Marketing Letters (2026) 37:30
https://doi.org/10.1007/s11002-026-09824-7

ORIGINAL RESEARCH

# Using LLMs as sentiment analyzers to predict review helpfulness: first insights to open the black box

**Christian G. H. Winter**[1] · **Nicolas A. Zacharias**[1] · **Mattis Hartwig**[2,3]

Received: 22 April 2025 / Accepted: 3 May 2026
© The Author(s) 2026

## Abstract

This study examines the potential of large language models for sentiment analysis in marketing. Using the empirical setting of online customer reviews, we further explore implications for prediction of review helpfulness. Relying on a dataset of 28,900 product reviews from a consumer platform and an experiment with $N = 1063$ participants, we find that the LLM's accuracy in assessing intended meaning (as in the star-rating) in customer-written text depends on the degree of emotionality, as in purchases of hedonic (vs. utilitarian) goods. We further demonstrate that deviations between LLM classification and original star rating predict lower review helpfulness. This effect is mediated by the deviation of human readers' assumption on the intended star rating from the actual star rating and moderated by the degree of information asymmetry before the purchase; that is, a greater deviation between the LLM classification and the original star rating indicates lower review helpfulness for search goods than for experience goods.

**Keywords** Large language model · Online customer reviews · Product type · Review helpfulness · Sentiment analysis

✉ Christian G. H. Winter
christian.winter@uni-jena.de

Nicolas A. Zacharias
nicolas.zacharias@uni-jena.de

Mattis Hartwig
mattis.hartwig@dfki.de; mattis.hartwig@singular-it.de

[1] Faculty of Economics and Business Administration, Chair of Marketing, Friedrich Schiller University Jena, Carl-Zeiß-Straße 3, 07743 Jena, Germany

[2] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Maria-Goeppert-Straße 15, 23562 Lübeck, Germany

[3] singularIT GmbH, Inselstraße 27, 04103 Leipzig, Germany

Published online: 22 May 2026

## 1 Introduction

Advances in large language models (LLMs) have ushered in a new state-of-the-art in the field of artificial intelligence. An area likely to benefit from this technology is sentiment analysis, which is used for a variety of applications in marketing research and practice (e.g., Guler et al., 2024). Previously, to assess the sentiment expressed in text (e.g., created by customers), researchers and companies had to rely on specifically designed dictionaries (e.g., Short et al., 2010), which are resource-intensive to create and face difficulties interpreting context, sarcasm, and ambiguity (Khoo & Johnkhan, 2018), and often require large sample sizes to allow for meaningful conclusions (Chapman, 2020). Some machine learning methods have been demonstrated to perform better, but they typically require resource-intensive calibration and feature engineering for each specific use case (Tang et al., 2015). Early approaches show general potential for LLMs to conduct sentiment analysis without any resource-intensive specifications (i.e., zero-shot classification) (e.g., Yang et al., 2024). However, boundary conditions for LLMs to perform such analyses are still unclear.

One promising business application for easy-to-implement sentiment analysis is handling customer interactions (e.g., Xiao & Kumar, 2021). In this context, assessing customers' satisfaction and emotions precisely is crucial for an appropriate response (Menon & Dubé, 2000). AI systems in marketing are less trusted to handle situations with high levels of customer emotions (Huang & Rust, 2024); thus, the emotional context of the interaction is a likely boundary condition for sentiment analysis effectiveness. Such emotional context depends on the purpose for which a good is bought; for example, hedonic goods purchases are driven by affect based on expected sensory pleasure (Longoni & Cian, 2022), whereas purchases of utilitarian goods are more cognitively driven, based on the degree to which a product is a means to an end (Longoni & Cian, 2022). Previous research has demonstrated that AI systems generally perform well in a utilitarian context but tend to struggle in hedonic contexts (Longoni & Cian, 2022; Xiao & Kumar, 2021). We therefore investigate the degree of emotionality as represented by product type (i.e., hedonic vs. utilitarian goods) as a boundary condition for LLM-driven sentiment analysis, leading to the following research question:

RQ1: Does LLMs' accuracy in assessing the intended meaning (as in the star-rating) in customer-written text depend on the degree of emotionality of the context?

Answering this question can help businesses deal with a type of text-based customer interaction especially important for the success of online platforms, shop operators, and third-party sellers: online customer reviews (Ceylan et al., 2024). Such reviews mitigate customers' limitations in physically experiencing the product of interest and are often trusted as non-company-sponsored authentic information (Choi & Leon, 2020). However, not all reviews are equal; to influence

customers' perceptions and decision making, a review needs to be written such that readers perceive it as helpful (Mudambi & Schuff, 2010).

Any given product on customer platforms likely has too many reviews to comprehend, causing adverse "review overload" (Carlson et al., 2023). Because only a small number of users vote on review helpfulness, helpful reviews quickly lose visibility. This reduction in available high-quality information lowers customer experience, which in turn harms key economic outcomes for platform operators (Ceylan et al., 2024). Therefore, ranking reviews so that customers see helpful reviews first is important, and the ability to predict helpfulness immediately after the review has been published is critical (Guha Majumder et al., 2022).

Reviews are helpful to customers when they reduce information asymmetry (Kübler et al., 2024), especially when they provide cognitively easy-to-process and internally consistent information (Mudambi et al., 2014; Wang et al., 2025). However, less consistent and more ambiguous review texts may not only increase cognitive processing costs for human readers but also pose challenges for LLM processing (Liu et al., 2023). This suggests that reviews that are difficult for LLMs to process may also be less helpful for humans.

However, the importance of cognitively easy-to-process review information is likely to vary, as information asymmetry is not evenly distributed across products. For some products, core qualities can be easily assessed using information provided by the seller (search goods), whereas other products must be experienced firsthand to evaluate their quality (experience goods) (Hsieh et al., 2005). Consequently, readers' information needs and search strategies differ (Kübler et al., 2024), making them more reliant on consistent and cognitively easy-to-process information when researching experience goods. Against this backdrop, we investigate the following second research question:

RQ2: Do deviations in LLMs' classification of a review's star rating from the actual review's star rating indicate lower review helpfulness, and is this effect contingent on information asymmetry before the purchase?

By answering our research questions, we provide two main contributions to marketing research and practice. First, we explore the application of LLMs in sentiment analysis, investigating their potential and boundary conditions to address challenges in the field (e.g., resource-intensive dictionary validation, context dependency, susceptibility to small sample sizes; Chapman, 2020), informing both researchers aiming to include sentiment analysis in their research and (small) businesses aiming to measure their customers' opinions on certain aspects of their business without intensive configuration. Second, we investigate whether LLMs can predict review helpfulness. We thereby inform the literature identifying antecedents and predictors of review helpfulness (e.g., Park et al., 2021) and provide platform and shop operators guidance on how to identify helpful and non-helpful customer reviews early. This guidance is particularly relevant for small and medium-sized enterprises, as our highly accessible but effective approach deliberately relies on an existing model in a zero-shot configuration and does not require task-specific training or fine-tuning.

## 2 Overview of studies

We investigate our research questions by conducting a set of three studies. Study 1 addresses RQ1 by identifying whether LLMs and humans share a common "sense of satisfaction"—that is, whether a LLM rater and human reviewers agree on the valence of a customer text and if deviations are related to the hedonic–utilitarian context of the product. To do so, we use a set of 28,900 customer reviews on products whose star rating we compare with a LLM's prediction on the respective star rating. Studies 2 and 3 address RQ2. Connecting the dataset from study 1 with data on review helpfulness, study 2 examines when high deviations between human author and LLM rater indicate review helpfulness, and whether this helpfulness prediction is contingent on the ease of researching core product properties (experience vs. search goods). Study 3 experimentally validates the results of study 2 and reveals the underlying mechanism, demonstrating that deviations between human authors' and the LLM's predictions of the star rating indicate lower review helpfulness (Fig. 1).

## 3 Data collection and measures

We rely on two data sources. First, we constructed a comprehensive dataset of 28,900 reviews from a consumer platform, which serves as the empirical basis for all three studies. We drew from previous work categorizing products alongside the context they are bought in (hedonic vs. utilitarian goods) and the ease of researching

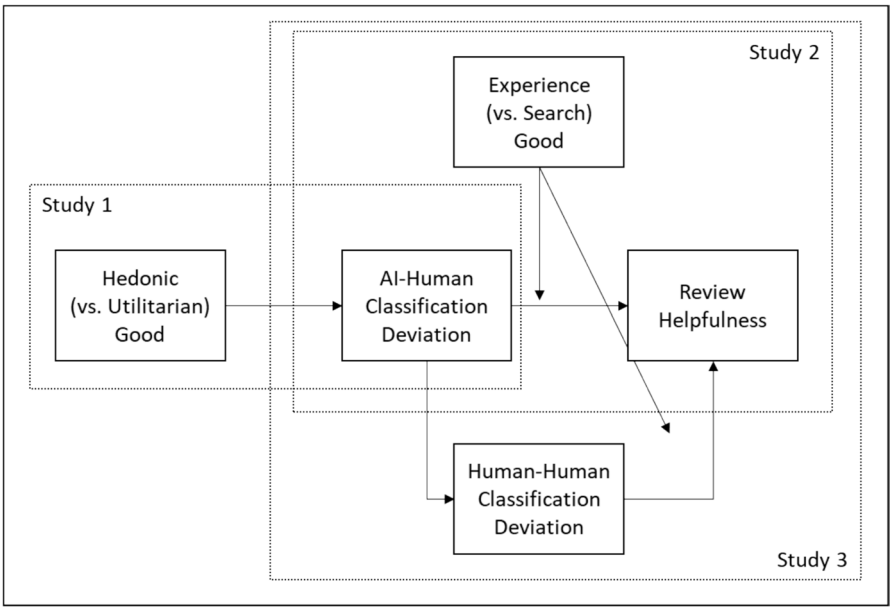


**Fig. 1** Overview of studies

their properties (search vs. experience goods; Kübler et al., 2024). Examples are smartphone (search good, hedonic), printer (search good, utilitarian), mascara (experience good, hedonic), and vacuum cleaner (experience good, utilitarian). To collect our sample, we used Kübler et al. and and's (2024, p. 7) list of 12 products as search terms on a global consumer platform, where we collected all publicly available product reviews including text, rating, and meta information for all products on the first page of results for each search term. Removing reviews without any text resulted in a comprehensive final sample of 28,900 reviews. We then dummy-coded reviews as *experience good* (vs. search good) and as *hedonic good* (vs. utilitarian good) according to Kübler et al. (2024). Further, we measured *review length* as the number of characters of the review text (Hong et al., 2017), *review age* as the number of days the review had been online on the platform at the day of data collection, *rating* of the review as dummy variables for each of the five possible star ratings (Ceylan et al., 2024), and *review helpfulness* as the number of helpful votes on a particular review at the time of data collection. Additionally, we dummy-coded the presence of *images* or *videos* in the reviews (Kübler et al., 2024).

We then fed the reviews into the latest version (at the time of data collection) of the most widely disseminated LLM tool globally (GPT-4), using the OpenAI API (OpenAI 2024). For every review, we generated a client session and gave it a system-level prompt instructing it to the review's star rating based on the review's text. This input provided *LLM star classification*. Next, we recorded the absolute deviation between the rating assigned by the review author and the LLM star classification as *AI–human classification deviation.*

To ensure rating consistency, we set a relatively low temperature of.2. The temperature parameter of an LLM controls the degree of randomness and is associated with creativity of results (Peeperkorn et al., 2024). A high temperature results in less probable results in the generated output. Other studies show that for tasks that require individual answers (as in our case, a single scoring), low temperature values produce better results (Xu et al., 2022). The Technical Appendix provides the R-script used to call the OpenAI API, including the system-level prompt.

For study 3, we additionally collected experimental evidence; based on the 28,900 reviews in our first dataset, we created a stratified random sample (Kovács, 2024). To ensure an unbiased experimental setting, we accounted for the presence of each category of star rating and experience (vs. search) goods and hedonic (vs. utilitarian) goods by creating a total of 5 (star ratings) × 2 (experience/search goods) ×2 (hedonic/utilitarian goods) = 20 bins. From each bin, we drew 20 reviews at random, resulting in an experimental subsample of 400 out of the 28,900 reviews.

We recruited 1100 participants from Prolific for our experiment. After we eliminated 37 individuals who failed an attention check or spent more than three standard deviations over the mean time on the task, $N = 1063$ participants (54.9% female, average age 41.42 years, all native English speakers living in the USA) remained in the final sample. They viewed one of those 400 reviews and information on the type of product (e.g., mascara) being reviewed, but no information on the review's star rating. They then indicated which star rating they believed the author had assigned to the review, which allowed us to calculate the absolute deviation between human reader classification and human author classification as *human–human classification*

*deviation*. This task does not represent a decision situation that naturally occurs in practice. Rather, it serves as a deliberately controlled experimental device to isolate the mechanism linking AI–human classification deviation to review helpfulness by allowing us to compare how human participants and the LLM classify the same review content under the same informational constraints.

We furthermore assessed *review helpfulness* by asking participants to rate perceived review helpfulness. Corresponding to the five-level star rating of a review, we measured all responses on five-point Likert scales.

## 4 Study 1

Study 1's objective was to learn whether LLMs can correctly identify a customer review's sentiment as expressed by its star rating. Further, we investigated whether deviations in AI prediction from the actual star rating are greater for hedonic than for utilitarian contexts.

### 4.1 Method

We investigated the relationship between the rating assigned by the human review author and *LLM star classification* relying on correlational analysis. We further conducted a simple regression analysis with *LLM star classification* as the independent variable and *rating* as the dependent variable.

Next, to identify factors driving these deviations, we performed regression analysis with *AI–human classification deviation* as the dependent variable. In hedonic (vs. utilitarian) contexts, product reviews might be driven by more (less) heterogeneous customer preference structures; therefore, we included *hedonic* as our central explanatory variable. Further, because for different degrees of valence, review structure and text might differ systematically, we also added dummy variables for each star rating, with a star rating of 1 as the reference category. Following Mudambi & Schuff, (2010), we also added *review length* (see Web Appendix 1 for details on model specifications).

### 4.2 Results and discussion

To predict review helpfulness from AI–human classification deviation, LLM ratings must both generally align with human judgments and exhibit sufficient disagreement to allow meaningful variation. A pre-study confirms both conditions: human and LLM ratings are highly correlated ($r = .763$, $p < .001$), indicating a shared "sense of satisfaction," while 37% of reviews are misclassified and about 5% differ by two or more rating points. While zero-shot LLM classification is substantially easier to implement, benchmarking results show that this approach achieves accuracy comparable to, or even exceeding, that of four commonly used fine-tuned open-source machine learning models, namely TF-IDF in combination with logistic regression,

XGBoost, or random forest, and a sentence transformer in combination with CatBoost (see Web Appendix 2).

Building on these findings, we next investigated the hedonic (vs. utilitarian) context of the product as a predictor of the rating deviations noted in the first part of this study (see Table 1 for the respective regression analysis). Interestingly, despite the higher degree of emotionality (see Sect. 1), AI–human classification deviation is smaller for hedonic than for utilitarian products ($\beta = -.058$, $p < .001$). A possible explanation for this surprising result might be that authors reviewing hedonic products more strongly emphasize the subjective nature of their consumption experience, using explicitly evaluative language (e.g., "great!"). To further examine this interpretation, we conducted an additional post hoc analysis, revealing that reviews of hedonic products indeed are significantly more subjective than reviews of utilitarian products ($\beta = -.018$, $p < .001$, measure obtained with *TextBlob*), whereas no significant difference emerges for reading complexity (Flesch-score, $\beta = -.020$, $p = .938$, Flesch, 1948). This pattern suggests that the observed differences are driven by subjectivity rather than by general linguistic complexity.

Results also reveal that AI-human-classification deviation is more likely for reviews associated with lower ratings (particularly one-star reviews), for longer reviews, and for more recently published reviews. These findings are in line with recent literature indicating temporal shifts affecting the meaning of star-rating scales (Fang et al., 2024).

## 5 Study 2

Study 2's objective was to investigate whether the extent of AI–human classification deviation negatively predicts the amount of helpfulness votes a review receives. We further researched whether this negative effect is stronger when the degree of information asymmetry before the purchase is high, as in experience (vs. search) goods.

**Table 1** Regression results study 1

| Variables | Coefficient | SE | Beta | *t* | *p* |
|---|---|---|---|---|---|
| Controls | | | | | |
| Intercept*** | .739*** | .023 | | 32.207 | <.001 |
| Two-star-rating*** | −.189*** | .032 | −.048 | −5.982 | <.001 |
| Three-star rating*** | −.295*** | .025 | −.138 | −11.742 | <.001 |
| Four-star rating** | −.066** | .023 | −.045 | −2.797 | .005 |
| Five-star rating*** | −.341*** | .023 | −.268 | −15.029 | <.001 |
| Review length*** | .000*** | .000 | −.052 | −8.907 | <.001 |
| Review age | −.000 | .000 | −.013 | 2.080 | .380 |
| Main effect | | | | | |
| Hedonic good*** | −.058*** | .007 | −.047 | −8.156 | <.001 |

Notes: ** $p < .01$; *** $p < .001$. $N = 28{,}900$. Dependent variable: AI–human classification deviation. *SE*, standard error

### 5.1 Method

We linked the dataset from study 1 with data on review helpfulness votes, which serves as our dependent variable. As mentioned previously, internally inconsistent language should increase readers' cognitive load, thereby likely reducing a review's helpfulness. Assuming that high cognitive load statements are of little help to customers, we included AI–human classification deviation as our independent variable to explain review helpfulness. We further expect this negative effect to be especially pronounced for experience (vs. search) goods: Because researching these goods' core qualities before purchase is difficult, we assume customers to be especially dependent on cognitively easy to process reviews. We therefore included the interaction term of review helpfulness and experience. We further include the rating dummies (Ceylan et al., 2024), review length (Hong et al., 2017) and the presence of images or videos (Ceylan et al., 2024) and review age as controls.

Our dataset likely is affected by data censorship and selection bias. Selection bias might occur because customers encountering extreme rather than moderate experiences are more likely to publish online customer reviews (Brandes et al., 2022). Further, readers can only process a limited amount of reviews, making it unclear whether readers did not vote on a particular review because they found it unhelpful or because they did not read it (Kübler et al., 2024). Therefore, being part of the sample might correlate with explanatory variables (Mudambi & Schuff, 2010).

Data censorship can occur when the dependent variable is bound in its range (i.e., it cannot fall below zero) and is therefore limited to the extremes (Srivastava & Kalro, 2019). Helpfulness cannot be voted down, even if it is perceived as completely unhelpful (Kübler et al., 2024), causing a heavily skewed distribution. Tobit models can account for both these biases (Mudambi & Schuff, 2010). We therefore estimate a Tobit model with left-censoring at zero, with standard errors clustered at the product (ASIN) level (see Web Appendix 3 for more details on model specification).

### 5.2 Results

Table 2 contains the results of the respective stepwise Tobit regression. Model 1 estimates the direct effect of AI–human classification deviation on review helpfulness alongside the control variables. Results do not reveal a direct effect ($\beta = -.650$, $p = .428$). However, including the interaction effect of AI–human classification deviation and experience good in model 2 reveals that review helpfulness can be expected to be significantly lower per unit of AI–human classification deviation for experience goods than for search goods ($\beta = -7.239$, $p < .001$), indicating that for experience goods readers indeed are more reliant on easy-to-process, internally consistent information. As an additional robustness check, the analysis was replicated based on AI-human classification obtained with a later version of the LLM (GPT-5-nano). All results remain stable (see Web Appendix 4).

**Table 2** Tobit regression results

| Variable | Model 1 | | | Model 2 | | |
|---|---|---|---|---|---|---|
| | Coefficient | SE | *p* | Coefficient | SE | *p* |
| Controls | | | | | | |
| Intercept | − 62.681*** | 13.679 | <.001 | − 69.896*** | 15.215 | <.001 |
| Review length | .059*** | .011 | <.001 | .061*** | .012 | <.001 |
| Two-star rating | − 22.739*** | 5.203 | <.001 | − 24.777*** | 5.458 | <.001 |
| Three-star rating | − 2.902 | 4.784 | .544 | −4.780 | 4.793 | .319 |
| Four-star rating | 10.179 | 5.423 | .061 | 8.672 | 5.306 | .102 |
| Five-star rating | 1.519 | 4.962 | .759 | −.535 | 4.863 | .969 |
| Images | 27.436*** | 5.106 | <.001 | 26.006*** | 4.812 | <.001 |
| Videos | −.210 | 3.921 | .957 | .502 | 3.734 | .893 |
| Review age | .015*** | .003 | <.001 | .015*** | .004 | <.001 |
| Review count[a] | −.156* | .070 | .026 | −.163* | .068 | .017 |
| Direct effects | | | | | | |
| AI–human classification deviation | −.650 | .821 | .428 | 4.345** | 1.776 | .014 |
| Experience good | | | | − 12.825*** | 3.982 | <.001 |
| Interaction effect | | | | | | |
| AI–human classification deviation×experience good | | | | − 7.239*** | 2.148 | <.001 |

Note: Dependent variable: review helpfulness. * $p<.05$, ** $p<.01$; *** $p<.001$. $N=28{,}900$. Left-censored at 0: 15,529. *SE*, standard error. One-star rating serves as a reference category. SEs clustered at the product level. [a]in thousands

## 6 Study 3

Study 3's objective is to investigate whether human–human classification deviation serves as a mediator for the negative relationship of AI–human classification deviation and review helpfulness established in study 2, which is moderated by experience (vs. search) goods. Further, study 3 replicates the results of study 2 in an experimental setting not suffering from data censorship.

### 6.1 Method

AI–human classification deviation might be due to the review text's inconsistent, hard-to-process information, which should also make it harder for human readers to extract the true level of satisfaction a review author intended to convey. This should cause a greater mental load, causing the review's helpfulness to diminish. For experience (vs. search) goods, for which researching the product's core qualities is harder, this effect should be stronger than for search goods.

In alignment with Studies 1 and 2, we added the respective review rating and review length as control variables. In a similar experimental setting, Kovács (2024)

shows that age plays a role in how participants interact with reviews; therefore, we added *age* of participants in years as an additional control variable. Equations for the estimated moderated mediation can be found in Web Appendix 5.

### 6.2 Results and discussion

Table 3 summarizes the results. AI–human classification deviation indicates greater human–human classification deviation ($\beta$=.394, $p$<.001), which in turn reduces review helpfulness ($\beta$= −.150, $p$=.015). When controlled for human–human classification deviation, AI–human classification does not predict review helpfulness ($\beta$= −.001, $p$<.996), indicating a full mediation of the effect (Zhao et al., 2010). Although a significant interaction term indicates that AI–human classification deviation reduces helpfulness more for experience goods than for search goods ($\beta$= −.268, $p$=.021), human–human classification deviation indicates greater review helpfulness for experience goods ($\beta$= −.173, $p$=.049). These ambiguous findings raise questions on the nature of the total effect of AI–human classification deviation on review helpfulness for experience and search goods.

Bootstrapping with 5000 samples confirms that AI–human classification deviation indeed indicates lower review helpfulness for search goods (−.059,

**Table 3** Regression coefficients: study 3

| | Dependent variable | | | | | |
|---|---|---|---|---|---|---|
| | Human–human classification deviation | | | Review helpfulness | | |
| | Coefficient | SE | *P* | Coefficient | SE | *p* |
| Controls | | | | | | |
| Intercept | .287** | .091 | .002 | 3.881*** | .140 | <.001 |
| Age | .002 | .002 | .193 | .001 | .002 | .645 |
| Review length | .000 | .000 | .194 | .001*** | .001 | <.001 |
| Two-star rating | .262*** | .071 | <.001 | .117 | .103 | .256 |
| Three-star rating | .230** | .072 | .001 | .006 | .104 | .952 |
| Four-star rating | .103 | .071 | .151 | −.119 | .102 | .248 |
| Five-star rating | .068 | .074 | .363 | −.131 | .107 | .229 |
| Direct effects | | | | | | |
| AI–human classification deviation | .394*** | .040 | <.001 | −.001 | .082 | .996 |
| Human–human classification deviation | | | | −.150* | .062 | .015 |
| Experience good | | | | −.045 | .094 | .628 |
| Interaction effects | | | | | | |
| AI–human classification deviation×experience good | | | | −.268* | .116 | .021 |
| Human–human classification deviation× experience good | | | | .173* | .088 | .049 |

Notes: * $p$<.05; ** $p$<.01; *** $p$<.001. $N$=1063. One-star rating serves as a reference category

95CI[−.111; −.013]), while for experience goods the data show no effect (−.009, 95CI [−.040;.058]). The effects significantly differ from each other (index of moderated mediation:.068, 95CI [.001;.141]), indicating a moderated mediation of the effect of AI–human classification deviation on review helpfulness, mediated by human–human classification deviation and moderated by experience (vs. search) goods.

## 7 Implications for theory and practice

Three studies demonstrate that an LLM generally can predict customer review valence. However, deviations between LLM estimates and actual star ratings are more likely for utilitarian than for hedonic goods, likely due to less subjective and evaluative language. Deviations are also more prevalent for reviews with low ratings (1-star ratings), which are longer rather than short, or have been published relatively recently. We further demonstrate that such deviations systematically indicate lower review helpfulness. A follow-up experiment demonstrates that deviations in LLM classification from the actual review rating predict lower review helpfulness. Interestingly, while analysis of 28,900 reviews from a global consumer platform indicates that this effect is especially strong for experience goods, the follow-up experiment indicates the consequences of a deviation of the LLM's classification from the actual star rating for review helpfulness are stronger for search goods. These results yield several implications for practitioners and researchers:

First, we contribute to establishing LLMs as tools for sentiment analysis in marketing theory and practice. LLMs offer significant advantages over established procedures for sentiment analysis (e.g., validated dictionaries) because they are less susceptible to confounding aspects (e.g., small sample sizes, context, slang words, abbreviations; Krugmann & Hartmann, 2024) and do not require resource- and time-intensive dictionary validation. We demonstrate that, even without any calibrations, easily accessible LLMs can achieve good results in assessing a customer's sentiment. By deliberately relying on a zero-shot application of an existing model without task-specific training or fine-tuning, our approach keeps data, computational, and infrastructural requirements low, making it broadly applicable across contexts. For scholars in marketing, strategy, and beyond, we provide an easy-to-access additional methodological perspective on various research questions, potentially making data sources accessible for sentiment analysis that were previously deemed unsuitable for such analysis. Marketing managers could employ LLM-driven sentiment analysis to gather their customers' opinions on aspects of their brand, analyze marketing campaign effectiveness, or gain insights on the technical roots of customer complaints (Chapman, 2020).

Second, we demonstrate that although LLM-driven sentiment analysis generally performs well, its effectiveness might depend on the context of the text in terms of its degree of emotionality. Our finding that LLM star classification is less accurate when analyzing reviews for utilitarian goods than hedonic goods indicates that LLM applications relying on sentiment analysis of customer texts, like synthesis reviews (Carlson et al., 2023), or automated responses to customer

reviews, might be more suitable in hedonic contexts. In any case, accounting for systematic biases is essential to prevent responses that are inappropriate for the specific customer situation (e.g., being overly apologetic when not warranted). Further, our finding that LLMs process certain types of information less accurately than others underscores the importance of accounting for content categories when assessing classification performance of algorithms, as well as for temporal shifts affecting the meaning and interpretation of scales (Fang et al., 2024).

Third, we show that deviations between LLMs' classification of a review's star rating and its actual star rating indicate lower review helpfulness because they predict human reader misclassifications. Such misclassifications likely are due to higher cognitive processing cost of inconsistent, hard-to-process information. We thereby uncover a new mechanism through which review helpfulness manifests. For shop and platform operators, our results provide a new and resource-efficient method to assess a review's helpfulness immediately after publication, allowing sorting algorithms to present customers with recent and helpful reviews first. As only a few customers vote on review helpfulness, an early indication on review helpfulness could disseminate relevant information more quickly, which has the potential to significantly affect financial performance (Guha Majumder et al., 2022). Because this assessment relies on a lightweight, zero-shot application of an existing LLM and does not require task-specific training, fine-tuning, or historical helpfulness votes, it remains scalable and easy to implement even for small and medium-sized firms with limited data and technical resources.

As with all studies, this paper has limitations that offer avenues for future research. First, although we control for review age, temporal shifts in the meaning and interpretation of review scales may affect AI–human classification deviations differently over time (Fang et al., 2024). Future research could examine whether more time-sensitive measures improve the prediction of review helpfulness. Second, this study relies on a specific LLM version. Our results remained stable when retested with a later model, while more advanced LLMs' classification rationales increasingly resemble human reasoning. Still, changes in available technology may influence the practical relevance of our findings. Future research should therefore replicate our approach using newly released models. Third, we argue that AI-human classification deviation and review helpfulness for humans may be linked due to inconsistent, hard-to-process language. Future research might empirically test this assumption specifically. Lastly, this paper deliberately focuses on a lightweight approach that relies on an existing model in a zero-shot configuration, requiring no task-specific training or fine-tuning. However, future research might investigate the effectiveness of more complex, fine-tuned multimodal models as a complementary extension.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11002-026-09824-7.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Data Availability** The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Ethical approval** Not applicable.

**Informed consent** Not applicable.

**Competing interests** All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Brandes, L., Godes, D., & Mayzlin, D. (2022). Extremity bias in online reviews: The role of attrition. *Journal of Marketing Research, 59*, 675–695.

Carlson, K., Kopalle, P. K., Riddell, A., Rockmore, D., & Vana, P. (2023). Complementing human effort in online reviews: A deep learning approach to automatic content generation and review synthesis. *International Journal of Research in Marketing, 40*, 54–74.

Ceylan, G., Diehl, K., & Proserpio, D. (2024). Words meet photos: When and why photos increase review helpfulness. *Journal of Marketing Research, 61*, 5–26.

Chapman, C. (2020). Commentary: Mind your text in marketing practice. *Journal of Marketing, 84*, 26–31.

Choi, H. S., & Leon, S. (2020). An empirical investigation of online review helpfulness: A big data perspective. *Decision Support Systems, 139*, Article 113403.

Fang, L., Wu, C., & Sun, B. (2024). *Shifting standards or changing experiences? Unraveling review polarization via LLMs*. Available at SSRN: https://doi.org/10.2139/ssrn.5004249

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Guha Majumder, M., Dutta Gupta, S., & Paul, J. (2022). Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. *Journal of Business Research, 150*, 147–164.

Guler, A. U., Misra, K., & Singh, V. (2024). Local market reaction to brand acquisitions: Evidence from the craft beer industry. *Marketing Science*. https://doi.org/10.1287/mksc.2022.0383

Hong, H., Xu, D., Wang, G. A., & Fan, W. (2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems, 102*, 1–11.

Hsieh, Y.-C., Chiu, H.-C., & Chiang, M.-Y. (2005). Maintaining a committed online customer: A study across search-experience-credence products. *Journal of Retailing, 81*, 75–82.

Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing*. https://doi.org/10.1177/00222429231224748

Khoo, C. S. G., & Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science, 44*, 491–511.

Kovács, B. (2024). The Turing test of online reviews: Can we tell the difference between human-written and GPT-4-written online reviews? *Marketing Letters, 35*(4), 651–666.

Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of generative AI. *Customer Needs and Solutions, 11*(1), 3.

Kübler, R. V., Lobschat, L., Welke, L., & van der Meij, H. (2024). The effect of review images on review helpfulness: A contingency approach. *Journal of Retailing, 100*, 5–23.

Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N., & Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 790–807). Singapore. Association for Computational Linguistics.

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing, 86*, 91–108.

Menon, K., & Dubé, L. (2000). Ensuring greater satisfaction by engineering salesperson response to customer emotions. *Journal of Retailing, 76*, 285–307.

Mudambi, S., & Schuff, D. (2010). Research note: What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*, 185–200.

Mudambi, S. M., Schuff, D., & Zhang, Z. (2014). Why aren't the stars aligned? An analysis of online review content and star ratings. *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, 2014,pp. 3139-3147. https://doi.org/10.1109/HICSS.2014.389

Park, S., Shin, W., & Xie, J. (2021). The fateful first consumer review. *Marketing Science, 40*, 481–507.

Peeperkorn, M., Kouwenhoven, T., Brown, D., & Jordanous, A. (2024). Is temperature the creativity parameter of large language models?. https://doi.org/10.48550/arXiv.2405.00492

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA). *Organizational Research Methods, 13*, 320–347.

Srivastava, V., & Kalro, A. D. (2019). Enhancing the helpfulness of online consumer reviews: The role of latent (content) factors. *Journal of Interactive Marketing, 48*, 33–50.

Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: Successful approaches and future challenges. *Wires Data Mining and Knowledge Discovery, 5*, 292–303.

Wang, D., Xia, Q., Feng, Y., & Cheng, T. (2025). Unravelling the effects of two inconsistencies on online review helpfulness: Evidence from TripAdvisor. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2025.114450

Yang, L., Wang, Z., Li, Z., Na, J.-C., & Yu, J. (2024). An empirical study of multimodal entity-based sentiment analysis with ChatGPT: Improving in-context learning via entity-aware contrastive learning. *Information Processing & Management*, *61*(4), 103724, ISSN 0306-4573.https://doi.org/10.1016/j.ipm.2024.103724

Xiao, L., & Kumar, V. (2021). Robotics for customer service: A useful complement or an ultimate substitute? *Journal of Service Research, 24*, 9–29.

Xu, F. F., Alon, U., Neubig, G., & Hellendoorn, V. J. (2022). A systematic evaluation of large language models of code. In S. Chaudhuri & C. Sutton (Eds.), *MAPS '22: 6th ACM SIGPLAN International Symposium on Machine Programming, San Diego CA USA, 13 06 2022 13 06 2022* (pp. 1–10). ACM.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*, 197–206.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.