LOST IN TRANSLATION: MEASURING THE EFFECT OF NON-NATIVE ENGLISH ON END USER PERFORMANCE OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used by people whose first language is not English, yet these users have been shown to receive systematically lower-quality responses than fluent speakers. Which specific features of non-native English drive this gap remains unclear, because fluency is itself a composite of mechanical accuracy, vocabulary use, organization, and discourse coherence. Here, we introduce FABLE, a controlled dataset of 190,911 English prompt variants derived from 174K real user prompts for writing-related tasks. Evaluating responses from 34 open-weight LLMs, we find a clear asymmetry; while models do not propagate surface errors such as misspellings into their outputs, models do mirror higher-level rhetorical and lexical qualities present in the user's prompt. Further, the overall quality of responses differs substantially between the least- and most-fluent prompts. These results highlight a key LLM performance disparity for non-native English LLM users, resulting in both lower-quality and less-fluent answers.

1 INTRODUCTION

Non-native English speakers represent a critical and rapidly growing user base for large language models (LLMs) (Liu et al., 2025a). However, LLMs have demonstrated performance degradation for non-native English speakers (Reusens et al., 2025). Intuitively, this performance degradation may stem from differences in the language produced by non-native speakers of different fluency levels, compared to the native English that LLMs are trained on. Fluency, however, is not a single property: it is a composite of features ranging from surface mechanics such as spelling and grammar to higher-level qualities such as organization and discourse coherence (Chambers, 1997). Without knowing which of these features matter for LLM behavior, it is difficult to diagnose why non-native speakers are disadvantaged and to design effective interventions to improve LLM performance for these users. Here, we systematically test which features of non-native English drive performance degradation for non-native speakers.

Existing work on prompt sensitivity has examined how surface-level variations such as formatting (He et al., 2024), example ordering (Guo et al., 2024; Lu et al., 2022), length (Liu et al., 2025b), tone (Dobariya & Kumar, 2025), and politeness (Yin et al., 2024) affect model behavior. A smaller line of work has begun to consider specific linguistic properties such as tense, mood, and voice (Leidinger et al., 2023) and the broad effect of speaker nativeness (Reusens et al., 2025). However, these studies typically treat fluency as a single binary or scalar property and rely on benchmark-style prompts that do not reflect how everyday users actually interact with LLMs. As a result, it remains unclear which specific aspects of non-native English are most consequential for downstream task performance, or whether their effects are consistent across realistic, open-ended user tasks.

We address this gap by introducing FABLE (Fluency-Adjusted Benchmark for LLM Generation Evaluation), a controlled dataset of 190,911 English prompt variants derived from real-world user interactions for writing-related tasks. Starting from 174K filtered prompts in English and Chinese collected from WildChat (Zhao et al., 2024) and ShareLM (Don-Yehiya et al., 2025)—a language pair representing one of the largest populations of bilingual LLM users (Yang, 2006)—we generate multiple English translations of each Chinese prompt that conform to specified profiles along the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

five dimensions of the ESL Composition Profile (Jacobs et al., 1981). This design disentangles the contribution of each fluency dimension while holding the prompt's semantic intent approximately constant. We evaluate 34 open-weight LLMs on a shared sample of 5,000 prompt variants, and evaluate their responses using both rule-based linguistic metrics and LLM-as-judge ratings of fluency and overall response quality.

Our analysis reveals a striking asymmetry between Surface Form Features and higher-order qualities. LLMs do not substantially propagate surface errors (e.g., misspellings, grammatical mistakes) from prompts into their responses, but they do mirror higher-level rhetorical and lexical qualities: prompts with weaker Content, Organization, or Vocabulary yield responses with correspondingly weaker proficiency along those same dimensions. More importantly, prompt fluency has a substantial effect on response quality: a one-point increase in each retained prompt-fluency dimension on the four-point scale leads to a 0.04–0.14 point increase in overall response quality on a five-point scale. This effect is consistent across writing-focused task categories, although the relative importance of individual fluency dimensions varies by task. Together, these findings suggest that users who turn to LLMs to compensate for limited English fluency may still receive systematically lower-quality output, with implications for equitable access to LLM technology.

Our paper makes the following three contributions. (1) We introduce FABLE, a controlled dataset of 190,911 prompt translations spanning the ESL Composition Profile that enables fine-grained study of how prompt fluency affects LLM behavior. (2) We provide an evaluation methodology that jointly quantifies surface-level and higher-order linguistic properties of both prompts and responses. (3) We demonstrate empirical evidence that LLM response quality is shaped more by the rhetorical and lexical qualities of a prompt and that, across 34 models, LLM responses to less-fluent prompts are themselves less fluent and lower quality. Our results point to an important source of inequality in access to fluent, high-quality outputs by all users. All data and code will be made available upon publication (CC-by-SA-4.0).

2 RELATED WORK

Language Fluency and Language Technology Language fluency has been shown to play a critical role across multiple NLP fields. In Machine Translation, prior work demonstrates that noisy input significantly degrades translation quality for both neural MT systems and large language models (Popović et al., 2024; Pan et al., 2024). This phenomenon can be viewed as a manifestation of reduced language fluency of input. Meanwhile, "fluency" is also recognized as a fundamental dimension in the evaluation of translation quality, with researchers proposing structured frameworks, such as MQM Park & Padó (2024), to systematically assess fluency alongside other aspects of translation quality. Similarly, in Information Retrieval, the effectiveness of retrieval systems is closely tied to the quality of user queries: well-formed queries tend to achieve better performance Chikkamath et al. (2024), while typos or other linguistic errors can substantially reduce system accuracy Zhuang & Zuccon (2022). In this work, we seek to understand the impact of language fluency on LLM responses to non-native English speakers.

Native v. Non-native Use of LLMs Prior studies about contrastive rhetoric have shown that first language and cultural background influence how people write in a second language Kaplan et al. (1966). In particular, native and non-native English speakers show systematically different usage of linguistic structure beyond semantic content Rabinovich et al. (2016). Such variation naturally carries over to interactions with LLMs; e.g., Reusens et al. (2025) demonstrate that the nativeness of prompts significantly influences model performance. Our work identifies the specific aspects of non-native prompt fluency that cause differences in model behavior.

Effects of Prompt Design Prompts that are semantically equivalent can exhibit a substantial model performance gap solely due to variations in their surface form (Cao et al., 2024). Even when the semantic content remains identical, different formatting can lead to different performance (He et al., 2024; Sclar et al., 2023). Example ordering also plays an important role in designing prompts for in-context learning (Lu et al., 2022; Guo et al., 2024). Prompt length also serves as a decisive factor in domain-specific tasks (Liu et al., 2025b). Additionally, in real-world interactions, research indicates that the tone (Dobariya & Kumar, 2025) and politeness (Yin et al., 2024) of a prompt significantly influence how models respond. This extensive evidence of LLM sensitivity to

variation in prompts' surface form provides the motivation for our own comparison of semantically equivalent prompts at different fluency levels.

While these studies provide valuable insights into different dimensions of prompts, a systematic evaluation of how the language fluency level of prompts influences model behaviors is still missing. Moreover, prior studies often evaluate prompt effectiveness using benchmark tasks that may not reflect real-world user interactions. In contrast, our work systematically investigates how variations in prompt fluency influence LLM responses on real-world interaction data.

3 Curating a Dataset of Realistic, Translated User Prompts

To investigate the impact of fluency in prompts on model responses and downstream task performance, we curate a set of prompts from real-world interactions in the WildChat (Zhao et al., 2024) and ShareLM (Don-Yehiya et al., 2025) datasets. Both datasets consist of conversations of diverse domains collected from real human-AI interactions. These prompts allow us to better match the experience of everyday users in realistic workflows (e.g., summarizing an email) when compared with prompts for multiple choice questions traditionally seen in NLP.

How Are People Using LLMs? Adapting the methodology and taxonomy proposed in Chatterji et al. (2025), we categorize each prompt into one of 24 predefined task categories. The distribution of prompts across task categories is shown in Fig 1. We observed that among the ten most common user task categories, writing-related queries account for the majority, with "Computer Programming" being the only category not directly related to writing. Moreover, we believe that LLM responses in these writing-related tasks such as "argument or summary generation," "creative ideation," and "edit or critique provided text" may be influenced by users' fluency.

Data Filtering To ensure data quality and experimental consistency, we apply a sequence of data filtering steps. For multi-turn conversations, we only keep the first user prompt. While we recognize the importance of multi-turn conversations, our focus here is on how differences in linguistic proficiency affect the initial model response, rather than whether users can *eventually* achieve comparable outcomes to native speakers after multiple rounds of repair. The first-turn model response is consequential because it determines what the user must subsequently correct, clarify, or refine. A systematic quality deficit at the first turn means that non-native users begin the interaction at a disadvantage and must expend additional turns, time, and effort to reach parity; this additional interaction burden is itself an equity cost. Additionally, effectively simulating realistic interactions across multiple turns remains a challenging problem (Ivey et al., 2024), and there are currently no curated datasets that are easily amenable to testing multi-turn effects of fluency. Therefore, we leave exploration of multi-turn scenarios to future work.

Additionally, since it is difficult to evaluate the fluency of very short prompts, and extremely long prompts are less representative of typical user queries, we perform length-based filtering. We remove both the shortest and longest 5% of prompts and randomly retain 30% of the prompts shorter than 30 tokens to control the prompt length distribution.

To filter out prompts that contain multiple languages, we perform sentence-level language identification using the fastText language identification model Joulin et al. (2017). We segment each prompt into sentences and keep prompts which have all sentences classified as the same language with confidence scores above 0.6.

In the following experiments, we focus on prompts in English and Chinese. English is the dominant language used in LLM training and inference (Qin et al., 2024), while Chinese represents one of the most-spoken non-English languages used with LLMs (Liu et al., 2024). Further, Chinese-English bilinguals worldwide make up a large, natural population of speakers with varying degrees of English fluency who are therefore affected by models' sensitivity to English fluency. Studying these two languages also reflects common real-world scenarios where bilingual users may tend to interact with LLMs in English even when it is not their first language (Dedema et al., 2026). We do recognize that non-native English speakers have many other first languages, each producing characteristic interference patterns (Kobayashi & Rinnert, 1992). The specific fluency effects we observe may not generalize to other source languages, particularly those with substantially different typological properties (e.g., morphologically rich languages, or languages with writing systems beyond Han

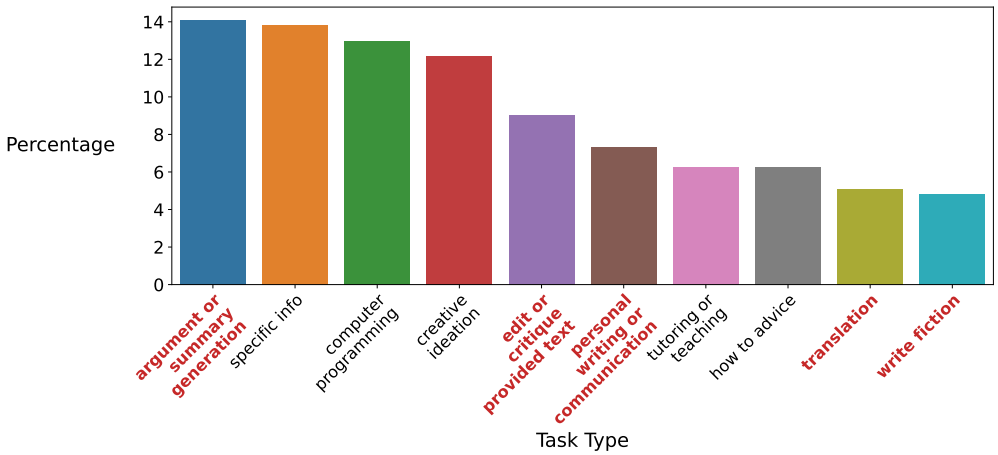

Figure 1: The top-10 task categories for Chinese prompts in WildChat and ShareLM, categorized using the taxonomy proposed by Chatterji et al. (2025), shows that many real-world user prompts involve writing and information-seeking tasks. Bold denotes categories of prompts we use.

characters). Nonetheless, the fluency taxonomy we use is language-agnostic. We therefore expect aspects of our findings to generalize beyond Chinese-speaking users.

We restrict the prompts to those focusing on writing, as these represent the most common forms of real-world LLM usage according to our analysis of WildChat and ShareLM and prior studies on real-world usage data, e.g. Chatterji et al. (2025). Specifically, we only include prompts from the following task types: "argument or summary generation," "edit or critique provided text," "personal writing or communication," "translation," and "write fiction." These prompts are well-suited for studying the relationship between prompt fluency and model behavior. After the above processing steps, our final dataset consists of 173,726 prompts. WildChat contributes 21,734 English and 27,532 Chinese prompts, while ShareLM contributes 90,121 English and 34,339 Chinese prompts. To provide a reference for downstream response quality evaluation, we use DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) to translate the original Chinese prompts into English, and treat the resulting outputs as ground-truth translations.

4 Introducing Translation Artifacts

Language fluency is a multidimensional spectrum, with many features of language contributing to the perception of fluency, such as grammatical structure, word choice, discourse coherence, or even capitalization. To more precisely control for these different factors, we introduce FABLE, which systematically constructs prompts using heterogeneous sources of disfluencies and provides multiple variants of the same prompt intent across the fluency spectrum. FABLE allows testing how specific types of linguistic behavior influence a model's downstream behavior. Here, we first describe how we measure fluency and then how we generate FABLE.

4.1 Quantifying Fluency

We quantify language fluency from two complementary perspectives: **Surface Form Features** that measure properties like mechanical competence (spelling, capitalization), and **Rhetorical and Lexical Qualities** features that measure higher-level constructs like discourse coherence and vocabulary. Surface Form Features mark the ability to control the form of the language (Witte & Faigley, 1981), while rhetorical and lexical qualities measure compositional competence (Siekmann et al., 2022).

For **Surface Form Features**, we build a rule-based evaluation pipeline based on LanguageTool (Miłkowski, 2010; LanguageTool contributors, 2026) to detect spelling, grammatical, lexical, and stylistic errors. We also compute Flesch Reading Ease to evaluate readability (Flesch, 1948).

Dimension	Description
Content	Topic relevance and depth of idea development
Organization	Logical sequencing, coherence, and paragraph structure
Vocabulary	Lexical richness, word choice, and idiomatic usage
Language Use	Grammar and sentence construction
Mechanics	Spelling, punctuation, capitalization, and formatting

Table 1: The five dimensions in the adapted ESL Composition Profile (Jacobs et al., 1981).

For **Rhetorical and Lexical Qualities**, we adopt a widely used analytic ESL writing rubric which evaluates a given text along five dimensions: Content, Organization, Vocabulary, Language Use, and Mechanics (Jacobs et al., 1981). A brief description of all dimensions is presented in Table 1. We collect these proficiency dimensions using an LLM-as-a-judge framework, where the judging model assigns an integer score from 1 to 4 for each dimension independently according to the rubric definitions. We use `google/gemma-4-26B-A4B` as judge. To validate our pipeline, we use the ICNALE dataset (Ishikawa, 2018), consisting of 656 learner-written texts annotated with four-level CEFR-aligned writer proficiency labels. We apply our judge template and take the average score from all dimensions; this score has a correlation of 0.47, suggesting the judge is moderately well calibrated on what is a hard task for humans.

4.2 Controlled Generation of Translation

To systematically study the effect of prompt fluency on models' responses, we construct FABLE to contain a range of translations with controlled fluency levels. Specifically, we create prompts that instruct according to the five dimensions of **Rhetorical and Lexical Qualities**. Each dimension is assigned a discrete level from 1 to 4, corresponding to increasing levels of proficiency in this specific dimension: Level 1 denotes *very poor* proficiency, level 2 represents *fair to poor* proficiency, level 3 means *good to average* proficiency, and level 4 stands for *excellent to very good proficiency*. Each combination of these levels defines a *translation profile*. For each reference Chinese-language prompt, we include (i) 10 translations produced by random samples of valid profiles and (ii) translations from four uniform proficiency profiles where all dimensions have the same level. We then use Qwen3-32B to generate translations that conform to the specified ESL composition profile. Appendix Table 2 presents an example of translated prompts under different instructed levels. This process yields multiple translation variants per prompt, enabling a systematic exploration of translation-induced prompt variation across different dimensions of fluency.

Because LLM-generated translations may contain extraneous text or generation artifacts, we apply post-generation quality control before constructing the final dataset. This procedure combines rule-based checks for generation-related text, repeated trigram detection, language identification, and paragraph-count checks. After post-generation quality control, FABLE contains 190,911 English prompt variants spanning the five selected writing-related task categories.

FABLE is not intended to reproduce every idiosyncratic pattern found in non-native writing. Rather, it operationalizes fluency using a theory-grounded taxonomy that decomposes this inherently difficult-to-define construct into five established and complementary dimensions: Content, Organization, Vocabulary, Language Use, and Mechanics. Taken together, these dimensions provide a robust and comprehensive representation of fluency that extends beyond surface-level grammatical correctness to include higher-level rhetorical, lexical, and organizational properties.

4.3 Evaluating Translated Prompts

To analyze the effectiveness of our approach for generating prompts with controlled fluency levels, we evaluate the translated prompts along two dimensions: language fluency and semantic preservation. For language fluency, we apply the evaluation pipeline described above in Section 4.1 to assess the **Surface Form Features** and **Rhetorical and Lexical Qualities** of each translated prompt—i.e., does instructing a model to introduce specific aspects of fluency lead to those appearing in the translation? In addition, we consider semantic preservation during translation, as semantic distortion may affect model behavior and downstream task performance.

We found that higher instructed proficiency levels generally yield corresponding improvements in measured fluency, although cross-dimensional effects remain. Semantic similarity is only weakly associated with instructed fluency, and human evaluation of 100 randomly sampled translations yields a mean semantic-preservation score of 4.13 out of 5, suggesting that the translations generally preserve the original intent. Additionally, instructed levels across all five fluency dimensions show little association with semantic preservation. Detailed methods and results are provided in Appendix A.3.

5 Quantifying the Impact of Prompt Fluency

Given two prompts with the same intent, but which differ in English fluency, how different will LLMs' responses be? We evaluate *responses* for their fluency (surface form features and rhetorical and lexical qualities), as well as how well they address the user query (response quality).

5.1 Experimental Setup

To generate responses to our fluency-controlled prompts, we evaluate 34 LLMs spanning multiple model families and sizes, including Qwen, Gemma, Llama, Nemotron, OLMo, and others. The complete list of models is provided in Appendix A.4. We randomly sample 5,000 English prompt variants from FABLE and use the same sampled set for all models.

Drawing on prior work on LLM-as-judge evaluation (Zheng et al., 2023; Liu et al., 2023), we develop a response-quality evaluation protocol using Qwen3.8-27B (Qwen Team, 2026). To mitigate the potential inconsistencies caused by varying translation quality in the user prompts, we provide the golden translations obtained earlier in Section 3 to the judging LLM. This design allows the evaluation to focus on differences in model responses rather than artifacts introduced by translation noise. (See Appendix A.6 for more details.)

To examine the evaluation protocol, we measure the consistency between the scores of the LLM judge and two human annotators on a random sample of 100 model responses. For overall response quality, human interannotator agreement is Krippendorff's $\alpha = 0.71$, and the judge's scores achieve a Spearman correlation of $\rho = 0.55$ with the mean human scores. Given the challenging and subjective nature of evaluating open-ended responses, this moderate positive correlation provides encouraging evidence that the judge captures meaningful aspects of human assessments of response quality.

5.2 Results

The effects of prompt fluency are assessed in terms of the generated language and task accuracy.

5.2.1 Response Fluency

LLMs are known to accommodate to users' style (Durandard et al., 2025). For less fluent users, such accommodation is potentially disadvantageous in writing tasks as LLMs may mirror undesired errors. For example, a user asking to help draft an email with a less-fluent prompt may receive a version that has grammatical mistakes.

Response Surface Form Features To assess the impact of the prompt's fluency, we first fit separate linear regressions to predict the Surface Form Features for each response from the input prompt's rhetorical and lexical quality levels.

Prompt fluency shows relatively weak associations with the surface form features of LLM responses. As shown in Figure 2, LLMs did not substantially accommodate to reproduce errors from the prompt. The largest changes were in the reading ease, but changes to other qualities did not produce a systematic trend of introducing more/fewer mechanistic errors. These results suggest that the LLMs we test exhibit relatively stable performance at the lower linguistic level (e.g., few grammar errors) and are largely robust to variations in prompt fluency. Details of the experiment are provided in Appendix A.7.1.

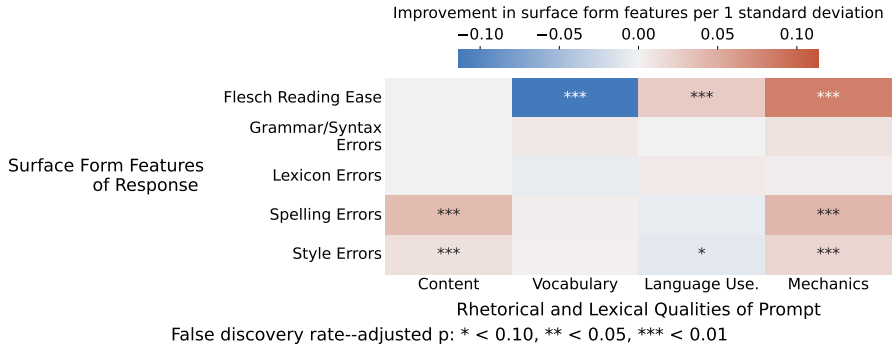

Figure 2: Prompt rhetorical and lexical qualities do not exhibit a systematic impact on model responses' surface form features. This figure presents the regression results between different linguistic properties of responses and different fluency metrics of prompts.

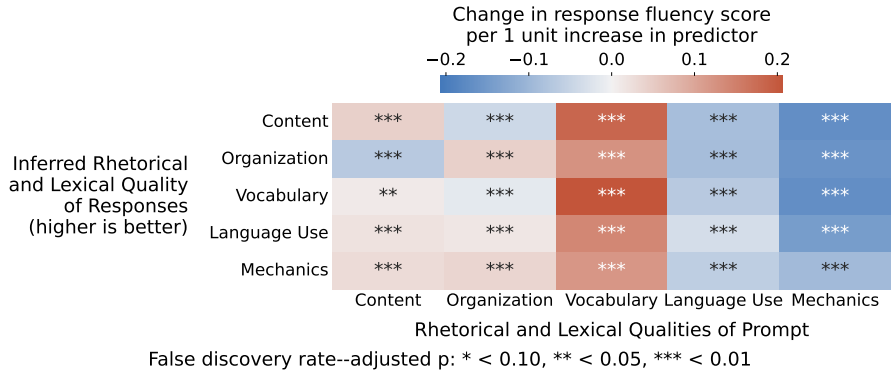

Figure 3: The prompt's qualities significantly influence the LLM's response qualities, shown by regression coefficients of the prompt's rhetorical/lexical quality levels (cols) on the response's (row).

Response Rhetorical and Lexical Qualities Do the rhetorical and lexical qualities of the prompt influence these qualities in the LLM's response? To answer this, we score responses using the same LLM-as-judge and then, similar to Surface Form Features, we fit a linear regression to predict each of the response's quality levels from the prompt's fluency profile.

LLM responses tend to mirror the rhetorical and lexical qualities of users' prompts, as shown in Figure 3. In particular, higher levels of Content, Organization, and Vocabulary show positive associations with response proficiency, whereas Language Use and Mechanics exhibit negative associations. Thus, users who write with lower levels of Organization or less diverse Vocabulary receive LLM responses matching those qualities. Perceived language proficiency is known to be associated with peer esteem (Dev & Qiqieh, 2016) and positive social outcomes (e.g., hiring; Pandey & Pandey, 2014). Our results suggest that users who are less fluent and attempting to use these models to make up for their gap in fluency will still suffer a penalty due to these LLMs' propagation of errors. Details of the experiment are provided in Appendix A.7.2.

5.2.2 Response Correctness

Response Quality Is the correctness of the models' output sensitive to the rhetorical and lexical qualities of the prompt, i.e., do models give less-accurate outputs when the prompt is less fluent? To answer this question, we evaluate using the protocol described in Section 5.1 and obtain an overall quality score for each response and analyze its relationship with prompt fluency using regression.

Prompt fluency has a notable impact on model responses, more fluent prompts generally leading to higher quality outputs, as seen in Figure 4. More specifically, a one-point increase in a prompt fluency dimension on the four-point scale is associated with an estimated 0.04–0.14 point increase in

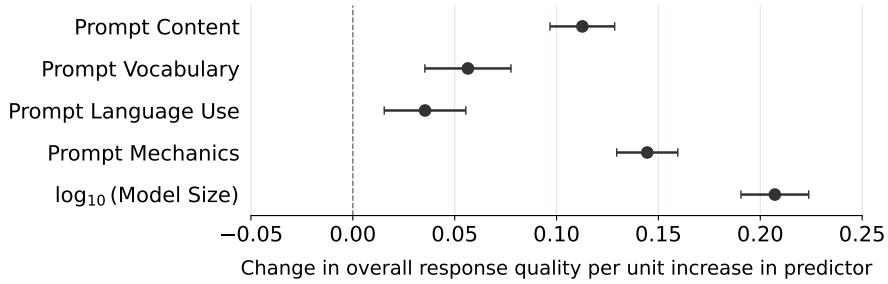

Figure 4: Higher prompt fluency leads to better response quality, with Content and Mechanics showing larger positive correlation than Vocabulary and Language Use. Points represent regression coefficients, and error bars indicate 95% confidence intervals.

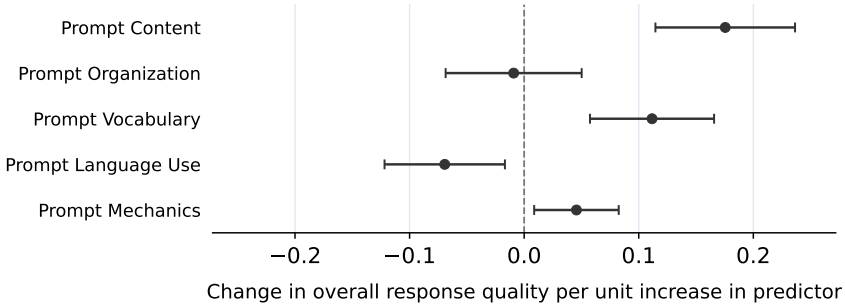

Figure 5: Different prompt fluency dimensions exhibit different impact on response quality in semantically matched comparisons. Points represent regression coefficients relating within-pair differences in each fluency dimension to differences in overall response quality.

overall response quality on a five-point scale. For reference, a tenfold increase in model size (going from a 1B to 10B model) is associated with a $\sim$0.21 increase, which suggests less fluent LLM users experience substantially worse responses, akin to those given by much less capable models. Among all metrics, Content and Mechanics show a more notable impact on response quality compared to Vocabulary or Language Use. One possible explanation is that Content and Mechanics capture how clearly a request is communicated, whereas richer vocabulary and more proficient grammar may provide smaller additional benefits when the intended request is already understandable. Details of the experiment are provided in Appendix A.7.3.

Semantic Matched Pair Analysis To further examine the relationship between prompt fluency and response quality while reducing variation in semantic content, we conduct a complementary analysis using semantically matched pairs of Chinese prompts collected from real-world user prompts. For each pair, we translate two prompts with different fluency profiles and then regress differences in overall response quality on differences in the five prompt fluency dimensions.

Even with similar semantic intent, more fluent prompts generally receive higher-quality responses, as shown in Figure 5. Content, Vocabulary, and Mechanics retain positive impact on response quality. Organization has a coefficient close to zero, whereas Language Use shows a negative association. These results further support our previous findings by being broadly consistent with the preceding analysis. For prompts with similar semantic intent, simply improving the fluency level can significantly improve the response quality. This suggests that refining how a request is expressed can help users obtain better responses in everyday interactions with LLMs. Details of the experiment are provided in Appendix A.7.4.

Does prompt fluency impact the output of some tasks more? Given that models produce less fluent and lower quality responses for less fluent prompts, here we assess whether this behavior is

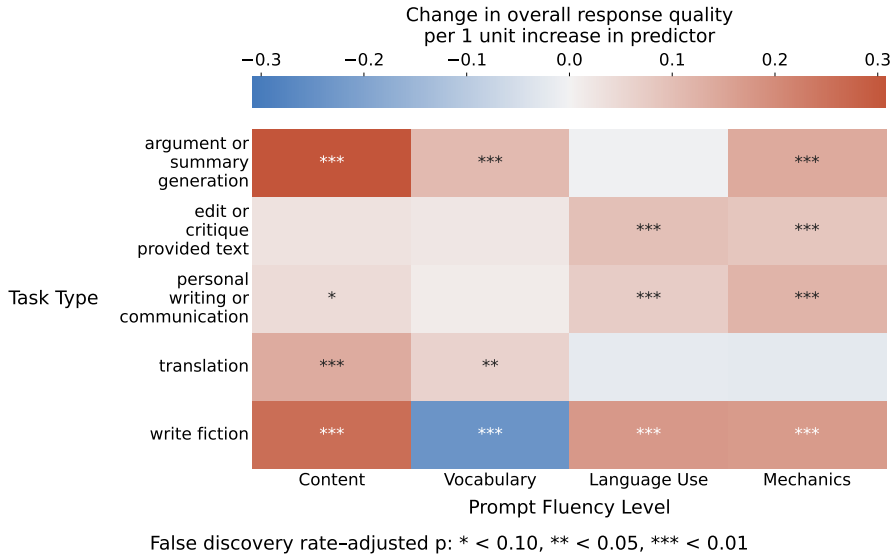

Figure 6: Prompt fluency shows broadly consistent positive associations with response quality across task types, with variation in individual dimensions. Colors indicate coefficients from separate regressions within each task.

more pronounced in certain tasks. We apply a similar regression analysis to measure the relationship between response quality and different prompt fluency dimensions by task type, shown in Figure 6.

More fluent prompts generally receive higher-quality responses across writing tasks, although the strength of this relationship varies by task types. Specifically, "translation" exhibits relatively weak associations with Language Use and Mechanics, while "editing or critiquing provided text" and "personal writing or communication" show less sensitivity to Content and Vocabulary. "Fiction writing" presents a notable exception to the generally positive pattern, with Vocabulary negatively associated with response quality. Details of the experiment are provided in Appendix A.7.5.

Overall, our results suggest that prompt fluency has a substantial influence on model response quality. More fluent prompts generally lead to responses with higher quality across different task types, although the magnitude of this effect varies slightly depending on the task. Among different fluency dimensions, Content and Mechanics generally exhibit stronger influences on response quality. These findings suggest that the effectiveness of prompts depends not only on their semantic intent, but also on how fluently and coherently users express their intentions. More broadly, our results provide practical implications for real-world interactions with LLMs, indicating that improving prompt fluency may serve as a simple yet effective strategy for obtaining high quality model responses.

6 Conclusion

We investigate how the fluency of non-native English prompts shapes the responses that large language models produce. To enable controlled experimentation, we curate FABLE, a collection of 190,911 writing task–related prompts that vary systematically in fluency while preserving semantic intent. In experiments with responses from 34 popular open-source models, we find that LLMs mirror the rhetorical and lexical qualities of the prompt, and their overall response quality is consistently lower for less fluent prompts. More specifically, a one-point increase in a retained fluency dimension on the four-point scale corresponds to an estimated 0.04–0.14 point increase in response quality on the five-point scale. These effects are broadly consistent across writing tasks. Taken together, our findings indicate that non-native English-speaking users of LLMs are likely disadvantaged through their degree of fluency in writing prompts. This disparity in quality is driven less by surface mistakes than by the higher-level discourse and lexical qualities of their prompts, and we hope FABLE enables future work on diagnosing and mitigating this disparity.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets, assisting with translation, cleaning and reformatting data, designing and providing feedback on research methodology and experiments, implementing methods, supporting qualitative and thematic data analysis, creation of artifacts, discovering research topics and identifying gaps, brainstorming, sourcing/searching for information, identifying relevant literature, summarizing and analyzing existing literature, and proposing a title and keywords for the research paper. We have not used generative AI tools for help developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, interpreting results, formulating questions for surveys or interviews, creating or modifying scientific figures or images, suggesting experimental parameters, formatting references, suggesting a structure for the research paper, or transcribing recordings of research material. Additionally, we used generative AI tools for creating and editing software code, drafting parts of this research paper, and editing to improve readability. We have reviewed all AI-assisted work, including manually reviewing model outputs and proposed edits to writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility of our results, we will release downloadable source code and the FABLE data upon publication. We also provide a complete description of data processing steps in the main text of the paper and Appendix, including prompts used for data synthesis and response rating.

REFERENCES

Bowen Cao, Deng Cai, Zhisong Zhang, Yuexian Zou, and Wai Lam. On the worst prompt performance of large language models. *Advances in Neural Information Processing Systems*, 37: 69022–69042, 2024.

Francine Chambers. What do we mean by fluency? *System*, 25(4):535–544, 1997. ISSN 0346-251X. doi: https://doi.org/10.1016/S0346-251X(97)00046-8. URL https://www.sciencedirect.com/science/article/pii/S0346251X97000468.

Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. NBER Working Paper 34255, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w34255.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5), 2024.

Renukswamy Chikkamath, Deepak Rastogi, Mahesh Maan, and Markus Endres. Is your search query well-formed? a natural query understanding for patent prior art search. *World Patent Information*, 76:102254, 2024.

Meredith Dedema, Chiao Xuen Goh, and Pengyi Zhang. When generative ai mixes languages: Multilingual users' code-switching behavior in human-llm interaction. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI EA '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722813. doi: 10.1145/3772363.3798492. URL https://doi.org/10.1145/3772363.3798492.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression, 2026. URL https://arxiv.org/abs/2609.19969.

Smitha Dev and Sura Qiqieh. The relationship between english language proficiency, academic achievement and self-esteem of non-native-english-speaking students. *International Education Studies*, 9(5):147–155, 2016.

Om Dobariya and Akhil Kumar. Mind your tone: Investigating how prompt politeness affects llm accuracy (short paper). *arXiv preprint arXiv:2510.04950*, 2025.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. The sharelm collection and plugin: Contributing human-model chats for the benefit of the community. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 167–177, Vienna, Austria, 2025. Association for Computational Linguistics.

Noé Durandard, Saurabh Dhawan, and Thierry Poibeau. Language style matching in large language models. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 620–636, 2025.

Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. What makes a good order of examples in in-context learning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14892–14904, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.884. URL https://aclanthology.org/2024.findings-acl.884/.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*, 2024.

Shinichiro Ishikawa. The icnale edited essays; a dataset for analysis of l2 english learner essays based on a new integrative viewpoint. *English Corpus Studies*, 25:117–130, 2018.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, Dustin Wright, Abraham Israeli, Anders Giovanni Møller, Lechen Zhang, and David Jurgens. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*, 2024.

Holly L. Jacobs, Stephen A. Zinkgraf, Deanna R. Wormuth, V. Faye Hartfiel, and Jane B. Hughey. *Testing ESL Composition: A Practical Approach*. Newbury House, 1981.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics, April 2017.

Robert B Kaplan et al. Cultural thought patterns in inter-cultural education. *Language learning*, 16 (1), 1966.

Hiroe Kobayashi and Carol Rinnert. Effects of first language on second language writing: Translation versus direct composition. *Language learning*, 42(2):183–209, 1992.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

LanguageTool contributors. Languagetool: Style and grammar checker. https://github.com/languagetool-org/languagetool, 2026. Version 6.8, accessed 2026-09-26.

Alina Leidinger, Robert Van Rooij, and Ekaterina Shutova. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 9210–9232, 2023.

Chuang Liu, Linhao Yu, Jiaxuan Li, Renren Jin, Yufei Huang, Ling Shi, Junhui Zhang, Xinmeng Ji, Tingting Cui, Tao Liu, Jinwang Song, Hongying Zan, Sun Li, and Deyi Xiong. OpenEval: Benchmarking Chinese LLMs across capability, alignment and safety. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 190–210, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.19. URL https://aclanthology.org/2024.acl-demos.19/.

Jialin Liu, Yongyuan He, Zhihan Zheng, Yi Bu, and Chaoqun Ni. Ai-assisted writing is growing fastest among non-english-speaking and less established scientists. *arXiv preprint arXiv:2511.15872*, 2025a.

Qibang Liu, Wenzhe Wang, and Jeffrey Willard. Effects of prompt length on domain-specific tasks for large language models, 2025b. URL https://arxiv.org/abs/2502.14255.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Marcin Miłkowski. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience*, 40(7):543–566, 2010. doi: 10.1002/spe.971.

Leiyu Pan, Yongqi Leng, and Deyi Xiong. Can large language models learn translation robustness from noisy-source in-context demonstrations? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2798–2808, 2024.

Meenu Pandey and Prabhat Pandey. Better english for better employment opportunities. *International journal of multidisciplinary approach and studies*, 1(4):93–100, 2014.

Dojun Park and Sebastian Padó. Multi-dimensional machine translation evaluation: model evaluation and resource for korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 11723–11744, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Maja Popović, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. Effects of different types of noise in user-generated reviews on human and machine translations including chatgpt. In *Proceedings of the Ninth Workshop on Noisy and User-Generated Text (W-NUT 2024)*, pp. 17–30, 2024.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers, 2024. URL https://arxiv.org/abs/2404.04925.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1881, 2016.

Manon Reusens, Philipp Borchert, Jochen De Weerdt, and Bart Baesens. Native design bias: Studying the impact of english nativeness on language model performance. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1195–1215, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Lea Siekmann, Judy M Parr, and Vera Busse. Structure and coherence as challenges in composition: A study of assessing less proficient efl writers' text quality. *Assessing Writing*, 54:100672, 2022.

Stephen P Witte and Lester Faigley. Coherence, cohesion, and writing quality. *College Composition & Communication*, 32(2):189–204, 1981.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Jian Yang. Learners and users of english in china. *English today*, 22(2):3–10, 2006.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pp. 9–35, 2024.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild, 2024. URL https://arxiv.org/abs/2405.01470.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Shengyao Zhuang and Guido Zuccon. Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 1444–1454, 2022.

A Appendix

A.1 Translated Prompt Synthesizing Details

We use the below prompt to synthesize prompt translations according to specific fluency levels in the ESL Composition Profile.

```
CONTENT_SPECS = {
4: "Content is rich and fully developed with specific, relevant details
    that clearly address the topic.",
3: "Content shows adequate knowledge of the topic and generally relevant
    ideas, though some points lack depth.",
2: "Content shows limited knowledge of the topic, with few supporting
    details and underdeveloped ideas.",
1: "Content shows very little grasp of the topic; ideas are vague,
    off-topic, or insufficient to evaluate."}

ORGANIZATION_SPECS = {
4: "Ideas are clearly organized with logical sequencing, clear
    paragraphing, and strong cohesion.",
3: "Organization is generally clear, though some parts may be loosely
    connected or uneven in sequencing.",
2: "Organization is weak; ideas are sometimes confused or disconnected
    and lack clear progression.",
1: "Organization is minimal; ideas are hard to follow and overall
    structure is unclear or missing."}
```

```
VOCAB_SPECS = {
4: "Uses a wide range of vocabulary with accurate and appropriate word
    choice and idiomatic expressions.",
3: "Uses an adequate range of vocabulary with occasional word-choice
    errors, but meaning is usually clear.",
2: "Uses a limited range of vocabulary with frequent word-choice
    problems; meaning is sometimes obscured.",
1: "Vocabulary is very limited, with frequent inappropriate or incorrect
    word choices; meaning is often unclear."}

LANGUSE_SPECS = {
4: "Uses effective simple and complex sentence constructions with few
    errors in tense, agreement, and word order.",
3: "Uses mostly correct simple sentences and some complex ones; grammar
    errors occur but rarely block understanding.",
2: "Shows major problems with grammar and sentence structure; errors are
    frequent and sometimes obscure meaning.",
1: "Has very little control of sentence construction; grammar errors
    dominate and communication often breaks down."}

MECH_SPECS = {
4: "Demonstrates mastery of spelling, punctuation, capitalization, and
    paragraphing with almost no errors.",
3: "Shows occasional errors in spelling, punctuation, capitalization, or
    paragraphing, but meaning is not obscured.",
2: "Shows frequent errors in spelling, punctuation, capitalization, or
    paragraphing; meaning may be affected.",
1: "Shows pervasive errors in spelling, punctuation, capitalization, and
    paragraphing; text is hard to read."}


user_message = f'''
    When producing text, you must simulate the following ESL composition
        profile:
    Content (level): {content_desc}
    Organization (level): {org_desc}
    Vocabulary (level): {vocab_desc}
    Language use (level): {lang_desc}
    Mechanics (level): {mech_desc}
Translate the following question into English that matches this profile:
    {original_prompt}
'''
```

A.2 Translation Examples

Table 2 shows examples of the same prompt translated with different uniform rating levels for each quality in instructions (e.g., all rated 4). Prompts retain much of their semantic equivalence, while prompts from instructions with lower ratings have visibly more errors.

A.3 Evaluation of Translated Prompts

Surface Form Features Higher instructed proficiency levels generally yield fewer linguistic errors and improved readability as shown in Figure 7. In particular, higher Mechanics levels are associated with fewer spelling and grammar–syntax errors, while higher Language Use levels are associated with improved readability. These relationships are not uniform: higher Vocabulary levels are associated with more detected spelling errors.

Rhetorical and Lexical Qualities Higher instructed levels generally correspond to higher judged proficiency scores as shown in Figure 8. The instructed level of a dimension generally has the strongest association with its corresponding measured score, particularly for Vocabulary, Language Use, and Mechanics. Cross-dimensional associations are also present, indicating that the instructions produce targeted fluency variation without fully independent control of the five dimensions.

Level	Prompt
1	Your foreign friend want to introduce Chinese food to classmate, send email ask about information. Please write a reply email, content include: 1. intoduction to Chinese food; 2. recommend a food and explain reason. Word count about 120.
2	Your foreign friend wants to introduce Chinese food to the class and sent an email asking for some information. Please write a reply email, including: 1. an introduction to Chinese food; 2. recommend a dish and explain the reason. Word count around 120.
4	Your foreign friend is going to introduce Chinese cuisine to classmates and has sent an email inquiring about relevant information. Please write a reply email to him, including: 1. An introduction to Chinese cuisine; 2. Recommend a dish and explain the reasons. Word count: around 120 words.

Table 2: Example translations generated from the same Chinese prompt under different instructed levels. While the semantic intent remains largely preserved, the fluency shows a notable gap across different levels.

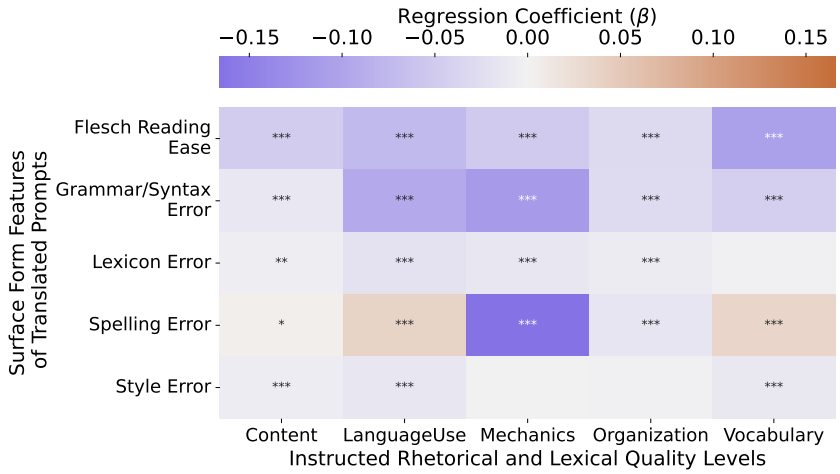

Figure 7: Standardized regression coefficients relating instructed proficiency levels to surface form features of translated prompts. Negative coefficients indicate fewer errors for error metrics; positive coefficients indicate greater reading ease for readability. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Semantic Preservation Cross-lingual cosine similarity, measured using `BAAI/bge-m3` (Chen et al., 2024), varies little with instructed fluency. The Spearman correlation between the average instructed level and similarity to the original Chinese prompt is 0.06.

Human evaluation of 100 randomly sampled translations yields a mean semantic-preservation score of 4.13 on a five-point scale, where 1 denotes severe information loss and 5 denotes no loss. This suggests that the translations generally preserve the original semantics with limited information loss on average. Additionally, Pearson correlations between individual instructed levels and the same-intent score are presented in Table 3. Together, these results support generally preserved intent and little association between instructed fluency and semantic preservation, while not ruling out semantic changes in individual translations.

Table 3: Pearson correlations between instructed proficiency levels and the same-intent score.

Instructed dimension	Correlation
Content	−0.112
Organization	0.073
Vocabulary	−0.063
Language Use	−0.010
Mechanics	−0.015

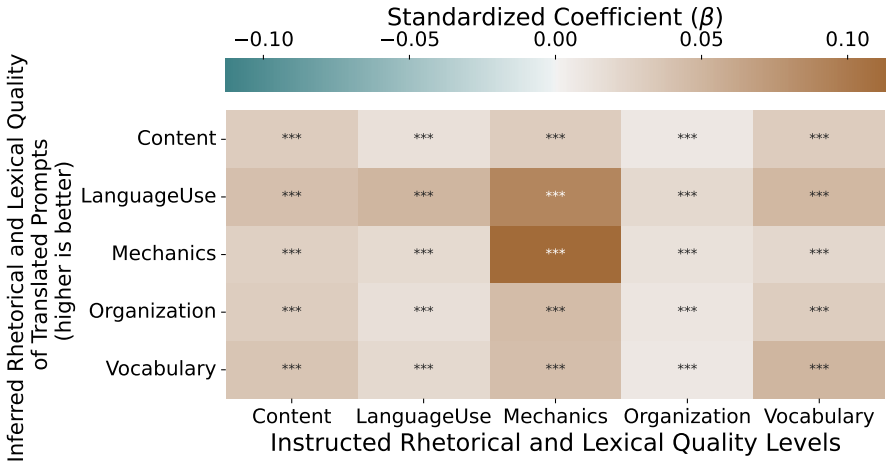

Figure 8: Standardized regression coefficients relating the five instructed proficiency levels to judged proficiency scores of translated prompts. Results show both targeted and cross-dimensional associations. Stars denote statistical significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.4 Evaluated Models

Table 4 lists the 34 models evaluated in our experiments. All models receive the same set of 5,000 English prompt variants randomly sampled from FABLE, and their responses are assessed using the same evaluation pipeline.

A.5 Fluency Evaluation

We use the below prompt to evaluate the fluency of LLM responses using the ESL Composition Profile.

```
user_message = f'''
Evaluate writing strictly according to the provided rubric:{rubric}.
Evaluate the following text on 5 dimensions: Content, Organization,
    Vocabulary, LanguageUse, Mechanics.
Each score must be an integer from 1 to 4.
'''
```

A.6 Response Quality Evaluation

We use the below prompt to evaluate the quality of LLM responses to instructions of varying fluency.

```
You are an impartial evaluator of AI assistant responses.
Evaluate the assistant response relative to the user's question.
Score each evaluation dimension independently.
Be objective, consistent, and appropriately critical.
Do not let the score for one dimension automatically determine another
    dimension.
Output only valid JSON.
For each dimension, provide a brief justification before assigning the
    score.
Each justification must be concise and no more than 30 words.
```

User message

```
Evaluate the assistant response on the following five dimensions:

1. Helpfulness: How useful the response is in helping the user achieve
    their goal.
```

Table 4: The 34 evaluated models, grouped by model family. Version and release identifiers are retained where applicable. IT denotes instruction-tuned models.

Model family	Model name
Nemotron	NVIDIA Nemotron Nano 9B v2
	Llama 3.3 Nemotron Super 49B v1.5
OLMo	OLMo 2 13B Instruct (1124)
	Olmo 3 7B Instruct
	Olmo 3.1 32B Instruct
Phi	Phi-4
Qwen	Qwen3-30B-A3B-Instruct-2507
	Qwen3-32B
	Qwen3-4B
	Qwen3-8B
	Qwen3.5-0.8B
	Qwen3.5-35B-A3B
	Qwen3.5-4B
	Qwen3.5-9B
	Qwen3.6-27B
	Qwen3.6-35B-A3B
	Qwen3.8-27B
Seed-OSS	Seed-OSS-36B-Instruct
SmolLM	SmolLM3-3B
Apertus	Apertus 70B Instruct (2509)
	Apertus 8B Instruct (2509)
Aya	Aya Expanse 32B
EXAONE	EXAONE 4.0 32B
Gemma	Gemma 4 12B IT
	Gemma 4 26B-A4B IT
	Gemma 4 31B IT
	Gemma 4 E2B IT
	Gemma 4 E4B IT
Granite	Granite 4.2 30B
	Granite 4.2 3B
	Granite 4.2 8B
Llama	Llama 3.3 70B Instruct
Ministral	Ministral 8B Instruct (2410)
Mistral	Mistral Small 3.2 24B Instruct (2506)

```
2. Relevance: How directly the response addresses the user's question
    without unnecessary or off-topic content.
3. Accuracy: How factually correct and logically sound the response is.
4. Creativity: How effectively the response demonstrates originality,
    insight, or creative problem-solving when appropriate.
5. LevelOfDetail: Whether the response provides an appropriate amount of
    specific and useful detail.

Also provide an Overall score reflecting the overall quality of the
    response.
The Overall score should be a holistic judgment and should not simply be
    the arithmetic average of the five dimension scores.

For every dimension and the Overall score, use the same rating scale:
1 = Very Poor
```

```
2 = Poor
3 = Average
4 = Good
5 = Very Good

Evaluate each dimension independently based only on that dimension.

Output exactly in the following JSON format:
{
  "Helpfulness": {
    "reason": "brief justification",
    "score": 1
  },
  "Relevance": {
    "reason": "brief justification",
    "score": 1
  },
  "Accuracy": {
    "reason": "brief justification",
    "score": 1
  },
  "Creativity": {
    "reason": "brief justification",
    "score": 1
  },
  "LevelOfDetail": {
    "reason": "brief justification",
    "score": 1
  },
  "Overall": {
    "reason": "brief justification",
    "score": 1
  }
}

[Question]
{reference_translation}

[The Start of Assistant's Answer]
{model_response}
[The End of Assistant's Answer]
```

A.7 Additional Experimental Details

A.7.1 Response Surface Form Features

For each response surface form measure, we estimate

$$Y = \alpha + \beta_C C + \beta_V V + \beta_L L + \beta_M M + \boldsymbol{\gamma}^\top \mathbf{X} + \epsilon,$$

where C, V, L, and M denote prompt Content, Vocabulary, Language Use, and Mechanics scores. $\mathbf{X}$ includes log-transformed model size, prompt length, response length, and model-family fixed effects. The outcome and all continuous predictors are standardized. We exclude Organization because its VIF exceeded 5, to reduce collinearity. The responses used in this experiment were generated by 34 models mentioned in Appendix A.4.

A.7.2 Rhetorical and Lexical Quality of Response

We fit a separate regression for each metric of response rhetorical and lexical quality:

$$Y_k = \alpha_k + \beta_{k,C} C + \beta_{k,O} O + \beta_{k,V} V + \beta_{k,L} L + \beta_{k,M} M + \epsilon_k,$$

where Y_k is the response score for metric k, and C, O, V, L, and M are the prompt's Content, Organization, Vocabulary, Language Use, and Mechanics scores, respectively. The responses used in this experiment were generated by Qwen3-14B, Llama-3.1-8B, and GPT-OSS-20B.

A.7.3 Response Quality

We fit an OLS regression:

$$Q = \alpha + \beta_C C + \beta_V V + \beta_L L + \beta_M M + \beta_S \log_{10}(S) + \boldsymbol{\gamma}^\top \mathbf{X} + \epsilon,$$

where Q is overall response quality on a five-point scale; C, V, L, and M are the prompt's Content, Vocabulary, Language Use, and Mechanics scores on a four-point scale; and S is model size in billions of parameters. $\mathbf{X}$ includes prompt length, response length, and model-family fixed effects. We exclude Organization to reduce collinearity because we identify its VIF > 5. The responses used in this experiment were generated by 34 models mentioned in Appendix A.4.

A.7.4 Semantically Matched Pair Analysis

For semantically matched prompt pairs, we fit:

$$\Delta Q = \alpha + \beta_C \Delta C + \beta_O \Delta O + \beta_V \Delta V + \beta_L \Delta L + \beta_M \Delta M + \epsilon,$$

where Δ denotes the difference within each pair, Q is overall response quality, and C, O, V, L, and M are the five prompt fluency scores. We estimate the regression using OLS on unstandardized differences and report 95% confidence intervals. The responses used in this experiment were generated by Qwen3-14B, Llama-3.1-8B, and GPT-OSS-20B.

A.7.5 Analysis by Task Type

For each task category t, we fit a separate OLS regression:

$$Q = \alpha_t + \beta_{t,C} C + \beta_{t,V} V + \beta_{t,L} L + \beta_{t,M} M + \boldsymbol{\gamma}_t^\top \mathbf{X} + \epsilon,$$

where Q is overall response quality; C, V, L, and M are the four retained prompt fluency scores; and $\mathbf{X}$ contains prompt and response lengths. All variables retain their original scales and we exclude Organization to reduce collinearity because we identify its VIF > 5. The responses used in this experiment were generated by 34 models mentioned in Appendix A.4.

A.8 Computational Environment

We run all our experiments on 4 NVIDIA-L40S-48GB GPUs. All LLM inferences are powered by vLLM 0.5.4 Kwon et al. (2023), Huggingface Transformers 4.43.3 Wolf et al. (2020) and PyTorch 2.4.0 Paszke et al. (2019) on a CUDA 12.4 environment. Temperatures are set to 0.0 to minimize the effect of randomness.

A.9 Licenses

All data and code will be publicly released under the CC BY-SA 4.0 license.