A Multidimensional LLM Evaluation Pipeline for Automatic Responses to Restaurant Reviews in Brazilian Portuguese

Danilo Bruno[1], **Carlos M. S. Figueiredo**[1], **Danielle Pontes**[1], **Tiago de Melo**[1]

[1]Universidade do Estado do Amazonas, Manaus, 69050-020, Brasil

{dbds.eng21,cfigueiredo,dnoronha,tmelo}@uea.edu.br

Resumo. *Este trabalho avalia GPT-5.5, Gemini 3.1 e Llama 4 na geração automática de respostas a avaliações de restaurantes em português brasileiro, tarefa que exige empatia, adequação de tom, fluência e consistência de marca. Um experimento fatorial ($3 \times 2 \times 2 \times 2 = 2.400$ respostas) combina 15 avaliadores humanos (Likert 1–5; concordância robusta nas dimensões pragmáticas, Fleiss' $\kappa \in [0{,}76;\ 0{,}81]$, e moderada em fluência, $\kappa = 0{,}48$) com quatro métricas automáticas. Nenhum modelo é universalmente superior: Llama 4 lidera em empatia, Gemini 3.1 em adequação de tom e consistência de marca, e GPT-5.5 em fluência. Não há evidência de superioridade consistente da estratégia zero-shot sobre a few-shot nas dimensões pragmáticas ($p = 0{,}47$ em empatia). Mais relevante, há discordância de ranking entre as métricas automáticas e o julgamento humano nas dimensões pragmáticas. Em adequação de tom, o modelo mais bem avaliado pelos humanos é o pior pelo proxy automático, o que desaconselha seu uso isolado em tarefas com nuances pragmáticas.*

Abstract. *This work evaluates GPT-5.5, Gemini 3.1, and Llama 4 in the automatic generation of responses to restaurant reviews in Brazilian Portuguese, a task that requires empathy, tone adequacy, fluency, and brand consistency. A factorial experiment ($3 \times 2 \times 2 \times 2 = 2{,}400$ responses) combines 15 human raters (Likert 1–5; strong agreement in the pragmatic dimensions, Fleiss' $\kappa \in [0.76, 0.81]$, and moderate agreement in fluency, $\kappa = 0.48$) with four automatic metrics. No model is universally superior: Llama 4 leads in empathy, Gemini 3.1 in tone adequacy and brand consistency, and GPT-5.5 in fluency. We find no evidence of a consistent advantage of zero-shot over few-shot prompting in the pragmatic dimensions ($p = 0.47$ for empathy). More importantly, automatic metrics show ranking disagreement with human judgment in pragmatic dimensions, in tone adequacy, the model rated best by humans is the worst according to the automatic proxy, which discourages their isolated use in tasks involving pragmatic nuances.*

1. Introdução

Avaliações em plataformas como Google Reviews e TripAdvisor[1] influenciam diretamente a reputação de restaurantes [Silva et al. 2019]. Nesse contexto, os estabelecimentos precisam acompanhar e responder às avaliações dos clientes, tarefa que se torna difícil de realizar manualmente em grande escala. Os Modelos de Linguagem de Grande Escala,

[1]https://ir.tripadvisor.com/investor-faqs

do inglês *Large Language Models* (LLMs), surgem como uma alternativa para automatizar a geração dessas respostas [Brown et al. 2020]. Entretanto, respostas adequadas não dependem apenas da correção gramatical, mas também de aspectos como *empatia*, *adequação de tom*, *fluência* e *consistência de marca*. Essas dimensões têm impacto direto na satisfação do cliente [Rohden and Espartel 2026].

Diferentemente de tarefas tradicionais de geração textual, respostas a avaliações gastronômicas exigem sensibilidade pragmática e compreensão contextual. Uma resposta sintaticamente correta pode ser percebida como fria, genérica ou incompatível com a identidade comunicacional do restaurante. Esse descompasso entre fluência linguística e adequação pragmática representa um desafio ainda pouco explorado em português brasileiro.

Considere, por exemplo, um comentário negativo relacionado a uma comemoração especial. Respostas humanas tendem a contextualizar a experiência emocional do cliente, enquanto respostas automáticas genéricas frequentemente se limitam a fórmulas como "agradecemos seu feedback", sem reconhecer os elementos específicos da situação. A literatura recente demonstra que consumidores penalizam interações impessoais ou excessivamente robotizadas [Crolic et al. 2022], especialmente em contextos de serviço e hospitalidade [Rohden and Espartel 2026].

O problema envolve ainda o gerenciamento de dados textuais: extração e normalização de avaliações multiplataforma, anonimização (conformidade com a LGPD), armazenamento estruturado em esquema reprodutível e *pipelines* de avaliação escaláveis. A literatura carece de *frameworks* que integrem métricas automáticas com avaliação humana para o português brasileiro no domínio gastronômico [Liang et al. 2023, Howcroft and Rieser 2021, de Araújo et al. 2024].

Diante deste cenário, este trabalho busca responder a três questões de pesquisa: (RQ1) Qual LLM produz respostas de maior qualidade percebida e em quais dimensões? (RQ2) *Zero-shot* supera *few-shot* em dimensões pragmáticas? (RQ3) Métricas automáticas convergem com a avaliação humana?

Este trabalho apresenta três contribuições principais. **(C1)** Um *framework*[2] de avaliação multidimensional reprodutível para respostas automáticas a avaliações gastronômicas em português brasileiro, integrando avaliação humana (escala Likert sobre quatro dimensões de avaliação) e quatro métricas automáticas complementares. **(C2)** A execução de um experimento fatorial completo ($3 \times 2 \times 2 \times 2$), totalizando 2.400 respostas e 24 cenários distintos, com avaliação humana estratificada por painel heterogêneo de 15 avaliadores e concordância robusta nas dimensões pragmáticas (Fleiss' $\kappa \in [0{,}76;\ 0{,}81]$). **(C3)** Evidência empírica de discordância de ranking entre métricas automáticas e julgamento humano em dimensões pragmáticas. Em adequação de tom, o modelo mais bem avaliado pelos humanos é o pior pelo proxy automático, demonstrando que tais proxies são insuficientes para a comparação relativa de qualidade pragmática sem triangulação com avaliação humana.

O restante do trabalho está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados, a seção 3 apresenta a metodologia proposta, a seção 4 seus resultados, que são melhor discutidos na seção 5, e conclusões e trabalhos futuros são

[2] `https://github.com/tmelo-uea/eniac2026-llm-restaurant-reviews-ptbr`

apontados na seção 6.

2. Trabalhos Relacionados

A aplicação de LLMs em atendimento automatizado levanta desafios que vão além da correção gramatical. Hudeček e Dušek [Hudeček and Dušek 2023] demonstraram que o controle fino de tom e empatia permanece um problema em aberto; Li et al. [Li et al. 2016] ressaltam que a diversidade de respostas é essencial para evitar a percepção robótica pelos clientes. Para operacionalizar empatia e tom computacionalmente, Sharma et al. [Sharma et al. 2020] propuseram um *framework* multidimensional de empatia em texto, avançado pelas *Shared Tasks* do WASSA [Barriere et al. 2022, Barriere et al. 2023], enquanto Danescu-Niculescu-Mizil et al. [Danescu-Niculescu-Mizil et al. 2013] formalizaram a polidez como construto mensurável. Em contexto comercial, Rohden e Espartel [Rohden and Espartel 2026] evidenciaram que o tom empático de chatbots impacta positivamente a satisfação do consumidor.

Quanto às estratégias de geração, Brown et al. [Brown et al. 2020] estabeleceram o *In-Context Learning* como uma capacidade central dos LLMs; Mello et al. [Mello et al. 2024] validaram abordagens *few-shot* para tarefas em português, e Wei et al. [Wei et al. 2022] mostraram que exemplos atuam como âncoras cognitivas, embora seu efeito em tarefas criativas e subjetivas permaneça pouco investigado. Na avaliação de texto gerado, Liang et al. [Liang et al. 2023] propuseram o HELM como referência holística, e Howcroft e Rieser [Howcroft and Rieser 2021] alertam que métricas automáticas isoladas não substituem a avaliação humana, achado que fundamenta diretamente o protocolo adotado neste trabalho.

Diferentemente dos trabalhos anteriores, este estudo triangula a avaliação humana e automática especificamente para respostas a avaliações gastronômicas em português brasileiro, oferecendo evidências empíricas de divergência sistemática entre as duas modalidades. A Tabela 1 posiciona o presente trabalho frente à literatura, em que nenhuma das abordagens anteriores reúne, simultaneamente, o domínio de atendimento em português brasileiro, a geração de texto e a triangulação entre avaliação humana e métricas automáticas.

Tabela 1. Posicionamento deste trabalho frente à literatura relacionada. Geração: geração de texto (vs. classificação ou teoria); Serviço: domínio de atendimento ao cliente; Humana/Autom.: avaliação humana/métricas automáticas; Triang.: triangulação humano×automático. ✓ atende; (∼) atende parcialmente.

Trabalho	PT-BR	Geração	Serviço	Humana	Autom.	Triang.
Hudeček & Dušek [Hudeček and Dušek 2023]		✓	✓		✓	
Li et al. [Li et al. 2016]		✓	(∼)		✓	
Sharma et al. [Sharma et al. 2020]				(∼)	✓	
Danescu et al. [Danescu-Niculescu-Mizil et al. 2013]					✓	
Rohden & Espartel [Rohden and Espartel 2026]			✓	✓		
Liang et al. [Liang et al. 2023]		✓		(∼)	✓	(∼)
Mello et al. [Mello et al. 2024]	✓	✓		(∼)	✓	
Araújo et al. [de Araújo et al. 2024]	✓				✓	
Este trabalho	✓	✓	✓	✓	✓	✓

3. Metodologia

3.1. Corpus, Design e Framework de Avaliação

A Figura 1 apresenta a arquitetura experimental adotada, integrando a coleta e curadoria do corpus, geração das respostas pelos LLMs, avaliação humana e cálculo das métricas automáticas em um *pipeline* unificado.

O corpus de origem compreende 200 comentários reais coletados do Google Maps, TripAdvisor e Instagram, normalizados, anonimizados e armazenados em esquema estruturado (texto, plataforma, período, escores). Os comentários foram divididos em dois períodos com base na data de publicação, tomando como marco temporal o lançamento do ChatGPT em 30/11/2022: 100 do período pré-popularização das LLMs, associados a respostas humanas usadas como *baseline*, e 100 do período pós-popularização. Essa divisão não permite identificar o uso de IA generativa em casos individuais, sendo utilizada como uma aproximação temporal para analisar possíveis diferenças entre os períodos.

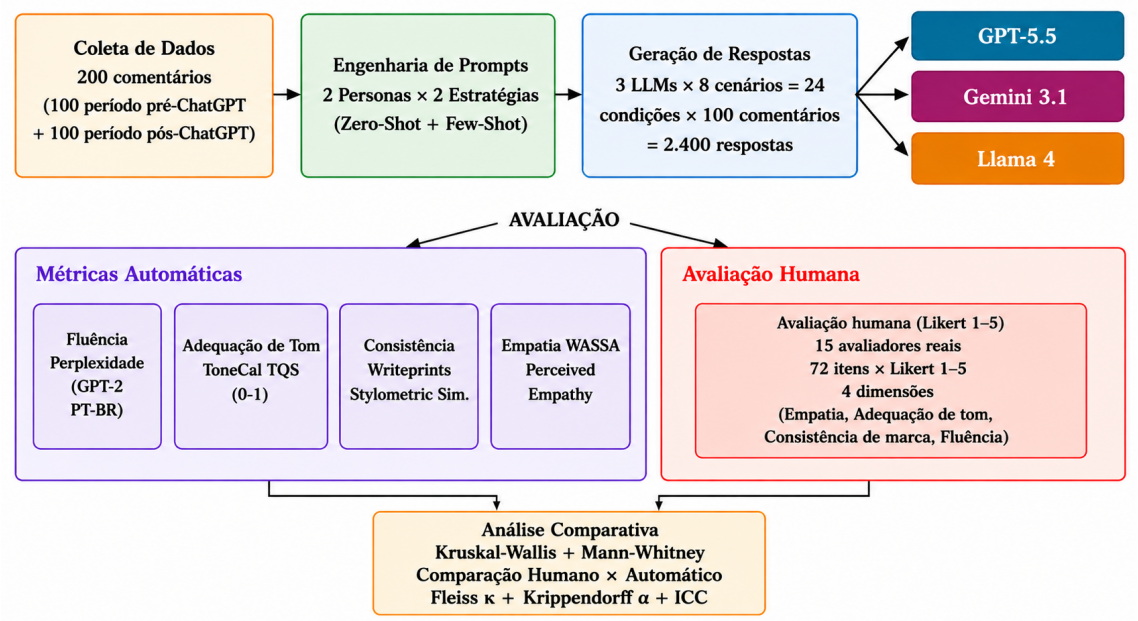


Figura 1. Arquitetura experimental: *pipeline* de avaliação.

O corpus resultante contém 2.400 respostas, representadas pelo cruzamento fatorial de 3 modelos $\times$ 2 personas $\times$ 2 estratégias $\times$ 2 períodos (24 cenários, 100 respostas cada). Foram empregados GPT-5.5 (OpenAI) e Gemini 3.1 (Google), acessados via APIs oficiais, e Llama 4 (Meta), executado localmente em ambiente controlado (2025), sem *fine-tuning*, sob temperatura $T = 0{,}7$ e *top-p* $= 0{,}9$, valores amplamente adotados em geração de texto [Brown et al. 2020]; o número máximo de *tokens* foi mantido no padrão de cada implementação, e a extensão foi controlada por instrução no *prompt* (duas a quatro frases). Cada chamada gerou uma única resposta, sem regeneração nem descarte seletivo: as 2.400 saídas foram preservadas integralmente, evitando viés de seleção *a posteriori*. As personas avaliadas são *Formal* (vocabulário elevado, "Prezado(a) cliente") e *Informal* (registro conversacional); as estratégias, *zero-shot* (apenas o comentário) e *few-shot* (três pares exemplares; *In-Context Learning* [Brown et al. 2020]).

Avaliação humana (padrão-ouro): 15 avaliadores avaliaram 72 itens por amostragem estratificada (3 itens $\times$ 24 cenários; 36 do período pré e 36 do período pós), em delineamento cego ao modelo e ao período, utilizando escala Likert de 1–5 para empatia, adequação de tom, fluência e consistência de marca. As dimensões foram avaliadas segundo a percepção dos participantes, sem o fornecimento de uma rubrica detalhada ou de critérios operacionais específicos para cada nível da escala.

Foram selecionados 15 avaliadores independentes, com idades entre 21 e 71 anos (média de 40), distribuídos em quatro faixas etárias para assegurar representação intergeracional: quatro entre 21–29 anos, cinco entre 30–39, quatro entre 40–59 e dois com 60 ou mais. Quanto à atuação profissional, o painel reuniu perfis heterogêneos: quatro profissionais da saúde (dois médicos, um dentista e um educador físico), quatro engenheiros (dois de *software*, um florestal e um elétrico), três estudantes universitários, uma advogada, um corretor de imóveis, uma autônoma e um aposentado. Essa composição contempla diferentes níveis de letramento digital e contextos de consumo gastronômico, mitigando vieses associados a bolhas tecnológicas, ocupacionais ou geracionais e aproximando o painel da heterogeneidade da clientela real de restaurantes, em linha com recomendações recentes da literatura de avaliação humana em PLN [Howcroft and Rieser 2021].

As métricas automáticas utilizadas foram selecionadas por sua relevância para cada dimensão avaliada:

- **Perplexidade (GPT-2 PT-BR)** [Pistotti et al. 2025]: mede a fluência, indicando a previsibilidade do texto sob um modelo treinado em português brasileiro;
- **ToneCal** $[0, 1]$, baseado em [Danescu-Niculescu-Mizil et al. 2013]: avalia a adequação de tom, quantificando polidez e formalidade;
- **Estilometria (*Writeprints*)** [Abbasi and Chen 2008]: compara o perfil lexical e estilístico das respostas geradas com as respostas humanas originais, estimando a consistência de marca;
- **WASSA** [Barriere et al. 2023]: estima empatia em $[1, 5]$ por meio de um modelo pré-treinado em tarefas de empatia textual.

Cada métrica foi aplicada de forma independente a todas as respostas, permitindo a comparação direta com os escores atribuídos pelos avaliadores humanos.

4. Experimentos e Resultados

A confiabilidade da avaliação humana foi analisada por meio de três medidas complementares de concordância inter-avaliadores: o coeficiente de Fleiss (κ), adequado a escalas categóricas com múltiplos avaliadores; o alfa de Krippendorff (α), robusto a diferentes tipos de dados; e o Coeficiente de Correlação Intraclasse (ICC, *Intraclass Correlation Coefficient*), empregado para estimar a consistência dos julgamentos. Todos os 15 avaliadores pontuaram os mesmos 72 itens, e os três coeficientes foram calculados separadamente por dimensão. Os resultados indicaram concordância robusta nas três dimensões pragmáticas, com Fleiss' $\kappa \in [0{,}76; 0{,}81]$ (empatia 0,81; tom 0,80; marca 0,76), Krippendorff's $\alpha \in [0{,}84; 0{,}90]$ e ICC$(2, 1) \in [0{,}87; 0{,}90]$. A fluência apresentou concordância apenas moderada (Fleiss' $\kappa = 0{,}48$; $\alpha = 0{,}49$; ICC$(2, 1) = 0{,}56$), reflexo de um efeito de teto (quase todas as respostas são fluentes, com notas oscilando entre 4 e 5). Essas medidas embasam o uso das avaliações humanas pragmáticas como referência neste estudo. O fator período foi incluído no desenho como uma dimensão de controle e análise de

sensibilidade. Embora tenha apresentado efeito significativo em empatia ($\Delta = +0{,}226$, $p = 0{,}017$) e adequação de tom ($\Delta = +0{,}348$, $p = 0{,}004$), os 100 comentários do período pré são distintos dos 100 do período pós. Portanto, esses resultados não permitem inferência causal sobre o efeito do período e são interpretados apenas de forma descritiva. Por essa razão, os resultados a seguir são apresentados de forma agregada sobre os dois períodos.

Procedimentos estatísticos. As comparações entre modelos foram conduzidas por meio do teste de Kruskal-Wallis, seguido de comparações pareadas *post-hoc* pelo Mann-Whitney U. Em virtude do número de comparações múltiplas (três pares de modelos para cada uma das quatro dimensões humanas), os *p*-valores pareados foram ajustados pelo procedimento de Holm-Bonferroni dentro de cada dimensão. As magnitudes das diferenças são interpretadas pelas próprias médias Likert reportadas nas tabelas, dado que diferenças da ordem de 0,3–0,8 ponto em uma escala 1–5 correspondem a efeitos de magnitude pequena a moderada em escalas ordinais.

4.1. Resultados por Modelo (RQ1)

A Tabela 2 e a Figura 2 consolidam os resultados. Nenhum modelo domina todas as dimensões. Llama 4 lidera em empatia, Gemini 3.1 prevalece em adequação de tom e consistência de marca, e GPT-5.5 em fluência; cada modelo é superior em ao menos uma dimensão e nenhum é uniformemente inferior.

Tabela 2. Resultados globais por modelo: avaliação humana (Likert 1–5) e métricas automáticas. Negrito: melhor valor por coluna.

	Avaliação Humana				Métricas Automáticas			
Modelo	**Empatia**	**Tom**	**Fluência**	**Marca**	**WASSA↑**	**ToneCal↑**	**PPL↓**	**Estil.↑**
Llama 4	**4,247**	4,275	4,714	3,103	3,280	**0,582**	229,4	0,052
GPT-5.5	3,983	3,975	**4,997**	3,203	**3,360**	0,544	**118,3**	0,061
Gemini 3.1	3,953	**4,594**	4,475	**3,831**	3,113	0,513	179,5	**0,070**

O teste de Kruskal-Wallis indica diferenças significativas entre os modelos em adequação de tom ($H = 9{,}96$, $p = 0{,}007$), fluência ($H = 39{,}41$, $p < 0{,}001$) e consistência de marca ($H = 16{,}18$, $p < 0{,}001$); em empatia, as diferenças não alcançam significância ($H = 4{,}39$, $p = 0{,}11$). As comparações pareadas Mann-Whitney U, ajustadas por Holm-Bonferroni dentro de cada dimensão, localizam os contrastes significativos: em adequação de tom, Gemini 3.1 $>$ GPT-5.5 ($p_{\text{adj}} = 0{,}004$); em consistência de marca, Gemini 3.1 supera tanto Llama 4 ($p_{\text{adj}} < 0{,}001$) quanto GPT-5.5 ($p_{\text{adj}} = 0{,}004$); em fluência, os três pares diferem (GPT-5.5 $>$ Gemini 3.1, $p_{\text{adj}} < 0{,}001$; GPT-5.5 $>$ Llama 4, $p_{\text{adj}} = 0{,}004$; Llama 4 $>$ Gemini 3.1, $p_{\text{adj}} = 0{,}002$). Em empatia, nenhum par sobrevive à correção (Llama 4 $>$ Gemini 3.1, $p = 0{,}053$ bruto). O tamanho da amostra humana (72 itens) limita o poder para discriminar as diferenças menores, sobretudo em empatia, mas os contrastes em tom, marca e fluência são robustos ao controle do erro tipo I.

4.2. Efeitos de Persona e Estratégia (RQ2)

A Tabela 3 apresenta as médias humanas decompostas por modelo e persona, enquanto a Tabela 4 consolida a comparação entre as estratégias *zero-shot* e *few-shot* para cada

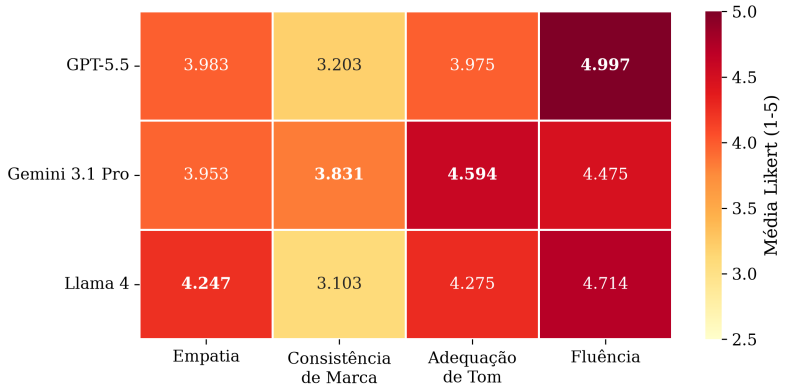


Figura 2. Médias Likert por modelo e dimensão na avaliação humana.

modelo.

Tabela 3. Avaliação humana por modelo e persona (Likert 1–5).

Modelo	Persona	Empatia	Tom	Marca	Fluência
Llama 4	Informal	**4,306**	**4,289**	3,044	4,511
Llama 4	Formal	4,189	4,261	**3,161**	**4,917**
GPT-5.5	Informal	**4,039**	**4,039**	**3,378**	**5,000**
GPT-5.5	Formal	3,928	3,911	3,028	4,994
Gemini 3.1	Informal	**4,017**	4,372	3,456	4,456
Gemini 3.1	Formal	3,889	**4,817**	**4,206**	**4,494**

Tabela 4. Avaliação humana por modelo e estratégia de prompting (Likert 1–5).

Modelo	Estratégia	Empatia	Tom	Marca	Fluência
Llama 4	Zero-shot	**4,339**	**4,367**	**3,128**	4,611
Llama 4	Few-shot	4,156	4,183	3,078	**4,817**
GPT-5.5	Zero-shot	**4,100**	**4,111**	3,128	4,994
GPT-5.5	Few-shot	3,867	3,839	**3,278**	**5,000**
Gemini 3.1	Zero-shot	3,889	**4,622**	3,783	**4,489**
Gemini 3.1	Few-shot	**4,017**	4,567	**3,878**	4,461

Persona. A persona informal supera a formal em empatia em *todos* os modelos (4,120 vs. 4,002); em adequação de tom e consistência de marca, porém, a forte persona formal do Gemini 3.1 (tom 4,817; marca 4,206) inverte a média entre modelos a favor da formal. As métricas automáticas também favorecem a formal (WASSA: 3,329 vs. 3,205; ToneCal: 0,555 vs. 0,538); assim, a divergência humano$\times$automático no eixo de registro concentra-se na empatia.

Estratégia. A estratégia *zero-shot* apresentou médias ligeiramente superiores à *few-shot* em empatia ($\Delta = +0{,}096$) e adequação de tom ($\Delta = +0{,}170$), mas a diferença não é estatisticamente significativa ($p = 0{,}47$ em empatia) nem consistente entre os modelos. O efeito é positivo e na mesma direção para GPT-5.5 ($\Delta = +0{,}233$ em empatia

e $+0{,}272$ em tom) e Llama 4 ($\Delta = +0{,}183$ em ambas), mas o Gemini 3.1 inverte a tendência em empatia ($\Delta = -0{,}128$), anulando o efeito agregado. As métricas automáticas tampouco distinguem as estratégias (diferença em empatia WASSA praticamente nula, $\Delta = 0{,}003$). Em conjunto, não há evidência de que o *zero-shot* supere sistematicamente o *few-shot* nas dimensões pragmáticas; o efeito, quando presente, é específico do modelo.

4.3. Análise Qualitativa das Respostas

A análise qualitativa identificou padrões pragmáticos não capturados pelas métricas automáticas. As respostas mais bem avaliadas tendiam a contextualizar explicitamente o comentário, reconhecer a dimensão emocional do cliente e referenciar diretamente os elementos mencionados.

Para o comentário "*O prato veio morno e isso comprometeu a experiência*", a versão *zero-shot* do Llama 4 (persona formal) contextualizou a situação, "*lamentamos que o prato tenha chegado morno e comprometido o momento da comemoração*", enquanto a versão *few-shot* replicou a abertura dos exemplos demonstrativos ("*Prezado(a) cliente, agradecemos seu feedback*"), sugerindo que os exemplos podem restringir a expressividade. De forma análoga, diante de elogios específicos, a persona informal do GPT-5.5 personalizou a resposta agradecendo nominalmente os elementos elogiados, enquanto a persona formal do Gemini 3.1 reconheceu o cliente de modo elaborado e alinhado à marca, coerente com suas médias superiores em tom e consistência de marca. Em comentários ambíguos (e.g., "*comida ótima, mas o garçom demorou*"), as respostas mais bem avaliadas reconheceram ambos os polos de forma proporcional, ao passo que aberturas formulaicas tenderam a enfatizar o aspecto positivo e minimizar a crítica.

Observaram-se também casos de desalinhamento pragmático em que respostas gramaticalmente corretas iniciavam com construções positivas mesmo diante de comentários negativos. Em empatia, os avaliadores tenderam a favorecer respostas acolhedoras e conversacionais, penalizando fórmulas impessoais; já em adequação de tom e consistência de marca, respostas formais bem construídas e ricas em reconhecimento contextual também alcançaram avaliação elevada, indicando que o fator decisivo não é o registro em si, mas a adequação pragmática ao comentário.

4.4. Convergências e Divergências (RQ3)

A Tabela 5 sintetiza a comparação de rankings entre as duas modalidades de avaliação, obtidos pela ordenação dos modelos segundo as médias da avaliação humana e os respectivos escores das métricas automáticas. A consistência de marca apresenta convergência forte: avaliação humana e estilometria coincidem ao posicionar o Gemini 3.1 em primeiro. Em empatia e fluência há convergência apenas parcial: as modalidades concordam em um dos extremos do ranking (Gemini 3.1 como o mais fraco em empatia; GPT-5.5 como o mais forte em fluência), mas discordam nas demais posições. A divergência mais expressiva está na adequação de tom, em que ocorre uma *inversão* de ranking: o Gemini 3.1, mais bem avaliado pelos humanos, é classificado em último pela ToneCal, enquanto o Llama 4, líder do proxy automático, aparece em segundo na avaliação humana, evidenciando que os marcadores de polidez captados pela ToneCal não acompanham a qualidade de tom percebida.

O padrão evidencia uma limitação estrutural dos *proxies*: por dependerem de modelos e corpora de distribuição predominantemente formal e de traços de superfície, po-

Tabela 5. Avaliação humana vs. métricas automáticas.

Dimensão	Ranking Humano	Ranking Automático	Convergência
Tom	Gemini $>$ Llama $>$ GPT	Llama $>$ GPT $>$ Gemini	Divergente
Fluência	GPT $>$ Llama $>$ Gemini	GPT $>$ Gemini $>$ Llama	Parcial
Empatia	Llama $>$ GPT $>$ Gemini	GPT $>$ Llama $>$ Gemini	Parcial
Marca	Gemini $>$ GPT $>$ Llama	Gemini $>$ GPT $>$ Llama	Forte

dem não apenas atenuar, mas *inverter* o ordenamento humano em dimensões pragmáticas, como ocorre em adequação de tom. Embora perplexidade e estilometria preservem convergência útil em fluência e consistência de marca, a possibilidade de inversão desaconselha o uso isolado de métricas automáticas para decidir qual modelo oferece melhor qualidade pragmática; conclusões finais requerem triangulação com avaliação humana.

5. Discussão

5.1. Viés de Registro Formal e Limitações dos *Proxies*

Os resultados evidenciam que as métricas automáticas podem divergir do julgamento humano a ponto de *inverter* o ordenamento entre modelos em dimensões pragmáticas. O caso mais nítido é a adequação de tom: o modelo mais bem avaliado pelos humanos é classificado em último pela ToneCal. Uma possível explicação é que os modelos de perplexidade, polidez e empatia foram treinados majoritariamente em textos jornalísticos ou institucionais; assim, a sintaxe canônica e os marcadores explícitos de cortesia recebem pontuações superiores independentemente da adequação pragmática ao comentário. No eixo de persona, esse viés formal manifesta-se sobretudo na empatia, em que os avaliadores valorizam o acolhimento da persona informal enquanto os *proxies* premiam o registro formal. Cada métrica captura apenas uma faceta: a perplexidade mede previsibilidade estatística e penaliza textos espontâneos; o WASSA foi desenvolvido em inglês para domínios distintos, limitando sua transferência ao português comercial; a estilometria *Writeprints* mede similaridade lexical com um corpus de referência, sem considerar a adequação situacional. Métricas automáticas são, portanto, aproximações operacionais úteis para comparações relativas de fluência e consistência de marca, mas insuficientes, e potencialmente enganosas, para ordenar modelos em tom e empatia sem triangulação com julgamento humano.

5.2. Implicações para Engenharia de *Prompt*

O efeito da estratégia de *prompting* mostrou-se dependente do modelo, e não universal. Para GPT-5.5 e Llama 4, a abordagem *zero-shot* apresentou leve vantagem nas dimensões pragmáticas, compatível com a hipótese de que exemplos *few-shot* atuam simultaneamente como aprendizado em contexto e como restrição estilística, induzindo respostas mais formulaicas; o Gemini 3.1, contudo, beneficiou-se do *few-shot* em empatia, invertendo a tendência. Como o efeito agregado não é significativo ($p = 0{,}47$), não se sustenta uma recomendação geral de *zero-shot* sobre *few-shot*; o desenho de *prompt* mais eficaz parece depender do modelo, e a combinação de descrições claras de persona com exemplos bem escolhidos deve ser calibrada caso a caso.

5.3. Implicações Práticas para Sistemas de Atendimento

A escolha do modelo deve ser orientada pela dimensão prioritária do negócio: estabelecimentos que valorizam adequação de tom e consistência de marca tendem a se beneficiar do Gemini 3.1; contextos que priorizam acolhimento empático, do Llama 4; e aqueles em que fluência é crítica, do GPT-5.5. A estratégia de *prompting* não apresenta vantagem universal e deve ser calibrada por modelo. Por fim, os resultados desaconselham ciclos de avaliação baseados apenas em métricas automáticas em produção: como esses *proxies* podem inverter o ranking humano em dimensões pragmáticas, equipes de qualidade devem prever rodadas periódicas de avaliação humana, ainda que em amostras reduzidas, para detectar regressões pragmáticas. O *framework* é adaptável a outros domínios de serviço (hospitalidade, varejo, suporte ao cliente).

5.4. Limitações Estatísticas e de Escopo

Os resultados devem ser interpretados à luz de algumas limitações. O tamanho da amostra humana (72 itens estratificados pelos 24 cenários) implica poder estatístico limitado: as diferenças entre modelos em empatia não alcançam significância, ainda que os contrastes em tom, marca e fluência sejam robustos à correção de Holm-Bonferroni. O efeito da estratégia de *prompting* é pequeno, não significativo ($p = 0{,}47$ em empatia) e de direção dependente do modelo, exigindo replicação em amostras maiores. O uso de testes não paramétricos limita a estimação direta de intervalos de confiança, embora as magnitudes possam ser inferidas das tabelas. O efeito do período é interpretado como descritivo, dado que os comentários-base diferem entre os dois períodos. Além disso, as diferenças observadas são relativas entre os modelos avaliados, pois o estudo não estabelece um limiar validado de desempenho que permita determinar sua prontidão para uso em produção. Por fim, os achados refletem o corpus de um único estabelecimento, um painel de 15 avaliadores e versões específicas dos modelos (2025), exigindo replicação para generalização a outros domínios, populações ou versões futuras das LLMs.

6. Conclusão

Este trabalho apresentou e aplicou um *framework* de avaliação multidimensional para respostas automáticas a avaliações gastronômicas em português brasileiro, comparando 2.400 respostas (GPT-5.5, Gemini 3.1 e Llama 4) em um experimento fatorial completo. (RQ1) Nenhum modelo é universalmente superior: Llama 4 lidera em empatia, Gemini 3.1 em adequação de tom e consistência de marca, e GPT-5.5 em fluência; cada modelo prevalece em ao menos uma dimensão. O Kruskal-Wallis indica diferenças significativas em tom ($p = 0{,}007$), marca ($p < 0{,}001$) e fluência ($p < 0{,}001$), mas não em empatia ($p = 0{,}11$); os contrastes pareados em tom, marca e fluência sobrevivem à correção de Holm-Bonferroni (e.g., Gemini 3.1 $>$ GPT-5.5 em tom, $p_{\text{adj}} = 0{,}004$; Gemini 3.1 $>$ Llama 4 em marca, $p_{\text{adj}} < 0{,}001$). (RQ2) Não se observou vantagem consistente da estratégia *zero-shot* sobre a *few-shot* nas dimensões pragmáticas: o efeito é pequeno, não significativo ($p = 0{,}47$ em empatia) e de direção dependente do modelo. (RQ3) As métricas automáticas convergem com a avaliação humana em consistência de marca e, parcialmente, em empatia e fluência, mas apresentam forte discordância de ranking em adequação de tom, o modelo mais bem avaliado pelos humanos é o pior pela métrica automática, evidenciando que *proxies* de superfície são insuficientes para ordenar a qualidade pragmática.

A principal contribuição é evidenciar a importância da triangulação entre avaliação humana e automática. Métricas automáticas isoladas podem discordar fortemente do ordenamento humano de modelos em dimensões pragmáticas, na adequação de tom, o modelo mais bem avaliado pelos humanos é o pior pelo proxy, além de premiar o registro formal onde os humanos valorizam o acolhimento informal (empatia). Esse achado tem implicações diretas para o desenho de ciclos de avaliação de sistemas de resposta automática em português brasileiro.

Limitações e trabalhos futuros. O corpus restrito a um único estabelecimento e a amostragem humana (72 de 2.400 itens, estratificada pelos 24 cenários) limitam a generalização e o poder estatístico para detectar diferenças de magnitude moderada; a evolução contínua dos modelos restringe a validade temporal. Extensões naturais incluem a expansão para múltiplos restaurantes, a adoção de modelos de efeitos mistos, a calibração de métricas automáticas para o registro informal em português e a investigação de *fine-tuning* específico de domínio.

Agradecimentos

Os autores agradecem o apoio da Universidade do Estado do Amazonas (UEA), por meio da Gratificação de Produtividade Acadêmica (GPA), conforme a Portaria nº 1177/2025-GR/UEA. Este trabalho contou com apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional e Tratamento e Disseminação de Informação (INCT-TILD-IAR), financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo nº 408490/2024-1.

Declaração de Uso de Inteligência Artificial

Ferramentas de Inteligência Artificial Generativa foram utilizadas como apoio à revisão linguística, organização textual e melhoria de clareza do artigo. GPT-5.5, Gemini 3.1 e Llama 4 também foram empregados como objetos de estudo experimental, conforme descrito na metodologia. A responsabilidade integral pelo conteúdo, resultados, análises e referências permanece com os autores.

Referências

Abbasi, A. and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7.

Barriere, V., Sedoc, J., Tafreshi, S., and Giorgi, S. (2023). 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*, pages 511–525.

Barriere, V., Tafreshi, S., Sedoc, J., and Alqahtani, S. (2022). 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*, pages 214–227.

Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Crolic, C., Thomaz, F., Hadi, R., and Stephen, A. T. (2022). Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, 86(1):132–148.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the ACL*, pages 250–259.

de Araújo, G., de Melo, T., and Figueiredo, C. M. S. (2024). Explorando a eficácia das linguagens generativas em tarefas de análise de sentimentos no português brasileiro. *Linguamática*, 16(2):41–58.

Howcroft, D. M. and Rieser, V. (2021). 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*, pages 8932–8939.

Hudeček, V. and Dušek, O. (2023). Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, W. B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., et al. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Mello, R. F., Rodrigues, L., Cabral, L., Pereira, F. D., et al. (2024). Prompt engineering for automatic short answer grading in brazilian portuguese. In *SBIE*, pages 1730–1743. SBC.

Pistotti, T., Witbrock, M., O'Leary, P. A. T., and Brown, J. (2025). 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 (UncertaiNLP 2025)*, pages 61–65.

Rohden, S. F. and Espartel, L. B. (2026). Emotional artificial intelligence: The impact of chatbot empathy and emotional tone on consumer satisfaction and word of mouth. *International Journal of Human-Computer Studies*, 210:103764.

Sharma, A., Miner, A. S., Atkins, D. C., and Althoff, T. (2020). 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)*, pages 5263–5276.

Silva, L. A., Biancolino, C. A., and Spers, E. E. (2019). 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*, 12:1–15.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.