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

Zixin Tang[1] **Chieh-Yang Huang**[2] **Tsung-Chi Li**[3] **Ho Yin Sam Ng**[1]
Hen-Hsen Huang[3] **Ting-Hao 'Kenneth' Huang**[1]
[1]College of Information Sciences and Technology, The Pennsylvania State University
[2]MetaMetrics Inc. [3]Institute of Information Science, Academia Sinica
[1]{zxtang,sam.ng,txh710}@psu.edu [2]cyhuang@lexile.com
[3]{george,hhhuang}@iis.sinica.edu.tw

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

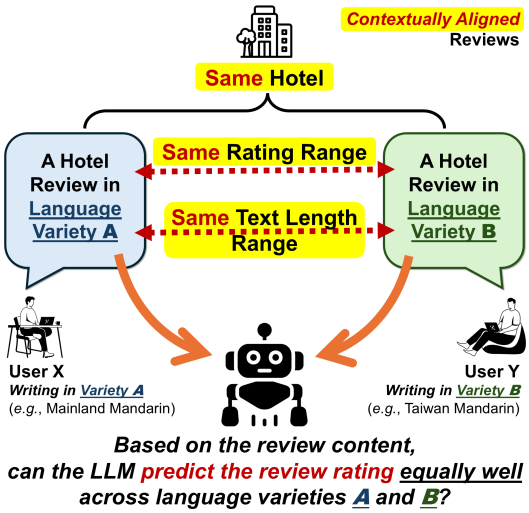

Figure 1: Online review platforms can be data sources to build datasets that capture comments in different language varieties from similar real-world scenarios. These *contextually aligned* datasets can then be used to benchmark LLMs' performance across language varieties.

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.

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

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 342–355

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). These datasets, being **contextually aligned**, can then be used to benchmark LLMs' performance across language varieties for tasks like sentiment analysis and text generation (Figure 1). 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`,[1] 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 from Booking.com to determine the reviews' language varieties. In total, we collected 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 (TW)** and **Mainland Mandarin (CN)** reviews, we paired them based on 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.

[1]Data processing code: https://github.com/Crowd-AI-Lab/Contextually-Aligned-Online-Reviews

Text Length (#Character)	Model	Accuracy (Acc)↑								
		structured			plain			shuffled		
		tw	cn	ΔAcc (cn-tw)	tw	cn	ΔAcc (cn-tw)	tw	cn	ΔAcc (cn-tw)
Short (1-49)	GPT-4o	26.52	27.43	0.91	19.16	20.78	1.62***	18.57	20.16	1.60***
	Llama3 8b	27.40	26.39	-1.01	19.21	19.08	-0.13	17.43	17.71	0.28
	Llama3 70b	35.43	35.00	-0.43	28.21	29.60	1.39**	27.54	29.51	1.97***
	Llama3 405b	37.96	40.51	2.55***	27.42	30.12	2.70***	27.59	30.17	2.58***
	Gemma2 9b	15.69	14.45	-1.24**	17.01	17.26	0.25	15.81	16.35	0.54
	Gemma2 27b	15.34	14.27	-1.07**	13.94	14.03	0.09	13.91	14.29	0.37
Long (50+)	GPT-4o	35.59	38.39	2.79***	28.15	33.16	5.01***	26.73	31.36	4.64***
	Llama3 8b	25.31	27.01	1.70*	19.53	21.24	1.71**	18.92	21.11	2.19***
	Llama3 70b	34.66	38.24	3.59***	35.02	37.45	2.43**	33.66	36.43	2.77***
	Llama3 405b	37.20	40.52	3.31***	36.09	38.00	1.91*	34.38	36.60	2.22**
	Gemma2 9b	14.84	15.66	0.82	18.22	20.00	1.78**	16.59	17.98	1.38*
	Gemma2 27b	13.44	14.52	1.08	15.48	16.99	1.51*	15.16	17.16	2.00***
Overall	GPT-4o	29.61	31.16	1.55***	22.22	24.99	2.78***	21.35	23.98	2.63***
	Llama3 8b	26.69	26.61	-0.08	19.32	19.82	0.50	17.94	18.88	0.94*
	Llama3 70b	35.16	36.10	0.94*	30.53	32.27	1.75***	29.62	31.87	2.24***
	Llama3 405b	37.70	40.51	2.81***	30.39	32.82	2.43***	29.92	32.38	2.46***
	Gemma2 9b	15.40	14.86	-0.54	17.42	18.19	0.77*	16.07	16.90	0.83*
	Gemma2 27b	14.69	14.35	-0.34	14.47	15.04	0.57	14.34	15.27	0.93**

Table 1: Accuracy (Acc ↑) by length for GPT-4o, Llama3 (8b, 70b, 405b), and Gemma2 (9b, 27b) models. Red (green) indicates better (worse) performance in CN, with darker shades representing larger gaps. (Statistical group differences are indicated as * (p<.05), ** (p<.01), and *** (p<.001) regarding the model performance.)

- **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 ensure paired reviews have similar text lengths, we grouped reviews into 10-token bins before pairing and required both reviews in each pair to fall within the same length bin. Reviews longer than 500 tokens were excluded (see Appendix E.)

The final dataset contained 22,918 review pairs, each with one TW and one CN user review.

3.2 Data Quality Validation

Five native speakers of Taiwan Mandarin reviewed 200 random Taiwan Mandarin reviews; the same process applied to Mainland Mandarin. The focus was on two key aspects: *(i)* **writing quality** and *(ii)* **content-rating agreement**, evaluated on a 5-point Likert scale (see Appendix B.1.) 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 examine biases from review structure, we tested three settings: *(i)* **Structured review** retains the original format with title, positive, and negative feedback. *(ii)* **Plain review** concatenates all elements into a single paragraph. *(iii)* **Shuffled review** includes all elements but in random order. 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 (Appendix C) includes the task description, the review content, and the prediction scale (1-10). Table 1 and Table 2 show the prediction accuracy (Acc) and mean squared error (MSE) across models and settings (see Appendix D for valid prediction counts.)

LLMs performed significantly worse in Taiwan Mandarin compared to Mainland Mandarin. Among all 54 experiments with different models and prompt settings, 38 of them had significant group differences in accuracy (Table 1), and 47 had significant group differences in MSE (Table 2). Among all significant accuracy differences, LLMs

Text Length (#Character)	Model	Mean Squared Error (MSE) ↓								
		structured			plain			shuffled		
		tw	cn	ΔMSE (cn-tw)	tw	cn	ΔMSE (cn-tw)	tw	cn	ΔMSE (cn-tw)
Short (1-49)	GPT-4o	3.563	3.769	0.206***	4.091	3.385	-0.706***	4.347	3.561	-0.786***
	Llama3 8b	2.187	2.268	0.082	2.999	2.801	-0.199***	3.377	3.016	-0.361***
	Llama3 70b	1.732	1.626	-0.107**	2.977	2.534	-0.443***	3.006	2.605	-0.401***
	Llama3 405b	2.782	2.635	-0.147	4.624	3.685	-0.939***	4.620	3.740	-0.880***
	Gemma2 9b	3.026	3.164	0.138*	4.483	3.828	-0.655***	4.928	4.131	-0.797***
	Gemma2 27b	2.945	3.028	0.083	4.888	4.191	-0.697***	4.944	4.250	-0.693***
Long (50+)	GPT-4o	1.846	1.577	-0.269***	1.834	1.57	-0.264***	2.070	1.743	-0.327***
	Llama3 8b	1.674	1.548	-0.127***	2.046	1.895	-0.152***	2.127	1.906	-0.220***
	Llama3 70b	1.473	1.302	-0.171***	1.534	1.406	-0.128**	1.671	1.495	-0.176***
	Llama3 405b	1.910	1.674	-0.236***	1.909	1.766	-0.143*	2.085	1.892	-0.194**
	Gemma2 9b	2.479	2.337	-0.142**	2.199	2.024	-0.175***	2.511	2.294	-0.217***
	Gemma2 27b	2.703	2.519	-0.184***	2.680	2.500	-0.180***	2.649	2.496	-0.153**
Overall	GPT-4o	2.978	3.022	0.044	3.323	2.767	-0.555***	3.571	2.942	-0.630***
	Llama3 8b	2.011	2.021	0.010	2.672	2.490	-0.182***	2.948	2.635	-0.313***
	Llama3 70b	1.644	1.515	-0.129***	2.486	2.150	-0.335***	2.551	2.227	-0.324***
	Llama3 405b	2.483	2.306	-0.177***	3.695	3.028	-0.667***	3.752	3.107	-0.645***
	Gemma2 9b	2.840	2.882	0.043	3.705	3.213	-0.491***	4.105	3.505	-0.600***
	Gemma2 27b	2.863	2.855	-0.008	4.136	3.615	-0.521***	4.162	3.653	-0.509***

Table 2: Mean squared error (MSE ↓) by length for GPT-4o, Llama3 (8b, 70b, 405b), and Gemma2 (9b, 27b) models. Statistical significance notations and color coding follow the same conventions as in Table 2.

made less accurate sentiment predictions toward Taiwan Mandarin users (36 out of 38 in Acc, and 45 out of 47 in MSE).

When the reviews' structures are disrupted, the performance gap increases. Table 1 and Table 2 show that structured input reduces performance gaps and generally improves model performance. Without knowing the structure inside reviews (*i.e.*, plain or shuffled cases), bias toward Taiwan Mandarin and Mainland Mandarin increases.

Shorter reviews tend to produce larger MSE gaps. Our pilot study (Appendix E) found that shorter texts may lack information and often affect model performance and behavior. We thus categorized our dataset into two groups based on review's text length: short (1-49 Chinese characters) and long (50+ Chinese characters). Table 2 shows that the MSE gap between Taiwan Mandarin and Mainland Mandarin widens in the short text group (also see Figure 2 in Appendix E), while this trend is less clear for Acc (Table 1).

4.1 Can We Just Use Machine Translation?

A natural question is whether we could 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 (*e.g.*, [a review in TW, its translation into CN].) The results (Table 3) show an **asymmetry between the two translation directions**. Translating Taiwan Mandarin data to Mainland Mandarin increased accuracy and decreased MSE (Table 3's 1st, 3rd, and 5th rows). However, translating Mainland Mandarin to Taiwan Mandarin produced mixed results: it decreased accuracy but improved MSE. These results suggest that while using machine translation to create review pairs between language varieties is technically feasible, it can introduce an additional layer of bias, as machine translation itself is a language technology that is not immune from biases across

	Ori.	Acc↑			MSE↓		
		tw	cn	ΔAcc (cn-tw)	tw	cn	ΔMSE (cn-tw)
stru.	tw	29.60	30.20	0.60*	2.985	2.036	-0.948***
	cn	30.31	31.16	0.85**	1.969	3.026	1.056***
plain	tw	22.26	23.06	0.80***	3.262	2.577	-0.686***
	cn	24.03	25.02	0.99***	2.267	2.727	0.460***
shuf.	tw	21.40	22.10	0.70***	3.489	2.688	-0.802***
	cn	23.48	24.01	0.53**	2.393	2.901	0.508***

Table 3: GPT-4o performance on original (Ori.) and machine-translated texts. TW-to-CN translation improved Acc and MSE; CN-to-TW showed mixed results. Statistical significance notations and color coding follow the same conventions as in Table 2.

language varieties. In our case, machine translation might be better at Taiwan Mandarin to Mainland Mandarin than the other way around (Kantharuban et al., 2023). Furthermore, mature machine translation systems for specific language varieties are not always readily available (Ziems et al., 2023; Kumar et al., 2021).

5 Examining Confounding Variables

Could the performance gap be due to Mainland Mandarin reviews having better writing quality or better alignment between content and ratings? *Rationale:* Better writing quality or better content-rating alignment could make it easier for LLMs to predict ratings. *Analysis & Findings:* **No.** Our human validation (Section 3.2) shows that Mainland Mandarin reviews had slightly worse writing quality and content-rating alignment.

Could the performance gap be due to more code-mixed usage in Taiwan Mandarin? *Rationale:* NLP models often struggle with code-mixed data (Zhang et al., 2023; Ochieng et al., 2024). *Analysis & Findings:* **No.** The Mainland Mandarin reviews contain more mixed-language input (30.99%) than the Taiwan Mandarin reviews (25.26%, see Appendix G and Table 8).

Could the performance gap be due to Mainland Mandarin users systematically giving higher scores, which align better with LLM-generated scores? *Rationale:* LLMs tend to assign higher scores (Stureborg et al., 2024; Kobayashi et al., 2024; Golchin et al., 2025). *Analysis & Findings:* **Unlikely.** In our dataset, Taiwan Mandarin and Mainland Mandarin reviews show no significant difference in scores ($t(22917) = .160$, $p = .873$).

Are Mainland Mandarin reviews easier for humans to guess ratings? *Rationale:* Human performance is sometimes used as an indicator of a task's difficulty for LLMs (Sakamoto et al., 2025; Ding et al., 2024). *Analysis & Findings:* **Plausible.** We conducted a user study with 10 participants (5 native speakers from each variety) who reviewed 50 random CN-TW review pairs (100 total reviews) and predicted their rating scores. Participants performed significantly better at predicting ratings for reviews in Mainland Mandarin. After excluding two TW native speakers whose accuracy was more than two standard deviations below the mean, 6 out of the 8 participants had better accuracy on CN reviews than TW reviews, and 7 had better (lower) MSE on CN reviews than TW reviews (see Appendix B.2 for more details).

These results should be interpreted with caution. Unlike question-answering, predicting hundreds of review scores from content is not a typical human task, and most NLP papers on sentiment analysis do not compare model performance to human performance. Thus, it is unclear whether human performance gaps in such tasks reliably indicate task difficulty for LLMs, especially given the small differences between the two varieties. Additionally, our participants may not represent the average Mandarin speaker's ability in sentiment analysis, as the two participants performed notably poorly. Finally, despite our efforts to examine confounding variables such as text length, code-mixing, and writing quality, we still **lack a clear understanding of what causes the observed LLMs' performance gaps across language varieties**.

6 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, analysis using predefined character sets revealed that 30.99% of samples in the CN group contained characters beyond simplified Chinese, and 25.26% 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 8 shows the distribution of Chinese script variants among users.

7 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.

8 Limitations

As the study that is among the first to benchmark LLMs' performance across language varieties using contextually aligned data, this study and the data pairing method we introduced have several limitations.

- The first limitation is that, despite the contextual alignment, unknown confounding factors might contribute to 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. Relaxing strict semantic alignment between paired text items inevitably introduces confounding variables. We believe that this trade-off is worth exploring because it enables researchers to compare model behaviors across language varieties in new ways.

- 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.

- A third limitation concerns our machine translation-based analysis. We recognize that the observed performance differences when translating between Taiwan Mandarin and Mainland Mandarin may arise from a combination of morphosyntactic variations, script differences, and normalization of non-Chinese script elements. More importantly, while MT-based approaches are technically feasible, they can introduce additional biases, as MT systems themselves exhibit performance disparities across language varieties. Further analyses are required to better isolate and address these compounding factors.

9 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.

Acknowledgement

We thank the anonymous reviewers for their feedback and the participants for their contributions to our human studies. This work was partially supported by the 2024-2025 Seed Grant from the College of Information Sciences and Technology at Pennsylvania State University. We also acknowledge Dr. Janet G. van Hell, Co-PI of the seed grant, for her support and valuable input. Additionally, this work was partially supported by the National Science and Technology Council (NSTC), Taiwan, under the project "*Taiwan's 113th Year Endeavoring in the Promotion of a Trustworthy Generative AI Large Language Model and the Cultivation of Literacy Capabilities (Trustworthy AI Dialog Engine, TAIDE)*."

References

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*.

carpedm20. emoji : emoji terminal output for python. https://github.com/carpedm20/emoji.

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*.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, and Furong Huang. 2024. Easy2hard-bench: Standardized difficulty labels for profiling LLM performance and generalization. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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*.

Shahriar Golchin, Nikhil Garuda, Christopher Impey, and Matthew Wenger. 2025. Grading massive open online courses using large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3899–3912, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 68–77, Mexico City, Mexico. Association for Computational Linguistics.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

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*.

Taku Sakamoto, Saku Sugawara, and Akiko Aizawa. 2025. Development of numerical error detection tasks to analyze the numerical capabilities of language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9957–9976, Abu Dhabi, UAE. Association for Computational Linguistics.

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*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

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*.

tsroten. Zhon: Constants used in chinese text processing. `https://github.com/tsroten/zhon`.

Unicode. Unicode character database. `https://www.unicode.org/reports/tr44/`.

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.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Fikri Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

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 4 shows a sample of the collected Booking.com review.

B Human Validation

B.1 Questions for data quality 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。

B.2 Score prediction

We used the following questions to further investigate potential content differences in review pairs, which can further lead to gaps in LLMs' performance differences. In this study, participants were asked to rate 1) the readability of the review, 2) the overall nativeness of the review, and 3) the score of the review. For the convenience of reading, all reviews were converted into either traditional or simplified Chinese characters so that all participants could process them in the writing style of their native language variety. Both English and its translation, in either Mainland Mandarin or Taiwan Mandarin based on the participants' language background, were provided in the instruction.

1. Readability (1-5), where: 1 = The writing doesn't contain any literal information; 3 = The writing requires additional effort to process/comprehend; 5 = The writing is fluent and clear in terms of content delivery
   - Taiwan Mandarin: 評論可讀性(1-5分)，其中：1分表示評論不具備可讀性，或其語句無任何實際意義；3分表示評論存在語句不通的情況，且該情況會導致歧義或理解困難；5分表示評論語句通順，表達連貫，語義明確且清晰。
   - Mainland Mandarin: 评论可读性(1-5分)，其中：1分表示评论不具备可读性，或其语句无任何实质意义；3分表示评论存在语句不通的情况或语病，且该情况会影响阅读或理解；5分表示评论语句通顺，表达连贯，语义明确且清晰。
2. Nativeness - the review is generated by: 1. a less proficient non-native Chinese speaker; 2. a highly proficient non-native Chinese speaker or a native Chinese speaker; 3. machine translation from another language; or 4. not sure/inconclusive
   - Taiwan Mandarin: 你覺得該評論可能出自：1. 低水平的中文非母語者；2. 高水平的中文非母語者或中文母語者；3. 來

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

自其他語言的機器翻譯；4. 不確定/無法判斷。

- Mainland Mandarin: 你觉得该评论可能出自：1. 低水平中文非母语者；2. 高水平中文非母语者或中文母语者；3. 来自其他语言的机器翻译；4. 不确定/无法判断。

3. Score Rating (1-10, 1 is the lowest, 10 is the highest)

   - Taiwan Mandarin: 旅館評分(1-10，1為最差，10為最好)
   - Mainland Mandarin: 酒店评分(1-10，1为最差，10为最好)

We further excluded two participants' responses due to the lack of score agreement against other participants and their significantly lower performance in prediction accuracy. Among the other 8 participants, there are no significant differences in score predictions among the data pairs, indicating raters have no biases in reading and understanding reviews from either group of speakers/writers. However, results showed statistical significance in both Accuracy (37.00% vs. 28.75%, p=.016) and MSE (2.795 vs. 3.510, p=.036), showing that native speakers might have more difficulties in correctly guessing the review scores for reviews in Taiwan Mandarin.

C Prompts

The following prompt is used for the 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 the plain and shuffled conditions.

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.

D Distribution of Valid and Invalid Predictions

Table 5 and Table 6 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. We only included pairs with completely valid data entries for the prediction analysis (Table 1 and Table 2), referring to the smallest number of each model in Table 5.

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 2 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 3 and Figure 4.

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 3), 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 4), points closer to the top-left indicate better performance in Mainland Mandarin.

Our analysis of Figure 3 does not reveal a significant difference between the short and long text groups in terms of accuracy. However, Figure 4 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

model	All			Short			Long		
	plain	shuffled	structured	plain	shuffled	structured	plain	shuffled	structured
GPT-4o	45,828	45,830	45,836	45,828	45,830	45,836	45,836	45,836	45,836
LLaMA-3.1 8B	45,668	45,707	45,697	45,694	45,726	45,712	45,810	45,817	45,821
LLaMA-3.1 70B	45,835	45,835	45,834	45,835	45,835	45,834	45,836	45,836	45,836
LLaMA 3.1 405B	45,805	45,795	45,706	45,808	45,801	45,710	45,833	45,830	45,832
Gemma-2 9B	45,836	45,836	45,819	45,836	45,836	45,820	45,836	45,836	45,835
Gemma-2 27B	45,833	45,833	45,824	45,833	45,833	45,824	45,836	45,836	45,836
GPT-4o+Translation	45,682	45,644	45,836	-	-	-	-	-	-

Table 5: Number of valid prediction samples in the study across different models and data configurations.

model	All			Short			Long		
	plain	shuffled	structured	plain	shuffled	structured	plain	shuffled	structured
GPT-4o	−8	−6	0	−8	−6	0	0	0	0
LLaMA-3.1 8B	−168	−129	−139	−142	−110	−124	−26	−19	−15
LLaMA-3.1 70B	−1	−1	−2	−1	−1	−2	0	0	0
LLaMA 3.1 405B	−31	−41	−130	−28	−35	−126	−3	−6	−4
Gemma-2 9B	0	0	−17	0	0	−16	0	0	−1
Gemma-2 27B	−3	−3	−12	−3	−3	−12	0	0	0
GPT-4o+Translation	−154	−192	0	-	-	-	-	-	-

Table 6: 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.

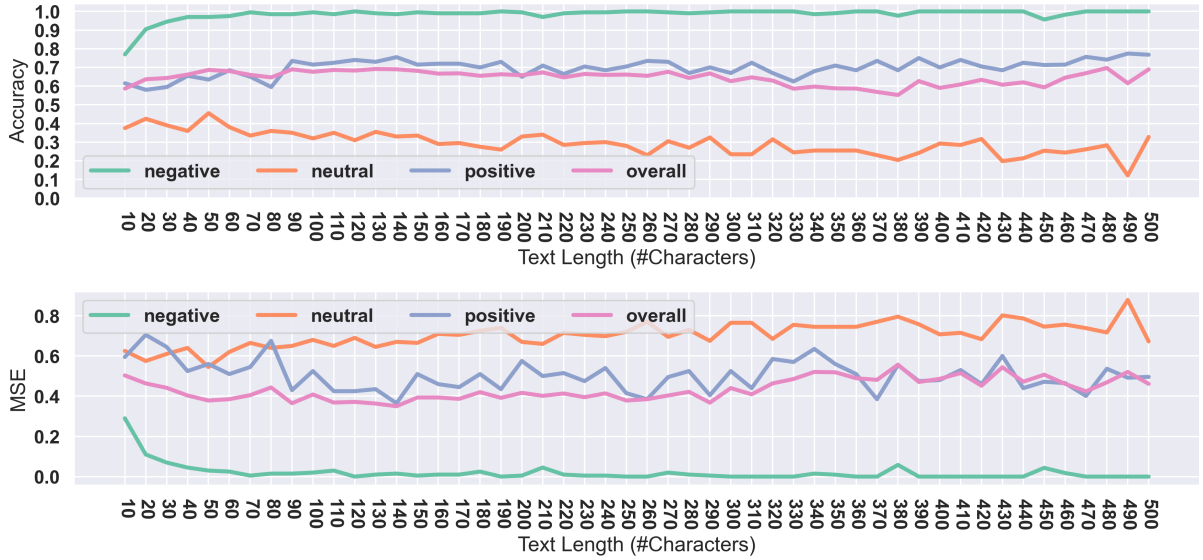

Figure 2: Impact of text length on sentiment classification performance. The top graph shows accuracy, and the bottom graph shows MSE for negative, neutral, positive, and overall sentiments across different text lengths (0-500 characters). While overall performance remains relatively stable, individual sentiment categories show varying levels of accuracy and error, particularly for shorter texts.

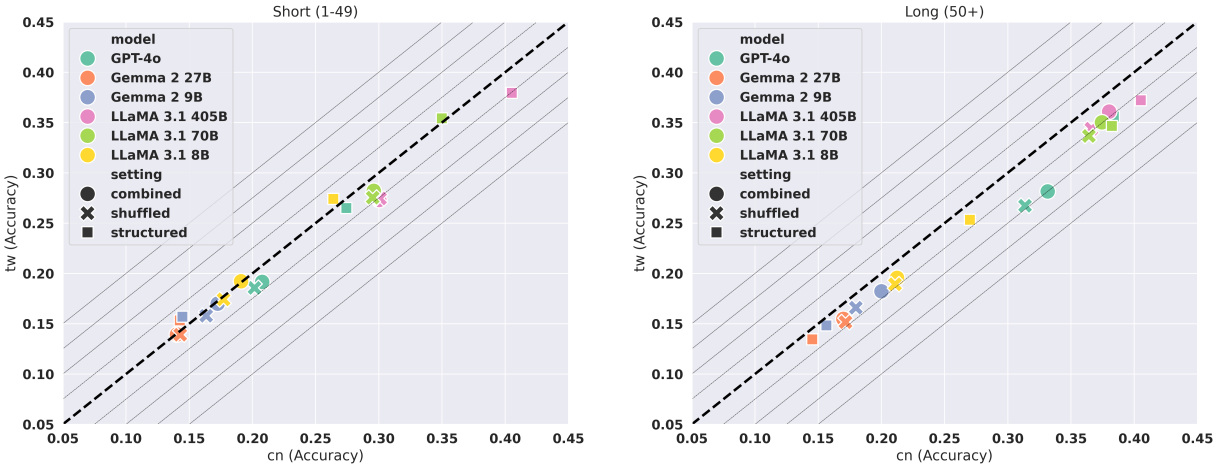

Figure 3: Comparison of accuracy between Mainland Mandarin and Taiwan Mandarin for short (left) and long (right) texts. Each point represents a [model, setting]'s performance. The diagonal line ($x = y$) indicates equal performance. Points above the line suggest better performance in Taiwan Mandarin, while points below suggest better performance in Mainland Mandarin. We do not see a big difference between the short and long texts.

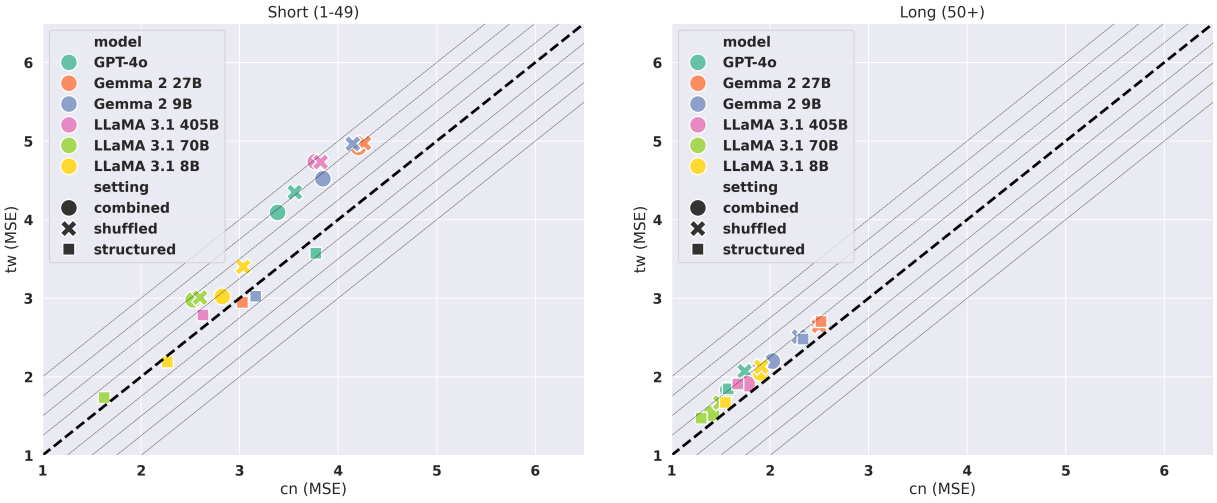

Figure 4: Comparison of MSE between Mainland Mandarin and Taiwan Mandarin for short (left) and long (right) texts. Each point represents a model's performance. The diagonal line ($x = y$) indicates equal performance. Points below the line suggest better performance in Taiwan Mandarin, while points above suggest better performance in Mainland Mandarin. Note the larger performance gap for short texts compared to long texts.

Category	Example
(A) Rare Chinese characters	夲
(B) Fullwidth Latin letters	ＪＲＯＫ
(C) Emoticon-based Special Characters	ദ്ദി(˵ •̀ ᴗ - ˵)

Table 7: Example of special characters found in our dataset.

Category	CN		TW	
	Count	Ratio	Count	Ratio
Only Traditional	2,000	8.73%	17,130	74.74%
Only Simplified	15,816	69.01%	90	0.39%
Only English	107	0.47%	119	0.52%
Only Emoji	1	0.00%	4	0.02%
Only Symbol	1	0.00%	5	0.02%
Only Bopomofo	4	0.02%	35	0.15%
Only JP/KR	0	0.00%	0	0.00%
Only Punctuation	4	0.02%	5	0.02%
Only Unknown	0	0.00%	0	0.00%
Traditional + English	251	1.10%	3,022	13.19%
Traditional + Emoji	75	0.33%	666	2.91%
Traditional + Symbol	79	0.34%	894	3.90%
Traditional + Bopomofo	8	0.03%	66	0.29%
Traditional + JP/KR	0	0.00%	9	0.04%
Traditional + Unknown	30	0.13%	246	1.07%
Simplified + English	2,681	11.70%	12	0.05%
Simplified + Emoji	383	1.67%	1	0.00%
Simplified + Symbol	323	1.41%	0	0.00%
Simplified + Bopomofo	0	0.00%	0	0.00%
Simplified + JP/KR	22	0.10%	0	0.00%
Simplified + Unknown	90	0.39%	1	0.00%

Table 8: Language distribution. CN and TW users similarly mix non-Chinese elements with their primary writing systems (Simplified or Traditional Chinese). However, CN users incorporate Traditional characters more frequently than TW users use Simplified ones.

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

G Language Detection Analysis

To have a better understanding of Chinese and non-Chinese script elements in reviews, we conducted a detailed character-level analysis across our dataset. Using predefined vocabulary sets from zhon (tsroten), the Unicode Character Database (Unicode), and emoji (carpedm20), we categorized characters into the following groups: traditional Chinese characters, simplified Chinese characters, English letters, emojis, bopomofo, Japanese characters, Korean characters, mathematical symbols, punctuation, and numbers. The table below presents the distribution of these elements across CN and TW users' reviews.

Our analysis revealed that CN and TW users exhibit similar patterns when incorporating non-Chinese elements into their primary writing system (Simplified Chinese with other elements for CN users, Traditional Chinese with other elements for TW users). The key difference lies in cross-script usage: CN users demonstrate a higher frequency of Traditional character usage compared to TW users' usage of Simplified characters.

Beyond the identified script elements, we found 103 characters in an "Unknown" category, appearing across 388 samples. Further investigation revealed these primarily consist of (1) rare Chinese characters not included in the zhon (tsroten) vocabulary list (7 (A)), (2) fullwidth Latin letters (7 (B)), and (3) characters from other languages, with the latter mainly used in emoticons (7 (C)). As our current analysis is conducted at the character level, we cannot identify complete pinyin words or emoticon compositions. We will acknowledge this limitation and encourage future research to explore these aspects more comprehensively.

How Non-Chinese Elements Affect LLM Performance? To investigate how non-Chinese elements affect LLM performance, we analyzed GPT-4o's performance on review pairs under different language constraints. We define "Chinese" as the primary writing system for each user group (Traditional for Taiwan Mandarin users, Simplified for Mainland Mandarin users). We included only pairs where both reviews strictly adhered to these constraints. For instance, Mainland Mandarin reviews must contain only Simplified Chinese characters, while Taiwan Mandarin reviews must contain only Traditional Chinese characters. "Chinese+English" refers to reviews containing only the primary Chinese writing system plus English letters.

The results are presented in 9. When restricting the analysis to primary Chinese characters only (the Chinese row), the performance gap between Taiwan Mandarin and Mainland Mandarin widened (see [plain, ΔMSE] and [shuffled, ΔMSE]), indicating a potential bias in processing Traditional versus Simplified Chinese characters. In the code-switching scenario with English letters, both groups showed relatively closer performance, with a smaller gap between them. This suggests that English elements may help normalize the performance across both language groups.

Setting	Char. Set	#Pairs	Acc↑			MSE↓		
			tw	cn	ΔAcc (cn-tw)	tw	cn	ΔMSE (cn-tw)
structured	**All**	22,918	29.614	31.172	1.558***	2.985	3.026	0.206
structured	**Chinese**	12,237	28.193	29.901	1.708**	2.965	3.013	0.082
structured	**Chinese+English**	917	37.514	37.077	-0.436	1.762	1.700	-0.107
plain	**All**	22,914	22.231	25.011	2.780***	3.323	2.768	-0.147***
plain	**Chinese**	12,237	21.051	24.197	3.146***	3.335	2.642	0.138***
plain	**Chinese+English**	917	28.571	30.862	2.290	1.943	1.799	0.083
shuffled	**All**	22,915	21.353	24.002	2.649***	3.573	2.941	-0.269***
shuffled	**Chinese**	12,237	20.315	22.857	2.542***	3.580	2.808	-0.772***
shuffled	**Chinese+English**	917	26.609	28.680	2.072	2.196	1.937	-0.260

Table 9: Analysis of LLM performance across different character sets.