How Important is 'Perfect' English for Machine Translation Prompts?

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved top results in recent machine translation evaluations, but they are also known to be sensitive to errors and perturbations in their prompts. We systematically evaluate how both humanly plausible and synthetic errors in user prompts affect LLMs' performance on two related tasks: Machine translation and machine translation evaluation. We provide both a quantitative analysis and qualitative insights into how the models respond to increasing noise in the user prompt. The prompt quality strongly affects the translation performance: With many errors, even a good prompt can underperform a minimal or poor prompt without errors. However, different noise types impact translation quality differently, with character-level and combined noisers degrading performance more than phrasal perturbations. Qualitative analysis reveals that lower prompt quality largely leads to poorer instruction following, rather than directly affecting translation quality itself. Further, LLMs can still translate in scenarios with overwhelming random noise that would make the prompt illegible to humans.

1 Introduction

LLMs, particularly closed-source systems accessible via APIs, have recently dominated machine translation benchmarks (Kocmi et al., 2024a), and users increasingly turn to them over other systems. However, LLMs are notoriously sensitive to prompt perturbations (Qiang et al., 2024, inter alia).

Research publications tend to contain well-crafted prompts. The users, i.e., the actual target audience of the models, are less meticulous than researchers motivated to achieve state-of-the-art results with their models. Thus, the evaluation mode is misaligned with the way the models are used and evaluations in research might misrepresent the true model performance in the wild.

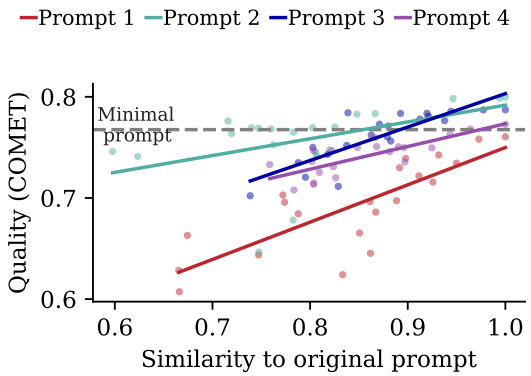


Figure 1: Changing model performance, as measured by COMET score (y-axis), across all noised prompts. The similarity of each noised prompt to the original is measured by the inner product of their sentence embeddings (x-axis).

Our work aims to fill this practice-evaluation gap. We evaluate LLM robustness to mistakes and perturbations in prompts on two machine translation-related tasks: machine translation itself, and LLM-based machine translation evaluation (Kocmi and Federmann, 2023b).

We simulate real-world LLM usage by adding noise to the prompts. Some of the noisers imitate L2 practitioners using prompts in "imperfect" English, which allows us to gauge the potential performance loss incurred by such users. The controllable noisers help us to quantify the impact of information loss through perturbations on the model performance, and observe error modes at more extreme noise levels. We focus specifically on noise in the *user prompt*, as opposed to other parts of the input, such as the system prompt. This is to mimic the typical user/researcher use-case, who typically do not have access to the system prompt of state-of-the-art LLMs.

[0]We release the 2.2M translations in 3 language pairs from 6 state-of-the-art (closed & open) models and 7 noisers at `[anonymized]` to enable further research into the effects of prompt quality on LLM performance.

Contributions. We provide a systematic evaluation of the effect of errors in the user prompt on LLM performance in machine translation and MT evaluation. By considering different error types, we establish the severity of natural and synthetic errors, both at the orthographic and lexical-phrasal level. Our analysis is both quantitative and qualitative, showing how models try to recover, as well as when off-target language responses become more common.

Findings. Through a large-scale quantitative evaluation (followed by smaller-scale qualitative evaluation) on three language pairs, six state-of-the-art models, and seven noisers, we find that:

- All models and prompts show similar sensitivity to various levels of prompt noise (see Figure 1).
- Various noise types impact translation quality differently, with synthetic and phonetic noises degrading performance more than orthographic or phrasal perturbations.
- High prompt noise also increases the rate of off-target language outputs, with models often responding in the wrong language, especially when translating between unsupported language pairs.
- Lower prompt quality does largely lead to lower instruction following, rather than lowering the translation quality itself.
- LLMs are capable of providing translations even when the prompt is illegible to humans.
- The suitability of a prompt dramatically varies by model, implying that comparing the performance of two models using a single prompt is unfair.

Similar findings also hold true for the sibling task of translation quality estimation: Lower-quality prompts show a weakly detrimental effect on the automatic quality assessment, as meta-evaluated by system-level correlation with human judgments.

2 Related Work

LLMs for Machine Translation. General-purpose decoder-only LLMs have demonstrated state-of-the-art performance in machine translation with zero- and few-shot prompts (Kocmi et al., 2024a). However, LLMs may refuse to answer or generate unwanted text surrounding the translation which adversely affects automatic evaluation (Briakou et al., 2024). Further, performance has been shown to vary depending on the chosen prompt (Bawden and Yvon, 2023).

LLMs show strong translation performance with zero-shot prompting (Hendy et al., 2023). This is especially true for explicitly multilingual models such as EuroLLM (Martins et al., 2024). Both fine-tuning (Xu et al., 2024) and instruction-tuning on the translation task can further boost performance (Alves et al., 2023). For example, TowerLLM (Rei et al., 2024), which is instruction-tuned for multilingual translation and related tasks, achieved leading results on the WMT24 general translation task (Kocmi et al., 2024a).

Robustness of LLMs. Robustness of language models has been explored in the context of adaptation to low-resource settings and user-generated text (e.g., Bafna et al., 2024; Srivastava and Chiang, 2025). Both these papers model multiple types of variation through automated means, similar to our approach. Srivastava and Chiang (2025) focus on modelling variation in English and provide a tool to reproduce their interventions. Bafna et al. (2024) rather introduce noise for multiple languages and evaluate models for robustness against noise in the input segments of classification tasks.

Belinkov and Bisk (2018) diagnosed NMT models to be sensitive to both synthetic and natural noise in the input text. More recently, Peters and Martins (2024) found GPT-3.5 to be surprisingly resilient against synthetic noise. Unlike us, they looked at the input segment and not the user prompt. They also only applied synthetic typos in 10-100% of input tokens, while we define a much broader set of noise types.

Relatively little work has addressed noise in the prompt specifically. Zhu et al. (2024) generate 'adversarial' prompts containing possible typos and semantic errors, as generated by several different tools. However, they seem to be working with rather mild noising and do not address different noise levels. Additionally, while they cover a number of tasks, these are mostly classification tasks. Their evaluation for translation is superficial compared to ours.

LLM-as-a-judge for Translation. LLMs have been shown to be effective evaluators of models' instruction-following abilities (Zheng et al., 2023), and have since been successfully applied to translation evaluation. Kocmi and Federmann (2023a,b) introduce GEMBA, a prompt-based metric using GPT-4 to produce direct assessments (DA), multidimensional quality metric (MQM) analysis, or error span annotations (ESA; Kocmi et al., 2024b).

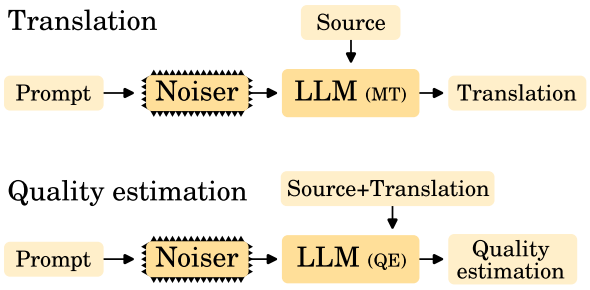


Figure 2: **Top:** Machine translation pipeline. The original prompt is noised, then augmented with source language, target language, and source sentence, before being translated by an LLM. **Bottom:** Quality estimation pipeline. Translations are evaluated using the GPT Estimation Metric Based Assessment (GEMBA) tool.

In this work, we use the zero-shot GEMBA-DA prompt which achieves high system-level correlations with human judgments for both reference-based evaluation and reference-free quality estimation, competitive with fine-tuned metrics (Freitag et al., 2023, 2024) such as CometKiwi (Rei et al., 2022b) and XCOMET-QE (Guerreiro et al., 2024). Improvements to LLM-based quality estimation are observed with chain-of-thought prompting for error analysis (Lu et al., 2024) and fine-tuning on human judgments, which boosts poor segment-level correlations of LLMs-as-judges (Fernandes et al., 2023). While Huang et al. (2024) investigate the effect of including source and references on evaluation performance, to our knowledge, our work is the first to investigate the prompt robustness of LLMs-as-a-judge for translation.

3 Prompt Noising

We define a number of noise types and noising scenarios, covering both plausible human errors and random synthetic noise. A full description of the individual noising functions is given in Appendix A. See examples of the noised prompts in Table 1.

3.1 Modeling Scenarios of Interest

Given the noising functions, we model the following scenarios:

1. **Random noise.** Our random noiser (A.1) creates character-level typos, parameterized by a probability p. We use it to stress-test model tolerance to noise with 10 choices of p uniformly spaced between $[0, 1]$.
2. **Natural orthographic noise** (spelling errors) due to imperfect proficiency is modelled by our orthographic noiser (A.2). This also introduces character-level noise with a parameter p. The specific perturbations are motivated by documented errors from L1 and L2 speakers (Cook, 1997). We manually choose a range for $p \in [0, 0.4]$ as representing a natural spectrum for the intensity of this type of error, and generate noised prompts for 10 uniformly spaced values of p within this range. We may imagine that latter parts of the range represent less proficient L2 writers, and L1 and L2 children.
3. **Phonetic LLM-generated noise** (A.3) is motivated by different phonetic transcription systems that speakers of different first languages may have. This noiser also produces primarily character-level edits but is not parameterized.
4. **Phrasal noise** (A.4) mimics phrasal substitutions and simplifications as made by beginner and intermediate speakers of English.
5. **Register noise** (A.5) also operates on a phrase level, but rather transforms the prompts to an informal register. We generate noised prompts corresponding to two different intensities.
6. **Low-proficiency writers, such as L2 learners** of English, presumably commit lexical/phrasal errors as well as spelling errors. We model this as a combined scenario by applying orthographic noise (with the same settings as above) over both levels of simplification as applied by the phrasal noiser. This results in $10 \cdot 2 = 20$ noiser parameterizations per prompt. Different compositions of the two noisers can be imagined to represent the diversity of proficiency in English, i.e. users with syntactic proficiency but imperfect spelling or vice versa.
7. **Lazy users** use informal registers, and presumably also make spelling errors. As above, we compose the orthographic noiser in the selected range over both levels of the register noiser to generate prompts with varied noise levels.

For all scenarios involving the random and orthographic noisers, we generate 20 noised prompts per noiser parametrization.

3.2 Noise Level and Sampling

Recall that we want to obtain noised prompts over a range of error intensities in order to measure the effect of errors on LLM performance. Since not all of our noisers are straightforwardly parameterized, we define two metrics of prompt similarity to measure noise level:

- **chrF** between the base prompt and the noised version gives a surface measure of prompt similarity, where a lower chrF score indicates a higher noise level.
- **Inner product of embeddings** of the base prompt and the noise prompts gives a more semantic measure of prompt similarity. The embeddings are derived from `all-MiniLM-L6-v2` from SentenceTransformers (Reimers and Gurevych, 2019a).

By using various noiser parameterizations per scenario as described above, we are able to obtain noised prompts modeling our scenarios of interest over a range of noise levels as measured by chrF score. We sort all noised prompts for a particular scenario and prompt into $k = 10$ buckets of increasing error intensity, and study model performance over these buckets. Given a bucket, the prompt used for a particular input is sampled randomly over noised prompts in that bucket. This helps in providing stable estimates of model performance by reducing vulnerability to outlier prompts.

4 Experimental Settings

Prompts. We choose four zero-shot prompts used by LLM-based systems at the WMT24 General Translation task (Kocmi et al., 2024a) as our base prompts. As a sanity check, we include results for an additional minimalistic baseline that was shown to perform well by Zhang et al. (2023). We do not apply any perturbations to this baseline. See Table 4 for the full baseline prompts, Table 1 for examples of perturbed prompts, and Appendix B for implementation details.

Setup. We select two closed-source API models and four open-weight models:

- GPT-4o-mini (OpenAI, 2024)
- Gemini-2.0-flash (Google, 2024)
- Llama-3.1-8B-Instruct (Dubey et al., 2024)
- Qwen2.5-7B-Instruct (Yang et al., 2025)
- EuroLLM-9B-Instruct (Martins et al., 2024)
- TowerInstruct-7B-v0.2 (Rei et al., 2024)

The models are selected so that they support the languages used for the experiments, and, for the open-weight models, that we are able to run them on our infrastructure. See Appendix Table 7 for the list of models we considered and the languages they support.

Prompt 3: Translate this from {src_lang} to {tgt_lang}:\n {src_lang}: {src_text}\n {tgt_lang}:
Orthographic (0.1) Trranslate ti from {src_lang} too {tgt_lang:}: \n {src_lang}: {src_text} \n {tgt_lang}:
Orthographic (0.4): Tranzlate dhiss from {src_lang} to {tgt_lang}: \n {src_lang}: {src_text} \n {tgt_lang}:
Orthographic (0.5): Granslaas yhii fftom {src_lang} tto {tgt_lang}: \n {src_lang}: {src_text} \n {tgt_lang}:
Orthographic (1.0): Reaajaky fgo trormm {src_lang} ttk {tgt_lang}:: \n {src_lang} {src_text} \n {tgt_lang}::
Lexical/Phrasal (1): Make this text in {tgt_lang} from {src_lang}: \n {src_lang}: {src_text} \n {tgt_lang}:
Lexical/Phrasal (2): You translate this text to {tgt_lang} fromm {src_lang}: \n {src_lang}: {src_text} \n {tgt_lang}:
Phonetic: Tranzlate thees from {src_lang} to {tgt_lang}: \n {src_lang}: {src_text} \n {tgt_lang}:
Register (1): {tgt_lang} version of this pls: \n {src_lang}: {src_text} \n {tgt_lang}:
Register (2): change lang {src_lang} -> {tgt_lang}: \n {src_lang}: {src_text} \n {tgt_lang}:

Table 1: Various types and levels (denoted in parentheses) of noise applied to Prompt 3.

We use language pairs present in WMT 2024 (Kocmi et al., 2024a), specifically: Czech-Ukrainian, German-English, and English-Chinese. Qwen officially supports only English, while the other models either officially support or empirically show good performance on these languages (i.e., by taking part in WMT24). For each language pair, we randomly choose 500 segments, which is close to the total number in the test set. For evaluation, we use ChrF (Popović, 2015) and $\text{COMET}_{22}^{\text{DA}}$ (Rei et al., 2022a). We rely on both because COMET is known to struggle on out-of-distribution translations (Zouhar et al., 2024).

Quality Estimation with GEMBA. We use GPT-4o-mini for consistency with our translation experiments. We use two base prompts: GEMBA-DA (quality estimation Prompt 1) (Kocmi and Federmann, 2023b) and TMU-HIT's WMT24 quality estimation prompt (QE Prompt 2) (Sato et al., 2024); full prompts are shown in Appendix Table 5. We test on Czech-Ukrainian, German-English, and English-Chinese, as for translation. We meta-evaluate the quality estimation performance by computing system and segment level Pearson correlations with human scores on submitted WMT24 systems. We follow a strict setup with no retries; when GEMBA fails to output a correctly formatted score, we set the score for that segment to 0. We limit experiments on the quality estimation task to orthographic noise on the two base prompts, both to maintain a realistic scenario and to limit costs.

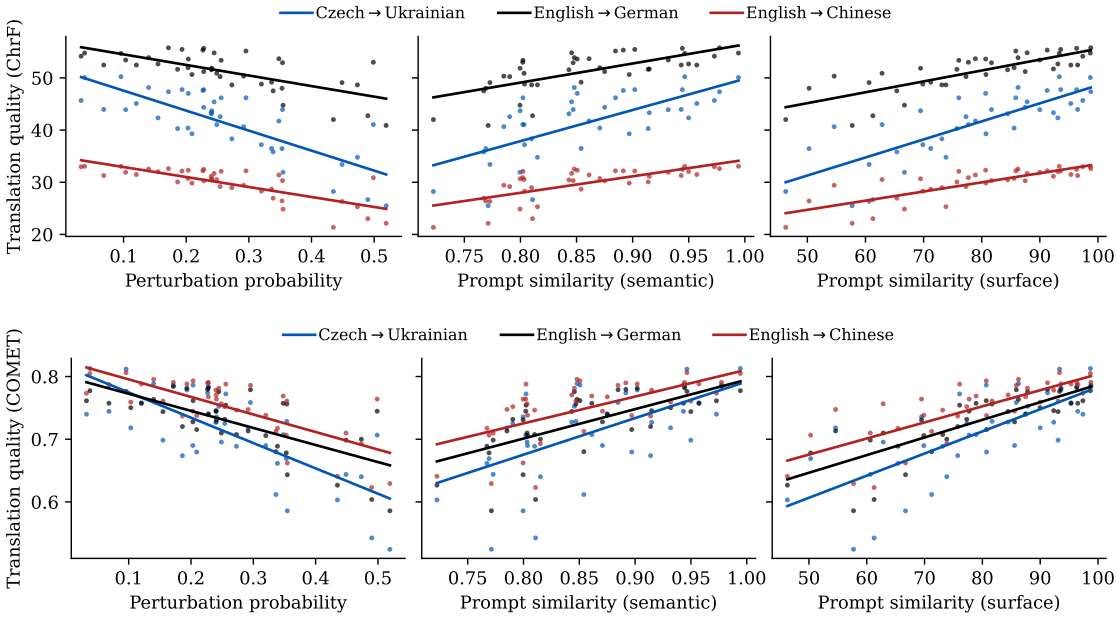


Figure 3: Sensitivity to perturbations by language pair. Translation quality measured by ChrF (top) or COMET (bottom), given a certain amount of perturbations (x-axes). Perturbation probability refers to the probability p of applying orthographic noise. Prompt similarity (semantic) refers to the inner product of sentence embeddings. Prompt similarity (surface) refers to the chrF score of the noised prompt against the base prompt.

5 Results and Discussion

5.1 Both Prompt Choice And Errors Matter

Figure 1 shows the changes in translation quality depending on the semantic similarity to the original prompt, per prompt. We see that all four base prompts are affected by applying noise. However, there are also marked differences in performance between the base prompts, showing that prompt choice matters, and state-of-the-art performance is more likely to be achieved with a better prompt.

The 'minimal' prompt yields a reasonable performance but stays behind the best base prompts. Its key benefit compared to the other prompts is the fact that the minimal prompt is essentially impossible to make mistakes with: Using an otherwise 'good' prompt with many mistakes leads to much worse performance, and may be worse than using a 'bad' prompt, or the minimal prompt. At the same time, a 'bad' prompt without errors can perform worse than a 'good' prompt with only few mistakes. These observations show clearly that both prompt choice and correctness matter for best results.

5.2 Effects Appear Across Metrics & Models

Figure 3 shows error sensitivity per language pair. We show six subplots, one for each combination of quality metric (ChrF, COMET) and measure of noise level (noising probability p, semantic prompt similarity, and surface prompt similarity). Thus, we validate that the perturbation probability p for the orthographic noiser is closely related to the two measures of prompt similarity. Both translation quality metrics show a similar degradation with the introduction of more errors.

Importantly, we also see that all language pairs are affected to a similar degree. Czech-Ukrainian appears slightly more sensitive than the other two, possibly due to less robust support of the models for this language pair, while translation into Chinese scores lower on ChrF. Note that TowerInstruct does not support Czech or Ukrainian, and Llama-3.1 officially does not support Czech, Ukrainian, or Chinese. Similarly, Appendix Figure 7 shows the sensitivity per language pair *and model*. All models are affected, for all languages, to a similar degree.

5.3 Comparing Noise Types

Table 2 shows Pearson correlations of translation quality with the prompt similarity to the base prompt, per noise type and per prompt. A high correlation implies that at low similarity, translation quality is also low, i.e, changes in this direction have more impact on output quality. A low correlation means that at low similarity, translation quality is not strongly affected.

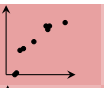
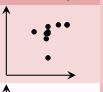
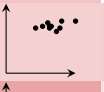
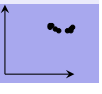
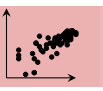
Noiser	Prompt 1	Prompt 2	Prompt 3	Prompt 4	All prompts
Random noise	0.86 	0.92 	0.68 	0.61 	0.77 
Phonetic LLM	0.84 	0.72 	0.86 	0.57 	0.75 
Lazy user	0.94 	0.77 	0.36 	0.52 	0.65 
Orthographic	0.71 	0.58 	-0.01 	0.67 	0.49 
L2	0.71 	0.46 	0.47 	0.14 	0.44 
Register	0.74 	0.88 	0.31 	-0.17 	0.44 
Phrasal	0.51 	0.59 	0.30 	-0.67 	0.18 
All noisers	0.76 	0.70 	0.42 	0.24 	0.53 

Table 2: Pearson correlation within each noiser and prompt (averaged across models and languages). See Appendix Figure 6 for visualization in a single plot.

On average across all prompts, the random noise shows the strongest effect. This may be partly because these errors are less realistic and therefore further out-of-distribution than the more plausible orthographic errors. Additionally, the random noise was applied in a wide range of intensity, and it shows a relatively steep drop-off in performance as the prompt similarity decreases.

The effect of the phonetic LLM-generated noise is also strong, with both low prompt similarity measures and low translation quality.

Phrasal perturbations show the least overall effect: For Prompt 4, we even see a negative correlation, meaning some substitutions performed better than the original prompt. Phrasal noise tends to simplify the prompt, and is rated less distant from the original prompt by the inner product metric. This increases robustness to phrasal noise overall, and can even help.

The 'L2' and 'Lazy User' scenarios combine the orthographic noiser with phrasal and register noise, respectively. While the orthographic noiser affects spelling, phrasal and register noise changes the wording of the prompt. Each combined scenario has more impact than the wording changes alone, though notably the 'L2' scenario is not worse overall than orthographic errors alone. This may be due to the simplifying effects of phrasal noise.

5.4 Comparing Prompts

Remarkably, there is no correlation between noise level and quality when Prompt 3 is noised with realistic orthographic errors. Additionally, as noted above, Prompt 4 even seems to benefit from phrasal noise. Prompts 3 and 4 appear more resilient to noising overall. We conjecture that this is primarily due to one factor: They are short, and we preserve the critical variables of source and target language, as well as the input segment. Therefore, noising the rest of the short prompts introduces relatively less confusion than when noising the more complex prompts 1 and 2. Additionally, prompts 1 and 4 do not outperform the minimal prompt baseline in terms of absolute quality (see Figure 1). While Prompt 1 is long and appears brittle, Prompt 4 benefits from phrasal substitution, perhaps because there is room for improvement.

Additionally, the best-performing prompt for one model can be the worst-performing prompt for another. We observed that GPT-4o tends to benefit from the structure of Prompt 1, however, the remaining models tend to copy parts of it regardless of noise type or intensity, leading to lower scores. Similarly, Prompt 4 is the best-performing prompt for EuroLLM and Qwen 2.5, but makes Gemini more likely to produce translations into languages other than the target. We thus conclude that using

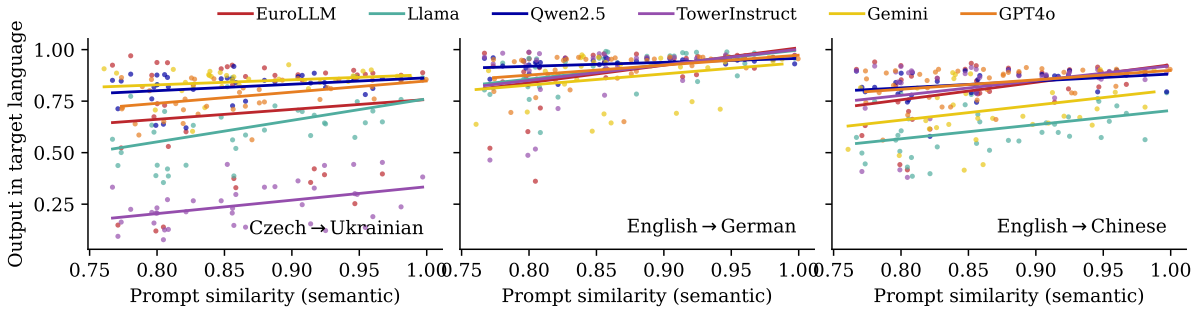


Figure 4: Percentage of outputs in the target language, by language pair and model. Note that TowerInstruct does not officially support Ukrainian or Czech.

the same prompt to compare the performance of multiple models is unfair, especially if the prompt tuning was done using one model.

5.5 Off-Target Outputs

We observe that a common error mode is responding in a non-target language. Figure 4 shows the proportion of outputs in the target language by model and language pair. For all models and language pairs, high noise levels decrease the proportion of on-target outputs. Of the target languages, German has the highest proportion of on-target outputs. For Czech-Ukrainian, TowerInstruct tends to output other languages even with an unnoised prompt, because the two languages are not well-supported by the model.

Note also that COMET may still score off-target outputs highly, for instance, if the model outputs an English or Russian (for Czech-Ukrainian) translation of the input, or copies the input to the output. These types of outputs have high semantic similarity to the correct translation, but COMET does not consider whether the translation is in the correct target language.

5.6 Transferability to Quality Estimation

Figure 5 shows that both quality estimation prompts are (weakly) affected by applying *realistic* orthographic noise, with differences in both the base prompts' performance and the effect of the noiser. This reinforces that prompt choice matters for quality estimation as for translation.

Table 3 shows Pearson correlations of system-level correlations with the prompt similarity, per prompt. A higher correlation implies that decreasing prompt similarity also decreases quality estimation correlation with human judgments. We observe a similar effect of orthographic noise on

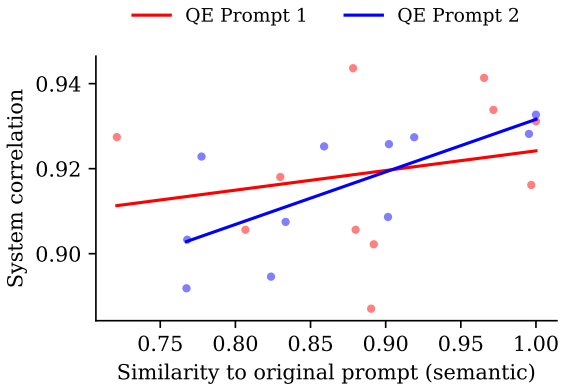


Figure 5: Changing model performance, as measured by system-level correlation (y-axis), across orthographically noised quality estimation prompts, against semantic similarity of the noised prompt to its original (x-axis). The results suggest only a weak trend.

the system-level correlation of GEMBA across languages, with an overall correlation of 0.43, compared to 0.49 for translation. This suggests the effect of orthographic noise is transferable and largely consistent across tasks. See Appendix Figure 9 for per-language results.

However, we observe a negative correlation for Quality Estimation Prompt 1 at the segment level. The unnoised prompt already achieves a poor segment-level correlation of 0.16. This may be an artefact of our strict setting which prohibits retries and artificially sets the resulting score to 0. Further, the shorter prompt may explain the reduced variance in outputs and therefore weaker correlations, though additional testing is required to elucidate this effect.

5.7 Qualitative Analysis

We performed qualitative analysis on the machine translation outputs by manually inspecting a sam-

Level	Prompt 1	Prompt 2	All prompts
System-	0.15 	0.71 	0.43 
Segment-	-0.38 	0.57 	0.09 

Table 3: Pearson correlation between system and segment-level correlations and prompt similarity, for the orthographic noiser and quality estimation prompts (averaged across languages).

ple of the lowest-scoring translations in each setting. We also sampled 10 source segments per language pair with their translations from every setting, to understand how various noise types and levels change the translation of a given sentence.

Gemini adds most supplementary information. In addition to providing the translation, Gemini frequently offers useful background information, such as multiple versions of the translation or the pronunciation of the Chinese translation explained in Latin script. These phenomena appear in the outputs regardless of the noise type or intensity. On the other hand, the model explains its choice of words more frequently with increasing noise intensity. This behavior may be beneficial for a user, but is difficult to parse in an automated setting.

Lower scores are often due to redundant text. The translations by a single model tend to remain relatively stable across various noise types and levels. The main explanation for the differing scores is the presence of redundant text, such as adding the name of the target language, saying "here is your translation", or repeating the source sentence.

GPT-4o is a notable exception: It generates more diverse translations and less redundant text, unless subjected to a high level of random noise ($p > 0.5$). Up to that threshold, a lower metric score for GPT-4o is more likely to correspond to genuinely lower translation quality.

This finding is also supported quantitatively in Figure 8. It shows that adding noise increases the average length of the LLM output and that there are frequently significant differences between target and reference length. GPT-4o produces the shortest texts while Gemini produces the longest. The average length of Qwen2.5 outputs is the least affected by noise.

LLMs can still translate with illegible prompts. Realistic noising scenarios produce prompts that are mostly legible to humans. However, applying random noise with $p > 0.46$, exceeding the natural error range, makes the prompts largely illegible to humans (compare Table 1).

LLMs sometimes produce an error message or request clarification without providing a translation in response to an illegible prompt. However, they frequently produce valid translations even when given prompts with a high $p = 0.78$, which are nonsensical to the human eye. These translations are frequently accompanied by strategies such as copying the prompt verbatim, attempting to translate or fix it, or treating the text as a cipher to decode. In rare cases, they ignore the noise altogether and provide the expected translation. We show examples of these outputs in Appendix Table 6.

6 Conclusions

In this work we describe the gap between academic evaluations and real world LLM usage for machine translation. We find that good prompts can be ruined by making many errors, leading to worse translations than an otherwise subpar prompt. The effect seems to be largest when spelling is severely affected, while phrasal substitutions may even help in some cases.

Fortunately for the users, 'imperfect' English in prompts most often does not lead to lower translation quality, but rather worse instruction following. The actual translations are thus still retrievable by humans and can be of appropriate quality, though this would fail in an automated pipeline. This finding reflects the gap between automated evaluation and individual usage of models.

Furthermore, we find that LLMs can produce translations even when prompted by instructions illegible to humans. This suggests that if an LLM is capable of performing a task, it can recognize the task and perform it even with an objectively bad prompt. We believe this finding may be helpful for future research, as it implies that if an LLM does not generalize to a task as demonstrated by a handful of prompts, further prompt engineering efforts are unlikely to change that outcome.

Limitations

This study was limited to only a small number of language pairs. However, we observe very consistent patterns across these language pairs, as well as across multiple models, and are thus reasonably

confident our findings will transfer to further language pairs.

We used automatically generated noise rather than error data from real learners. One important reason for this is the difficulty of sourcing real examples. Asking learners directly would produce observation effects which again distort the distribution. While using generated errors may mean that some of the examples are less realistic, it allows for a broader statistical analysis and provides us with better control over our experimental variables. Our noising implementation is carefully designed, based on attested observations of common errors, to be close to realistic.

Ethics Statement

We do not anticipate any negative ethical implications arising from this study. We took care to ensure realistic representations of errors without casting users in a negative light. The translation data we used is from WMT24 (Kocmi et al., 2024a), which is freely available to use for research purposes.

The total inference cost for the two proprietary models (GPT-4o-mini and Gemini-2.0-flash) is less than USD 100. While we did run the other models locally, the overall cost for all the models likely does not exceed USD 200.

The licenses for the open-weights models are: Llama 3.1 Community License for Llama 3.1; Apache 2.0 for EuroLLM and the Qwen model we used; and CC-BY-NC-4.0 for the Tower model we used (with its base model Llama 2 being licensed under the Llama 2 Community License). These licenses all permit our use of the model weights.

We used AI-assisted coding (i.e. Copilot) with the bulk being human-written. For writing, AI was used to check grammar mistakes.

References

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11127–11148, Singapore. Association for Computational Linguistics.

Niyati Bafna, Kenton Murray, and David Yarowsky. 2024. Evaluating large language models along dimensions of language variation: A systematik investigatiom uv cross-lingual generalization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 18742–18762. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 157–170. European Association for Machine Translation.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose LLM outputs: A case study in translation evaluation.

Vivian J Cook. 1997. L2 users and English spelling. *Journal of Multilingual and Multicultural Development*, 18(6):474–488.

Abhimanyu Dubey et al. 2024. The Llama 3 herd of models.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, 47–81. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, 578–628. Association for Computational Linguistics.

Google. 2024. Introducing gemini 2.0: Our new ai model for the agentic era.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 3546–3562, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, 1–46. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, 768–775. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, 1440–1453. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

Ben Peters and André F. T. Martins. 2024. Did translation models get more robust without anyone even noticing? *preprint*, arXiv:2403.03923 [cs.CL].

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. Association for Computational Linguistics.

Yao Qiang, Subhrangshu Nandi, Ninareh Mehrabi, Greg Ver Steeg, Anoop Kumar, Anna Rumshisky, and Aram Galstyan. 2024. Prompt perturbation consistency learning for robust language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, 1357–1370. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 578–585. Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, Jo textasciitilde ao Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, 185–204. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Association for Computational Linguistics.

Ayako Sato, Kyotaro Nakajima, Hwichan Kim, Zhousi Chen, and Mamoru Komachi. 2024. TMU-HIT's submission for the WMT24 quality estimation shared task: Is GPT-4 a good evaluator for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, 529–534. Association for Computational Linguistics.

Aarohi Srivastava and David Chiang. 2025. We're calling an intervention: Exploring fundamental hurdles in adapting language models to nonstandard text. In *Proceedings of the Tenth Workshop on Noisy and User-generated Text*, 45–56, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, 57–68, New York, NY, USA. Association for Computing Machinery.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, 1272–1288. Association for Computational Linguistics.

A Descriptions of Noising Functions

A.1 Random noise

This noiser, parameterized by probability p, introduces random perturbations into the prompt, modeling natural typos. The perturbations include random character transposition, omission, doubling, and substitution for neighbouring letters on the keyboard. The character-wise frequency of error is controlled by p, and the type of error is sampled uniformly from the above mentioned types.

A.2 Orthographic noise

This noiser models spelling errors, both due to imperfect proficiency in written English as well as random typos. Cook (1997) provide a classification of the types of spelling errors made by both L1 and L2 speakers, and report the relative frequency of these errors, finding higher error rates for L2 speakers, but similar distributions over error categories. Guided by this work, we define the follow classes of orthographic errors:

- **Natural typos**: We re-use our random noiser as described above. This corresponds to the category "other" as defined by Cook (1997).
- **Omission**: Omitting one of a non-word-initial consonant pair (e.g. `ck`→`k`), dropping `r` before a consonant, dropping `e` if it is word-final, or before `ly`.
- **Insertion**: Doubling non-word-initial consonant.
- **Substitution**: Confusing specific sets of consonants (such as `s`, `c`, `z`), confusing vowels with each other. For the latter, we generate errors consistently with the finding that confusions between `a`, `e`, `i` constitute 60% of vowel substitutions.
- **Transposition**: Transposing consecutive vowels (`ie`→`ei`), transposing certain bigrams (`er`, `ng`).

Similarly to the random noiser, the orthographic noiser is controlled by a parameter p, which corresponds to the probability of error on a given character. Varying p allows us therefore to generate prompts over different intensities of noise. Given a character to be noised, we sample a type of error from the above list, as per the natural distribution over these categories of error described in Cook (1997). Given a type of error (e.g. substitution), we uniformly sample a subtype of error from all subtypes applicable to the character and its context. For example, the "a-e"-confusion subtype is only relevant for "a's". Note that a character may have no relevant subtypes under a given type: In this case, we simply skip the character.

A.3 Phonetic LLM-generated noise

We also investigate the impact on LLM performance of errors made by non-native speakers writing English sentences based on phonetic transcriptions in their first languages. We prompt an LLM to mimic these errors in various languages (Arabic, Chinese, German, Polish and Spanish) spoken by beginner English learners. According to our tests, LLMs can simulate typical phonetic errors for a particular language, despite not being fully fluent in it. Example for a Polish person: *Translate the following line from English to Chinese.* → *Translejt de follouing lajn from English tu Chinese.*

A.4 Phrasal simplification

We would like to study the effect of alternate lexical/phrasal simplification, as possibly committed by L2 speakers. Note that prompts generally use largely restricted vocabulary, and potential phrasal errors are therefore limited. We consider two levels of L2 proficiency: Beginner and intermediate, and prompt an LLM to mimic such errors made by L2 speakers of each level, generating $k = 10$ noised candidates per prompt and level. We manually examine the generations and discard implausible options. We find that LLM-generated errors cover a reasonable range of plausible errors of this type.

A.5 Register changes

We are also interested in the effect of informal registers of users, who may query LLMs similarly to querying search engines, with non-standard casing, dropping of articles and function words, and reframing for conciseness. For example, *Translate from de to en*→*translate de - en*. This type of noise also offers a limited number of possible transformations of a base prompt. Similarly to above, we prompt an LLM to generate $k = 10$ informal versions of each base prompt with the above changes, for two levels (medium and high) of informality, and manually discard unlikely candidates.

B Implementation Details

For evaluation we use the following settings:

- nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1 (Post, 2018, sacrebleu)
- Python3.11.5|Comet2.2.5|fp32|Unbabel/wmt22-comet-da|r1 (Rei et al., 2022a, sacrecomet Zouhar et al., 2024)
- sentence-transformers/all-MiniLM-L6-v2 (Reimers and Gurevych, 2019b)

Prompt 1: ### Instruction:\n Translate Input from {src_lang} to {tgt_lang} \n ### Input:\n {src_text}\n ### Response:\n
Prompt 2: Translate the following line from\n {src_lang} to {tgt_lang}.\n Be very literal, and only translate the content of the line, do not add any explanations: {src_text}
Prompt 3: Translate this from {src_lang} to {tgt_lang}:\n {src_lang}: {src_text}\n {tgt_lang}:
Prompt 4: Translate the following text from {src_lang} to {tgt_lang}.\n {src_text}
Prompt minimal: {src_lang}: {src_text}\n {tgt_lang}:

Table 4: Base forms for investigated machine translation prompts.

QE Prompt 1: Score the following translation from {src_lang} to {tgt_lang} on a continuous scale from 0 to 100, where a score of zero means 'no meaning preserved' and score of one hundred means 'perfect meaning and grammar'.\n {src_lang} source: '{src_text}'\n {tgt_lang} translation: '{tgt_text}'\n Score:
QE Prompt 2: Please analyze the given source and translated sentences and output a translation quality score on a continuous scale ranging from 0 to 100. Translation quality should be evaluated based on both fluency and adequacy. A score close to 0 indicates a low quality translation, while a score close to 100 indicates a high quality translation. Do not provide any explanations or text apart from the score.\n {src_lang} Sentence: {src_text}\n {tgt_lang} Sentence: {tgt_text}\n Score:

Table 5: Base forms for investigating quality estimation prompts.

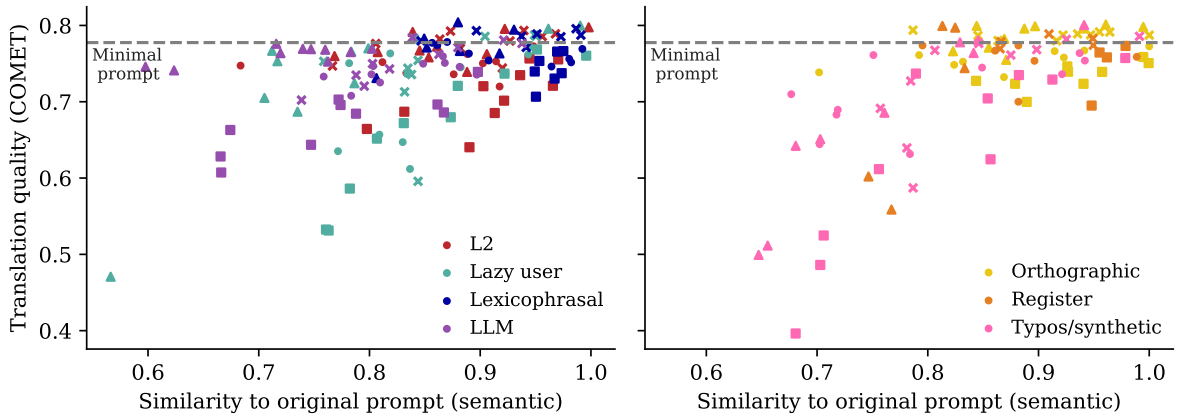


Figure 6: Average performance (across models and languages) with respect to individual prompts and noisers. Each shape is one of four prompts. Visualizes Table 2 in a single plot.

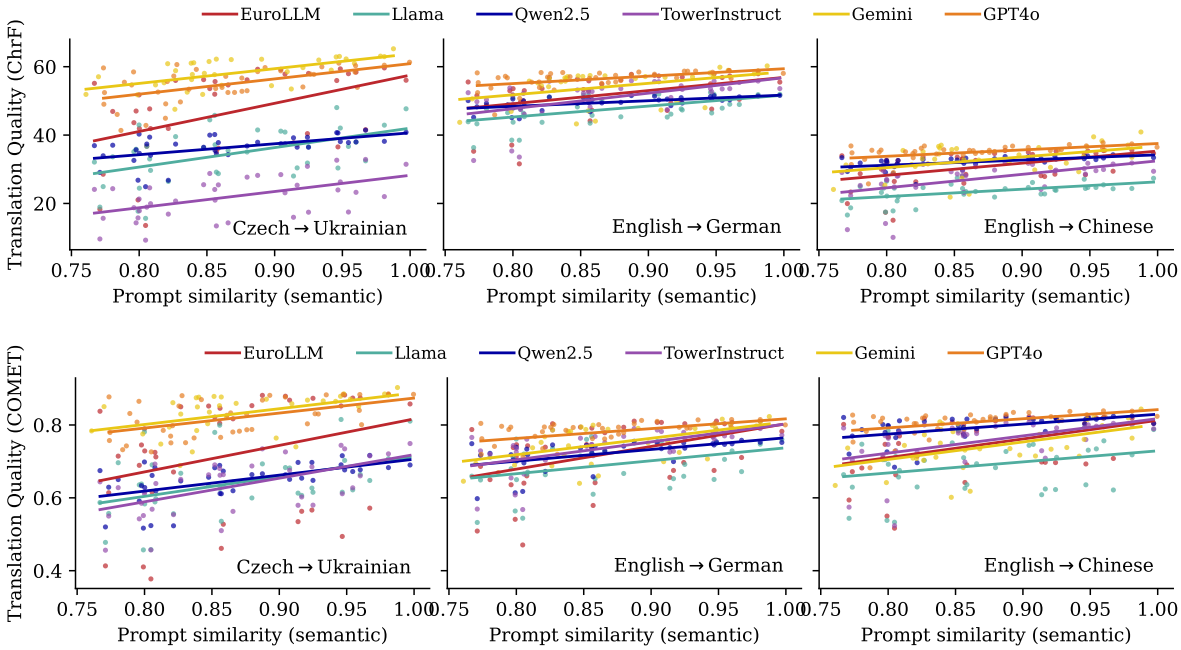


Figure 7: Sensitivity of individual models to prompt noising, for each language pair and by model. x-axis: Prompt similarity to base prompt (semantic). y-axes: Translation quality measured by ChrF (top) and COMET (bottom).

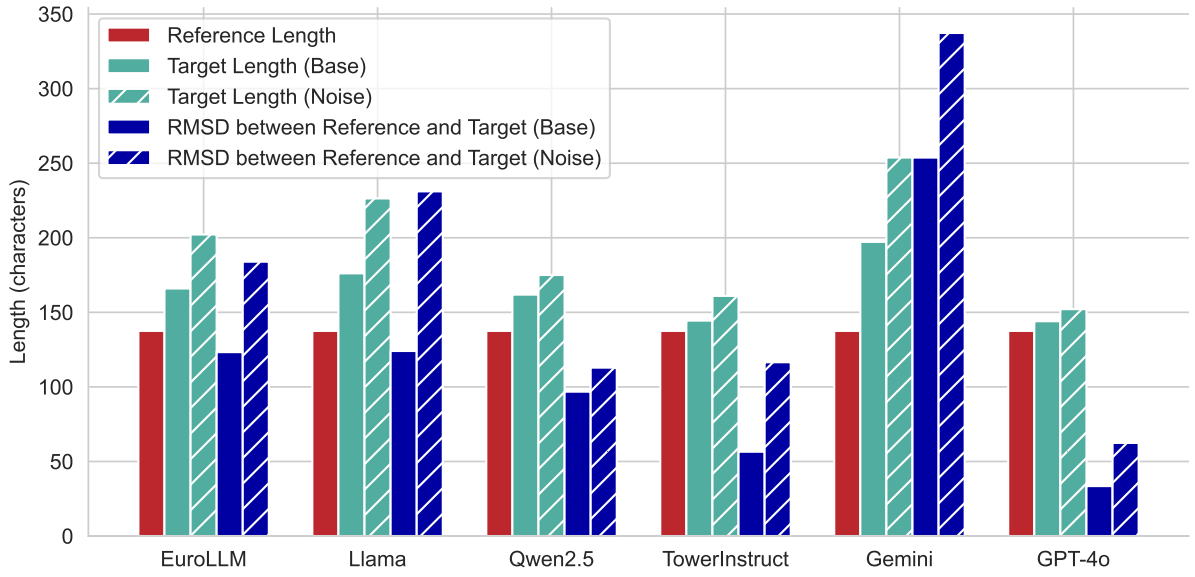


Figure 8: Average length of generated outputs compared to the reference length. We compare the output lengths given the base prompts to the aggregate of all noise variants, and observe that noised prompts lead to consistently longer texts.

EuroLLM, Orthographic (0.7): Rrwnslswfee tn gplowwgn linnee ffro \n English t German.\n Bee yg oiiteerarl, ann olyn rtnlqata thet cnotetn kff feg linns, ri ohtt adad anay exlqntaiiox:: *Ich finde es sehr lebensbejahend; aus so einfachen Zutaten etwas so Reichhaltiges und Komplexes zu machen. Wie bei der Geburt eines Kindes.*
Qwen2.5, Orthographic (0.8): Hier ist die Übersetzung von Englisch ins Deutsche: *Offensichtlich gibt es nur zwei Ebenen, die im Original-Super-Mario-Maker noch nicht abgeschlossen wurden.*
Gemini, Orthographic (0.8): Gerne, hier ist die Übersetzung des Satzes "I better get some motorbikes fixed, or else!" ins Deutsche:\n \n Ich sollte besser ein paar Motorräder reparieren lassen, oder aber!
Llama, Orthographic (0.8): The encoded message is a cipher. It appears to be a simple substitution cipher, where each letter has been shifted by a certain number of positions in the alphabet. After analyzing the text, I believe it is a Caesar cipher with a shift of 13 positions. Here's the decoded message: Original: Gafnsnlatae tet ffllwlif llnie gon Decoded: "Fantastical life line gone" [...]
GPT-4o, Orthographic (0.9): It seems like there are some typos and formatting issues in your message. However, I can help you translate the provided English text into German. Here's the translation: **English:** "I find it very life affirming; transforming such simple ingredients into something so rich and complex. Like witnessing a birth." **German:** *"Ich finde es sehr lebensbejahend; so einfache Zutaten in etwas so Reichhaltiges und Komplexes zu verwandeln. Wie das Zeugnis einer Geburt."* If you need further assistance or corrections, feel free to ask!
TowerLLM, Orthographic (0.8): Gafnsnlatae tet ffllwlif llnie gon \n Deutsch gehen Englisch.\n Sehr traurig, wenn man bedenkt, dass die Mieten in den letzten Jahren um durchschnittlich 6 Prozent pro Jahr gestiegen sind, aber diese Vorschläge würden die Mieten um bis zu 15 Prozent erhöhen, was ironischerweise höher ist als die historischen jährlichen Preissteigerungen. [...].

Table 6: Examples of LLM outputs when presented with perturbed prompts.

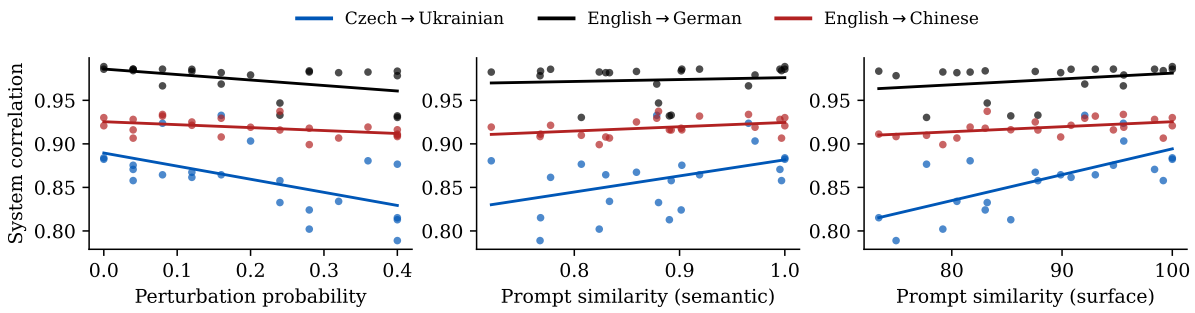


Figure 9: Sensitivity of QE outputs to perturbations by language pair. System-level correlation is measured against human evaluations on the test set, given a certain perturbation amount. Semantic prompt similarity measures the inner product of noised and base sentence embeddings, while surface similarity measures ChrF between noised and base prompts. The results show that effects are seen across language pairs, though the magnitude of the effect varies.

	English	Czech	Ukrainian	German	Chinese	Open
Claude-3.5 Sonnet[1]	✓	?	?	?	?	✗
CommandR+	✓	✓ [2]	✓ [2]	✓	✓	✗
GPT-4o	✓	✓	✓	✓	✓	✗
Gemini-1.5 Pro	✓	✓	✓	✓	✓	✗
Phi-3	✓	✓	✓	✓	✓	✓
Phi-4 14B	✓	✓	✓	✓	✓	✓
EuroLLM	✓	✓	✓	✓	✓	✓
Llama	✓	✗	✗	✓	✗	✓
Tower	✓	✗ [3]	✗ [3]	✓	✓	✓
Aya23	✓	✓	✓	✓	✓	✓
DeepSeek-V3[1]	✓	?	?	?	✓	✓
Qwen-2.5	✓	✗	✗	✗	✗	✓
Mistral	✓	✗	✗	✓	✓	✓

Table 7: List of models taken into consideration. The list of supported languages for the open-weight models is taken from their Hugging Face model cards.
[1]: The model is multilingual but the list of supported languages is not available;
[2]: Languages included in the pre-training but not post-training (Cohere documentation);
[3]: Tower70B took part to WMT2024 on the Czech→Ukrainian language pair (Kocmi et al., 2024a), but the model card for Unbabel/TowerInstruct-7B-v0.2 does not include it.