Using Contextually Aligned Online Reviews to Measure LLMs' Performance Disparities Across Language Vareities

Anonymous ACL submission

Abstract

A language can have different varieties. These varieties can affect the performance of natural language processing (NLP) models, including large language models (LLMs), which are often trained on data from widely spoken varieties. This paper introduces a novel and cost-effective approach to benchmark model performance across language varieties. We argue that international online review platforms, such as Booking.com, can serve as effective data sources for constructing datasets that capture **comments in different language varieties from similar real-world scenarios**, like reviews for the same hotel with the same rating using the same language (*e.g.*, Mandarin Chinese) but different language varieties (*e.g.*, Taiwan Mandarin, Mainland Mandarin). To prove this concept, we constructed a **contextually aligned** dataset comprising reviews in Taiwan Mandarin and Mainland Mandarin and tested six LLMs in a sentiment analysis task. Our results show that LLMs consistently underperform in Taiwan Mandarin.

1 Introduction

A language can have different varieties. Of the world's 7,000 languages, sixty (60) million people speak British English, 23 million speak Taiwan Mandarin, and 10 million speak European Portuguese, compared to 330 million, 900 million, and 200 million who speak American English, Mainland Mandarin, and Brazilian Portuguese, respectively. These varieties differ enough in accent, vocabulary, or syntax for native speakers to distinguish them. NLP technologies, including LLMs, are known to perform better in English varieties that are more widely represented in the internet data they are trained on, particularly Mainstream American English (MAE), compared to less represented varieties like African American English (AAE) (Ziems et al., 2022, 2023). Specifically, LLMs more accurately predict sentiment scores in MAE (Ziems et al., 2022), generate higher-quality texts in MAE (Ziems et al., 2022), and hold better conversations in MAE (Ziems et al., 2023). These comparisons were made possible by intensive, targeted efforts specific to each language variety, such as "translating" data instances from a standard variety (*e.g.*, MAE) to less widely represented varieties (*e.g.*, AAE), followed by validation from native speakers (Ziems et al., 2022, 2023). What is not known is whether these performance gaps and biases extend to a broader range of languages and their numerous varieties, such as Mainland Mandarin versus Taiwan Mandarin. Building effective benchmarking datasets for evaluating model performance across language varieties is expensive—creating "fair" comparisons between varieties often needs native speakers and language experts.

In this paper, using Mandarin Chinese as an example, we propose an approach that uses large-scale user-generated reviews to construct benchmarking datasets across varieties of a given language. We argue that the international online review platforms with millions of users, like Booking.com, when properly curated, can serve as effective data sources for constructing datasets that capture **comments in different language varieties from similar real-world scenarios**, like comments for the same hotel with the same rating using the same language (*e.g.*, Mandarin Chinese) but different language varieties (*e.g.*, Taiwan Mandarin, Mainland Mandarin, Singaporean Mandarin). These datasets, being **contextually aligned**, can then be used to benchmark LLMs' performance across language varieties for NLP tasks like sentiment analysis and text generation. Once a low-cost and generalizable approach becomes available, researchers can then compare model performance across a wide range of language varieties, enabling reliable benchmarking of progress in addressing performance gaps and moving toward an LLM that performs equally well across all language varieties.

2 Related Work

Beyond machine translation (Kantharuban et al., 2023), researchers tried to benchmark NLP models across language varieties (Zampieri et al., 2020; Joshi et al., 2024; Blodgett et al., 2020; Hovy and Johannsen, 2016; Zampieri et al., 2019), but the focus on identifying gaps between these varieties varies widely. Some prior work focused solely on a single less-representative variety, such as Taiwan Mandarin (Tam et al.; Chen et al., 2024), without measuring performance gaps across multiple varieties. Other studies that measured these gaps employed different levels of granularity. The most common approach, **task-level comparison**, benchmarks the same NLP task across language varieties (Faisal et al., 2024), such as sentiment analysis, but datasets often differ in source or genre across varieties, making the reported performance numbers not directly comparable. For instance, sentiment analysis datasets for Mainland Mandarin and Taiwan Mandarin often used different sources (Seki et al., 2007). A more refined approach, **scenario-level comparison**, evaluates performance within the same dataset or scenario, such as essay grading (Liang et al., 2023) or speech rating (Kwako et al., 2023), across data partitions of different language varieties (Lwowski et al., 2022; Blodgett and O'Connor, 2017). While this method eliminates biases caused by differing data sources, it cannot fully address biases introduced during dataset construction. The most rigorous method, **instance-level comparison**, involves constructing parallel datasets with an item-by-item alignment between varieties (Ziems et al., 2022, 2023; Groenwold et al., 2020; Kuzman et al., 2023), where each instance is converted between language varieties. However, creating such comparisons is very costly, requiring native speakers and language experts to ensure accuracy. Our approach achieves instance-level comparability with lower costs.

3 Constructing a Contextually-Aligned Review Dataset for Language Varieties

Data. We constructed a dataset of hotel reviews sourced from `Booking.com`, which has been used in prior research studies (Alderighi et al., 2022; Barnes et al., 2018). This dataset consists of 4,447,853 reviews labeled by the platform as written in Chinese. The reviews cover 149,879 hotels located in Japan, Mainland China, South Korea, Taiwan, Thailand, and Vietnam, and were collected from August 2021 to August 2024. These locations were selected to ensure a substantial volume of data, as they are popular destinations for Mandarin-speaking travelers. Each review comprises three main components: the review title, positive feedback, and negative feedback. Additionally, it includes review ratings (ranging from 1 to 10 stars) and metadata such as hotel ID, posting time, and more (see Appendix A for an actual sample). Booking.com claims to invest significant effort in ensuring that reviews are posted by real users and in maintaining review quality. We included only non-empty reviews, meaning reviewers provided input in at least one of the following: the review title, positive feedback, or negative feedback. In total, we collected 1,513,056 reviews written in Chinese.

3.1 Contextually Aligning Reviews

We used users' self-specified "nationality/region" labels, required by the Booking.com platform, to determine the language varieties used in this study. In total, we have 1,403,669 reviews written in Taiwan Mandarin and Mainland Mandarin, where 95.591% of them come from Taiwan Mandarin users. To ensure a balanced representation between Taiwan Mandarin and Mainland Mandarin reviews, we employed a sampling method that pairs reviews in both languages using the following criteria:

- **Same hotel for both reviews:** Both reviews in each pair are from the same hotel, ensuring that the reviewers are commenting on similar scenarios or objects—the hotel itself.
- **Similar ratings for both reviews:** To form comparable pairs with similar sentiments, we used a 3-class rating scheme (1-3 as negative, 4-7 as neutral, and 8-10 as positive) and paired reviews based on this classification. This approach maximizes the number of review pairs while maintaining comparable sentiment.
- **Similar text length for both reviews:** To minimize the effect of input size, we grouped reviews into 10-token bins before pairing and required both reviews in each pair to fall within the same length bin. This ensures paired reviews have similar text lengths. Reviews longer than 500 tokens were excluded (see Appendix E for details).

Following the criteria, the final paired dataset contained 22,918 pairs of reviews, with each pair containing one review from a Taiwan Mandarin user and one from a Mainland Mandarin user.

Text Length (#word)	Model	ACC ↑						MSE ↓					
		structured		plain		shuffled		structured		plain		shuffled	
		tw	cn	tw	cn	tw	cn	tw	cn	tw	cn	tw	cn
Short (1-49)	**GPT-4o**	.265	.274	.192***	.208	.186***	.202	3.563*	3.769	4.092***	3.385	4.347***	3.561
	Llama3 8b	.234***	.216	.176	.184	.162	.169	3.402***	3.931	3.059***	2.851	3.220**	3.071
	Llama3 70b	.392	.399	.297**	.310	.288***	.310	1.785**	1.661	3.465***	3.005	3.427***	2.941
	Llama3 405b	.386***	.412	.283***	.315	.290***	.322	2.605**	2.430	4.365***	3.598	4.355***	3.612
	Gemma2 9b	.150**	.138	.182*	.193	.171**	.184	2.987***	3.168	4.089***	3.367	4.462***	3.646
	Gemma2 27b	.109	.109	.122	.124	.122	.127	5.399***	5.120	6.150***	5.292	5.880***	5.183
Long (50+)	**GPT-4o**	.356***	.384	.281***	.332	.267***	.314	1.846***	1.577	1.834***	1.570	2.070***	1.743
	Llama3 8b	.224	.228	.189***	.214	.179***	.200	1.895	1.901	2.024***	1.835	2.111***	1.933
	Llama3 70b	.409***	.435	.373*	.391	.360***	.388	1.424**	1.284	1.667*	1.557	1.771**	1.608
	Llama3 405b	.380***	.410	.371**	.396	.357***	.384	1.866***	1.627	1.906**	1.719	2.052***	1.844
	Gemma2 9b	.142	.151	.192**	.209	.174*	.188	2.515***	2.350	2.144***	1.979	2.399***	2.193
	Gemma2 27b	.096	.092	.077	.075	.079	.083	5.135**	4.919	5.913**	5.578	5.606***	5.160
Overall	**GPT-4o**	.296***	.312	.222***	.250	.213***	.240	2.978	3.022	3.323***	2.767	3.571***	2.942
	Llama3 8b	.230*	.221	.181***	.195	.168***	.181	2.838***	3.170	2.672***	2.470	2.806***	2.644
	Llama3 70b	.398**	.412	.324***	.339	.313***	.337	1.658***	1.529	2.835***	2.496	2.847***	2.473
	Llama3 405b	.384***	.411	.314***	.344	.313***	.344	2.347***	2.148	3.505***	2.939	3.550***	2.992
	Gemma2 9b	.147	.143	.185***	.199	.172***	.186	2.823	2.882	3.410***	2.881	3.742***	3.137
	Gemma2 27b	.105	.103	.106	.107	.107	.111	5.307***	5.049	6.067***	5.392	5.785***	5.175

Table 1: Results by length for GPT-4o, Llama3 (8b, 70b, 405b), and Gemma2 (9b, 27b) models. Red cells indicate a worse performance than the other group. (Statistical group differences are indicated as * (p<.05), ** (p<.01), and *** (p<.001) regarding the model performance.)

3.2 Data Quality Validation

To validate the data quality, we recruited five native speakers of Taiwan Mandarin (TW) to review 200 randomly selected reviews written in Taiwan Mandarin. We followed the same process for reviews in Mainland Mandarin (CN). The focus was on two key aspects: *(i)* writing quality and *(ii)* content-rating agreement, evaluated on a 5-point Likert scale. Appendix B shows the material used. Each participant was paid $10.

As a result, for the writing quality ratings, the TW group had a mean of 4.18 (SD=0.44), and the CN group had a mean of 3.94 (SD=0.49). Regarding the rating-content agreement, the TW group had a mean of 4.00 (SD=0.46), and the CN group had a mean of 3.56 (SD=0.55).

4 Experimental Results

To study the potential biases introduced by the input review structure, we experimented with three different settings: *(i)* In the **structured review** setting, the content retained its original structure, consisting of the review title, positive feedback, and negative feedback. *(ii)* In the **plain review** setting, the content was concatenated into a single paragraph, starting with the review title followed by the positive and negative feedback. *(iii)* In the **shuffled review** setting, all three input elements were included but randomly shuffled before being concatenated. For the analysis, we excluded pairs that lacked complete predictions or received predictions that did not follow the specified format (see Appendix D). Once the contextually aligned dataset was constructed and available, we tested it using six LLMs: GPT-4o, Llama3 (8b, 70b, 405b), and Gemma2 (9b, 27b). The task involved predicting a rating score (from 1 to 10, where 1 is the worst and 10 is the best) based on the review content. The prompt includes the task description, the review content, and the prediction scale (1-10). See Appendix C for the prompt details. Table 1 shows the prediction accuracy and mean squared error (MSE) among different models and prompt settings (see Appendix D for the number of valid predictions in each setting). We are aware that Gemma2 27b underperforms significantly, yielding results similar to random guesses and more invalid predictions. Further research is needed to identify the root cause.

LLMs performed significantly worse in Taiwan Mandarin compared to Mainland Mandarin. Among all 54 experiments with different models and prompt settings, 37 of them had significant group differences in accuracy, and 51 had significant group differences in MSE (see Table 1). Among all those significant differences, LLMs typically made less accurate sentiment predictions toward Taiwan Mandarin users. Such an effect is robust and more consistent when the structure of

Original Language	Accuracy↑						MSE↓					
	structured		plain		shuffled		structured		plain		shuffled	
	TW	CN	TW	CN	TW	CN	TW	CN	TW	CN	TW	CN
TW	.296*	.302	.223***	.231	.214***	.221	2.985***	2.036	3.262***	2.577	3.489***	2.688
CN	.303**	.312	.240***	.250	.235**	.240	1.969***	3.026	2.267***	2.727	2.393***	2.901

Table 2: Comparison of GPT-4o performance on original and machine-translated texts. Red cells indicate the lower performing variant in each TW-CN pair. The results suggest that translating TW to CN generally improves performance, while CN to TW translation yields mixed results. (Statistical group differences are indicated as * (p<.05), ** (p<.01), and *** (p<.001) regarding the model performance.)

the input is removed (*i.e.*, shuffled text).

When the reviews' structures are disrupted, the performance gap increases. Table 1 shows that models provided with structured information exhibit a smaller performance gap between user groups and generally outperform those in plain and shuffled conditions. The knowledge that a sentence belongs to either a "positive-review" or "negative-review" category significantly enhances model performance. This observation suggests that when models must rely solely on the text for prediction (*i.e.*, in the plain and shuffled cases), bias towards Taiwan Mandarin and Mainland Mandarin becomes more pronounced.

Shorter reviews tend to produce larger MSE gaps. In our pilot exploration, we considered whether shorter texts should be excluded due to potentially insufficient information (see Appendix E) and found that text length does often influence model performance and behavior. As a result, we categorized our dataset into two groups based on length: short (1-49 characters) and long (50+ characters). Table 1 shows that the MSE gap between Taiwan Mandarin and Mainland Mandarin widens in the short text group (also see Figure 1 in Appendix E). One possible explanation is that shorter reviews offer less information for LLMs to base predictions on, which makes language variety biases more pronounced.

4.1 Machine Translation Baselines

A natural question is whether we could simply use machine translation to convert Taiwan Mandarin to Mainland Mandarin, and vice versa, to create a paired dataset for benchmarking. To explore this, we translated all texts to their opposite version (Taiwan Mandarin to Mainland Mandarin, or vice versa) using the Google Translate API. We then conducted sentiment analysis experiments using GPT-4o, comparing each original sample with its translated version. The results, shown in table 2, indicate that translating Taiwan Mandarin data to Mainland Mandarin generally improves performance. However, translating Mainland Mandarin to Taiwan Mandarin yielded mixed results: it decreased accuracy but improved MSE. While these experiments clearly demonstrate a model bias between Taiwan Mandarin and Mainland Mandarin, further research is needed to fully understand the effects of translation on model performance.

5 Discussion

Do users who self-label as being from Taiwan always use Taiwan Mandarin? In this study, we use users' self-reported nationality/region to infer whether they are speakers of Taiwan Mandarin or Mainland Mandarin. The convention is that Taiwan Mandarin employs traditional Chinese characters, while Mainland Mandarin uses simplified characters. However, after applying the Hanzi Identifier (git), we found that 15.72% of samples in the CN group contained characters beyond simplified Chinese, and 3.86% of samples in TW group included characters not limited to traditional Chinese. This suggests that the relationship between self-reported nationality/region, language variety, and character usage is more complex in real-world data. In Appendix G, Table 6 shows the distribution of Chinese script variants among users.

6 Conclusion and Future Work

This paper introduces a cost-effective method for benchmarking model performance across language varieties using international online reviews from similar contexts. To validate this, we built a contextually aligned dataset of Taiwan Mandarin and Mainland Mandarin reviews and tested six LLMs on sentiment analysis, finding that LLMs consistently underperform in Taiwan Mandarin. We aim to extend this approach to more language varieties, with the ultimate goal of creating LLMs that perform equally well across them.

7 Limitations

One limitation is that, despite the contextual alignment, there is no guarantee that unknown confounding factors do not contribute to the performance gaps. This is an inherent challenge when using user-generated data in the wild for apple-to-apple comparisons, as controlling all variables is almost impossible. Another limitation relates to the input prompts, which are code-mixed. Previous studies found that LLMs might still have deficits in dealing with cultural context and code-mixing input (Ochieng et al., 2024). We used English for instruction to exclude potential biases introduced if it is prompted in Chinese, regardless of its variety. However, such a setup may introduce additional confusion for LLMs to process, leading to lower performance results. The usage of English prompts regarding non-English tasks, or code-switching prompts, requires thorough studies to better investigate LLMs' capability of multilingualism and awareness of language and cultural diversity.

8 Ethics Statement

We assess that the general risks and ethical concerns of our work are no greater than those involved in using user-generated reviews to test sentiment analysis models.

References

GitHub - tsroten/hanzidentifier: Python module that identifies Chinese text as being Simplified or Traditional — github.com. https://github.com/tsroten/hanzidentifier. [Accessed 16-10-2024].

Marco Alderighi, Consuelo R. Nava, Matteo Calabrese, Jean-Marc Christille, and Chiara B. Salvemini. 2022. Consumer perception of price fairness and dynamic pricing: Evidence from booking.com. *Journal of Business Research*, 145:769–783.

Jeremy Barnes, Toni Badia, and Patrik Lambert. 2018. MultiBooked: A corpus of Basque and Catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*.

Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. Measuring taiwanese mandarin language understanding. In *First Conference on Language Modeling*.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages. *arXiv preprint arXiv:2403.11009*.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating african-american vernacular english in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883.

Dirk Hovy and Anders Johannsen. 2016. Exploring language variation across Europe - a web-based tool for computational sociolinguistics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2986–2989, Portorož, Slovenia. European Language Resources Association (ELRA).

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245.

Taja Kuzman, Peter Rupnik, and Nikola Ljubešić. 2023. Get to know your parallel data: Performing english variety and genre classification over macocu corpora. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 91–103.

Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang, and Li Cai. 2023. Does bert exacerbate gender or l1 biases in automated english speaking assessment? In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 668–681.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Brandon Lwowski, Paul Rad, and Anthony Rios. 2022. Measuring geographic performance disparities of offensive language classifiers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6600–6616.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2024. Beyond metrics: Evaluating llms' effectiveness in culturally nuanced, low-resource real-world scenarios. *arXiv preprint arXiv:2406.00343*.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi Chen, Noriko Kando, Chin-Yew Lin, et al. 2007. Overview of opinion analysis pilot task at ntcir-6. In *NTCIR*.

Zhi Rui Tam, Ya Ting Pai, Yen-Wei Lee, Hong-Han Shuai, Jun-Da Chen, Wei Min Chu, and Sega Cheng. Tmmlu+: An improved traditional chinese evaluation suite for foundation models. In *First Conference on Language Modeling*.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardzic, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, et al. 2019. A report on the third vardial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. Value: Understanding dialect disparity in nlu. *arXiv preprint arXiv:2204.03031*.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-value: A framework for cross-dialectal english nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768.

A Booking.com Data

Table 3 shows a sample of the collected Booking.com review.

B Questions Used in Human Validation

We used the following two questions in the human evaluation to assess data quality. For each part of the study, participants were shown both the English text and its translation, either into Taiwan Mandarin or Mainland Mandarin, depending on the context.

1. The review (including the title, positive, and negative sections) is easy to read, and the writing quality is comparable to online reviews written by native speakers, based on my experience.
   - Taiwan Mandarin: 根據我的經驗，這篇評論（包括標題、優點和缺點部分）很容易閱讀，且寫作品質與母語使用者撰寫的網路評論相當。
   - Mainland Mandarin: 根据我的经验，这篇评论（包括标题、优点和缺点部分）很容易阅读，而且写作质量与母语者撰写的网络评论相当。
2. The score (1-10, 1 is the worst, 10 is the best) assigned to this review accurately reflects the content of the review.
   - Taiwan Mandarin: 這篇評論的分數（1-10，1是最差，10是最好）準確反映了評論的內容。
   - Mainland Mandarin: 这篇评论的评分（1-10，1是最差，10是最好）准确反映了评论的内容。

C Prompts

The following prompt is used for structured condition.

System
```
You are a grading assistant for hotel
reviews
```
User
```
The following is a hotel review from
a user. Based on the title, positive
feedback, and negative feedback provided
below, give an overall score from 1
to 10, where 1 is the worst and 10 is
the best. DO NOT include any words
in your output, just provide the number.

Title: [title]
Positive Feedback: [positive_review]
Negative Feedback: [negative_review]
Overall Score (1-10):
```

The following prompt is used for both of the plain and shuffled condition.

System
```
You are a grading assistant for hotel
reviews
```
User
```
The following is a hotel review from
a user. Based on the input review
below, give an overall score from 1
to 10, where 1 is the worst and 10 is
the best. DO NOT include any words
in your output, just provide the number.

input: [text]
Overall Score (1-10):
```

For LLMs that don't have a system role setting (*e.g.* Gemma2), the system instruction is removed from the prompts.

Field	Value
hotel__booking_id	311092
hotel__ufi	-240213
user	——— (*Removed the user identity*)
user_nationality	tw
room_type	雙床房－附加床－禁煙 *(English Translation: Twin Room - Extra Bed - Non-Smoking)*
checking_date	2023-04-23
checkout_date	2023-04-26
length_of_stay	3
guest_type	null
score	10.0
review_title	null
positive_review	櫃檯很友善，有事情都很熱心協助，環境乾淨整潔，住的很舒適，還貼心附上各種充電頭，超級滿意！ *(English Translation: The front desk is very friendly and helpful. The environment is clean and tidy. The stay was comfortable. They thoughtfully provided various charging heads. Super satisfied!)*
negative_review	null
hotel_response	null
review_time	2023-05-15 10:55:59+00:00
created	2024-08-18 07:11:29.971276+00:00

**Note: English translations in italics are provided for readability and are not part of the actual data.*

Table 3: Sample data entry from the collected Booking.com. There are three review components: review_title, positive_review, and negative_review.

model	All			Short			Long		
	combined	shuffled	structured	combined	shuffled	structured	combined	shuffled	structured
GPT-4o	45,828	45,830	45,836	29,436	29,438	29,444	15,636	15,636	15,636
LLaMA-3.1 8B	45,748	45,746	42,784	29,356	29,356	26,392	15,636	15,634	15,636
LLaMA-3.1 70B	45,472	45,608	45,428	29,176	29,272	29,144	15,540	15,580	15,532
LLaMA 3.1 405B	45,682	45,618	45,468	29,348	29,344	29,098	15,580	15,518	15,618
Gemma-2 9B	45,824	45,824	45,814	29,436	29,436	29,422	15,632	15,632	15,636
Gemma-2 27B	43,104	43,134	43,464	27,544	27,590	27,926	14,836	14,802	14,834
GPT-4o+Translation	45,682	45,644	45,836	-	-	-	-	-	-

Table 4: Number of valid prediction samples in the study across different models and data configurations.

model	All			Short			Long		
	combined	shuffled	structured	combined	shuffled	structured	combined	shuffled	structured
GPT-4o	−8	−6	0	−8	−6	0	0	0	0
LLaMA-3.1 8B	−88	−90	−3,052	−88	−88	−3,052	0	−2	0
LLaMA-3.1 70B	−364	−228	−408	−268	−172	−300	−96	−56	−104
LLaMA 3.1 405B	−154	−218	−368	−96	−100	−346	−56	−118	−18
Gemma-2 9B	−12	−12	−22	−8	−8	−22	−4	−4	0
Gemma-2 27B	−2,732	−2,702	−2,372	−1,900	−1,854	−1,518	−800	−834	−802
GPT-4o+Translation	−154	−192	0	-	-	-	-	-	-

Table 5: Number of invalid predictions in the study across different models and data configurations. Negative values indicate the count of invalid samples. Results show that some models (e.g., Gemma-2 27B and LLaMA-3.1 8B) exhibit substantially higher numbers of invalid samples, particularly for structured data.

D Distribution of Valid and Invalid Predictions

Table 4 and Table 5 present the numbers of valid and invalid predictions obtained from our experimental procedures. Invalid predictions encompass instances where models deviated from the task requirements, such as providing explanations instead of numerical outputs, generating values outside the specified range of 1-10, or failing to engage with the task altogether. Our analysis reveals that two models, LLaMA-3.1 8B and Gemma-2 27B, exhibit a notably higher proportion of invalid predictions compared to their counterparts. We only include involved pairs with completely valid data entries for the prediction analysis (Table 1), referring to the smallest number of each model in Table 4.

E Pilot Study on Impact of Text Length

During our data exploration phase, we investigated whether short texts should be removed due to potentially insufficient information for accurate sentiment classification. To address this, we conducted a pilot experiment to analyze the relationship between text length and model performance.

Data We used the initial Booking.com dataset, assigning sentiment labels based on review scores: positive (8-10), neutral (4-7), and negative (1-3). The input text was created by concatenating three review components:

```
[review-title]
[positive-review]
[negative-review]
```

We categorized the texts into 50 bins of 10 characters each, up to 500 characters in length. For each bin, we selected a balanced set of 600 samples (200 per sentiment label) where possible. It's worth noting that for texts longer than 290 characters, maintaining this balance became challenging due to insufficient samples.

Predictions We employed GPT-4o (`gpt-4o-2024-08-06`) to classify each sample into one of the three sentiment categories using the following prompt (without a system prompt):

User

```
Predict the sentiment of the following text. Please answer one of the following label: (positive, negative, neutral). Do not reply anything like 'The sentiment is...'. Do not replay with any explanation. Directly output the answer.

Text: [text]
```

Predictions outside the specified labels were excluded from the analysis (only one sample was removed in this experiment).

Results Figure 1 illustrates the accuracy and MSE for each sentiment label and the overall performance across different text lengths. While the overall performance remains relatively stable across text lengths, we observed variations in performance for individual sentiment labels. This effect is particularly noticeable for negative sentiments in shorter texts. Our findings indicate that text length does influence model performance, though not to the extent of completely compromising the model's ability to classify sentiments. Based on these results, we decided against filtering samples based on text length. Instead, we report scores for different text length groups (short: 1-49 and long: 50+) to provide a comprehensive view of the model's performance across text lengths.

F Impact of Length on Model Performance

To further analyze the effect of text length on our main study results presented in Section 4, we plotted the performance on scatter plots. The x-axis represents the performance for Mainland Mandarin, while the y-axis represents the performance for Taiwan Mandarin. The results are displayed in Figure 2 and Figure 3.

In these plots, the diagonal line ($x = y$) represents equal performance between the two language variations. The distance of each point from this line indicates the performance gap. For the accuracy plot (Figure 2), points closer to the bottom-right indicate better performance in Mainland Mandarin, while points closer to the top-left indicate better performance in Taiwan Mandarin. Conversely, in the MSE plot (Figure 3), points closer to the top-left indicate better performance in Mainland Mandarin.

Our analysis of Figure 2 does not reveal a significant difference between the short and long text groups in terms of accuracy. However, Figure 3 shows a larger gap for the short text group compared to the long text group in terms of MSE. Based on these observations, we hypothesize that shorter reviews may introduce more bias. This could be due to insufficient contextual information in shorter

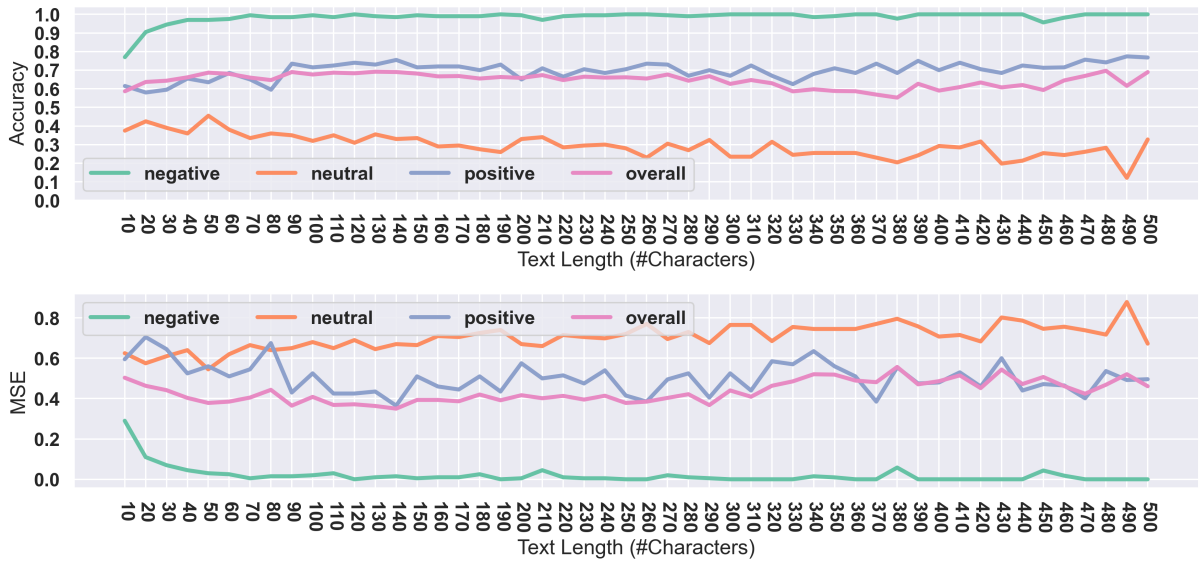

Figure 1: Impact of text length on sentiment classification performance. The top graph shows accuracy, and the bottom graph shows MSE for negative, neutral, positive, and overall sentiments across different text lengths (0-500 characters). While overall performance remains relatively stable, individual sentiment categories show varying levels of accuracy and error, particularly for shorter texts.

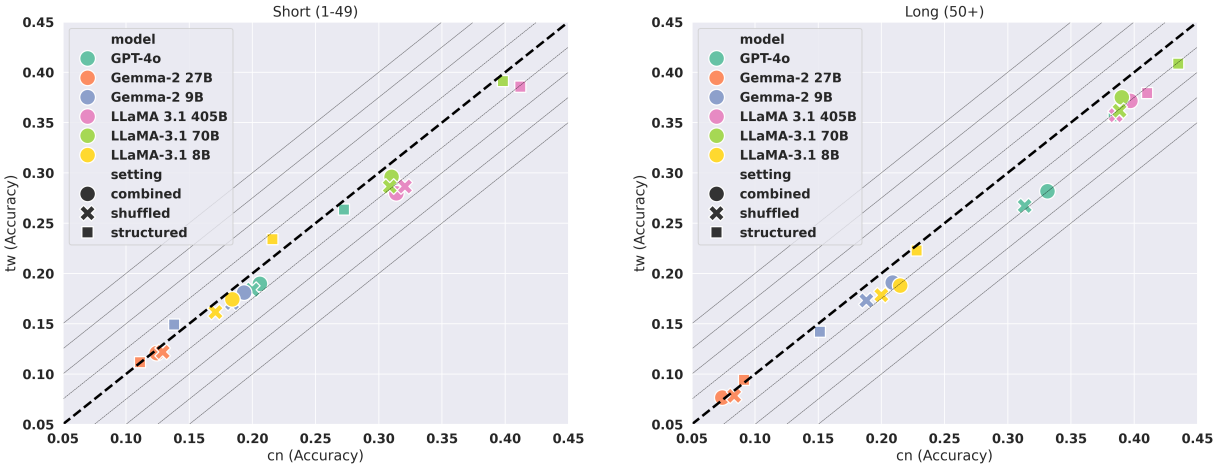

Figure 2: Comparison of accuracy between Mainland Mandarin and Taiwan Mandarin for short (left) and long (right) texts. Each point represents a [model, setting]'s performance. The diagonal line ($x = y$) indicates equal performance. Points above the line suggest better performance in Taiwan Mandarin, while points below suggest better performance in Mainland Mandarin. We do not see a big difference between the short and long texts.

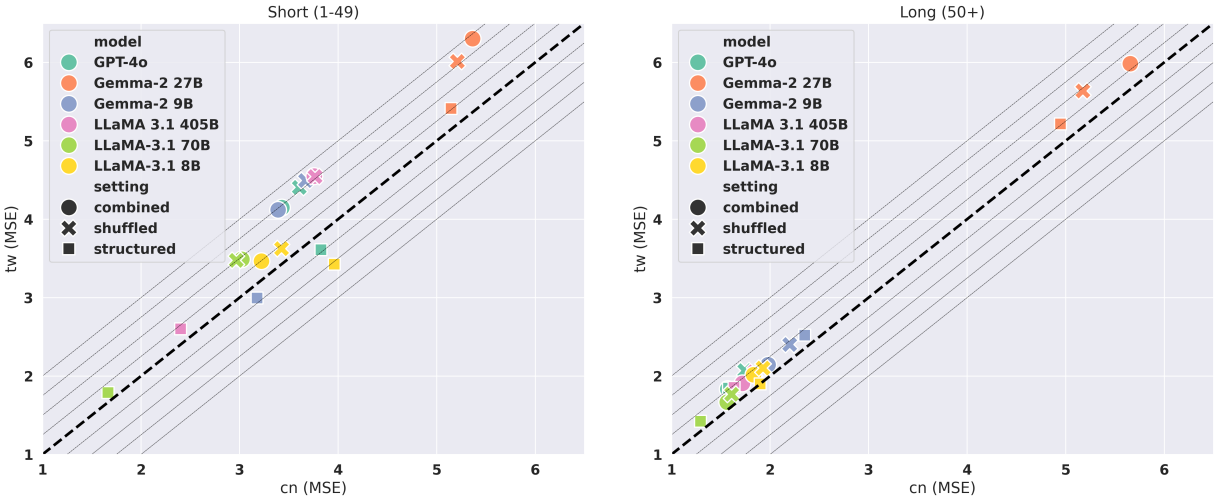

Figure 3: Comparison of MSE between Mainland Mandarin and Taiwan Mandarin for short (left) and long (right) texts. Each point represents a model's performance. The diagonal line ($x = y$) indicates equal performance. Points below the line suggest better performance in Taiwan Mandarin, while points above suggest better performance in Mainland Mandarin. Note the larger performance gap for short texts compared to long texts.

	Traditional	Simplified	Both	Others
TW	22,033	104	351	430
CN	2,443	19,315	742	418

Table 6: Language distribution. Both means that both the traditional and simplified Chinese appear in the text. Others mean that there are other languages involved or the language is unknown.

texts, where models have to judge based on its prior knowledge.

G Language Detection Analysis

To investigate the relation between self-reported nationality and language usage in reviews, we conducted an analysis using Hanzidentifier to identify Chinese language variations. We concatenated the review title, positive review, and negative review for each entry, separating them with newline characters, and then processed this combined text using Hanzidentifier. Table 6 presents the distribution of identified languages. In this context, Mainland Mandarin is referred to as Simplified Chinese, and Taiwan Mandarin as Traditional Chinese. The "Both" category indicates the presence of both Simplified and Traditional Chinese within a single text, while "Others" means texts containing other languages or those where the language could not be determined (possibly due to insufficient text length). Our findings reveal that the CN group exhibits a higher ratio (15.72%) to use languages other than Simplified Chinese. In contrast, the TW group shows a much lower rate (3.86%) of using languages other than Traditional Chinese.