Reference Points in LLM Sentiment Analysis: The Role of Structured Context

Junichiro Niimi

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Reference Points in LLM Sentiment Analysis: The Role of Structured Context

Junichiro Niimi[1,2]
Faculty of Business Management, Meijo University,
1-501, Tempaku-ku, Nagoya, Aichi 4688502, Japan
RIKEN AIP, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
jniimi@meijo-u.ac.jp

Abstract

Large language models (LLMs) are now widely used across many fields, including marketing research. Sentiment analysis, in particular, helps firms understand consumer preferences. While most NLP studies classify sentiment from review text alone, marketing theories, such as prospect theory and expectation–disconfirmation theory, point out that customer evaluations are shaped not only by the actual experience but also by additional reference points. This study therefore investigates how the content and format of such supplementary information affect sentiment analysis using LLMs. We compare natural language (NL) and JSON-formatted prompts using a lightweight 3B parameter model suitable for practical marketing applications. Experiments on two Yelp categories (Restaurant and Nightlife) show that the JSON prompt with additional information outperforms all baselines without fine-tuning: Macro-F1 rises by 1.6% and 4% while RMSE falls by 16% and 9.1%, respectively, making it deployable in resource-constrained edge devices. Furthermore, a follow-up analysis confirms that performance gains stem from genuine contextual reasoning rather than label proxying. This work demonstrates that structured prompting can enable smaller models to achieve competitive performance, offering a practical alternative to large-scale model deployment.

1 Introduction

1.1 Background

In recent years, with the rapid advancements in large language models (LLMs; Brown et al., 2020), both industrial and academic areas in wide range of domains have utilized LLMs for data analytics, automation, and decision support. In particular, due to their high applicability in textual data, many studies have implemented sentiment analysis using LLMs (cf. Krugmann and Hartmann, 2024).

LLMs indeed demonstrate remarkable capabilities in understanding textual context; however, the actual 'context' is referred to as a relationship between the tokens, which is captured through Transformer (Vaswani et al., 2017) and attention mechanisms (Bahdanau et al., 2015) and most existing approaches limit their analysis to the linguistic context within review texts alone. Regarding real-world marketing applications, the actual context of consumer evaluation contains the factors which extend far beyond the written review, such as past purchasing patterns, prior experiences with the business, comparative evaluations against competitors, and opinions from social media.

This gap is particularly relevant in customer relationship management (CRM; Oliver, 1999; Reinartz et al., 2004), where understanding customer sentiment accurately drives business decisions. Marketing research has long established through prospect theory (Kahneman and Tversky, 2013) and expectation-disconfirmation theory (EDT; Oliver, 1980) that consumers evaluate experiences relative to these broader reference points. This insight remains largely unexplored in LLM-based sentiment analysis.

Furthermore, practical deployment, particularly for real-time recommendation, has two critical challenges. First, regarding computational efficiency, many business cannot deploy large-scale models (over 70 billion parameters) due to latency and infrastructure constraints. Second, despite having rich contextual data (e.g., user contents, browsing history), current methods cannot efficiently incorporate this information into LLM-based sentiment analysis.

1.2 Research Gap

Despite LLMs' ability to process diverse input formats, sentiment analysis studies predominantly focus on review text alone (e.g., You et al., 2015; Flanagin and Metzger, 2013; Sparks and Brown-

ing, 2011). However, real-world platforms possess rich contextual information. From these gaps, we sequentially derive four research questions (RQs 1–4):

RQ1 Reference-point utilization: Does supplying user- and business-average ratings actually help an LLM classify sentiment more accurately?

RQ2 Prompt format: If the same information is presented in a machine-readable structure or plain texts, does the prompt format affect the model performance?

RQ3 Proxy effect: If supplying the reference points improve accuracy, is it due to their implicit encoding of the ground-truth labels?

RQ4 Reference interactions: How do interactions between multiple reference points affect prediction accuracy?

We address these RQs by three experimental studies. In Study 1, we set up two approaches for the prompts: natural-language (NL) and machine-readability (JSON), and several combinations of contextual factors: user average (U), business average (B), and other attributes (O). We compare the model performance across these models. In Study 2, we test whether such information reflect the label; average ratings may act as a proxy of ground truth. Finally, in Study 3, we further examine how accuracy changes according to the interactions of those two reference points. We progressively explore not only whether reference points improve performance, but also how they function within the model's inference process.

The remainder of this study is constituted as follows: Section 2 reviews related studies, Section 3 outlines the model construction, and Section 4 presents empirical analyses. We discuss the key findings and implications in Section 5. Finally, we list our research limitations in Section 6.

2 Related Study

2.1 Sentiment Analysis

Sentiment analysis has been conducted with various methodologies, including lexicon-based models (Hutto and Gilbert, 2014), machine learning approaches with embeddings (Mikolov et al., 2013; Bojanowski et al., 2017), and deep neural networks such as BERT and RoBERTa (Devlin et al., 2018; Liu et al., 2019).

Recently, LLM-based approaches have gained attention for sentiment analysis (Krugmann and Hartmann, 2024). Models like GPT (Brown et al., 2020; OpenAI, 2023a) and Llama (Touvron et al., 2023a,b; Grattafiori et al., 2024) demonstrate superior performance compared to fully-supervised models and even fine-tuned RoBERTa (Wang et al., 2024; Krugmann and Hartmann, 2024). Key advantages include broad applicability due to pre-training and ability to process raw text without extensive preprocessing, achieving high accuracy without fine-tuning.

Other approaches to sentiment analysis using LLMs include aspect-based sentiment analysis (AbSA; Do et al., 2019; Nazir et al., 2022), which simultaneously predicts multiple aspects in reviews such as price and service quality. Ensemble approaches combine the decisions of multiple LLMs to create the robust model (Xing, 2025; Huang et al., 2024; Niimi, 2025; Chen et al., 2025).

However, existing approaches, including AbSA and ensemble methods, predominantly focus on review text alone. While this concentration highlights interesting challenges from the viewpoint of NLP, it does not necessarily guarantee practical utility for marketing or business decision-making, where contextual information beyond the textual modality plays a crucial role. Although some multimodal sentiment analysis utilizes the other modalities such as images and audio (Das and Singh, 2023; Gandhi et al., 2023), they rarely address psychological reference points, such as prior expectations expressed through numeric values.

2.2 A Role of Reference Point in Service Evaluation

In the field of marketing, extensive research has examined how consumers evaluate the product and service quality. Notable frameworks include prospect theory and EDT. Prospect theory posits that consumers evaluate services by comparing their actual experience to a pre-established reference point (Kahneman and Tversky, 2013). If their experience falls short of this reference point, they tend to feel dissatisfied; conversely, if it exceeds the reference point, satisfaction is more likely. Furthermore, EDT explains the evaluation process from two perspectives. Absolute evaluation involves assessing whether the perceived quality meets a fixed standard, while relative evaluation is based on com-

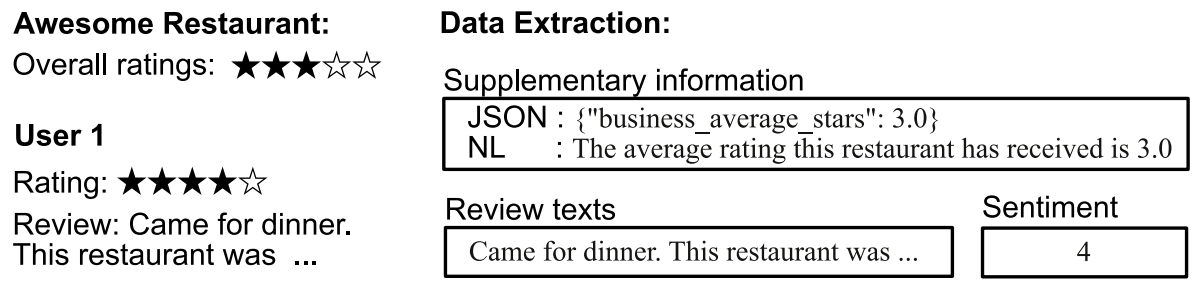

Figure 1: Sample extraction from the dataset.

paring prior expectations with the perceived quality (Oliver, 1980). In both cases, prior expectations play a significant role in determining overall customer satisfaction.

Prior studies on a wide range of products and services have examined the factors that affect or shape expectations. Some of the key factors include consumers' past experiences (Oliver, 1980; Kopalle and Lehmann, 2001; Bolton, 1998; Cooil et al., 2007), reputations provided by other customers (Keiningham et al., 2015; Ryu et al., 2008; Babin et al., 2005), and perceived value (Ryu et al., 2008; Qin and Prybutok, 2008). In particular, reputations—such as an average rating and the helpful reviews from other consumers—serve as important reference points when comparing prior expectations with actual experiences.

However, as noted above, because sentiment analysis is predominantly based on actual review texts, few existing studies have taken these supplementary factors into account. To more accurately capture customers' preferences, it is crucial to incorporate such information into sentiment analysis. We therefore propose a framework that effectively leverages this additional information within LLMs.

3 Proposed Model

3.1 Pre-trained Model

To implement LLM-based sentiment analysis, we adopt Llama 3.2 with 3 billion (3B) parameters and is instruction-tuned (Llama-3.2-3B-Instruct[1]). Llama family has been widely adopted in sentiment analysis (Mai et al., 2024; Roumeliotis et al., 2024; Gautam et al., 2025). In particular, the 3B architecture is a relatively lightweight, compared to mainstream 70–100B models, which is suitable for resource-constrained environments, such as on-device processing, with maintaining privacy.

For the immediate deployment, we do not apply fine-tuning. The hyper-parameters for the model are set in temperature=1.0. The model will stop inference after generating one token (equivalent to underbar in the prompt shown in Fig. 2).

3.2 Basic Prompt

To obtain sentiment values, we use text-completion function which the model completes the continuous texts of the given prompt.

As shown in (Zhang et al., 2024), first, one-shot model is significantly outperforms the zero-shot model. Additionally, few-shot prompt clearly outperforms the one-shot; however, In this study, using multiple examples introduces additional variables (e.g., the reference points of example reviews, their alignment patterns, example selection biases), making it difficult to interpret core findings regarding structured contextual information. Therefore, we use one-shot prompt (Fig. 2) to maintain experimental clarity.

```
### Instruction
You are a helpful assistant evaluating the review texts about
the restaurant. Please evaluate the review text and assign an
integer score ranging from 1 for the most negative comment
to 5 for the most positive comment. The output should be a
single integer from 1 to 5.

### Example
User review: {example_review}
Output: {example_label}

### Task
User review: {user_review}
Output: _
```

Figure 2: Basic Prompt

3.3 Displaying Supplementary Information

To display additional information, the method of presentation needs to be discussed. Learning struc-

[1]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

tural information is highly dependent on in-context learning, which means that locating the explanations before the tabular data improves the understanding of the structure (Sui et al., 2024). Therefore, we set up and compare two display methods, text and JSON format. We include the supplementary information, such as the average user rating and the average venue rating after the review section of the prompt.

Display Method (NL) First, we adopt text method, which displays the average evaluations with the explanations. With this method, although the prompt gets longer, any forms of information, including texts and numeric values, can be input into the model as long as the information can be explained in natural language. Thus, the necessary amount of computing resources becomes larger. Therefore, we set up the explanations and the input as Fig. 3.

When evaluating the review, consider both the textual sentiment and the supplementary information. Use user's average score which user has given in their past reviews to understand the user's typical rating behavior and restaurant's average score which the restaurant has received across all users to compare this restaurant's performance relative to others. Additionally, use restaurant's name, open hours (the total number of hours the restaurant is open in a week), and open days (the total number of days the restaurant is open in a week) to contextualize the review.

Example
User review: This restaurant was...
Supplementary Information: The average of this user's past ratings is 2.6. The average rating this restaurant has received is 3.0
Output: 5

Figure 3: NL Prompt for Supplementary Information

Display Method (JSON) Second, we adopt JSON method, which displays supplementary information with JSON format. This method can provide information in the structured form. One study (Sui et al., 2024) which investigates the impact of using table data, such as CSV, JSON, and HTML, on the understanding of the information indicates that LLMs can comprehend the contents of the table unless the data structure is not too complex and the ability is improved when explanations are added before the structured input. Therefore, we set up the explanations and the structured input as Fig. 4.

When evaluating the review, consider both the textual sentiment and the supplementary information in JSON format. Use 'user_average' (the average score this user has given in their past reviews) to understand the user's typical rating behavior and 'restaurant_average' (the average score this restaurant has received across all users) to compare this restaurant's performance relative to others. Additionally, use 'restaurant_name', 'open_hours' (the total number of hours the restaurant is open in a week), and 'open_days' (the total number of days the restaurant is open in a week) to contextualize the review.

Example
User review: This restaurant was...
Supplementary Information: {"user_average_stars": 2.6, "business_average_stars": 3.0}
Output: 5

Figure 4: JSON Prompt for Supplementary Information

3.4 Dataset

We adopt Yelp Open Dataset (Yelp, 2022) which contains the evaluations and reviews for a wide range of establishments. Compared with the popular benchmarks, such as IMDb movie reviews (Maas et al., 2011) and Amazon Reviews (Jianmo Ni, 2019), Yelp covers diverse user backgrounds and business conditions, resulting in higher heterogeneity across both users and businesses. This heterogeneity makes predictions more challenging since these latent differences are not fully captured within review texts alone.

For the comprehensive analysis, we set up the two different groups for the analysis: Restaurant and Nightlife, which are extracted using the category tags given for the establishments (Table 1). For each group, to prevent data leakage between train and test sets, we ensured that both user IDs and store IDs are mutually exclusive between the training and test sets of the entire Yelp dataset. Under this constraint, we use at most 500 unique user–business pairs for our evaluation set. Reviews are written in English.

For preprocessing the review texts, we at least remove the line break codes of the texts to maintain the format of the prompt. Table 2 shows the summary statistics of the dataset. A number of tokens is counted with tiktoken (OpenAI, 2023b) which is adopted in Llama 3. As shown in the statistics, some samples have significantly long texts.

Datasets	Selected	Excluded
Restaurant	Restaurant	Fast Food, Food Truck, Bar, Nightlife
Nightlife	Bar, Nightlife	Fast Food, Food Truck

Table 1: Selected and excluded category tags for each dataset. We avoid duplicate samples between datasets and select business with fixed addresses.

	Mean	Std	Min	Max
Restaurant				
Stars	3.724	1.515	1	5
Chars	431.726	368.464	42	2552
Tokens	98.744	85.544	9	606
Nightlife				
Stars	3.544	1.605	1	5
Chars	511.082	484.695	65	4998
Tokens	117.104	112.465	15	1118

Table 2: Summary statistics of each category in the dataset

4 Analyses and Results

4.1 Study 1: Impact of Reference Points and Display Methods

First, to address RQ1 (reference-point utilization) and RQ2 (prompt format), we implement sentiment analysis in restaurant and bar evaluations. We compare evaluation metrics across different models and assess the effectiveness of incorporating additional information in two categories. The supplementary information consists of following three elements. U: the user's average rating, indicating the mean rating of the past evaluations given by the user on Yelp, B: the business' average rating, indicating the mean rating which the restaurant has received from all users, and O: other contextual factors, indicating additional information both of textual and numerical attributes, such as the restaurant name, operating hours and the number of days the restaurant is open per week. Both U and B are expected to serve as reference points that affect user's prior expectations.

The LLM-based approach is evaluated with multiple variations, considering the type of supplementary information used and its machine-readability. Accordingly, the LLM-based models are categorized as follows: JSON-UBO / NL-UBO: Utilizing all supplementary information, presented in JSON format or natural language, respectively; JSON-UB / NL-UB: Incorporating only the average ratings; JSON-O / NL-O: Incorporating only contextual factors; and LLM (None): A baseline model without any supplementary information. We also establish four well-established baselines: BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al.). These pretrained models are fine-tuned for 5 epochs with additionally extracted 1000 training samples and the test performances are computed when the validation losses become the lowest. Since the proposed model employs 3B model which focused on lightweight and fast inferences, reference models are also base-sized (e.g., RoBERTa-Base-Uncased).

Sentiment analysis has not only the classification aspect but also the regression due to that the sentiment label is on the ordinal scale, which means that the magnitude of prediction error is important in addition to the concordance. Therefore, we adopt both Macro-F1 score and root mean square error (RMSE) for the evaluation metrics. For the baseline models we extracted an additional training set and fine-tuned each model.

Table 3 and 4 report the results for the Restaurant and Nightlife datasets, respectively. First, among the two datasets, JSON-UBO achieves the highest score in both datasets and improves significantly over LLM (None), which receives no supplementary information. Among the reference models, the strongest baseline differs by domain: LLM (None) ranks first in Restaurant, whereas RoBERTa-Base leads in Nightlife. In general, RoBERTa and DeBERTa outperform BERT, followed by DistilBERT.

Comparing the models within each display format, first, JSON prompts consistently improve performance as more information is added. These results indicate that, supplying a reference point in a machine-friendly format helps the model capture the complex relationships among factors, enabling effective inference. Increasing information from JSON-UB to JSON-UBO leads particularly

$n = 500$		UBO	UB	O
Macro-F1	LLM (JSON)	**0.612**†	**0.598**	**0.588**
	LLM (NL)	**0.593**†	**0.599**	0.524
	LLM (None)	0.587		
	DeBERTa	0.538		
	RoBERTa	0.533		
	BERT	0.474		
	DistilBERT	0.465		
RMSE	LLM (JSON)	**0.564**	**0.616**	**0.647**
	LLM (NL)	**0.620**	**0.624**	0.686
	LLM (None)	0.675		
	DeBERTa	0.703		
	RoBERTa	0.742		
	BERT	0.758		
	DistilBERT	0.804		

Table 3: Study 1 Results (Dataset 1: Restaurant). † indicates the statistically significant difference ($p < .05$) by two-sided McNemar test against LLM (None). Bold number indicates that the model surpasses all the reference models; shaded cells indicate the overall best value.

$n = 500$		UBO	UB	O
Macro-F1	LLM (JSON)	**0.635**†	0.622	0.592
	LLM (NL)	0.602	**0.628**	0.580
	LLM (None)	0.526		
	DeBERTa	0.523		
	RoBERTa	0.625		
	BERT	0.574		
	DistilBERT	0.481		
RMSE	LLM (JSON)	**0.597**	**0.613**	0.665
	LLM (NL)	0.672	**0.647**	0.666
	LLM (None)	0.709		
	DeBERTa	0.688		
	RoBERTa	0.657		
	BERT	0.668		
	DistilBERT	0.746		

Table 4: Study 1 Results (Dataset 2: Nightlife). † indicates the statistically significant difference ($p < .05$) by two-sided McNemar test against LLM (None). Bold number indicates that the model surpasses all the reference models; shaded cells indicate the overall best value.

to reduce RMSE in both datasets. As a result, prediction accuracy rises; for Restaurant, Macro-F1 rises by 4.3% (from 0.587 to 0.612) and RMSE reduces by 16.44% (from 0.675 to 0.564) relative to both LLM (None). For Nightlife, the improvements are even larger relative to LLM (None), +20.7% (from 0.526 to 0.635) and -15.8% (from 0.709 to 0.597), respectively, and still effective compared with RoBERTa (+1.6% / -9.1%). This result aligns with prospect theory and expectation-disconfirmation theory, providing the clear answer to RQ1. The user- and business-level average ratings act as reference points for the LLM and improve the prediction accuracy.

By contrast, the results of NL prompts do not follow this pattern; increasing information from NL-UB to NL-UBO does not contribute on the performance, and in Nightlife the accuracy even decreases below the best baseline. This suggests that, although LLMs can process natural language using the large context window, the models still struggle to capture their complex relationship particularly when large quantities of contextual factors are embedded as plain text.

These empirical results also provide the response to RQ2. Supplying additional information in a machine-readable structure allows the LLMs to effectively utilize those reference points and contextual factors for the prediction.

4.2 Study 2: Relationship with the Expectation

From the results of Study 1, a remaining concern is that these reference points may have worked as proxies for the labels. Therefore, to address RQ3 (proxy effect), Study 2 investigates whether model performance decreases as the review score diverges from these reference points according to the extent of the gaps. We use the JSON-UBO results for the examination.

Since prior expectations are formed based on the user's past behavior and the reputations, we treat such average score as indicators of prior expectations. We define expectation–evaluation gap for both user (U) and business (B) average from user i to store j as follows:

$$gap_{i,j}^{(U)} = rating_{i,j} - user\_average_i \quad (1)$$

$$gap_{i,j}^{(B)} = rating_{i,j} - business\_average_j \quad (2)$$

The data set is divided into five bins according to the extent of gaps, yielding groups that range

from "far below expectations" to "far above expectations." For each bin, we measure the Micro-F1 and RMSE. If the averages were merely proxies for the labels, performance would peak in the middle bin (where the gap is smallest) and decline sharply as the gap widens.

The results are shown in Table 5 (Restaurant) and Table 6 (Nightlife). The leftmost group includes cases where the actual rating falls below expectations, while the rightmost group consists of cases where the actual rating exceeds expectations.

	Expectation				
	below ←		met	→ beyond	
User Average					
$gap_{i,j}^{(U)}$	-1.788	-0.185	0.224	0.748	1.653
Micro-F1	**0.690**	0.636	0.667	**0.890**	**0.870**
RMSE	0.700	0.621	0.619	**0.374**	**0.436**
Business Average					
$gap_{i,j}^{(B)}$	-2.130	-0.640	0.415	0.945	1.565
Micro-F1	0.760	0.450	0.760	**0.860**	**0.910**
RMSE	0.624	0.819	0.490	**0.447**	**0.300**

Table 5: Study 2 Results (Dataset 1: Restaurant). Bold number indicates that the group surpasses the expectation-met group. Shaded cells indicate the overall best value.

First, in the Restaurant category, prediction performances increase in the upper two quantile groups for both the user's and store's average, compared to the middle group where the actual rating was close to the reference point. This suggests that the most accurate predictions were made when the actual experience exceeded prior expectations. In particular, when the experience was beyond the expectation, the performance improved by +25.1% for Micro-F1 and -39.6% for RMSE in user-average compared with the middle group while +13.2% for Micro-F1 and -8.8% for RMSE in business-average.

Next, in the Nightlife category, as in Study 1, merely meeting expectations did not consistently lead to better predictions. However, unlike the Restaurant category, prediction accuracy improved not only when experiences exceeded expectations, but also when they fell significantly short. Notably, the highest accuracy was observed in the group where the actual rating was far below the prior expectation (the leftmost group).

	Expectation				
	below ←		met	→ beyond	
User Average					
$gap_{i,j}^{(U)}$	-2.083	-0.312	0.133	0.523	1.375
Micro-F1	**0.743**	**0.709**	0.700	**0.752**	**0.798**
RMSE	**0.507**	0.660	0.548	0.554	0.611
Business Average					
$gap_{i,j}^{(B)}$	-2.465	-0.995	0.350	0.975	1.520
Micro-F1	**0.830**	0.520	0.710	**0.850**	**0.800**
RMSE	**0.412**	0.911	0.566	**0.510**	**0.447**

Table 6: Study 2 Results (Dataset 2: Nightlife). Bold number indicates that the group surpasses the expectation-met group. Shaded cells indicate the overall best value.

Although there are substantial differences in business nature and customer behavior between restaurants and nightlife venues, at least, we do not confirm that the performance metrics increase in the group with the closest reference points to actual labels. In both cases, reference points do not simply function as proxies for the correct labels, but rather as relative evaluation values in inference, meaning that they function literally as reference points, which is a strong answer to RQ3.

4.3 Study 3: Error Analysis

To further understand how the model interacts with the reference points, we finally combine user and business average scores to create a 5×5 matrix where each cell represents the Micro-F1 score.

Restaurant	BA				
UA	1	2	3	4	5
1	-	1.000	0.812	0.750	-
2	1.000	0.833	0.771	0.643	0.000
3	1.000	0.643	0.688	0.716	0.000
4	-	0.250	0.679	0.816	0.750
5	-	1.000	0.909	1.000	1.000

Table 7: Error analysis by user average (UA) and business average (BA) for restaurant dataset

Table 7 (Restaurant) and 8 (Nightlife) show the results. First, in both categories, the model achieves highest performance (100% in most cases) when UA is 5. Second, two different reference points

Nightlife	BA				
UA	1	2	3	4	5
1	-	1.000	0.920	1.000	-
2	-	0.714	0.680	0.545	-
3	-	0.750	0.623	0.756	-
4	-	0.556	0.682	0.746	1.000
5	-	1.000	1.000	1.000	1.000

Table 8: Error analysis by user average (UA) and business average (BA) for nightlife dataset

show a clear interaction for the prediction. The accuracy tends to improve when two reference points align (along the diagonal), indicating that, when user's past evaluation is close to other consumers' average ratings, the actual rating becomes easier to predict. Notably, in some combinations, the accuracy results in 0%. This indicates cases where conflicting references make prediction challenging. However, as shown in Study 1, this accuracy even outperforms other models. Therefore, our approach can identify unreliable or difficult samples to predict based on reference point conflicts. These results clearly answer RQ4.

These interaction patterns enable practical deployment strategies. Companies can employ adaptive inference where samples with aligned reference points (UA≈BA) are processed on-device environment, while conflicting cases are routed to larger cloud-based models. Additionally, low-confidence predictions can be systematically collected as training data for fine-tuning domain-specific models, enabling continuous performance improvement.

5 Conclusion

5.1 Key Findings

In this study, we enhance LLM-based sentiment analysis by incorporating supplementary information as reference points and other contextual factors, based on prospect theory and EDT.

Study 1 compared two prompting strategies (NL and JSON) with multiple combinations of contextual information. The JSON-UBO model significantly outperformed both NL prompts and four strong baselines. Notably, while JSON prompts showed consistent gains with increasing information, NL prompts failed to leverage the same contexts despite the same context window.

Study 2 addressed the potential concern about reference points serving as label proxies. Accuracy improved more for reviews whose ratings deviated from the average than for those close to the average, indicating that the model was not simply copying the reference points for JSON-UBO model.

Study 3 revealed that the interactions of two different reference points affect the model performance. Accuracy improved when those ratings align while conflicting references indicate inherently challenging cases.

These findings comprehensively answer our research questions. **RQ1 (Effect of reference points):** We demonstrated that the proposed model with U/B/O information significantly improves performance with 4.3–20.7% gains in Macro-F1 and 15.8–16.4% reductions in RMSE over baselines without fine-tuning. **RQ2 (Effect of machine readability):** NL prompts fail to leverage complex information, highlighting the importance of prompt design even for models with large context windows. **RQ3 (Effect of proxy labels):** Follow-up analysis confirms that model performance improves when ratings deviate from expectations, indicating that reference points assist contextual inference rather than serving as mere label proxies. **RQ4 (Effect of reference interactions):** We revealed that aligned reference points improve prediction accuracy, while conflicting reference points indicate inherently challenging prediction cases.

5.2 Implications

This study has both academic and practical implications.

Academic. First, by incorporating the theoretical approach into the LLM-based sentiment analysis, model performance significantly improved, indicating that LLM's rich capability to handle complex context contributes to the predictions. By using JSON format, it is possible to input various information into LLMs, and combining more abundant information may further improve prediction accuracy. Second, even if the amount of information is same across the several prompts, the results of the inferences, including the prediction and performance, vary depending on the display methods, despite the large context window of modern LLMs. Second, simple scalar values, such as 1–5 star averages, can be used directly in the JSON prompt; no discretization or embedding tricks are required. These suggest that we can flexibly employ various factors, including textual and numeric information,

into sentiment analysis. Furthermore, although we employ sentiment analysis for the model verification, the proposed approach using JSON-based contextual information is transferable in wide domain of document classification task, including marketing analysis.

Practical. Since the proposed method only relies on prompt construction, companies can feed existing database contents to LLMs with JSON prompt and immediately construct their own extended models. As shown in the results, our approach achieves RMSE of 0.564 (restaurant) and 0.597 (nightlife), meaning the average prediction error is less than 1-star on a 5-point scale. This level of accuracy is sufficient for practical applications such as the simple recommendation systems, where distinguishing between adjacent rating categories (e.g., 4 vs 5 stars) is often less critical than identifying overall sentiment polarity. Achieving this performance with a 3B parameter scale without fine-tuning indicates that the company can immediately deploy the recommendation agent on edge application environments combined with the rich customer database. Furthermore, our error analysis revealed that samples with aligned reference points can be accurately predicted while cases with conflicting references could be routed to larger models or LLM-based ensemble strategy (Xing, 2025; Huang et al., 2024; Niimi, 2025). This enables the energy efficient processing where computational resources are allocated based on prediction difficulty.

6 Limitations

This study has several limitations. First, while our approach is grounded in prospect theory and EDT, we did not empirically test whether the psychological mechanisms underlying these theories actually explain the model's improved performance. Therefore, we need to further validate those relationships by the psychological experiments to support our findings.

Second, our experiments were conducted using only a single model architecture (Llama-3.2-3B-Instruct). The effectiveness of structured prompting may vary across different model families and scales, limiting the generalizability of our findings. Particularly, we cannot conclude whether the benefits of JSON formatting extend to larger models or different architectures.

Third, our evaluation is restricted to English reviews from two categories within the Yelp Open Dataset (Yelp, 2022). Additional verifications for other domains, languages, benchmarks are also required to support the effectiveness.

In addition, we did not compare our method with simple post-hoc calibration baselines (e.g., shifting predictions according to user or business averages). Such heuristics may adjust main effects but cannot capture interaction effects between users and venues, which our structured prompting approach is designed to address. Future research may provide a systematic comparison between these alternatives, incorporating effect-size measures for more robust evaluation.

Finally, while we argue that our approach is computationally efficient due to the absence of fine-tuning, we did not provide quantitative measurements of inference time or memory usage across different prompt formats for the actual inferences.

Acknowledgment

We are grateful to the two anonymous reviewers for their insightful comments. Their feedbacks have greatly improved our study.

Both the dataset and model were managed and used in an appropriate environments that comply with the terms of use. We do not collect additional information that could lead to the identification of individuals.

This study is supported by JSPS KAKENHI (Grant Number: 24K16472).

References

Barry J Babin, Yong-Ki Lee, Eun-Ju Kim, and Mitch Griffin. 2005. Modeling consumer satisfaction and word-of-mouth: restaurant patronage in korea. *Journal of Services Marketing*, 19(3):133–139.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Ruth N Bolton. 1998. A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction. *Marketing science*, 17(1):45–65.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhijun Chen, Jingzheng Li, Pengpeng Chen, Zhuoran Li, Kai Sun, Yuankai Luo, Qianren Mao, Dingqi Yang, Hailong Sun, and Philip S Yu. 2025. Harnessing multiple large language models: A survey on llm ensemble. *arXiv preprint arXiv:2502.18036*.

Bruce Cooil, Timothy L Keiningham, Lerzan Aksoy, and Michael Hsu. 2007. A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. *Journal of marketing*, 71(1):67–83.

Ringki Das and Thoudam Doren Singh. 2023. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, 118:272–299.

Andrew J Flanagin and Miriam J Metzger. 2013. Trusting expert-versus user-generated ratings online: The role of information volume, valence, and consumer characteristics. *Computers in Human Behavior*, 29(4):1626–1634.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Himanshu Gautam, Abhishek Gaur, and Dharmendra Kumar Yadav. 2025. A survey on the impact of pre-trained language models in sentiment classification task. *International Journal of Data Science and Analytics*, pages 1–39.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yichong Huang, Xiaocheng Feng, Baohang Li, Yang Xiang, Hui Wang, Ting Liu, and Bing Qin. 2024. Ensemble learning for heterogeneous large language models with deep parallel collaboration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Julian McAuley Jianmo Ni, Jiacheng Li. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 188–197.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific.

Timothy Lee Keiningham, Bruce Cooil, Edward C Malthouse, Bart Lariviere, Alexander Buoye, Lerzan Aksoy, and Arne De Keyser. 2015. Perceptions are relative: an examination of the relationship between relative satisfaction metrics and share of wallet. *Journal of Service Management*, 26(1):2–43.

Praveen K Kopalle and Donald R Lehmann. 2001. Strategic management of expectations: The role of disconfirmation sensitivity and perfectionism. *Journal of Marketing Research*, 38(3):386–394.

Jan Ole Krugmann and Jochen Hartmann. 2024. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. ArXiv preprint arXiv:1907.11692.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Zhelu Mai, Jinran Zhang, Zhuoer Xu, and Zhaomin Xiao. 2024. Financial sentiment analysis meets llama 3: A comprehensive analysis. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, MLMI '24, pages 171–175, New York, NY, USA. Association for Computing Machinery.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. ArXiv preprint arXiv:1301.3781.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2022. Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Junichiro Niimi. 2025. A simple ensemble strategy for llm inference: Towards more stable text classification. In *Proceedings of the 30th International Conference on Natural Language & Information Systems (NLDB 2025)*, Lecture Notes in Computer Science. Springer.

Richard L Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research*, 17(4):460–469.

Richard L Oliver. 1999. Whence consumer loyalty? *Journal of marketing*, 63(4):33–44.

OpenAI. 2023a. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. *tiktoken: a fast BPE tokeniser for use with OpenAI's models*.

G Qin and Victor R Prybutok. 2008. Determinants of customer-perceived service quality in fast-food restaurants and their relationship to customer satisfaction and behavioral intentions. *Quality Management Journal*, 15(2):35–50.

Werner Reinartz, Manfred Krafft, and Wayne D Hoyer. 2004. The customer relationship management process: Its measurement and impact on performance. *Journal of marketing research*, 41(3):293–305.

Konstantinos I Roumeliotis, Nikolaos D Tselikas, and Dimitrios K Nasiopoulos. 2024. Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6:100056.

Kisang Ryu, Heesup Han, and Tae-Hee Kim. 2008. The relationships among overall quick-casual restaurant image, perceived value, customer satisfaction, and behavioral intentions. *International journal of hospitality management*, 27(3):459–469.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Beverley A Sparks and Victoria Browning. 2011. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism management*, 32(6):1310–1323.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. ArXiv preprint arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. ArXiv preprint arXiv:2307.09288.

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*, 30:5998–6008.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is chatgpt a good sentiment analyzer? In *First Conference on Language Modeling (COLM 2024)*.

Frank Xing. 2025. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Yelp. 2022. *Yelp Open Dataset, An all-purpose dataset for learning*. Yelp.

Ya You, Gautham G Vadakkepatt, and Amit M Joshi. 2015. A meta-analysis of electronic word-of-mouth elasticity. *Journal of Marketing*, 79(2):19–39.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.