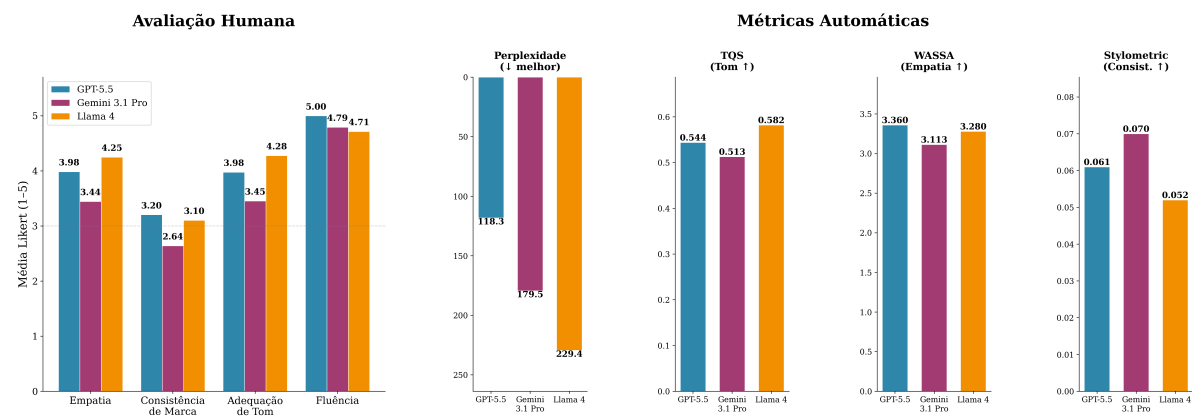


Figura 12 – Comparação entre os três modelos nas avaliações humana (esquerda, escala Likert 1–5) e automática (direita, escalas individuais por métrica).

A Figura 13 sintetiza o perfil multidimensional de cada modelo, integrando as oito métricas (quatro humanas e quatro automáticas) em uma representação normalizada.

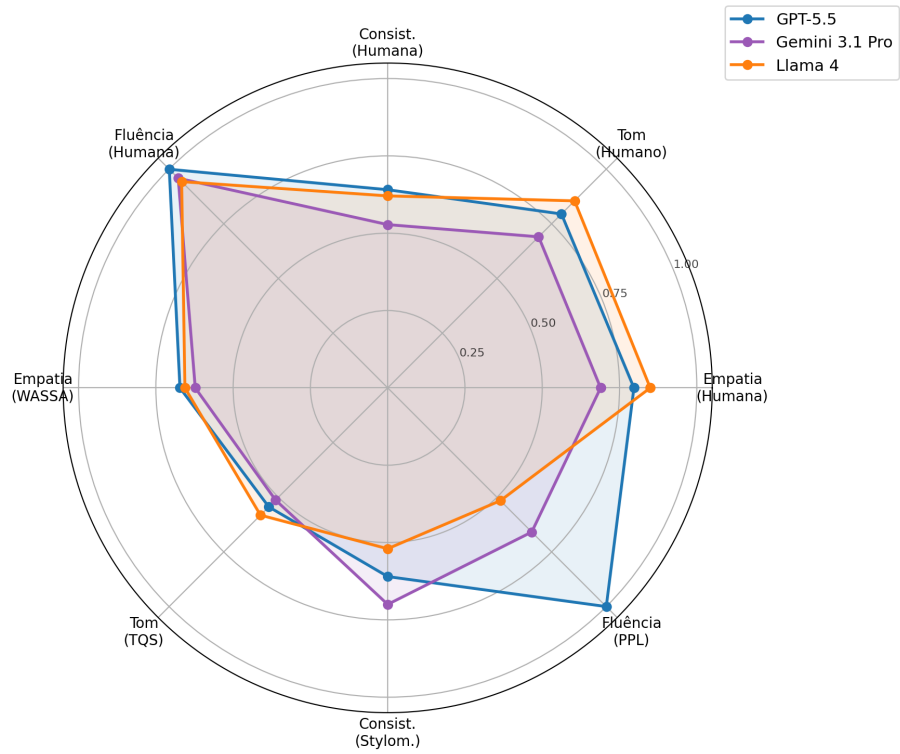


Figura 13 – Perfil multidimensional dos modelos combinando métricas humanas e automáticas (valores normalizados 0–1).

5.6 ANÁLISE ESTRATIFICADA

5.6.1 Interação Modelo × Persona

A Tabela 10 apresenta a avaliação humana estratificada por modelo e persona.

Tabela 10 – Avaliação Humana por Modelo e Persona (escala Likert 1–5)

Modelo	Persona	Empatia	Consist. Marca	Adeq. Tom	Fluência
Llama 4	P2 (informal)	4,306	3,044	4,289	4,511
Llama 4	P1 (formal)	4,189	3,161	4,261	4,917
GPT-5.5	P2 (informal)	4,039	3,378	4,039	5,000
GPT-5.5	P1 (formal)	3,928	3,028	3,911	4,994
Gemini 3.1	P2 (informal)	3,606	2,706	3,667	4,667
Gemini 3.1	P1 (formal)	3,283	2,572	3,239	4,917

Na avaliação humana, a persona informal (P2, acolhedora) recebe pontuações superiores em empatia, consistência de marca e adequação de tom para *todos* os modelos. Em contrapartida, as métricas automáticas (WASSA: P1 = 3,329 vs. P2 = 3,205; ToneCal: P1 = 0,555 vs. P2 = 0,538) favorecem sistematicamente a persona formal (P1, sofisticada).

Essa inversão direcional, discutida na Seção 5.8, constitui evidência de viés compartilhado dos *proxies* automáticos para registro formal. A persona formal é superior apenas em fluência, dimensão na qual avaliação humana e métricas automáticas convergem.

A sensibilidade ao efeito de persona varia entre modelos. O *Gemini* apresenta a maior sensibilidade: a diferença entre P2 e P1 em adequação de tom atinge +0,428 pontos e em empatia +0,322 pontos. No *Llama 4*, essas diferenças são menores (+0,028 em tom e +0,117 em empatia), indicando que seu desempenho é mais estável entre personas.

5.6.2 Interação Modelo $\times$ Estratégia de *Prompt*

A Tabela 11 estratifica os resultados humanos por modelo e estratégia de *prompt*.

Tabela 11 – Avaliação Humana por Modelo e Estratégia de *Prompt* (escala Likert 1–5)

Modelo	Estratégia	Empatia	Consist. Marca	Adeq. Tom	Fluência
<i>Llama 4</i>	<i>zero-shot</i>	4,339	3,128	4,367	4,611
<i>Llama 4</i>	<i>few-shot</i>	4,156	3,078	4,183	4,817
<i>GPT-5.5</i>	<i>zero-shot</i>	4,100	3,128	4,111	4,994
<i>GPT-5.5</i>	<i>few-shot</i>	3,867	3,278	3,839	5,000
<i>Gemini 3.1</i>	<i>zero-shot</i>	3,772	2,656	3,806	4,944
<i>Gemini 3.1</i>	<i>few-shot</i>	3,117	2,622	3,100	4,639

Na avaliação humana, a estratégia *zero-shot* tende a superar a *few-shot* em empatia ($\Delta = +0,357$, $p = 0,074$, marginalmente significativo) e adequação de tom ($\Delta = +0,387$) para *todos* os modelos, com magnitudes que variam conforme o modelo. Embora o efeito global de estratégia não atinja significância estatística ao nível convencional de 5%, a consistência direcional em todos os modelos reforça a tendência observada. O efeito é particularmente acentuado no *Gemini*, onde o *zero-shot* supera o *few-shot* em +0,655 pontos de empatia e +0,706 pontos de adequação de tom. Para o *GPT-5.5*, as diferenças são de +0,233 (empatia) e +0,272 (tom); para o *Llama 4*, de +0,183 em ambas as dimensões.

Esse padrão sugere que exemplos *few-shot* podem restringir a expressividade dos modelos, gerando respostas mais templáticas que os avaliadores humanos percebem como menos empáticas e menos adequadas ao tom. A estratégia *few-shot* apresenta vantagem apenas em consistência de marca para o *GPT-5.5* (+0,150) e em fluência para o *Llama 4* (+0,206). As métricas automáticas, por outro lado, não detectam a superioridade do *zero-shot*: a diferença em empatia WASSA entre as duas estratégias é de apenas $\Delta = 0,003$ (*few-shot* = 3,253; *zero-shot* = 3,250), sugerindo que os avaliadores humanos são sensíveis a nuances de expressividade que os *proxies* computacionais não capturam.

5.6.3 Interação Modelo $\times$ Período Temporal

A Tabela 12 apresenta os resultados estratificados por modelo e período temporal.

Tabela 12 – Avaliação Humana por Modelo e Período (escala Likert 1–5)

Modelo	Período	Empatia	Consist. Marca	Adeq. Tom	Fluência
<i>Llama 4</i>	<i>after</i>	4,489	3,200	4,561	4,633
<i>GPT-5.5</i>	<i>after</i>	4,061	3,222	4,156	4,994
<i>Llama 4</i>	<i>before</i>	4,006	3,006	3,989	4,794
<i>GPT-5.5</i>	<i>before</i>	3,906	3,183	3,794	5,000
<i>Gemini 3.1</i>	<i>after</i>	3,589	2,450	3,600	4,978
<i>Gemini 3.1</i>	<i>before</i>	3,300	2,828	3,306	4,606

O período *after* (após a popularização das LLMs) está associado a maiores pontuações de empatia ($\Delta = +0,309$, $p = 0,029$) e adequação de tom ($\Delta = +0,409$, $p = 0,019$) na avaliação humana. O *Llama 4* apresenta o maior benefício do período *after* ($\Delta = +0,483$ em empatia e $+0,572$ em tom), enquanto o *GPT-5.5* é o modelo menos sensível ao período ($\Delta = +0,156$ em empatia). O *Gemini* apresenta uma anomalia na dimensão de consistência de marca: obtém resultado superior no período *before* (2,828) em comparação com o *after* (2,450), contrariando a tendência geral dos demais modelos.

Cabe notar que os comentários-base diferem entre os dois períodos (100 distintos em cada), de modo que os resultados desta subseção representam tendências descritivas associadas ao período, e não efeitos causais isolados.

5.6.4 Melhores e Piores Cenários

A Tabela 13 apresenta os cinco cenários com maiores e menores médias de empatia na avaliação humana, permitindo identificar as combinações experimentais ótimas e subótimas.

Tabela 13 – Top 5 e Bottom 5 Cenários por Empatia na Avaliação Humana

Rank	Cenário	Empatia	Adeq. Tom	Consist. Marca	Fluência
<i>Top 5</i>					
1	Llama 4 — P1, <i>few-shot</i> , <i>after</i>	4,644	4,622	3,733	5,000
2	Llama 4 — P2, <i>few-shot</i> , <i>after</i>	4,600	4,822	3,333	4,711
3	Gemini — P2, <i>zero-shot</i> , <i>before</i>	4,600	4,667	3,222	5,000
4	GPT-5.5 — P1, <i>few-shot</i> , <i>after</i>	4,489	4,400	3,244	5,000
5	GPT-5.5 — P2, <i>zero-shot</i> , <i>after</i>	4,422	4,333	3,022	5,000
<i>Bottom 5</i>					
20	Gemini — P2, <i>few-shot</i> , <i>after</i>	3,333	3,333	2,711	5,000
21	GPT-5.5 — P1, <i>few-shot</i> , <i>before</i>	3,267	3,133	2,533	5,000
22	Gemini — P1, <i>few-shot</i> , <i>after</i>	3,111	3,044	2,067	4,911
23	Gemini — P2, <i>few-shot</i> , <i>before</i>	2,667	2,689	2,400	3,667
24	Gemini — P1, <i>zero-shot</i> , <i>before</i>	2,578	2,533	2,378	4,778

Os melhores cenários concentram-se no período *after* (4 dos 5 primeiros), com *Llama 4* e *GPT-5.5* ocupando a maioria das posições de destaque. Os piores cenários são dominados pelo *Gemini*, que ocupa 4 das 5 últimas posições, predominantemente com a estratégia *few-shot*. Notavelmente, o *Gemini* com persona informal e *zero-shot* no período *before* atinge empatia de 4,600 (2º lugar), **demonstrando que, sob condições específicas, o modelo alcança resultados competitivos com os demais. Esse achado reforça a hipótese de que parte do desempenho desfavorável observado em outros cenários pode estar relacionada à coleta via interface web, sem o controle dos parâmetros de *decoding* disponível para o *GPT-5.5* (API oficial) e para o *Llama 4* (execução local), conforme detalhado na Seção 7.4.**

5.7 ANÁLISE QUALITATIVA: EXEMPLOS ILUSTRATIVOS

Para complementar a análise quantitativa apresentada nas seções anteriores e oferecer uma compreensão concreta dos padrões observados, esta seção apresenta exemplos representativos extraídos do *Corpus Resultante*. Os trechos foram selecionados de forma a ilustrar diferentes cenários identificados na análise estratificada, com ênfase na justaposição entre comentário original e resposta gerada.

Exemplo 1: Resposta bem avaliada (alta empatia e adequação ao contexto)

Comentário do cliente:

“Tenho carinho pelo restaurante, mas nesta visita saí com sensação mista. Fomos no comemoração de namoro e pedimos camarão Iracema. O prato

estava bom, embora eu esperasse mais pelo valor cobrado. A climatização boa ajuda bastante, mas o prato veio morno e isso comprometeu a experiência.”

Resposta gerada (*Llama 4*, persona formal, *zero-shot*):

“Muito obrigado por compartilhar sua experiência conosco. Ficamos felizes que parte da experiência foi positiva, mas lamentamos que o prato tenha chegado morno e comprometido o momento da comemoração. Estamos sempre trabalhando para melhorar e esperamos recebê-lo novamente em uma próxima oportunidade.”

Análise: a resposta reconhece explicitamente os dois polos do comentário (positivo e negativo), referencia o aspecto específico mencionado (prato morno), contextualiza o evento (comemoração) e encerra com uma expectativa concreta de retorno. Esse padrão de *acknowledgement balanceado* é característico das respostas com pontuações humanas elevadas em empatia e adequação de tom.

Exemplo 2: Resposta fluente, mas pouco empática

Comentário do cliente:

“Tivemos uma ótima experiência no aniversário. Pedimos bolinho de bacalhau e feijão baião, ambos bem servidos e saborosos. A decoração sofisticada fez diferença e o atendimento do Lucas foi atencioso do início ao fim.”

Resposta gerada (*Gemini 3.1*, persona formal, *zero-shot*):

“É com satisfação que recebemos seu elogio ao nosso ambiente e ao serviço prestado. Nosso compromisso é oferecer uma atmosfera acolhedora aliada a um padrão gastronômico de elite, garantindo experiências únicas a cada visita.”

Análise: embora a resposta seja gramaticalmente correta e fluente, ela ignora os elementos específicos mencionados pelo cliente: o aniversário, o nome do atendente (Lucas) e os pratos referenciados. A construção é genérica e poderia ser aplicada a qualquer comentário positivo, configurando o padrão *templático* associado a baixas pontuações humanas de empatia, ainda que com fluência elevada nas métricas automáticas.

Exemplo 3: Resposta excessivamente formal

Comentário do cliente:

“Ambiente agradável e experiência muito positiva no happy hour. Pedimos lagosta grelhada e arroz dos mares, ambos bem servidos e saborosos. A música ao vivo agradável fez diferença e o atendimento do Wellington foi atencioso do início ao fim.”

Resposta gerada (*GPT-5.5*, persona formal, *few-shot*):

“Ficamos honrados com seu registro e com a generosidade de sua avaliação. É especialmente gratificante saber que sua experiência no happy hour foi marcada de forma tão positiva. Será uma satisfação compartilhar seu reconhecimento ao atendimento de Wellington com a equipe.”

Análise: a resposta utiliza um registro elevado, “honrados”, “generosidade de sua avaliação”, “especialmente gratificante” que, em um contexto de *review* público de restaurante, foi consistentemente percebido pelos avaliadores humanos como rebuscado e distante. Esse tipo de construção é favorecido pelas métricas automáticas (perplexidade e *ToneCal*) por sua aderência ao registro formal, mas penalizado na avaliação humana por reduzir a sensação de proximidade e acolhimento.

Exemplo 4: Resposta inconsistente com o contexto do comentário

Comentário do cliente:

“A proposta do restaurante é boa, mas o serviço não acompanhou. Fomos no comemoração de namoro e enfrentamos música muito alta logo no início. Quando o moqueca de camarão chegou, veio com aspecto abaixo do esperado e o prato estava morno. Houve ainda tempo de espera acima do esperado, o que comprometeu a noite.”

Resposta gerada (*GPT-5.5*, persona formal, *zero-shot*):

“É uma alegria saber que sua experiência foi tão positiva. Nossa proposta gastronômica é cuidadosamente pensada para proporcionar uma experiência marcante. Compreendemos que o tempo de espera impacta diretamente a experiência e levamos essa observação em consideração.”

Análise: este caso ilustra uma falha de leitura do sentimento do comentário: o cliente relata uma experiência majoritariamente negativa (música alta, prato morno, demora), mas a resposta abre com uma afirmação de comemoração (“É uma alegria saber que sua experiência foi tão positiva”), perceptivelmente desalinhada com o conteúdo recebido. Apesar de gramaticalmente correta e estilisticamente coerente com a persona formal, a resposta evidencia o que Liang et al. (2023) caracterizam como *descolamento pragmático*: aderência à forma sem aderência ao significado.

Exemplo 5: Comparação Zero-Shot vs. Few-Shot para o mesmo comentário

Comentário do cliente:

“Tenho carinho pelo restaurante, mas nesta visita saí com sensação mista. [...] o prato veio morno e isso comprometeu a experiência.”

Resposta *Zero-Shot* (*Llama 4*, persona formal):

“Muito obrigado por compartilhar sua experiência conosco. Ficamos felizes que parte da experiência foi positiva, mas lamentamos que o prato tenha chegado morno e comprometido o momento da comemoração. Estamos sempre trabalhando para melhorar e esperamos recebê-lo novamente.”

Resposta *Few-Shot* (Llama 4, persona formal):

“Prezado(a) cliente, agradecemos por compartilhar sua experiência. Ficamos satisfeitos que parte da experiência tenha sido agradável, e lamentamos pelo prato veio morno. Esperamos recebê-lo(a) novamente.”

Análise: a comparação direta entre as duas estratégias para o mesmo comentário ilustra o padrão quantitativo identificado na Seção 5.6: a resposta *zero-shot* é mais extensa, contextualiza o evento (“comemoração”), inclui um compromisso (“estamos sempre trabalhando para melhorar”) e evita fórmulas de tratamento rígidas. A resposta *few-shot*, em contraste, adota a estrutura dos exemplos demonstrativos, abertura com “Prezado(a) cliente”, construções mais formulaicas e encerramento curto, evidenciando como exemplos podem reduzir a expressividade do modelo. Este padrão, observado de forma consistente nos três modelos avaliados, fundamenta a recomendação operacional pela abordagem *zero-shot* com persona bem definida (Seção 6.3).

5.8 CONVERGÊNCIAS E DIVERGÊNCIAS ENTRE AVALIAÇÃO HUMANA E AUTOMÁTICA

A triangulação entre avaliação humana e métricas automáticas constitui um dos achados metodológicos centrais deste estudo. A Tabela 14 sintetiza a comparação dos rankings por dimensão, enquanto a Figura 14 apresenta a dispersão entre as duas modalidades de avaliação, estratificada por modelo e persona.

Tabela 14 – Convergência de Rankings: Avaliação Humana vs. Métricas Automáticas

Dimensão	Ranking Humano	Ranking Automático	Convergência
Adequação de tom	Llama 4 > GPT-5.5 > Gemini	Llama 4 > GPT-5.5 > Gemini (TQS)	Forte
Fluência	GPT-5.5 > Gemini > Llama 4	GPT-5.5 > Gemini > Llama 4 (PPL)	Forte
Empatia	Llama 4 > GPT-5.5 > Gemini	GPT-5.5 > Llama 4 > Gemini (WASSA)	Parcial
Consist. de marca	GPT-5.5 > Llama 4 > Gemini	Gemini > GPT-5.5 > Llama 4 (Estil.)	Divergente

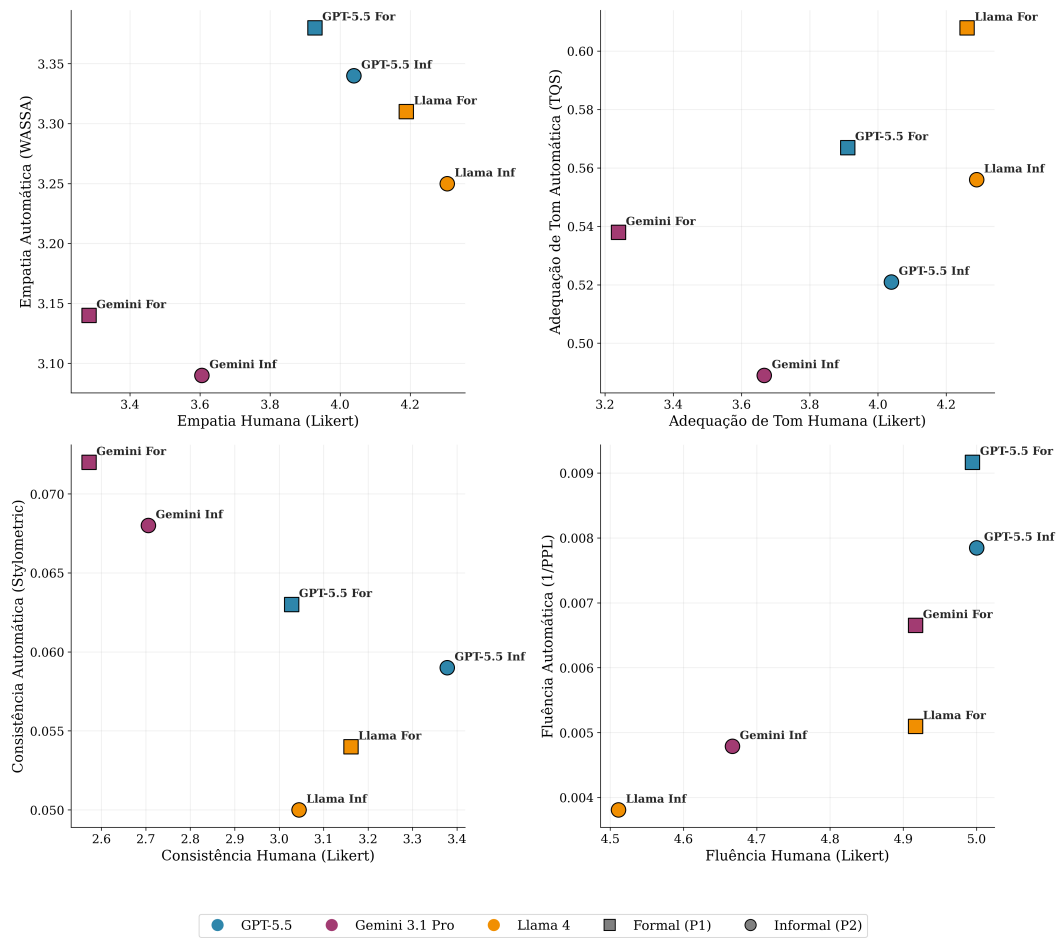


Figura 14 – Dispersão entre avaliações humana (eixo X) e automática (eixo Y) por modelo e persona. Quadrados representam a persona formal (P1), círculos a informal (P2).

5.8.1 Convergências

Os rankings de *adequação de tom* e *fluência* são idênticos nas duas modalidades de avaliação. Na adequação de tom, tanto a avaliação humana quanto o ToneCal posicionam *Llama 4* > *GPT-5.5* > *Gemini*, o que fortalece a validade dessa métrica como *proxy* confiável. Na fluência, tanto a percepção humana quanto a perplexidade convergem no ranking *GPT-5.5* > *Gemini* > *Llama 4*. Essas convergências sustentam a utilidade do ToneCal e da perplexidade como indicadores automáticos para suas respectivas dimensões.

5.8.2 Divergências

Na dimensão de *empatia*, ocorre uma inversão no topo do ranking: os 15 avaliadores humanos classificam o cenário *Llama 4* como o mais empático (4,247), enquanto a métrica WASSA posiciona o *GPT-5.5* em primeiro lugar (3,360). Essa divergência é atribuível às limitações do modelo WASSA, treinado em inglês para o domínio de notícias, com compressão de escala acentuada ($CV = 4,91\%$) que limita seu poder discriminativo em português brasileiro no domínio de atendimento ao cliente.

A divergência mais expressiva ocorre na *consistência de marca*: a avaliação humana posiciona *GPT-5.5* como mais consistente (3,203), enquanto a similaridade estilométrica favorece o *Gemini* (0,070), uma inversão completa de ranking. A estilometria mede similaridade superficial de estilo (léxico e sintaxe) com base nas 200 respostas originais do restaurante (100 do período *before* e 100 do período *after*), ao passo que os avaliadores humanos avaliam a adequação pragmática e contextual da resposta à identidade da marca.

5.8.3 Viés de Registro Linguístico

A inversão direcional do efeito de persona entre avaliação humana e métricas automáticas constitui um achado transversal relevante. Na avaliação humana, a persona informal (P2) recebe pontuações superiores em empatia, consistência de marca e adequação de tom para todos os modelos (Tabela 10). Nas métricas automáticas, a persona formal (P1) é sistematicamente favorecida: WASSA (P1 = 3,329 vs. P2 = 3,205), ToneCal (P1 = 0,555 vs. P2 = 0,538) e perplexidade (P1 = 128,4 vs. P2 = 223,1). Essa inversão constitui evidência de que métricas automáticas baseadas em modelos de linguagem, treinados predominantemente em texto formal, apresentam viés compartilhado para registro formal, podendo conduzir a conclusões opostas às da percepção humana real.

5.8.4 Natureza Aproximativa das Métricas Automáticas

Os achados de convergência parcial e divergência sistemática entre avaliação humana e métricas automáticas precisam ser interpretados à luz da *natureza proxy* desses indicadores. As métricas *Perplexidade*, *ToneCal*, *Writeprints* e *WASSA*, ainda que cientificamente fundamentadas, constituem *aproximações computacionais* de fenômenos linguísticos e pragmáticos complexos, e não verdades absolutas sobre a qualidade do texto avaliado. Cada métrica captura uma faceta operacionalizável de uma dimensão multidimensional da comunicação humana, definida pelas escolhas de modelagem, pelo corpus de treinamento e pelos pressupostos do método que a originou.

A **perplexidade** mede a probabilidade que um modelo de linguagem específico atribui a uma sequência textual. Como o *GPT-2 PT-BR* foi pré-treinado em corpora majoritariamente compostos por texto formal, notícias jornalísticas, artigos enciclopédicos e literatura editada, suas estimativas de fluência tendem a favorecer respostas com sintaxe canônica, léxico padrão e estruturas oracionais previsíveis. Construções típicas do registro informal brasileiro (contrações, marcas de oralidade, sintaxe paratática) recebem, por essa razão, perplexidades sistematicamente mais altas, ainda que sejam percebidas pelos avaliadores humanos como mais naturais no contexto de atendimento ao cliente em redes sociais.

O **ToneCal** opera sobre marcadores lexicais explícitos de cortesia, deferência e formalidade. A métrica é sensível a saudações formais, fórmulas de tratamento e uso do

modo subjuntivo, elementos predominantemente associados ao registro formal. Por consequência, respostas que mobilizam empatia através de proximidade afetiva, acolhimento informal e linguagem coloquial podem receber pontuações inferiores, mesmo quando a percepção humana as classifica como mais adequadas ao contexto pragmático de uma resposta a um cliente. A métrica captura, portanto, polidez lexical de superfície, não adequação pragmática contextualizada.

O **modelo WASSA** foi treinado em *datasets* anotados majoritariamente em inglês e originados de domínios jornalísticos. Ainda que sua arquitetura baseada em *transformers* multilíngues permita transferência parcial para outras línguas e domínios, a representação interna do conceito de empatia que o modelo aprendeu está ancorada nos padrões lexicais e estruturais do material de treinamento. Aplicado a respostas em português brasileiro no domínio de *review* de restaurantes, o modelo produz estimativas comparativamente úteis para diferenciar modelos e estratégias, mas com valores absolutos cuja calibração não pode ser presumida.

A **similaridade estilométrica** (*Writeprints*) mensura a aproximação entre o perfil estilístico de uma resposta gerada e o de um corpus de referência humano (distribuição lexical, complexidade sintática, *n-gramas* de caracteres, padrões de pontuação). A métrica é, por construção, agnóstica ao conteúdo semântico e à adequação pragmática: uma resposta pode replicar fielmente o estilo da marca sem responder de forma apropriada ao comentário recebido, e o inverso também é possível. A divergência observada nesta dimensão (ranking estilométrico oposto ao ranking humano) ilustra com clareza essa limitação conceitual. A explicação unificadora para o padrão recorrente “*métricas automáticas favorecem o formal, humanos preferem o acolhedor*” é, portanto, estrutural: todas as métricas adotadas dependem, direta ou indiretamente, de modelos de linguagem ou de corpora de referência cuja distribuição amostral pende para o registro formal. Esse *viés compartilhado* não invalida as métricas, elas continuam fornecendo sinais comparativos consistentes e reproduzíveis, mas delimitam o seu campo legítimo de aplicação: *proxies* úteis para comparações relativas e triagem em larga escala, jamais substitutos da avaliação humana em decisões finais sobre qualidade comunicacional. Daí a centralidade metodológica da triangulação adotada neste estudo: a integração sistemática entre as duas modalidades é o que permite distinguir achados robustos (indicados por ambas) de artefatos da medição automática (presentes apenas em uma).

6 CONCLUSÕES

6.1 SÍNTESE DOS RESULTADOS

O achado central desta investigação, nos limites do corpus, dos modelos e das configurações avaliadas, é que *nenhum dos modelos analisados foi universalmente superior*

em todas as dimensões consideradas. A análise de 2.400 respostas geradas, com avaliação automática integral e avaliação humana em amostra estratificada de 72 itens por 15 avaliadores independentes e quatro métricas automáticas especializadas, revela um cenário de *trade-offs* inerentes entre as dimensões de qualidade. O cenário do *Llama 4* apresenta as maiores pontuações na percepção humana de empatia (4,247/5) e adequação de tom (4,275/5), enquanto o *GPT-5.5* é superior em fluência (4,997/5) e consistência de marca (3,203/5). O *Gemini 3.1* apresenta as menores pontuações em três das quatro dimensões humanas, com diferenças estatisticamente significativas indicadas pelo teste de *Kruskal-Wallis* ($p < 0,05$ em todas as dimensões). **Esse resultado, contudo, deve ser interpretado à luz da assimetria de acesso aos modelos: o *Gemini 3.1* foi consultado via interface web, sem controle dos parâmetros de geração, enquanto o *GPT-5.5* e o *Llama 4* operaram com temperatura e *top-p* explicitamente configurados (cf. Seções 4.5 e 7.4).** Não é possível dissociar, neste estudo, quanto da diferença observada decorre da capacidade intrínseca do modelo e quanto reflete a configuração padrão da interface. Esse resultado corrobora a perspectiva de avaliação holística de Liang et al. (2023), segundo a qual modelos de linguagem exibem perfis de competência heterogêneos que somente avaliações multidimensionais conseguem capturar adequadamente.

A estratégia *zero-shot*, apoiada em descrições robustas de persona, tende a superar a *few-shot* na percepção humana de empatia ($\Delta = +0,357$, $p = 0,074$) e adequação de tom ($\Delta = +0,387$) para todos os modelos testados, com efeito especialmente pronunciado no *Gemini* ($\Delta > +0,65$) — **comportamento possivelmente amplificado pela configuração padrão da interface web, que não permite ajuste dos parâmetros de *decoding* (cf. Seção 7.4).** Embora o efeito global não atinja significância estatística convencional ($\alpha = 0,05$), a consistência direcional entre os três modelos sugere uma tendência substantiva. Exemplos *few-shot* parecem restringir a expressividade dos modelos, gerando respostas percebidas como mais templáticas e menos empáticas. Esse achado contraria a expectativa usual de que exemplos demonstrativos melhoram o desempenho, sugerindo que, para tarefas que requerem expressividade e engajamento emocional, a instrução descritiva pode ser mais eficaz do que a demonstração por exemplos.

6.2 CONTRIBUIÇÕES DO TRABALHO

Em sintonia com a organização apresentada na Introdução, as contribuições deste trabalho consolidam-se em três eixos:

(i) Proposta de um protocolo de avaliação multidimensional. A principal contribuição metodológica deste estudo é a proposição de um *framework* de avaliação multidimensional que integra avaliação humana e métricas automáticas para a geração de texto

em português brasileiro, com indícios empíricos preliminares de sua utilidade no cenário avaliado. Esse *framework* combina quatro dimensões de avaliação humana (empatia, consistência de marca, adequação de tom e fluência) com quatro *proxies* computacionais (WASSA, ToneCal, perplexidade GPT-2 e similaridade estilométrica), permitindo a triangulação sistemática entre percepção humana e medição automática.

A evidência empírica de que essa triangulação contribui substantivamente para a robustez das conclusões, e não apenas a complementa marginalmente, é um dos achados centrais do trabalho no cenário avaliado. As divergências sistemáticas entre avaliação humana e métricas automáticas seguem um padrão previsível: métricas treinadas predominantemente em texto formal favorecem sistematicamente o registro formal. A inversão direcional do efeito de persona (informal preferido por humanos; formal favorecido por métricas automáticas) é a evidência mais robusta desse viés compartilhado. Um estudo baseado exclusivamente em métricas automáticas teria chegado a conclusões parcialmente opostas às da percepção humana real, conforme alertado por Howcroft e Rieser (2021) ao documentar os riscos de avaliações subdimensionadas em processamento de linguagem natural.

A concordância inter-avaliadores obtida ($\kappa = 0,76\text{--}0,86$; $\alpha = 0,82\text{--}0,94$; ICC = $0,88\text{--}0,94$) atesta que a avaliação humana com painel diversificado é viável e reproduzível, reforçando seu papel como referência para a calibração de *proxies* automáticos.

(ii) Comparação sistemática entre modelos, personas e estratégias. O cruzamento fatorial completo ($3 \text{ modelos} \times 2 \text{ personas} \times 2 \text{ estratégias} \times 2 \text{ períodos}$) permitiu identificar *trade-offs* relevantes entre os modelos avaliados: o *Llama 4 Maverick* destaca-se em empatia e adequação de tom percebidas, enquanto o *GPT-5.5* prevalece em fluência e consistência de marca, com *Gemini 3.1* apresentando as menores pontuações nas três dimensões pragmáticas após correção para múltiplas comparações (cf. Tabela 8) — **resultado que deve ser lido em conjunto com a limitação metodológica de acesso ao modelo via interface web (cf. Seção 7.4)**. A estratégia *zero-shot* apoiada em descrição robusta de persona mostrou-se tendencialmente superior à *few-shot* nas dimensões pragmáticas para todos os modelos, sugerindo que exemplos demonstrativos podem restringir a expressividade em tarefas que demandam engajamento emocional.

(iii) Análise da convergência e divergência entre avaliação humana e métricas automáticas. A triangulação sistemática revelou convergência forte entre as duas modalidades nas dimensões de fluência e adequação de tom (rankings idênticos), convergência parcial em empatia e divergência completa em consistência de marca. A inversão direcional do efeito de persona, informal preferido por humanos *vs.* formal favorecido pelos *proxies* automáticos, evidencia um viés compartilhado das métricas automáticas para o

registro formal, com implicações diretas para o uso desses indicadores como ferramentas isoladas de tomada de decisão sobre qualidade comunicacional.

6.3 IMPLICAÇÕES PRÁTICAS

Para aplicações comerciais no setor gastronômico, os resultados oferecem orientações concretas. A configuração *zero-shot* com persona bem definida é indicada como *default* operacional, dada a tendência consistente de superioridade em empatia e adequação de tom na avaliação humana. O *GPT-5.5* oferece o menor risco e maior previsibilidade (menor variância entre cenários), sendo indicado quando fluência e profissionalismo são prioritários. O *Llama 4* destaca-se como a opção mais indicada para aplicações em que a empatia percebida é o objetivo central.

A persona informal (acolhedora) é preferida pelos avaliadores humanos em três de quatro dimensões, indicando que, no contexto de SAC de restaurante, o tom acolhedor é mais valorizado do que o formal. Esse achado alinha-se com a evidência de Rohden e Espartel (2024), que sugere o impacto positivo do tom empático de *chatbots* na satisfação e no *boca a boca* de consumidores.

Do ponto de vista ético, a implantação de respostas automatizadas em larga escala deve considerar os riscos de representação, viés linguístico e geração de conteúdo potencialmente nocivo documentados na literatura (WEIDINGER et al., 2021; GEHMAN et al., 2020; HARTVIGSEN et al., 2022), particularmente o viés para registro formal identificado neste estudo, que pode desfavorecer interações com públicos habituados ao registro informal. Os procedimentos de anonimização e conformidade com a *LGPD* (Lei n. 13.709/2018) adotados nesta pesquisa são detalhados na Seção 4.2.7.

6.4 TRABALHOS FUTUROS

A continuidade desta investigação aponta para quatro direções complementares.

(i) Avaliação ampliada de modelos. A inclusão de modelos não avaliados neste estudo — em particular as famílias *Claude (Anthropic)* e *Mistral Large* — sob condições experimentais padronizadas (acesso via *API* com parâmetros explicitamente controlados) permitiria um panorama comparativo mais abrangente do ecossistema atual de *LLMs* para o português brasileiro. Igualmente, reavaliar o *Gemini 3.1* sob condições simétricas às dos demais modelos seria desejável, dado que os achados atuais estão sujeitos à limitação de acesso via interface discutida na Seção 7.4.

(ii) Expansão do corpus e diversidade de domínios. A replicação do *framework* em múltiplos estabelecimentos, redes gastronômicas distintas e outros setores de serviço

(hospitalidade, varejo, suporte técnico) testaria a generalização dos achados e refinaria o entendimento sobre o impacto da persona em diferentes contextos comerciais.

(iii) Calibração de métricas automáticas para o português informal. O viés sistemático dos *proxies* automáticos para o registro formal motiva o desenvolvimento ou adaptação de métricas treinadas em corpora representativos do registro informal brasileiro (redes sociais, atendimento ao cliente, conversação espontânea).

(iv) Análise estatística mais robusta. A adoção de modelos de efeitos mistos, que respeitam a estrutura hierárquica dos dados (item dentro de cenário dentro de modelo), substituiria os testes não paramétricos isolados utilizados neste estudo, potencialmente aumentando o poder estatístico para detectar efeitos marginais.

Um título potencial para a próxima fase desta pesquisa é: “*Calibração de Métricas Automáticas para Avaliação de Geração Textual em Português Brasileiro Informal: Um Estudo Multidomínio.*”

7 LIMITAÇÕES DO ESTUDO

A despeito do rigor metodológico adotado, este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados e que apontam caminhos para investigações futuras.

7.1 CORPUS, AMOSTRA E AVALIAÇÃO HUMANA

O *Corpus de Origem* é composto por 200 comentários reais provenientes de um único estabelecimento (Coco Bambu Manaus), balanceados entre os períodos pré e pós-popularização das *LLMs*. Embora tenha gerado as 2.400 respostas que sustentam a análise, o tamanho amostral e a coleta concentrada em um restaurante específico restringem a generalização dos achados para outras unidades, redes e contextos regionais, dado que traços contextuais (tipo de culinária, perfil de clientela, padrão de queixas e elogios) podem não se replicar em outros segmentos.

A avaliação humana foi conduzida sobre uma amostra estratificada (72 itens, 15 avaliadores cobrindo os 24 cenários), com concordância robusta ($\kappa = 0,76\text{--}0,86$; $\alpha = 0,82\text{--}0,94$; ICC = 0,88–0,94). A não-avaliação humana integral do *Corpus Resultante*, contudo, impede afirmações pontuais sobre respostas individuais, restringindo as conclusões ao nível agregado dos cenários. Ademais, avaliações subjetivas em escala *Likert* estão sujeitas a vieses individuais de preferência estilística e expectativa comunicacional, mitigados — mas não eliminados — pela diversificação do painel.

7.2 LIMITAÇÕES DAS MÉTRICAS AUTOMÁTICAS

As métricas automáticas (Perplexidade, *ToneCal*, *Writeprints* e WASSA) constituem *proxies* computacionais que aproximam dimensões complexas da qualidade textual, mas não substituem a avaliação humana. A Perplexidade reflete a probabilidade do texto sob um modelo de língua específico (*GPT-2 PT-BR*), podendo favorecer padrões linguísticos próximos aos seus dados de pré-treinamento. O *ToneCal* opera sobre marcadores lexicais de superfície, com sensibilidade maior a indicadores explícitos de polidez do que a nuances pragmáticas. O *Writeprints* captura padrões estilométricos agregados sem mensurar diretamente conteúdo semântico. O modelo WASSA, originalmente treinado em inglês jornalístico, é empregado como indicador comparativo entre modelos, mais informativo para diferenças relativas do que para valores absolutos. A triangulação com avaliação humana adotada neste estudo é justamente a resposta metodológica a essas limitações.

7.3 ANÁLISE ESTATÍSTICA E ESTRUTURA DE DEPENDÊNCIA

A análise inferencial apoiou-se em testes não paramétricos (*Kruskal-Wallis* e *Mann-Whitney U*), adequados à natureza ordinal das pontuações *Likert*, com correções por *Holm-Bonferroni* e *Benjamini-Hochberg (FDR)* para múltiplas comparações. Reconhecem-se três refinamentos para trabalhos futuros: (i) reportar tamanhos de efeito padronizados (e.g., ϵ^2 , r de *rank-biserial*) e intervalos de confiança, ampliando a interpretação prática das magnitudes; (ii) aplicar correções estratificadas por subgrupo em estudos com amostras maiores; e (iii) adotar modelos de efeitos mistos ou modelos cumulativos para variáveis ordinais, com efeitos aleatórios para avaliador, comentário e modelo, que respeitem a estrutura hierárquica dos dados e permitiriam decompor a variância atribuível a cada fonte.

7.4 CONDIÇÕES EXPERIMENTAIS E CONFIGURAÇÃO DOS MODELOS

As conclusões deste estudo refletem o desenho específico adotado, com personas, instruções e exemplos calibrados para o domínio gastronômico — a transferência para outros domínios de atendimento ao cliente (varejo, hotelaria, saúde, serviços financeiros) requer validação empírica adicional, dado que cada contexto impõe expectativas próprias de tom, formalidade e densidade informacional.

Quanto à configuração de geração, os experimentos adotaram temperatura $T = 0,7$ e *top-p* = 0,9 sempre que esses parâmetros eram acessíveis ao usuário. **Para o *Gemini 3.1*, em razão de a *API* programática estável não estar disponível no período de execução, a coleta foi realizada via interface web oficial, sem controle sobre os parâmetros de *decoding*.** Essa assimetria de acesso constitui uma limitação assumida do estudo e foi explicitamente sinalizada nas seções de Resultados e Conclusões: parte do desempenho desfavorável do *Gemini* em relação aos demais modelos pode estar

associada a essa configuração padrão. Adicionalmente, os modelos avaliados são produtos em evolução contínua; atualizações de versão podem introduzir mudanças comportamentais não controláveis externamente, de modo que a comparação reportada é uma fotografia válida para a janela temporal em que os experimentos foram conduzidos.

Por fim, embora a separação temporal do *Corpus de Origem* (pré e pós 30 de novembro de 2022) tenha sido estabelecida para mitigar a contaminação por respostas geradas por IA, não é possível assegurar, para o subconjunto pós-popularização, que todas as respostas humanas tenham sido elaboradas sem qualquer assistência de ferramentas generativas. Essa possibilidade afeta especialmente o perfil estilométrico de referência da métrica *Writeprints*, e pode ser tratada em estudos futuros restringindo a referência ao período pré-2022 ou adotando procedimentos automatizados de detecção de texto gerado por IA.

REFERÊNCIAS

- ABBASI, A.; CHEN, H. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, v. 26, n. 2, p. 7, 2008.
- ACHIAM, J. et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- ALI, L. et al. Beyond the hype: Evaluating the impact of generative AI on brand authenticity, image, and consumer behavior in the restaurant industry. *International Journal of Hospitality Management*, Elsevier, v. 131, p. 104318, 2025.
- ALI, M.; KIM, W. S.; RYU, K. H. Effect of online reviews and crowd cues on restaurant choice of customer: Moderating role of gender and perceived crowding. *Frontiers in Psychology*, v. 12, p. 763519, 2021.
- ALMEIDA, T. S. et al. Sabiá-2: A new generation of Portuguese large language models. *arXiv preprint arXiv:2403.09887*, 2024. Disponível em: <<https://arxiv.org/abs/2403.09887>>.
- AMIDEI, J.; PIWEK, P.; WILLIS, A. The use of rating and likert scales in natural language generation human evaluation tasks: A review and some recommendations. In: *Proceedings of The 12th International Conference on Natural Language Generation*. [S.l.: s.n.], 2019. p. 397–402.
- ARAÚJO, G.; MELO, T. de; FIGUEIREDO, C. Explorando a eficácia dos modelos generativos em tarefas de análise de sentimentos no português brasileiro. *Linguamatica*, 2024.
- AURELIANO-SILVA, L.; LEUNG, X.; SPERS, E. E. The effect of online reviews on restaurant visit intentions: applying signaling and involvement theories. *Journal of Hospitality and Tourism Technology*, 2021.
- BAEK, T. H.; KIM, H. J.; KIM, J. AI-generated recommendations: Roles of language style, perceived AI human-likeness, and recommendation agent. *International Journal of Hospitality Management*, Elsevier, v. 126, p. 104106, 2025.
- BALMER, J. M. Strategic corporate brand alignment. *European Journal of Marketing*, v. 46, n. 7/8, p. 1064–1092, 2012.
- BALSHETWAR, S. V.; TUGANAYAT, R. M.; REGULWAR, G. B. Frame tone and sentiment analysis. *International Journal of Computer Sciences and Engineering*, v. 7, n. 6, 2019.
- BARRIERE, V. et al. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. [S.l.: s.n.], 2023. p. 511–525.
- BARRIERE, V. et al. Wassa 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In: *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. [S.l.: s.n.], 2022. p. 214–227.

- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 57, n. 1, p. 289–300, 1995.
- BROWN, T. et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, v. 33, p. 1877–1901, 2020.
- BUECHEL, S. et al. Modeling empathy and distress in reaction to news stories. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2019.
- CAO, H. et al. *TOOL-ED: Enhancing Empathetic Response Generation with the Tool Calling Capability of LLM*. 2024. Disponível em: <<https://arxiv.org/abs/2412.03096>>.
- CARMO, D. et al. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *arXiv preprint arXiv:2008.09144*, 2020. Disponível em: <<https://arxiv.org/abs/2008.09144>>.
- CELIKILMAZ, A.; CLARK, E.; GAO, J. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2021.
- CHOI, M. et al. ChatGPT or online review: Which is a better determinant of customers' trust in Airbnb listings and stay intention? *International Journal of Hospitality Management*, Elsevier, v. 122, p. 103852, 2024.
- CHOWDHARY, K. R. Natural language processing. In: *Fundamentals of artificial intelligence*. [S.l.]: Springer, 2020. p. 603–649.
- CRESWELL, J. W.; CLARK, V. L. P. *Designing and Conducting Mixed Methods Research*. 3. ed. [S.l.]: SAGE Publications, 2018.
- CROLIC, I. A. C. et al. Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, v. 87, n. 3, p. 478–498, 2023.
- DANESCU-NICULESCU-MIZIL, C. et al. A computational approach to politeness with application to social factors. In: *Proceedings of the 51st Annual Meeting of the ACL*. [S.l.: s.n.], 2013. p. 250–259.
- DANG, H. et al. How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models. In: *CHI'22 Workshops: GenAICHI*. [S.l.]: ACM, 2022.
- DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- DIAO, S. et al. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*, 2024.
- DUBEY, A. et al. The llama 3 herd of models. *arXiv e-prints*, p. arXiv–2407, 2024.
- DWIVEDI, Y. K. et al. Leveraging chatgpt and other generative ai-based applications in the hospitality and tourism industry. *International Journal of Contemporary Hospitality Management*, v. 36, n. 1, p. 1–12, 2024.
- FABBRI, A. R. et al. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, v. 9, p. 391–409, 2021.
- FILHO, J. A. W. et al. The brWaC corpus: A new open resource for Brazilian Portuguese. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*. [S.l.: s.n.], 2018. p. 4339–4344.

- FU, Y.; FENG, Y. Context-aware response generation with constructive triple learning. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2020. p. 1–11.
- GAMBOA-CRUZADO, J. et al. Use of chatbots in e-commerce: A comprehensive systematic review. *Journal of Theoretical and Applied Information Technology*, v. 101, p. 1172–1183, 2023.
- GAO, T. et al. *Benchmarking and Learning Real-World Customer Service Dialogue*. 2025. Disponível em: <<https://arxiv.org/abs/2510.22143>>.
- GEHMAN, S. et al. Realtotoxicityprompts: Evaluating neural toxic degeneration in language models. In: *Findings of EMNLP*. [S.l.: s.n.], 2020.
- GIORGI, S. et al. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In: *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. [S.l.: s.n.], 2024. p. 369–379.
- Google. *Gemini 3.1 Pro*. 2025. <<https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/>>. Anúncio oficial. Acesso em: 15 jun. 2026.
- HALPERN, J. What is clinical empathy? *Journal of General Internal Medicine*, v. 18, n. 8, p. 670–674, 2003.
- HARTVIGSEN, T. et al. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: *Proceedings of the 60th Annual Meeting of the ACL*. [S.l.: s.n.], 2022. p. 3309–3326.
- HIRSCHBERG, J.; MANNING, C. D. Advances in natural language processing. *Science*, v. 349, n. 6245, p. 261–266, 2015.
- HOLM, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, v. 6, n. 2, p. 65–70, 1979.
- HOLTZMAN, A. et al. The curious case of neural text degeneration. In: *International Conference on Learning Representations (ICLR)*. [s.n.], 2020. Disponível em: <<https://arxiv.org/abs/1904.09751>>.
- HONG, M. et al. Dial-In LLM: Human-aligned LLM-in-the-loop intent clustering for customer service dialogues. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Suzhou, China: Association for Computational Linguistics, 2025. p. 5885–5900. Disponível em: <<https://aclanthology.org/2025.emnlp-main.300/>>.
- HOWCROFT, D. M. et al. Twenty years of confusion in human evaluation. In: *Proceedings of the 13th INLG*. [S.l.: s.n.], 2020. p. 169–182.
- HOWCROFT, D. M.; RIESER, V. What happens if you treat ordinal ratings as interval data? human evaluations in nlp are even more under-powered than you think. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2021. p. 8932–8939.
- HU, Y.; CHEN, X. et al. A survey on large language models for task-oriented dialogue. *arXiv preprint arXiv:2308.04030*, 2023.
- JI, Z.; LEE, N. et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, v. 55, n. 12, p. 1–38, 2023.

- JIA, S.; CHI, O. H.; CHI, C. G. Unpacking the impact of AI vs. human-generated review summary on hotel booking intentions. *International Journal of Hospitality Management*, Elsevier, v. 126, p. 104030, 2025.
- JURAFSKY, D.; MARTIN, J. H. *Speech and language processing*. 3rd draft. ed. [S.l.]: Pearson, 2024.
- KAPLAN, J.; MCCANDLISH, S. et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- KAPOOR, S.; NARAYANAN, A. Reproducibility in machine learning: An open science perspective. *arXiv preprint arXiv:2307.01936*, 2023.
- KOJIMA, T. et al. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, v. 35, p. 22199–22213, 2022.
- KONG, A. et al. Better zero-shot reasoning with role-play prompting. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. [S.l.: s.n.], 2024. p. 4099–4113.
- LEE, H.-j.; LEE, K. K.-h.; CHOI, J. A structural model for unity of experience: Connecting user experience, customer experience, and brand experience. *Journal of Usability Studies*, v. 14, n. 1, p. 8–34, 2018.
- LEE, Y. et al. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2025. p. 15771–15798.
- LI, J. et al. A diversity-promoting objective function for neural conversation models. In: *Proceedings of NAACL-HLT*. [S.l.: s.n.], 2016. p. 110–119.
- LIANG, P. et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.
- LIU-THOMPCKINS, Y.; OKAZAKI, S.; LI, H. Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, v. 50, p. 1198–1218, 2022.
- LIU, X.; SCHUCKERT, M.; LAW, R. Can response management benefit hotels? evidence from hong kong hotels online. *Journal of Travel & Tourism Marketing*, v. 35, n. 7, p. 959–973, 2018.
- MADAAN, A. et al. Politeness transfer: A tag and generate approach. In: *Proceedings of the 58th Annual Meeting of the ACL*. [S.l.: s.n.], 2020. p. 1869–1881.
- MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT press, 1999.
- MAURI, A. G.; MINAZZI, R. Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management*, v. 34, p. 99–107, 2013.
- MELLO, R. F. et al. Prompt engineering for automatic short answer grading in brazilian portuguese. In: SBC. *SBIE*. [S.l.], 2024. p. 1730–1743.
- MESKÓ, B. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, v. 25, p. e50638, 2023.

- Meta AI. *Llama 4: Multimodal Intelligence*. 2025. <<https://ai.meta.com/blog/llama-4-multimodal-intelligence/>>. Anúncio oficial. Acesso em: 15 jun. 2026.
- MIKOLOV, T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- MIN, H.; LIM, Y.; MAGNINI, V. P. Factors affecting customer satisfaction in responses to negative online hotel reviews: The impact of empathy, paraphrasing, and speed. *Cornell Hospitality Quarterly*, v. 56, n. 2, p. 223–231, 2015.
- MINAEE, S. et al. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2025.
- MONTIEL-VÁZQUEZ, E. C.; URESTI, J. A. R.; LOYOLA-GONZÁLEZ, O. An explainable artificial intelligence approach for detecting empathy in textual communication. *Applied Sciences*, v. 12, n. 19, p. 9407, 2022.
- NORI, H. et al. Capabilities of GPT-4 on medical competency examinations. *arXiv preprint arXiv:2303.13375*, 2023.
- NOVIKOVA, J. et al. Why we need new evaluation metrics for nlq. *arXiv preprint arXiv:1707.06875*, 2017.
- ORTIZ-GARCES, I. et al. Optimizing chatbot effectiveness through advanced syntactic analysis. *Applied Sciences*, v. 14, n. 5, p. 1737, 2024.
- OUYANG, L.; WU, J. et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, v. 35, p. 27730–27744, 2022.
- PEEPERKORN, M. et al. Is temperature the creativity parameter of large language models? In: *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*. [s.n.], 2024. Disponível em: <<https://arxiv.org/abs/2405.00492>>.
- PISTOTTI, T. et al. The benefits of being uncertain: Perplexity as a signal for naturalness in multilingual machine translation. In: *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*. [S.l.: s.n.], 2025. p. 61–65.
- PRIYA, P.; FIRDAUS, M.; EKBAL, A. Computational politeness in natural language processing: A survey. *ACM Computing Surveys*, 2024.
- PROSERPIO, D.; ZERVAS, G. Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science*, v. 36, n. 5, p. 645–665, 2017.
- RAFFEL, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, v. 21, n. 140, p. 1–67, 2020.
- RASHKIN, H. et al. Towards empathetic open-domain conversation models: a new benchmark and dataset. In: *Proceedings of the 57th Annual Meeting of the ACL*. [S.l.: s.n.], 2019. p. 5370–5381.
- REAL, L.; FONSECA, E.; OLIVEIRA, H. G. The ASSIN 2 shared task: A quick overview. In: *Computational Processing of the Portuguese Language (PROPOR)*. [S.l.: Springer, 2020. p. 406–412.
- REIF, E. et al. A recipe for arbitrary text style transfer with large language models. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. [S.l.: s.n.], 2022. p. 837–848.

- RENZE, M.; GUVEN, E. The effect of sampling temperature on problem solving in large language models. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. [S.l.: s.n.], 2024.
- ROHDEN, S. F.; ESPARTEL, L. B. Emotional artificial intelligence: The impact of chatbot empathy and emotional tone on consumer satisfaction and word of mouth. *Journal of Retailing and Consumer Services*, 2024.
- SANDS, S. et al. Artificial intelligence in retail: The role of anthropomorphism and awareness on consumer reactions. *Journal of Retailing and Consumer Services*, v. 53, p. 101999, 2020.
- SCHULHOFF, S. et al. The prompt report: A systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*, 2025.
- SEDOC, J. et al. Chateval: A tool for chatbot evaluation. In: *Proceedings of NAACL (Demonstrations)*. [S.l.: s.n.], 2019. p. 60–65.
- SERRANO, S.; BRUMBAUGH, Z.; SMITH, N. A. Language models: A guide for the perplexed. *arXiv preprint*, 2023.
- SHANAHAN, M.; MCDONELL, K.; REYNOLDS, L. Role play with large language models. *Nature*, v. 623, p. 493–498, 2023.
- SHARMA, A. et al. A computational approach to understanding empathy expressed in text-based mental health support. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2020. p. 5263–5276.
- SILVA, L. A.; BIANCOLINO, C. A.; SPERS, E. E. The effect of online reviews and comments on the intention of visiting restaurants: An experimental study. *Anais do CLAV 2019 - 19º Congresso Latino-Americano de Varejo*, v. 12, p. 1–15, 2019.
- SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: *Brazilian conference on intelligent systems*. [S.l.: s.n.], 2020. p. 403–417.
- SPARKS, B. A.; BRADLEY, G. L. A “triple a” typology of responding to negative consumer-generated online reviews. *Journal of Hospitality & Tourism Research*, v. 41, n. 6, p. 719–745, 2017.
- TANG, Y.; YANG, Y.; ABBASI, A. *PersonaFuse: A Personality Activation-Driven Framework for Enhancing Human-LLM Interactions*. 2025. Disponível em: <<https://arxiv.org/abs/2509.07370>>.
- TEAM, G. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Tripadvisor, Inc. *Tripadvisor Fact Sheet*. 2024. <https://tripadvisor.s3.amazonaws.com/press/TA_Fact_Sheet.pdf>.
- TSENG, Y.-M. et al. Two tales of persona in LLMs: A survey of role-playing and personalization. In: *Findings of the Association for Computational Linguistics (EMNLP)*. [s.n.], 2024. Disponível em: <<https://arxiv.org/abs/2406.01171>>.
- TUYEN, Q. D. et al. Research on ChatGPT in tourism and hospitality: A systematic review and future directions. *International Journal of Hospitality Management*, Elsevier, v. 131, p. 104287, 2025.

- VARGAS, F. et al. HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*. [S.l.: s.n.], 2022. p. 7174–7183.
- VÁSQUEZ, C. Complaints online: The case of TripAdvisor. *Journal of Pragmatics*, v. 47, n. 1, p. 18–31, 2013.
- VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.
- WANG, Y. et al. Building customer loyalty through emotional connection. *Sustainability*, v. 17, n. 6, p. 2396, 2025.
- WEI, J. et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.
- WEI, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- WEIDINGER, L. et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- YALÇIN, Ö. N. Evaluating empathy in artificial agents. *arXiv preprint arXiv:1908.05341*, 2019.
- YE, Q. et al. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, v. 27, n. 2, p. 634–639, 2011.
- ZHAO, W. X. et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- ZHOU, L.; FARAG, Y.; VLACHOS, A. An llm feature-based framework for dialogue constructiveness assessment. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2024. p. 5389–5409.

A REPOSITÓRIO DO PROJETO

O código-fonte desenvolvido para este trabalho, incluindo os arquivos de engenharia de *prompts*, geração de respostas e avaliação automática (cálculo de métricas de fluência, ToneCal, similaridade estilométrica e empatia percebida), está disponível publicamente no seguinte repositório *Git*:

<https://github.com/danilobsilv/tcc2_files>

B FORMULÁRIO DE AVALIAÇÃO HUMANA

Este apêndice apresenta a versão integral do formulário utilizado para a coleta dos dados qualitativos junto aos avaliadores, gerado via *Google Forms*.

Avaliação Qualitativa

Os avaliadores deverão avaliar cada resposta abaixo nos quesitos de:

- Adequação de tom: O tom da resposta está apropriado ao contexto da interação? A linguagem utilizada reflete o sentimento adequado à situação (por exemplo, cordialidade em agradecimentos ou seriedade em reclamações)?

- Percepção de Empatia: A resposta demonstra empatia em relação ao cliente? O modelo reconhece e valida as emoções expressas, especialmente em casos de avaliações negativas, mostrando compreensão e disposição em resolver o problema?

- Consistência de Marca: A resposta está alinhada com a persona e a identidade comunicacional da marca definida no prompt? O estilo, o vocabulário e o nível de formalidade mantêm coerência com a voz e o posicionamento da marca?

- Coerência e Fluência: O texto é natural, coeso e gramaticalmente correto em português?

* Indica uma pergunta obrigatória

1. E-mail *

Avalie a resposta gerada nos seguintes aspectos

2. Prezado(a) cliente, é uma grande alegria receber seu retorno tão positivo. Ficamos * honrados com seus elogios ao atendimento do <> e do <> , profissionais dedicados a proporcionar momentos de excelência. Agradecemos sua preferência e esperamos recebê-lo(a) novamente em breve.

Marcar apenas uma oval por linha.

	1 (Péssimo)	2 (Ruim)	3 (Regular)	4 (Bom)	5 (Excelente)
Adequação de Tom	☐	☐	☐	☐	☐
Percepção de Empatia	☐	☐	☐	☐	☐
Consistência de Marca	☐	☐	☐	☐	☐
Coerência e Fluência	☐	☐	☐	☐	☐

Avalie a resposta gerada nos seguintes aspectos

3. Prezado cliente, suas palavras nos enchem de orgulho e gratidão. É uma honra * saber que sua experiência conosco foi impecável e que nosso colaborador <> e nossa gastronomia contribuíram para um momento tão marcante. Será um prazer recebê-lo novamente para novas experiências de sabor e hospitalidade.

Marcar apenas uma oval por linha.

	1 (Péssimo)	2 (Ruim)	3 (Regular)	4 (Bom)	5 (Excelente)
Adequação de Tom	☐	☐	☐	☐	☐
Percepção de Empatia	☐	☐	☐	☐	☐
Consistência de Marca	☐	☐	☐	☐	☐
Coerência e Fluência	☐	☐	☐	☐	☐

Avalie a resposta gerada nos seguintes aspectos

4. Muito obrigado pela gentileza! Ficamos felizes em saber que você gostou do nosso ambiente e atendimento. Estamos sempre ansiosos para recebê-lo novamente!

*

Marcar apenas uma oval por linha.

	1 (Péssimo)	2 (Ruim)	3 (Regular)	4 (Bom)	5 (Excelente)
Adequação de Tom	☐	☐	☐	☐	☐
Percepção de Empatia	☐	☐	☐	☐	☐
Consistência de Marca	☐	☐	☐	☐	☐
Coerência e Fluência	☐	☐	☐	☐	☐

Avalie a resposta gerada nos seguintes aspectos

5. Prezada cliente,

★

Agradecemos imensamente por compartilhar os detalhes de sua experiência conosco. É gratificante saber que conseguimos atender às suas expectativas, desde o planejamento inicial até a execução de seu evento.

Sua menção à dedicação de nossos colaboradores <>, <> e <> enche-nos de orgulho. Repassaremos seus elogios à equipe, que se dedica a oferecer um serviço seguro e impecável.

Esperamos ter a honra de recebê-la para futuras celebrações.

Marcar apenas uma oval por linha.

	1 (Péssimo)	2 (Ruim)	3 (Regular)	4 (Bom)	5 (Excelente)
Adequação de Tom	☐	☐	☐	☐	☐
Percepção de Empatia	☐	☐	☐	☐	☐
Consistência de Marca	☐	☐	☐	☐	☐
Coerência e Fluência	☐	☐	☐	☐	☐

Avalie a resposta gerada nos seguintes aspectos