Detectando Incoerências Avaliativas em E-commerce com LLMs - Um Estudo de Caso na Amazon Brasil

Emanuelle Marreira
Universidade do Estado do Amazonas
erm.eng22@uea.edu.br

Tiago de Melo
Universidade do Estado do Amazonas
tmelo@uea.edu.br

Miguel Oliveira
Universidade Federal do Amazonas
miguel.oliveira@icomp.ufam.edu.br

Carlos Maurício
Universidade do Estado do Amazonas
cfigueiredo@uea.edu.br

ABSTRACT

This study evaluates the ability of large language models (LLMs) to detect incoherence between the text of product reviews and their assigned rating (1 or 5 stars). Using ChatGPT-o3 in five independent runs, we observed high variability in labeling and low overall agreement (Fleiss' κ = 0.177). A conservative approach selected 415 reviews unanimously labeled as incoherent, which were subsequently submitted for human evaluation. The agreement between human annotators was substantial (Cohen's κ = 0.709), allowing the isolation of 231 cases with a clear sentiment judgment. The comparison showed that only 28.1% of the LLM classifications matched the human judgment. These results suggest that, while promising, LLMs still require rigorous validation and careful calibration for critical semantic interpretation tasks.

KEYWORDS

Incoerência Semântica, Large Language Models, Avaliações Online

1 INTRODUÇÃO

A popularização das plataformas de comércio eletrônico transformou as avaliações de usuários em um recurso crucial para consumidores e lojistas [6]. Embora as notas numéricas, frequentemente representadas por estrelas (1 a 5), forneçam um resumo imediato da percepção do cliente, o texto que as acompanha frequentemente contém nuances que a avaliação numérica isolada não captura ou, em alguns casos, contradiz [1]. Isso é ilustrado na Figura 1, onde o comentário textual é positivo, mas a avaliação de 1 estrela. Detectar automaticamente essas contradições é relevante por três motivos principais: (i) aumenta a confiabilidade das métricas de satisfação exibidas publicamente, (ii) auxilia lojistas na identificação de potenciais fraudes ou erros de digitação e (iii) oferece um novo sinal para sistemas de recomendação e análise de sentimentos. Adicionalmente, essa inconsistência pode aumentar o esforço cognitivo do consumidor, resultar em decisões de compra menos precisas e reduzir a utilidade da plataforma de avaliações [13].

Historicamente, a verificação de coerência entre nota e texto tem sido considerada um subtipo de análise de sentimentos, demandando conjuntos de dados anotados manualmente e modelos supervisionados específicos. O advento dos *Large Language Models*

Figura 1: Discrepância entre o texto e a nota.

(LLMs), como o ChatGPT, introduziu a possibilidade de avaliar incoerências em regime *zero-shot* ou *few-shot*, dispensando engenharia de características e *fine-tuning* extensivo. Apesar do avanço, a literatura carece de estudos sistemáticos que quantifiquem (a) a precisão desses modelos para essa tarefa específica e (b) a variabilidade das respostas quando múltiplas execuções independentes do mesmo LLM são consideradas.

Este trabalho investiga se modelos de linguagem de grande escala, como o ChatGPT, são capazes de identificar, de forma consistente, incoerências semânticas entre a nota (1 ou 5 estrelas) e o conteúdo textual de avaliações. O estudo se restringiu a notas extremas, pois elas apresentam polaridade mais clara, facilitando a análise de contradições explícitas. Avaliações intermediárias, por outro lado, são mais ambíguas e dificultariam a validação. Cada comentário foi submetido ao ChatGPT (versão o3) em cinco execuções independentes, o que permitiu avaliar sua estabilidade intra-modelo diante da variabilidade das respostas.

As contribuições deste artigo são: (i) curadoria de dados: disponibilização de um conjunto balanceado e anotado de avaliações extremas em dez categorias de *e-commerce*, em português brasileiro, contendo a nota numérica e o texto original[1]; (ii) protocolo de avaliação *zero-shot*: descrição de um *prompt* genérico e reproduzível para detectar incoerências nota–texto, aplicável a qualquer LLM moderno; (iii) análise de variabilidade: quantificou-se a concordância intra-modelo em cinco execuções, usando métricas clássicas de confiabilidade (κ de Fleiss e κ de Cohen), pouco exploradas em pesquisas com LLMs; e (iv) discussão de implicações práticas, através de demonstração de como as inconsistências detectadas podem apoiar a moderação automática de conteúdo e aprimorar sistemas de reputação em ambientes de varejo digital.

2 TRABALHOS RELACIONADOS

2.1 Comentário versus nota

A partir de uma revisão de literatura, Almansour *et al.* [2] identificam que a correlação entre sentimentos expressos em texto e as notas numéricas costuma ser de baixa a moderada, indicando que confiar exclusivamente em uma dessas fontes pode levar a

[1] https://github.com/emanuellemarreira/avaliacao-de-coerencia-com-llm

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

conclusões imprecisas. A origem das discrepâncias entre comentários e notas foi investigada por Geierhos et al. [5] que apontam que um dos fatores que contribuem para essa inconsistência são erros aleatórios individuais. Além disso, estudos como os de [12] e [16] sugerem que os consumidores tendem a penalizar experiências negativas com mais severidade do que recompensar experiências positivas nos textos.

Fazzolari *et al.* [3] analisaram a incoerência em avaliações de hotéis, identificando que 12% das notas baixas foram classificadas como positivas e 5% das notas altas como negativas. Para reduzir essa ambiguidade, Islam [8] propôs um sistema que unifica a nota numérica com a polaridade extraída do texto. De forma semelhante, Aljrees*et al.* [1] usaram o TextBlob para identificar vieses em avaliações de aplicativos, classificando como enviesadas 24,72% das avaliações com polaridade abaixo de 0,5 e nota acima de 3.

Esses estudos, em conjunto, evidenciam que os comentários textuais podem ser mais representativos do sentimento real do usuário do que as próprias notas, reforçando a necessidade de abordagens que integrem ambas as fontes de informação.

2.2 *Large Language Models*

LLMs avançaram significativamente o Processamento de Linguagem Natural (PLN), particularmente na compreensão de texto e mineração de opiniões. Sua capacidade de generalizar tarefas por meio de aprendizado *zero-shot* possibilitou aplicações em diversas áreas, incluindo predição de *ratings* [9] e sistemas de recomendação [19].

No contexto de sistemas de recomendação explicáveis, Liu *et al.* [11] propõem uma solução que busca alinhar a nota prevista com a explicação gerada pelo sistema. O modelo utiliza um LLM para prever a nota de um produto com base em informações do usuário e do produto; em seguida, gera uma explicação textual coerente com essa nota. Para avaliar a coerência entre nota e explicação, os autores utilizam o modelo GPT-4o, destacando o potencial dos LLMs na detecção de incoerências entre comentários e notas atribuídas.

Contudo, não foram encontrados trabalhos que utilizem LLMs para analisar a discrepância entre a nota atribuída pelo usuário e seu respectivo comentário. Além disso, embora o português seja uma das cinco línguas mais utilizadas na internet [15], pesquisas sobre o tema focam, principalmente, em inglês. Essa lacuna destaca a necessidade de investigar essa análise em português brasileiro, aproveitando o potencial dos LLMs para aprimorar o entendimento das incoerências entre comentários e notas de avaliações.

3 METODOLOGIA

3.1 Conjunto de Dados

O conjunto de dados deste estudo consiste em comentários de produtos publicados no site de comércio eletrônico Amazon Brasil, abrangendo o período de 2021 a 2024 através de um coletor próprio. Foram considerados apenas comentários em língua portuguesa. A pesquisa focou exclusivamente no texto do comentário e na avaliação numérica fornecida pelo usuário. O objetivo central deste trabalho é identificar contradições entre a nota atribuída e o conteúdo textual da avaliação, restringindo a análise a avaliações extremas, ou seja, comentários com 1 ou 5 estrelas. A decisão de focar exclusivamente em avaliações com notas extremas (1 e 5 estrelas) foi uma escolha metodológica para garantir que o estudo analisasse casos com expectativas de polaridade inequívocas. Essa abordagem permitiu concentrar a investigação na detecção de contradições explícitas entre texto e nota, onde a incoerência e mais crítica. Avaliações intermediárias, com notas como 2 e 4 estrelas, frequentemente contêm sentimentos mistos ou neutros, o que introduziria um nível de ambiguidade que poderia comprometer a objetividade da análise de coerência. Explorar a natureza dessas avaliações intermediárias constitui uma valiosa direção para trabalhos futuros.

Para garantir equilíbrio entre as classes de nota, cada categoria contém o mesmo número de comentários com 1 e 5 estrelas. A categoria *Automotivo* possui 873 avaliações de 1 estrela e 873 de 5 estrelas. O mesmo se aplica às outras categorias: *Bebê* tem 1.057 avaliações de cada nota; *Celulares*, 867; *Alimentos*, 742; *Jogos*, 1.217; *Computadores*, 185; *Livros*, 2.259; *Moda*, 1.443; *Pets*, 445; e *Brinquedos*, 1.205 avaliações de 1 e 5 estrelas. Assim, o total é de 20.586 comentários. Observa-se um maior volume de dados nas categorias *Livros* e *Moda*, enquanto *Computadores* e *Pets* apresentam menor representatividade. Essa variação reflete a popularidade relativa entre categorias, com destaque para *Livros*, possivelmente impulsionada pelo hábito dos leitores de avaliarem suas experiências.

3.2 Modelo Utilizado

O ChatGPT-o3 é um LLM orientado por instrução, baseado na arquitetura Transformer [17]. Assim como outros modelos da família GPT, ele passa por pré-treinamento não supervisionado em larga escala, seguido de alinhamento supervisionado e *Reinforcement Learning from Human Feedback* (RLHF) [14]. Ele incorpora avanços recentes, como raciocínio prévio à geração (*reason-first decoding*) e otimizações que reduzem latência e custo por *token* em comparação ao GPT-4o [7]. Ele suporta até 128 mil *tokens* de contexto e apresenta desempenho competitivo em tarefas de classificação *zero-shot* e *few-shot* em português, conforme resultados em *benchmarks* multilíngues como BLEURT-22, XCOPA e AmericasNLI [18].

Optou-se pelo ChatGPT-o3 por sua robusta cobertura do português, incluindo gírias e ironias comuns em avaliações *online*, além de avanços em tarefas complexas devido ao seu mecanismo de raciocínio interno pré-geração. Considerando que modelos generativos são não-determinísticos, o modelo foi aplicado em cinco rodadas para observar consenso. Para tentar mitigar o não-determinismo inerente a modelos generativos, o parâmetro de temperatura foi configurado para 0.

3.3 Prompt Utilizado

O *prompt* utilizado no experimento, ilustrado na Figura 2, foi estruturado em cinco blocos com funções específicas. O primeiro (1) orienta o modelo a verificar a coerência entre avaliação e texto em português. O segundo (2) define a regra de decisão: uma avaliação é `INCOERENTE` se a nota 1 não corresponder a um texto negativo ou a nota 5 a um texto positivo. Textos neutros ou irônicos sem contradição são considerados `COERENTES`. O terceiro bloco (3) instrui o modelo a usar a coluna categoria apenas para contexto, como gírias, e a determinar o sentimento dominante em casos de opiniões mistas. O quarto bloco (4) especifica o formato de entrada e saída como CSV, com colunas de `id`, `avaliacao`, `categoria` e `texto`. A saída deve adicionar a coluna `classificacao` e manter a ordem original

dos registros. O quinto bloco (5) exige que a saída seja apenas o CSV, sem códigos, comentários ou raciocínio interno, garantindo a compatibilidade com o *pipeline* automatizado. Essa estrutura clara e concisa assegura a uniformidade das respostas e a robustez para uso em cenários *zero-shot*.

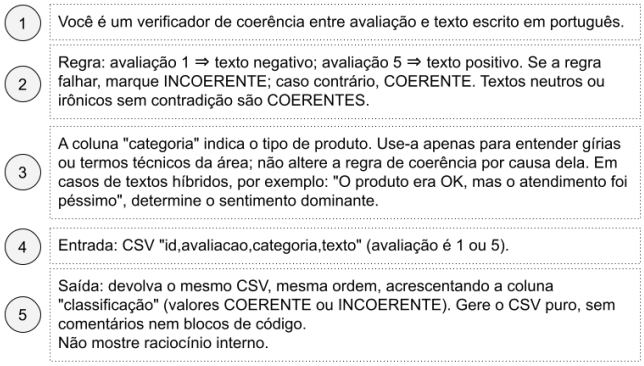

Figura 2: Prompt utilizado nos experimentos.

A combinação desses cinco elementos resulta em um *prompt* conciso, determinístico e operacionalmente robusto, adequado para uso em contexto *zero-shot*. A distinção clara entre a regra de decisão e o formato de resposta foi fundamental para garantir uniformidade nas execuções e permitir análise de consistência intra-modelo. O *prompt* foi definido a partir de experimentos exploratórios, nos quais diferentes formulações foram testadas até se obter uma versão clara, funcional e estável, compatível com o formato CSV exigido pelo protocolo de avaliação automatizada.

3.4 Métricas

Para quantificar a concordância entre as cinco execuções independentes do LLM, adotou-se o coeficiente κ de Fleiss [4]. Trata-se de uma generalização do κ de Cohen para $k > 2$ anotadores, adequada a categorias nominais e à ausência de dados faltantes, requisitos que correspondem exatamente ao cenário deste trabalho, onde cada comentário recebeu um rótulo binário (`INCOERENTE` ou `COERENTE`) proveniente de cinco execuções do ChatGPT-o3.

Definições formais: Seja N o número total de itens avaliados e k o número de avaliadores. Para cada item $i \in \{1, \ldots, N\}$ e categoria $c \in \{1, \ldots, C\}$, denotamos por n_{ic} a contagem de avaliadores que selecionaram a categoria c. No presente trabalho, $C = 2$ e $k = 5$. Com essas notações, a proporção média de acordos observados ($\bar{P}$) e a proporção de acordos esperados ao acaso ($\bar{P}_e$) são dadas por

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{c=1}^{C} n_{ic}(n_{ic} - 1)}{k(k-1)}, \qquad \bar{P}_e = \sum_{c=1}^{C} p_c^2,$$

onde $p_c = \frac{1}{N_k} \sum_{i=1}^{N} n_{ic}$ representa a frequência marginal da categoria c. O coeficiente de Fleiss é então definido por

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}.$$

Valores de κ variam entre -1 e 1, onde $\kappa = 1$ indica concordância perfeita, $\kappa = 0$ equivale a concordância ao nível do acaso e $\kappa < 0$ sugere discordância sistemática. Seguindo a escala de [10], considera-se os intervalos 0–0,20 (fraca), 0,21–0,40 (razoável), 0,41–0,60 (moderada), 0,61–0,80 (substancial) e 0,81–1,00 (quase perfeita).

4 EXPERIMENTOS

4.1 Identificação de Incoerência

A consistência interna do modelo ChatGPT-o3 foi avaliada por meio de cinco execuções independentes (v1-v5) de 20.586 comentários com o mesmo *prompt*. Observou-se uma variação significativa na contagem de votos `INCOERENTE`, com a v2 classificando apenas 5,2% dos comentários como tal e a v3 marcando 52,5%. Apenas 415 itens (2% do corpus) foram unanimemente considerados incoerentes (386 de nota 1 e 29 de nota 5), enquanto 1.754 itens (8,5%) obtiveram pelo menos três votos. Assim, observa-se um alto grau de incerteza do modelo em avaliações repetidas para o mesmo conjunto de dados.

Tabela 1: Total de comentários `INCOERENTE`.

Rodada	Quantidade	Percentual no corpus (%)
v1	2.505	12,2
v2	1.065	5,2
v3	10.812	52,5
v4	1.157	5,6
v5	2.386	11,6

Considerando as cinco execuções simultâneas, o coeficiente κ de Fleiss foi 0,177 ($IC_{95\,\%}$: 0,167–0,188), classificado como "concordância fraca" segundo a escala de Landis–Koch. A Tabela 2 apresenta os κ de Cohen para cada par de execuções; observam-se valores baixos nos pares envolvendo v3 (0,075–0,180), resultado direto de seu viés de prevalência. Em contraste, pares com prevalências mais próximas alcançam concordância moderada (v4–v5, κ = 0,487).

Tabela 2: Matriz de correlação Kappa de Cohen.

	v1	v2	v3	v4	v5
v1	–	0,343	0,180	0,355	0,438
v2	–	–	0,075	0,469	0,414
v3	–	–	–	0,096	0,173
v4	–	–	–	–	0,487
v5	–	–	–	–	–

Os resultados indicam que a grande disparidade nas prevalências — especialmente na execução v3, que classifica mais da metade dos comentários como incoerentes — aumenta significativamente o acordo esperado ao acaso, reduzindo tanto o κ global quanto os índices par-a-par. Esse fenômeno exemplifica o "paradoxo do κ", em que a distribuição das classes é severamente desequilibrada entre avaliadores. Além disso, os resultados mostram que o texto contradiz a nota, com pelo menos 415 itens (2% do corpus) rotulados como incoerentes por todas as cinco execuções do ChatGPT-o3, fornecendo sólida evidência de contradição. Contudo, como a detecção variou muito entre rodadas (prevalência de 5% a 53% e κ de Fleiss = 0,177), o modelo não apresenta estabilidade suficiente para confiar nos demais rótulos. Em síntese, contradições existem e podem ser identificadas pelo LLM, mas, para uso prático, deve-se limitar a análise aos casos de consenso ou recorrer a validação humana antes de qualquer decisão automática.

4.2 Concordância Humana

Adotou-se um critério conservador, selecionando para análise detalhada apenas os 415 comentários unanimemente classificados como `INCOERENTE`, por representarem o subconjunto de maior confiança estatística. Para validar essa avaliação, dois revisores humanos julgaram a polaridade textual utilizando os rótulos NEG (negativo), POS (positivo), NEU (neutro) e MIST (misto). A concordância entre eles foi de 80,5% (exata), com um Coeficiente Kappa de Cohen de 0,709, indicando uma concordância substancial segundo a escala de Landis e Koch. Isso sugere que, embora exista certa subjetividade na tarefa, os julgamentos humanos são relativamente consistentes.

Dos 415 comentários, 231 (55,7%) apresentaram julgamento humano unânime e polaridade clara (POS ou NEG), sendo esses casos usados para a comparação direta com a classificação do LLM. Os 184 casos restantes (44,3%) foram classificados como NEU ou MIST, refletindo ambiguidade, como em *"O conceito é perfeito, mas um livro que poderia ter 30 páginas"*. Esses casos foram excluídos do cálculo direto de coerência binária.

A comparação com os 231 casos de julgamento unânime revelou que apenas 28,1% das classificações do LLM concordaram com o juízo humano. Esse resultado sugere que, embora o LLM capture padrões de linguagem sofisticados, ele apresenta limitações significativas em face de nuances subjetivas, reforçando a importância da validação humana em tarefas de inferência de coerência semântica. Essa análise reforça a importância da validação humana em experimentos envolvendo inferência de coerência semântica, bem como a necessidade de incorporar mecanismos de calibração ou curadoria híbrida para aplicações críticas que dependem da interpretação precisa de opiniões textuais.

4.2.1 Análise da Distribuição de Incoerência por Nota. Para investigar a relação entre nota e a presença de incoerência, analisou-se 231 comentários com julgamento humano unânime e polaridade bem definida (positiva ou negativa). Desse total, 211 comentários (91,3%) eram de 1 estrela e 20 (8,7%) eram de 5 estrelas. Essa assimetria deve-se ao conjunto de dados que já possuía um desequilíbrio entre notas extremas. Após a filtragem por julgamento unânime, observou-se uma concentração de exemplos negativos, sugerindo maior concordância dos avaliadores nos casos de insatisfação. A assimetria também pode ter sido ampliada pelas respostas do LLM, que classificou mais notas baixas como inconsistentes.

Em relação à taxa de incoerência, 28,4% dos comentários de 1 estrela foram considerados incoerentes (texto positivo), enquanto 25% dos comentários de 5 estrelas foram classificados como incoerentes (texto negativo). Embora as taxas sejam próximas, a predominância de exemplos com nota 1 evidencia a fragilidade da correspondência entre texto e avaliação em contextos de insatisfação. Os resultados reforçam a importância de se considerar não apenas a polaridade do texto, mas também a distribuição e a natureza das avaliações ao se aplicar modelos automáticos de verificação de coerência.

4.2.2 Análise por Categoria. A Figura 3 mostra a distribuição de incoerências confirmadas entre texto e nota, segmentada por categoria de produto. A categoria *Pets* apresentou a maior taxa de incoerências (10,1%), superando significativamente as demais, que variaram de 0,9% (*Automotivo*) a 3,2% (*Computadores*).

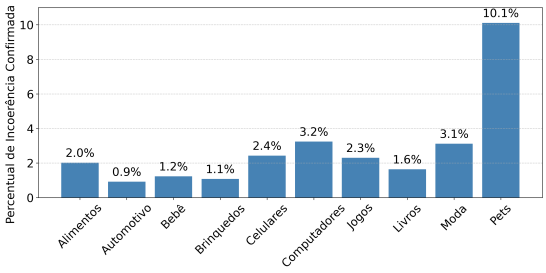

Figura 3: Distribuição de incoerência confirmada.

A discrepância na categoria *Pets* pode ser explicada por características específicas dos produtos ou pelo comportamento dos consumidores, que podem expressar maior complexidade emocional ou avaliações mistas. Isso gera interpretações ambíguas para modelos automatizados. Outra hipótese é a dificuldade em usar corretamente as escalas de *rating*, onde avaliações afetivas podem não refletir a satisfação objetiva. Nas demais categorias, os valores de incoerência são mais baixos e consistentes, sugerindo que o sistema de estrelas reflete adequadamente a polaridade textual.

A conclusão preliminar deste estudo indica uma clara variabilidade inter-categoria, com a categoria "*Pets*" apresentando a maior suscetibilidade à incoerência entre avaliações numéricas e textuais. Futuros estudos devem explorar mais detalhadamente as causas dessa variação, considerando particularidades linguísticas e emocionais associadas a categorias específicas.

5 CONCLUSÕES

Este estudo avaliou o ChatGPT-o3 na detecção de incoerências entre texto e notas (1 ou 5 estrelas) em 20.586 avaliações da Amazon Brasil. Cinco execuções independentes mostraram baixa concordância global (κ = 0,177), com unanimidade em apenas 2% do corpus. A validação humana dos 415 casos revelou concordância substancial (κ = 0,709), mas somente 28,1% dos rótulos do LLM coincidiam com o juízo humano. Incoerências confirmadas concentraram-se na categoria *Pets* (10,1%), sugerindo dependência de domínio, enquanto as demais categorias variaram de 0,9% a 3,2%. Os resultados indicam o potencial dos LLMs como filtro inicial, mas sua variabilidade e precisão limitada exigem curadoria híbrida ou regras de consenso antes de decisões automáticas.

A natureza "caixa-preta" de modelos proprietários como o ChatGPT-o3 impede a confirmação de que os comentários do *corpus* tenham sido parte do conjunto de treinamento, o que poderia introduzir viés. Embora o volume de dados torne esse impacto improvável, essa é uma limitação do estudo. Contudo, a associação entre os textos dos comentários e a tarefa de predição de *ratings* diretamente no conjunto de treino é bem menos provável, o que direciona o modelo a usar conhecimento generalizado para resolver a tarefa.

Para trabalhos futuros, propõe-se investigar três frentes para mitigar essa variabilidade: (i) ajuste fino de hiperparâmetros como temperatura e tamanho do contexto; (ii) desenvolvimento de *prompts* mais restritivos ou de cadeia de raciocínio; e (iii) exploração de abordagens de *ensemble* e curadoria humana para casos ambíguos, visando aumentar a robustez do sistema.

REFERENCES

[1] Turki Aljrees, Muhammad Umer, Oumaima Saidani, Latifah Almuqren, Abid Ishaq, Shtwai Alsubai, Imran Ashraf, et al. 2024. Contradiction in text review and apps rating: prediction using textual features and transfer learning. *PeerJ Computer Science* 10 (2024), e1722.

[2] Amal Almansour, Reem Alotaibi, and Hajar Alharbi. 2022. Text-rating review discrepancy (TRRD): an integrative review and implications for research. *Future Business Journal* 8, 1 (2022), 3.

[3] Michela Fazzolari, Vittoria Cozza, Marinella Petrocchi, and Angelo Spognardi. 2017. A study on text-score disagreement in online reviews. *Cognitive Computation* 9, 5 (2017), 689–701.

[4] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[5] Michaela Geierhos, Frederik Simon Bäumer, Sabine Schulze, and Valentina Stuß. 2015. " I grade what I get but write what I think." Inconsistency Analysis in Patients' Reviews.. In *ECIS*.

[6] Nan Hu, Noi Sian Koh, and Srinivas K Reddy. 2014. Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision support systems* 57 (2014), 42–53.

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[8] Mir Riyanul Islam. 2014. Numeric rating of Apps on Google Play Store by sentiment analysis on user reviews. In *2014 international conference on electrical engineering and information & communication technology*. IEEE, 1–4.

[9] Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474* (2023).

[10] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[11] Shijie Liu, Ruixin Ding, Weihai Lu, Jun Wang, Mo Yu, Xiaoming Shi, and Wei Zhang. 2025. Coherency Improved Explainable Recommendation via Large Language Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12201–12209.

[12] Juan Pedro Mellinas, Juan L Nicolau, and Sangwon Park. 2019. Inconsistent behavior in online consumer reviews: The effects of hotel attribute ratings on location. *Tourism Management* 71 (2019), 421–427.

[13] Susan M Mudambi, David Schuff, and Zhewei Zhang. 2014. Why aren't the stars aligned? An analysis of online review content and star ratings. In *2014 47th Hawaii International conference on system sciences*. IEEE, 3139–3147.

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[15] Denilson Alves Pereira. 2021. A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review* 54, 2 (2021), 1087–1115.

[16] Abhinav Sharma, Sangwon Park, and Juan L Nicolau. 2020. Testing loss aversion and diminishing sensitivity in review sentiment. *Tourism Management* 77 (2020), 104020.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[18] Hetong Wang, Pasquale Minervini, and Edoardo M Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. *arXiv preprint arXiv:2406.13229* (2024).

[19] Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large language models as evaluators for recommendation explanations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 33–42. https://doi.org/10.1145/3640457.3688075