When the Gold Standard Isn't Necessarily Standard: Challenges of Evaluating the Translation of User-Generated Content

Lydia Nishimwe **Benoît Sagot** **Rachel Bawden**
Inria
Paris, France
{firstname.lastname}@inria.fr

Abstract

User-generated content (UGC) is characterised by frequent use of non-standard language, from spelling errors to expressive choices such as slang, character repetitions, and emojis. This makes evaluating UGC translation challenging: what counts as a "good" translation depends on the desired standardness level of the output. To explore this, we examine the human translation guidelines of four UGC datasets, and derive a taxonomy of twelve non-standard phenomena and five translation actions (NORMALISE, COPY, TRANSFER, OMIT, CENSOR). Our analysis reveals notable differences in how UGC is treated, resulting in a spectrum of standardness in reference translations. We show that translation scores of large language models are highly sensitive to prompts with explicit UGC translation instructions, and that they improve when they align with the dataset guidelines. We argue that fair evaluation requires both models and metrics to be aware of translation guidelines. Finally, we call for clear guidelines during dataset creation and for the development of controllable, guideline-aware evaluation frameworks for UGC translation.[1]

What constitutes a "good" translation, and how can it be reliably assessed? Machine translation (MT) evaluation seeks to answer these fundamental questions. Human evaluation, performed by linguists, translators, or native speakers, traditionally assessed accuracy and fluency (Koehn and Monz, 2006; Callison-Burch et al., 2007), with recent approaches also considering criteria such as style, tone, and terminology (Lommel et al., 2014; Freitag et al., 2021). However, human annotations are costly and can be very subjective, motivating the need for automatic evaluation (AE). Reference-based AE, which compares MT outputs to human translations, remains the gold standard (Agrawal et al., 2024), though reference translations are costly, often scarce, and can vary in quality (Freitag et al., 2021). On the other hand, reference-less AE (or quality estimation) predicts translation quality without references. AE metrics have evolved considerably, from string-matching approaches such as BLEU (Papineni et al., 2002), which measure surface-level similarity, to neural models such as COMET (Rei et al., 2020), which focus on semantic similarity and are trained to align with human judgments. Neural metrics outperform traditional methods in ranking system outputs (Mathur et al., 2020; Specia et al., 2020) and, as a more recent phenomenon, large language models (LLMs) have

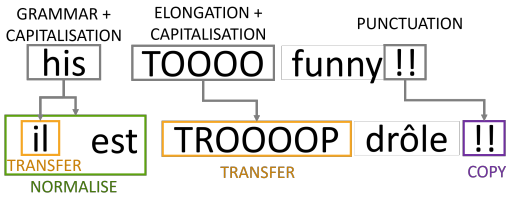


Figure 1: Example of non-standard phenomena in English–French translation with specific actions. The grammatical error is corrected (NORMALISE), the irregular capitalisation and word elongation are translated into their French equivalents (TRANSFER), and the repeated punctuation is copied (COPY).

[1] **TRIGGER WARNING:** UGC often contains texts that may be considered explicit, offensive, or vulgar. In this paper, we provide some examples containing profanity. We limit ourselves to using explicitly only two words: *f**k* and *sh**.

also demonstrated strong capabilities in reference-based and reference-less AE (Freitag et al., 2024; Zerva et al., 2024). Despite these advances, evaluating MT remains a complex and active research area, as evidenced by the long-running, annual WMT Metrics Shared Task (Callison-Burch et al., 2008; Lavie et al., 2025).

User-generated content (UGC) adds further complexity, as it is characterised by non-standard phenomena (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013; Baldwin and Li, 2015; van der Goot et al., 2018; Sanguinetti et al., 2020; Bawden and Sagot, 2023), including errors (e.g., grammatical, typographical, spelling) and literary devices that convey style, sentiment, and informality (e.g., acronyms, slang, phonetisation, code-switching, emojis and other marks of expressiveness). As a result, translating UGC presents unique challenges beyond those found in traditional text translation, raising questions about how much standardness should be preserved in translation: *Which non-standard phenomena should be corrected and normalised? Which should be maintained, and how?* (see Figure 1 for an example). Without explicit guidance, even human "gold" translations vary in their treatment of such features, making reference-based AE more tricky.

We address two research questions (**RQs**): **(1) What is a "good" translation of UGC?** and **(2) How can UGC translation be evaluated fairly?** First, we analyse four translation datasets of English and French UGC: RoCS-MT (Bawden and Sagot, 2023), FooTweets (Sluyter-Gäthje et al., 2018), MMTC (McNamee and Duh, 2022) and PFSMB (Rosales Núñez et al., 2019). We show that they apply different translation guidelines, a consequence of decisions about standardisation being influenced by the intended use case and context. Second, we evaluate six LLMs on UGC translation, under varying conditions: zero-shot without guidelines, with dataset-specific guidelines, and cross-application of guidelines between datasets. Through this controlled setup, we show that reference-based AE is highly sensitive to the underlying annotation standards, that providing explicit guidelines can significantly shift metric scores, and that fair evaluation of UGC translation requires guideline-aware models and metrics.[2]

[2]Code and guidelines can be found in `https://github.com/lydianish-phd/ugc-mt-eval-challenges`

1 Related Work

While most work on UGC translation robustness focuses on handling non-standard phenomena in the source text, Bolding et al. (2023) is one of the few to address non-standardness in the target side as well. They explore the use of LLMs to explicitly "*clean noisy translation data*," framing robustness primarily as the ability to normalise non-standard language. They use GPT-4 (OpenAI, 2023) to remove noise from the target side of the MTNT corpus (Michel and Neubig, 2018), producing a cleaned version intended to better evaluate robustness to non-standard input. Their underlying assumption is that "*an effective NMT model is capable of translating a noisy source sentence into a clean target sentence*." While this perspective aligns with many practical applications, it implicitly adopts a fully normalising view of UGC translation, where non-standard phenomena are treated as noise to be eliminated. In contrast, our work does not assume that the appropriate target of UGC translation is necessarily clean or standardised. We study how different datasets encode distinct translation guidelines, resulting in varying degrees of standardness in the reference translations. Rather than cleaning the data, we treat these guidelines as an explicit modelling and evaluation variable, and investigate how to control MT style by prompting LLMs to follow them—and how AE is affected when such guideline differences are ignored.

Early approaches to controlling MT model translation style (e.g., formality, politeness, dialect) included appending style-specific labels/tokens to the input sentences, as well as fine-tuning (Sennrich et al., 2016; Niu and Carpuat, 2020; Rippeth et al., 2022; Riley et al., 2023). However, these methods typically required retraining or fine-tuning models on style-specific data, limiting their flexibility. In contrast, the ability of LLMs to generalise across domains and follow instructions makes them particularly well-suited for controllable MT. Recent studies have shown that prompting LLMs with contextual cues can effectively steer the style and register of translations without additional training. Examples of guidelines include indicating specific terminology (Moslem et al., 2023; Lyu et al., 2024), the intended audience and purpose of the translation (Yamada, 2023), the domain of the text (Gao et al., 2024), the desired tone or style (Lyu et al., 2024; Liu et al., 2024), the language variant (Fleisig et al., 2024), and the

desired grammatical gender (Sánchez et al., 2024).

In addition to style transfer, LLMs have been used to improve the accuracy and fluency of translations. They demonstrate impressive zero-shot robustness to UGC translation, and tend to implicitly normalise and correct non-standard phenomena (Bawden and Sagot, 2023; Peters and Martins, 2024; Supryadi et al., 2024; Popović et al., 2024). Furthermore, Pan et al. (2024) showed that LLMs could learn MT robustness through in-context examples containing synthetic and natural UGC. Although LLMs have been applied to correction tasks such as grammatical error correction (Coyne et al., 2023; Fang et al., 2023; Maeng et al., 2023; Kwon et al., 2023a; Penteado and Perez, 2023; Katinskaia and Yangarber, 2024), spelling error correction (Zhang et al., 2023; Li et al., 2024), and dialectal normalisation (Alam and Anastasopoulos, 2025), models such as xTower (Treviso et al., 2024) show that LLMs can also be explicitly prompted to jointly translate and correct errors.

In our work, we use translation guidelines that relate to the style of the translation (e.g., UGC-specific phenomena and terminology) as well as its fluency (e.g., grammar, spelling, punctuation). In some cases, we ask the models to "transfer" the non-standardness to its equivalent in the target language, or to only expand abbreviations if the result is not unnatural. This is comparable to the *dynamic equivalence* (Nida, 1969) prompts of Yamada (2023), who asked ChatGPT to translate a Japanese sentence with cultural references into "something that would be understood in an English-speaking culture".

2 Methodology

First (**RQ1**), we extract and analyse the translation guidelines provided in existing UGC datasets, treating them as a proxy for the human preferences that shape reference translations. Second (**RQ2**), we conduct a controlled experimental study where LLMs generate translations under different prompting conditions, allowing us to assess the impact of explicit guidelines on AE scores.

2.1 Translation Guidelines as a Proxy for Human Preferences

We define a *translation guideline* as a prescribed action to be applied to a given non-standard linguistic phenomenon in the source text. In what follows, we first describe the main types of non-standard phenomena encountered in UGC, and then outline the possible actions that may be recommended in guidelines.

2.1.1 UGC Translation Datasets

To determine the various decisions that are made when translating non-standard text, we analyse four parallel datasets for the translation of UGC from social media sites and discussion forums: RoCS-MT (Bawden and Sagot, 2023) for English–French, FooTweets (Sluyter-Gäthje et al., 2018) for English–German, and MMTC (McNamee and Duh, 2022) and PFSMB (Rosales Núñez et al., 2019) for French–English translation.[3] In particular, the RoCS-MT dataset creation pipeline includes a lexical normalisation step, and it is the normalised versions that were then translated into the other languages. This was done to "optimise the quality of the translation and to reduce the arbitrariness that may be introduced when transferring non-standard variation to the target language". On the other hand, FooTweets is a corpus of *Twitter* posts about the 2014 FIFA World Cup created with the downstream task of sentiment analysis in mind: they showed a special focus on preserving the informal nature and the sentiment of the *tweets*.

MMTC and RoCS-MT have the most detailed lists of guidelines (i.e., the most phenomena with specific instructions), followed by FooTweets. Conversely, PFSMB has broader instructions that are summarised in two sentences: "Typographic and grammatical errors were corrected in the gold translations but the language register was kept. For instance, idiomatic expressions were mapped directly to the corresponding (e.g., `mdr` has been translated to `lol`) and letter repetitions are also kept (e.g., `ouiii` has been translated to `yesss`)".

2.1.2 Taxonomy of Non-standard Phenomena

We curate a list of twelve phenomena from the instructions that were given to human translators in the creation of the datasets. We do that specifically based on the MMTC and RoCS-MT guidelines (which are the most detailed ones): (1) Grammar; (2) Spelling; (3) Word elongation or letter repetitions; (4) Capitalisation (e.g., missing at the beginning of a sentence or a proper noun, using all caps or case swapping for emphasis);

[3] We left out other well-known datasets such as MTNT (Michel and Neubig, 2018) and Foursquare (Berard et al., 2019) because they provide no details of the guidelines (if any) that were given to the human translators.

Phenomenon	En–Fr RoCS-MT	En–De FooTweets	Fr–En MMTC	Fr–En PFSMB
1. Grammar	NORMALISE	NORMALISE	NORMALISE	NORMALISE
2. Spelling	NORMALISE	NORMALISE	NORMALISE	NORMALISE
3. Word elongation (e.g., `gooooaaaalllll`)	NORMALISE	TRANSFER	TRANSFER	TRANSFER
4. Capitalisation (e.g., `NOPE, SoRry`)	NORMALISE	TRANSFER	TRANSFER	TRANSFER
5. Informal abbreviations (e.g., `gonna, u`)	NORMALISE	NORMALISE	NORMALISE	TRANSFER
6. Informal acronyms (e.g., `LOL, TBH`)	NORMALISE*	NORMALISE*	TRANSFER	TRANSFER
7. Hashtags and subreddits	COPY	COPY	TRANSFER	TRANSFER**
8. URLs, user IDs, and retweet marks (`RT`)	COPY	COPY	COPY	COPY
9. Emoticons and emojis	COPY	COPY	COPY	COPY
10. Atypical punctuation	NORMALISE	COPY	COPY	COPY
11. Overt profanity (e.g., `fuck`)	TRANSFER	TRANSFER	TRANSFER	TRANSFER
12. Self-censored profanity (e.g., `f*ck`)	NORMALISE	NORMALISE	NORMALISE	TRANSFER

Table 1: Summary of guidelines for translating non-standard phenomena as used in the creation of four parallel UGC datasets. **: Acronyms are expanded (e.g., `TBH` → `to be honest`) unless doing so would sound unnatural (e.g., `LOL` is, in practice, more used in its abbreviated form than in its full form `laughing out loud`). **: Hashtags are translated only if they have a grammatical function in the sentence (e.g. `#ItAnnoysMeWhen people don't listen when I'm talking.`).*

(5) Informal abbreviations; (6) Informal acronyms; (7) Hashtags and subreddits (e.g., `#WorldCup`, `r/Nicegirls`); (8) URLs, @-mentions to user IDs, and retweet marks (RT); (9) Emoticons and emojis; (10) Atypical punctuation (e.g., missing or repeated); (11) Overt profanity; and (12) Self-censored profanity.

2.1.3 Taxonomy of Actions

From the guidelines that were given to human translators in the creation of these datasets, we define three major actions (NORMALISE, COPY and TRANSFER) to deal with non-standard phenomena while translating UGC (see Figure 1 for an example). We also define two more that do not appear in these guidelines, but can be expected from generated MT outputs: OMIT and CENSOR. The five are described below.

1. **NORMALISE**: the non-standard phenomenon is omitted or corrected in the source text, thus producing a standard translation. Examples of this are: correcting spelling and grammatical errors, standardising the use of capital letters and punctuation, expanding abbreviations, removing repeated characters, etc. Note that in the case of self-censored profanity, normalising self-censorship means to render the uncensored version: e.g., `f*ck` ↔ `fuck`.

2. **COPY**: the non-standard phenomenon is copied as it is in the translation. This is usually applied to special words and characters, such as social media entities, punctuation and emojis.

3. **TRANSFER**: the non-standard phenomenon is mapped to an equivalent non-standard phenomenon in the target language: e.g., `LOL (laughing out oud)` → `MDR (mort de rire)`. This is not to be confused with COPY, which does not perform any changes. For example, if a hashtag (e.g., `#WorldCup`) is kept intact in the output, it is *copied*, but if it is translated into a hashtag in the target language (e.g., `#CoupeDuMonde`), it is *transferred*.

4. **OMIT**: the non-standard phenomenon is ignored or skipped from the translation. This is not to be confused with cases where the omission is a result of standardisation (e.g., omitting repeated punctuation marks). Instead, the non-standardness is not dealt with at all (e.g., skipping a username mention or URL).

5. **CENSOR**: profanity and offensive language (e.g., swear words, insults) are replaced with less offensive terms. E.g. `fucked up` → `made a mistake`. Other examples include strong or potentially triggering words that are used figuratively: `only reason I haven't killed myself after that boring game is...` ↔ `only reason I haven't harmed myself after that uninteresting game is...`

2.1.4 Final List of Translation Guidelines

Table 1 summarises how the 12 non-standard phenomena are translated in the datasets using the taxonomy of actions previously defined.[4] We observe that RoCS-MT and PFSMB represent two

[4]For the phenomena without explicit mention in the FooTweets and PFSMB guidelines, we deduce the instructions by a qualitative analysis of the datasets. For example, we search for a word elongated in the source side (e.g., through a search for vowel repetitions) and see how it was translated.

ends of a continuum: RoCS-MT applies the highest degree of normalisation, while PFSMB preserves the most non-standard phenomena. PFSMB only has 5 out of 12 guidelines in common with RoCS-MT, while FooTweets and MMTC occupy an intermediate position between these two extremes, respectively with 9 and 7 guidelines in common with RoCS-MT. For better readability, we categorise the datasets into two groups corresponding to the level of normalisation of their guidelines and refer to them as: (i) RoCS-MT (the "most standardising"), and (ii) FooTweets, MMTC, and PFSMB, (the "least standardising").

Note that there are some exceptions provided in the guidelines. For example, RoCS-MT and FooTweets translators were suggested to normalise informal acronyms (i.e., expand them to their full form), provided that doing so would not sound unnatural in the target language. For instance, `LOL` is more naturally used in its abbreviated form than in its expanded form `Laughing Out Loud`. It has even become part of informal vocabulary, producing conjugated forms such as `I totally LOLed during the movie!`. On the handling of hashtags, we inferred that PFSMB seems to translate them when they serve a grammatical function in the sentence, e.g. `#CaMeVénèreQuand on m'écoute pas quand je parle.` → `#ItAnnoysMeWhen people don't listen when I'm talking.`

2.2 Experimental Study

We evaluate translation models on the four parallel UGC datasets and the effects of incorporating corpus-specific translation guidelines to control the generation of model outputs.[5]

Translation Models We use the state-of-the-art encoder-decoder model NLLB-3B (NLLB Team, 2022) as a baseline for evaluating MT performance. We also evaluate six instruction-tuned, decoder-only LLMs: Gemma-2-9B (Gemma Team, 2024), Granite-4.1-8B (IBM Research, 2026), LLaMA-3.1-8B (Llama Team, 2024), Mistral-0.3-7B (Mistral AI Team, 2024), Qwen-2.5-7B (Qwen Team, 2024) and Tower-0.2-7B (Alves et al., 2024). Note that the Tower model has been specifically fine-tuned for translation tasks. We generate outputs with a beam search of 5 for NLLB-3B and we use the vLLM toolkit (Kwon et al., 2023b) for LLM inference, with the following parameters: greedy sampling with BF16 mixed precision (Kalamkar et al., 2019), and a maximum model context length of 2,048 tokens. And we prompt the LLMs to translate one line of text at a time. We also run a post-processing script to extract the translated sentences from verbose outputs and identify refusals to translate (Briakou et al., 2024). The maximum output sequence length is set to 512 tokens for all models.

Controlled Generation In order to control the translation outputs of LLMs to match the style of a specific corpus's human references, we use a list of 12 translation guidelines derived from Table 1 as instructions in the LLM prompts. Appendix C details the LLM prompt templates and the list of translation guidelines for each corpus. We define a prompting configuration as a pair constituted of a model and a set of translation guidelines. We evaluate different prompting configurations for each LLM: one without any translation guidelines (the default), and one configuration for each of the specific translation guidelines for each corpus. In particular, we will compare two evaluation scenarios for each LLM and dataset:

1. **Matching guidelines**: the guidelines used in the prompts correspond to those originally defined for the dataset, e.g., using RoCS-MT guidelines when translating RoCS-MT texts;

2. **Mismatching guidelines**: the guidelines used in the prompts are taken from a different dataset, e.g., using PFSMB guidelines when translating RoCS-MT texts.

Evaluation Metrics We assess translation quality using neural semantic-based metrics for AE, specifically the reference-based COMET-22 (Rei et al., 2022a) and the reference-less COMET-Kiwi (Rei et al., 2022b). Following the SacreCOMET recommendations (Zouhar et al., 2024), we set sentence-level scores to zero for empty model outputs. We also use the surface-level metric BLEU (Papineni et al., 2002) to complement COMET-22's semantic-based scores.[6] For better readability, we report all scores as percentages, and

[5] See Appendix B for the links to the models and toolkits.

[6] We use SacreBLEU (Post, 2018) with the signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0` for BLEU and `Python:3.12|PyTorch:2.10.0|version:2.2.7` for the COMET models.

refer to COMET-22 simply as COMET. We evaluate statistical significance using paired bootstrap resampling (Koehn, 2004) with 300 samples and a sampling ratio of 0.4. For each metric, we compute the distribution of score differences between each configuration (model + guidelines) and the default no-guideline model. We report the mean difference, 95% confidence intervals, and p-values. A result is considered statistically significant if the 95% confidence interval of the difference does not include zero. For the metrics computed at the sentence level (COMET and COMET-Kiwi), bootstrap resampling is performed over sentence-level scores. Corpus-level scores are obtained by averaging the sampled sentence scores.

Human Evaluation We conducted a human evaluation on two UGC translation datasets: RoCS-MT (English–French) and PFSMB (French–English). Specifically, for PFSMB, we used the PMUMT subset (Rosales Núñez et al., 2021), which was manually normalised and annotated for UGC phenomena. To obtain a unified sampling framework across datasets, we first mapped the original dataset-specific UGC taxonomies to our proposed 12-category taxonomy. We then performed stratified sampling over the mapped categories to ensure coverage of diverse UGC phenomena, resulting in 50 sampled sentence pairs per dataset. We used the outputs of the Gemma-2-9B and LLaMA-3.1-8B models. For each sample, annotators were presented with the source sentence, its normalised version, the reference translation, the applicable guidelines, and two anonymised system outputs corresponding to the default and matching-guideline-based prompting conditions. The system pairs were randomly selected from the outputs of the two models. Annotators were asked to perform pairwise comparisons assessing both overall translation quality and adherence to the UGC guidelines. A total of four annotators, fluent in both English and French, participated in the study. Two annotators annotated both datasets, while two additional annotators annotated only the dataset for which the target language was their native language. Thus, each sample was annotated by three annotators. We measure inter-annotator agreement using Krippendorff's α (Krippendorff, 1970) and pairwise Cohen's κ (Cohen, 1960), and obtain final preferences using majority voting over the three annotations. To assess the agreement between human judgments and automatic metrics, we derive pairwise preferences from COMET and COMET-Kiwi by comparing the score difference $\Delta = s_{\text{guided}} - s_{\text{default}}$ against a threshold $\epsilon = 0.5$. Differences within $\pm\epsilon$ points are treated as ties.

3 Results and Analysis

3.1 Reference-based Quantitative Results

Table 2 illustrates the COMET and COMET-Kiwi scores for evaluating UGC translation across eleven prompting configurations and four datasets. BLEU scores are in the appendix Table 3.

3.1.1 No-guideline Results

When prompted without any specific guidelines, Tower-0.2-7B performs the best on all datasets except RoCS-MT, where it lags behind Gemma-2-9B. A qualitative analysis of the generated outputs shows that Tower-0.2-7B's outputs tend to be more formal than Gemma-2-9B's on RoCS-MT. Overall, most LLMs outperform the NLLB-3B baseline across datasets. The main exception is MMTC, where NLLB-3B retains a substantial advantage over most models (up to $\approx$ 10 COMET points in Table 2). The performance gap on MMTC is likely due to the dataset containing many *Twitter* posts starting with lists of username mentions, which several LLMs tend to ignore (an instance of the OMIT strategy discussed in §2.1.4), while NLLB-3B, Qwen-2.5-7B and Tower-0.2-7B generally preserve them.[7] In contrast, Mistral-0.3-7B and Qwen-2.5-7B underperform compared to NLLB-3B on RoCS-MT.

3.1.2 LLM Instruction Following

LLaMA Table 2 shows that LLaMA-3.1-8B results in the lowest COMET scores among the evaluated LLMs. Furthermore, its performance generally worsens when prompted with translation guidelines (e.g., up to a 6-point drop on FooTweets). One reason behind LLaMA-3.1-8B's poor performance is that the model refuses to generate translations when it considers the content to be offensive, explicit, hateful or harmful, a behaviour highlighted by Qian et al. (2024). It also fails to produce translations for texts that are too short or seem incomplete (e.g., stand-alone hashtags and usernames). The appendix Figure 2 illustrates the percentage of translation requests re-

[7] See the appendix Table 4 for the number of social media entities in the datasets and default model outputs.

Model	Guideline	COMET RoCS-MT	COMET FooTweets	COMET MMTC	COMET PFSMB	COMET-Kiwi RoCS-MT	COMET-Kiwi FooTweets	COMET-Kiwi MMTC	COMET-Kiwi PFSMB
NLLB-3B	—	73.21	79.90	86.54	73.17	75.52	76.51	79.52	69.83
Gemma-2-9B	None	76.36	83.22	77.78	78.88	77.89	79.41	77.65	75.17
	+RoCS-MT	**76.57** ↑	84.92 ↑ *	85.48 ↑ *	80.29 ↑ *	77.94 ↑	**79.97** ↑ *	80.17 ↑ *	**75.48** ↑
	+FooTweets	75.51 ↓ *	85.60 ↑ *	86.44 ↑ *	80.50 ↑ *	77.49 ↓ *	79.68 ↑	80.36 ↑ *	75.10 ↓
	+MMTC	75.67 ↓ *	85.27 ↑ *	86.59 ↑ *	**80.64** ↑ *	77.55 ↓	79.89 ↑ *	80.48 ↑ *	75.24 ↑
	+PFSMB	74.54 ↓ *	**86.20** ↑ *	86.24 ↑ *	80.51 ↑ *	77.06 ↓ *	78.81 ↓ *	80.14 ↑ *	74.95 ↓
Granite-4.1-8B	None	74.27	83.66	77.56	77.95	76.84	78.65	77.66	74.11
	+RoCS-MT	74.53 ↑	83.71 ↑	84.77 ↑ *	78.18 ↑	76.81 ↓	78.46 ↓	79.78 ↑ *	74.64 ↑
	+FooTweets	74.33 ↑	84.12 ↑ *	86.44 ↑ *	78.35 ↑	76.85 ↑	78.41 ↓	80.30 ↑ *	74.29 ↑
	+MMTC	74.16 ↓	83.74 ↑	87.77 ↑ *	78.34 ↑	76.75 ↓	78.44 ↓	80.73 ↑ *	74.28 ↑
	+PFSMB	73.75 ↓	83.88 ↑	87.99 ↑ *	78.32 ↑	76.51 ↓	78.19 ↓ *	80.76 ↑ *	74.04 ↓
LLaMA-3.1-8B	None	73.49	80.96	77.01	73.64	75.42	76.73	76.08	70.08
	+RoCS-MT	71.90 ↓ *	75.11 ↓ *	81.25 ↑ *	74.14 ↑	73.52 ↓ *	70.34 ↓ *	76.89 ↑	70.32 ↑
	+FooTweets	71.27 ↓ *	76.33 ↓ *	84.05 ↑ *	74.36 ↑	73.09 ↓ *	71.29 ↓ *	77.76 ↑ *	69.99 ↓
	+MMTC	70.94 ↓ *	75.19 ↓ *	82.72 ↑ *	71.79 ↓	72.60 ↓ *	70.62 ↓ *	76.46 ↑	67.50 ↓ *
	+PFSMB	71.04 ↓ *	77.59 ↓ *	83.38 ↑ *	73.07 ↓	73.10 ↓ *	72.61 ↓ *	77.15 ↑	68.45 ↓
Mistral-0.3-7B	None	72.33	81.18	78.46	75.73	75.00	77.02	77.17	73.22
	+RoCS-MT	72.96 ↑ *	81.71 ↑ *	83.43 ↑ *	76.06 ↑	75.28 ↑	77.61 ↑ *	78.89 ↑ *	73.83 ↑
	+FooTweets	72.85 ↑ *	81.99 ↑ *	84.23 ↑ *	76.36 ↑ *	75.28 ↑	77.56 ↑ *	79.19 ↑ *	74.04 ↑ *
	+MMTC	72.57 ↑	82.00 ↑ *	84.61 ↑ *	76.55 ↑ *	75.08 ↑	77.41 ↑ *	79.24 ↑ *	73.79 ↑ *
	+PFSMB	72.47 ↑	81.97 ↑ *	85.31 ↑ *	76.64 ↑ *	75.14 ↑	77.43 ↑ *	79.49 ↑ *	73.86 ↑ *
Qwen-2.5-7B	None	72.19	82.70	87.10	78.67	75.25	76.83	80.39	73.75
	+RoCS-MT	71.61 ↓	83.55 ↑ *	88.39 ↑ *	78.72 ↓	74.30 ↓ *	75.67 ↓ *	80.39	73.35 ↓
	+FooTweets	70.48 ↓ *	83.46 ↑ *	88.20 ↑ *	78.36 ↓	73.43 ↓ *	74.91 ↓ *	79.92 ↓ *	72.53 ↓ *
	+MMTC	70.76 ↓ *	82.52 ↓	88.42 ↑ *	78.63 ↓	73.77 ↓ *	75.93 ↓ *	80.05 ↓	72.58 ↓ *
	+PFSMB	70.60 ↓ *	82.81 ↑	88.18 ↑ *	78.33 ↓	73.80 ↓ *	75.72 ↓ *	80.01 ↓	72.38 ↓ *
Tower-0.2-7B	None	75.33	85.25	**88.94**	79.05	**78.16**	78.11	**81.04**	74.37
	+RoCS-MT	75.42 ↑	85.84 ↑ *	88.87 ↓	79.18 ↑	77.92 ↓	78.21 ↑	80.93 ↓	74.23 ↓
	+FooTweets	75.35 ↓	85.85 ↑ *	88.92 ↓	79.32 ↑	77.88 ↓	78.32 ↑	80.88 ↓	74.30 ↓
	+MMTC	75.31 ↓	85.82 ↑ *	**88.94**	79.03 ↓	77.90 ↓	78.12 ↑	80.94 ↓	74.13 ↓
	+PFSMB	75.26 ↓	85.82 ↑ *	88.86 ↓	79.04 ↑	77.89 ↓	77.97 ↓	80.92 ↓	74.21 ↓

Table 2: COMET and COMET-Kiwi scores for translating UGC with and without corpus-specific guidelines. *Compared to the no-guideline baseline within each model family, score improvements are marked by* ↑ *and decreases by* ↓. *Statistical significance with respect to the no-guideline baseline is indicated by * (95% confidence interval), while the best score within each model family is underlined. The best overall score for each dataset is shown in* ***bold****.*

fused by LLaMA-3.1-8B. We observe the highest percentage for the no-guideline scenario (3%) on PFSMB. Moreover, adding corpus-specific guidelines significantly increases the ratio of refused requests (up to 8% on PFSMB and MMTC), likely because all guidelines include explicit instructions to preserve profanity in the translation. In contrast, we notice lower rates of refusal on FooTweets and RoCS-MT, even with the guidelines (≤ 4%), possibly because FooTweets (based on football reactions) is more likely to have "safer" content. Meanwhile, explicit or triggering content are specifically filtered out in the RoCS-MT dataset creation pipeline.

Tower Tower-0.2-7B, which was built on top of a LLaMA-2 model (Touvron et al., 2023), does not display the same censored behaviour as LLaMA-3.1-8B. However, Table 2 shows that prompting Tower-0.2-7B with translation guidelines results in minimal gains compared to the no-guideline scenario. These results suggest that Tower-0.2-7B tends to stick to its own translation style, ignoring the instructions. To further confirm this, we compute BLEU scores to measure the lexical overlap between the model outputs of all configurations (with and without guidelines) and report the scores for Gemma-2-9B and Tower-0.2-7B in the appendix Figure 3. We see little lexical variation between Tower-0.2-7B's outputs, regardless of translation guidelines.

Other models In contrast, the other evaluated LLMs (Gemma-2-9B, Granite-4.1-8B, Mistral-0.3-7B and Qwen-2.5-7B) produce more noticeable variations across guideline configurations and, unlike LLaMA-3.1-8B, they do not exhibit systematic refusal behaviour. This suggests that these models are more willing to adapt their translation strategies to the requested

UGC handling instructions, an effect that we analyse further in the following sections.

3.1.3 Effects of Corpus-specific Guidelines

The following observations focus on Gemma-2-9B, Mistral-0.3-7B, Granite-4.1-8B and Qwen-2.5-7B. We note that when restricting LLaMA-3.1-8B to the subset of samples for which the model does not refuse to translate, we observe the same overall trends.

Matching guidelines on all datasets Across most models and datasets, adding corpus-specific translation guidelines improves COMET scores compared to the no-guideline setting, often with statistically significant gains (Table 2). In many cases, the best results are obtained with the matching dataset guidelines, particularly for Gemma-2-9B, Granite-4.1-8B and Mistral-0.3-7B. Overall, Gemma-2-9B with guidelines achieves the best results across datasets, except on MMTC, where the default Tower-0.2-7B model remains the strongest system.

Mismatching guidelines between the least standardising datasets Applying guidelines from the least standardising datasets (FooTweets, MMTC and PFSMB) generally improves performance across these datasets, often with statistically significant gains. In several cases, preservation-oriented guidelines from one dataset transfer effectively to another preservation-oriented dataset, and the score variations between these guideline configurations remain very small. In particular, guideline prompting substantially improves the handling of social media entities on MMTC, where several models tend to omit usernames and online entities by default. However, the least standardising guidelines, especially those of PFSMB, can also degrade translation quality by encouraging models to over-transfer non-standardness. This behaviour is particularly noticeable for Qwen-2.5-7B, which appears less robust to non-standard source inputs and more prone to generating hallucinated forms, ungrammatical constructions or noisy lexical variations when attempting to follow preservation-oriented instructions. On PFSMB, guideline prompting leads to statistically significant improvements for Gemma-2-9B and Mistral-0.3-7B, while the variations observed for the other models remain non-significant.

Mismatching guidelines between RoCS-MT and the least standardising datasets In contrast, applying the less standardising guidelines to RoCS-MT generally results in negligible variations or decreases in performance. Since RoCS-MT follows the most standardising translation strategy, these results suggest that several models already tend to normalise UGC phenomena by default. Consequently, introducing preservation-oriented instructions often moves the outputs away from the reference style. We observe statistically significant decreases for Gemma-2-9B and Qwen-2.5-7B, while the variations for Granite-4.1-8B and Mistral-0.3-7B remain non-significant. Conversely, applying RoCS-MT guidelines to the other datasets generally improves performance compared to the default setting, although these gains are usually smaller than those obtained with the less standardising guidelines. Overall, these findings indicate that guideline prompting is most effective when the prompting strategy matches the intended level of normalisation or preservation in the target dataset.

3.2 Reference-less Quantitative Results

The COMET-Kiwi results in Table 2 differ from the reference-based COMET results in that the best scores are generally obtained with either no guidelines (default Tower-0.2-7B) or with more standardising guideline configurations (RoCS-MT-guided Gemma-2-9B), rather than necessarily with the guidelines matching the dataset. However, similarly to COMET, the score variations across guideline configurations remain small and non-significant. Furthermore, when the guidelines lead to less standard outputs, the metric tends to assign lower scores. In particular, the least standardising guidelines, especially those of PFSMB, frequently result in the lowest COMET-Kiwi scores. This behaviour is particularly noticeable for Qwen-2.5-7B, whose outputs become substantially noisier under preservation-oriented prompting. These results suggest that the metric remains biased towards standardised outputs and is less robust to higher levels of non-standardness. This is consistent with Aepli et al. (2023)'s conclusion that COMET-Kiwi is not robust to non-standard orthographic variation in dialects and is biased towards standardised outputs.

3.3 Qualitative Analysis

The appendix Table 5 illustrates a few example sentences from the datasets and their output translations by Gemma-2-9B, with and without matching guidelines, as well as their sentence-level COMET scores. We observe that including specific translation guidelines in the prompt helps Gemma-2-9B to better deal with certain UGC-specific phenomena in a way that differs from its default behaviour, thus increasing the sentence-level scores overall (see examples 1, 2 and 3).

Note, however, that the model's behaviour may be inconsistent: Gemma-2-9B is not always able to apply the guidelines correctly. A prominent example of this is its tendency to translate (TRANSFER) hashtags on FooTweets, despite clear instructions to COPY, as seen in examples 4 and 5. Other instances include the failure to TRANSFER character repetitions (e.g., `niceeee` → `schööön` in example 4), and some attempts to do so even lead the model to repeat characters indefinitely. One possible reason for this is that this guideline contradicts the instruction to NORMALISE spelling, raising questions about the compositionality of the twelve defined guidelines and the overall feasibility of applying them consistently. Likewise, instructions to COPY irregular punctuation and capitalisation can contradict the guideline to NORMALISE grammar.

In addition, TRANSFERRING non-standardness can degrade translation quality, as evidenced by the ungrammatical formulation '`no even need`' in example 6 and the implicit repetition '`rn now`' in example 7 (which significantly lowered the COMET score).[8] We also observed that the profanity guidelines can cause Gemma-2-9B to hallucinate swear words or explicit terms.

Finally, example 8 shows that, while prompting Gemma-2-9B with guidelines helps control the translation output style, it does not increase the model's robustness to the non-standard phenomena that it inherently fails to translate.

3.4 Human Evaluation Results

The human evaluation results broadly confirm the trends observed with the automatic metrics (see the appendix Table 6). On both datasets, the guided outputs are preferred more often than the default outputs for both overall quality and guideline adherence. On PFSMB, the guided outputs are preferred in 38% of all samples for overall quality, compared to only 12% for the default outputs, while half of the samples are judged as ties. The effect is even stronger for guideline adherence, where the guided outputs are preferred in 36% of the samples compared to only 4% for the default outputs. When ties are ignored, these proportions rise to 76% and 90%, respectively. In contrast, the gains on RoCS-MT are smaller, with guided outputs preferred in 30% of the samples for quality and 12% for guideline adherence, while ties account for 54% and 80% of the judgments. This is consistent with the quantitative results; the models already tend to produce relatively standardised outputs by default, making the impact of additional prompting less noticeable on a dataset whose references are highly standardising.

The joint evaluation outcomes further highlight the differences between the two datasets. On PFSMB, the guided outputs are preferred on both dimensions in 24% of the evaluated samples, compared to only 10% on RoCS-MT. Furthermore, we observe virtually no cases where the guided outputs improve adherence at the expense of translation quality (0% on PFSMB and 2% on RoCS-MT). This indicates that preservation-oriented prompting can improve the handling of UGC phenomena without substantially degrading translation quality.

Inter-annotator agreement is moderate overall (Krippendorff's α = 0.25–0.45). We also observe moderate to substantial agreement between human preferences and automatic metrics. Overall quality judgments show the strongest agreement with COMET-Kiwi, reaching Cohen's κ values of 0.62 on RoCS-MT and 0.54 on PFSMB. In contrast, guideline adherence judgments are better captured by COMET, particularly on PFSMB (κ = 0.56 compared to 0.46 for COMET-Kiwi). This is consistent with the fact that COMET has access to the references and can therefore better capture adherence to the translation style they embody.

4 Discussion

Defining the "gold standard" for UGC translation (RQ1) By analysing the human references of different corpora and how they treat non-standard phenomena, we can infer style preferences for UGC translation. Our study of four datasets shows that what counts as a "good" translation varies with the reference style: some datasets are highly standardising (RoCS-MT),

[8] '`rn`' stands for '`right now`'.

and others minimally standardising (PFSMB), with FooTweets and MMTC in between. Using our 12-phenomena taxonomy and five actions (NORMALISE, COPY, TRANSFER, OMIT, CENSOR), we observe systematic differences in how acronyms, hashtags, letter repetitions, and social media entities are handled. Both the automatic and human evaluation results further confirm that these stylistic differences materially affect model performance and human preferences. This demonstrates that translation of UGC is inherently guideline-dependent and context-sensitive.

Guideline awareness in models and metrics for fair evaluation (RQ2) Fair evaluation requires putting models in comparable situations, taking into account the fact that UGC translation is guideline-dependent (see RQ1). Ideally, models should be guided by the same translation principles used to generate references. NLLB-3B, an encoder-decoder model that is not style-controlled, cannot be prompted with guidelines and thus produces a stable but standard style. On the other hand, instruction-tuned LLMs such as Gemma-2-9B, Granite-4.1-8B, Mistral-0.3-7B and Qwen-2.5-7B adapt their outputs to guideline-based prompting, while Tower-0.2-7B shows minimal sensitivity to the prompts, suggesting limited stylistic controllability. In contrast, LLaMA-3.1-8B frequently refuses to translate "unsafe" inputs, lowering corpus-level scores and complicating comparisons. We also observe important differences between evaluation metrics. Reference-based metrics (BLEU, COMET) are implicitly style-aware, rewarding outputs aligned with reference guidelines, but penalising mismatches, as with the least-standardising prompts applied to RoCS-MT. In contrast, the reference-less COMET-Kiwi metric tends to favour more standardised outputs, assigning lower scores to highly non-standard translations even when they better match the intended preservation-oriented guidelines, as seen with PFSMB. This highlights the challenge of reliably scoring extreme variation without guideline alignment.

Recommendations We make two practical recommendations for ensuring a fairer evaluation of UGC translation: (1) when style is not a priority and all linguistic variations should be treated equally, use either a reference-less metric or a reference-based metric with multiple versions of the references spanning different levels of standardness; (2) when control over a specific style is required, follow the approach proposed in this work by prompting an LLM with explicit guidelines and evaluating outputs with reference-based methods, ideally an LLM-as-a-judge (Zheng et al., 2023) configured with the same guidelines to provide a controllable, style-sensitive evaluation. More generally, our results suggest that future work on UGC evaluation would benefit from more fine-grained human evaluation protocols explicitly targeting individual UGC phenomena and translation actions. Rather than relying solely on overall quality judgments, future evaluation frameworks could include phenomenon-specific rubrics.

5 Conclusion

In this work, we investigated the role of translation guidelines in the evaluation of UGC machine translation. Through the analysis of four datasets, we showed that UGC translation is inherently style-dependent, with datasets reflecting different translation philosophies ranging from highly standardising to preservation-oriented approaches. Using a unified taxonomy of UGC phenomena and translation actions, we highlighted systematic differences in how non-standard language is treated across corpora.

Our experiments further demonstrated that modern LLMs differ substantially in their ability to follow stylistic translation instructions. While some models successfully adapt to corpus-specific prompting strategies, others either ignore the instructions or exhibit unstable behaviour when handling highly non-standard or explicit content. Both the automatic and human evaluation results show that guideline prompting is most effective when the prompting strategy matches the stylistic assumptions underlying the reference data.

Finally, our findings highlight important limitations of current evaluation practices for UGC translation. Reference-based metrics are strongly influenced by the stylistic properties of the references, while reference-less metrics tend to favour more standardised outputs and remain less robust to highly non-standard language. Overall, our results emphasise the need for more controllable, guideline-aware evaluation frameworks capable of accounting for multiple valid translation styles and varying levels of non-standardness in UGC.

Limitations

Implicit vs. Explicit Translation Guidelines We aimed to use explicit annotation guidelines as a proxy for human preferences. Inferring missing explicit guidelines from the implicit annotator choices can be seen as introducing circularity. However, we argue that in such cases, the guideline creators delegate the decision to the annotators, thus letting the latter's preferences take precedence over their own. In other words: explicit guidelines reflect the preferences of the creators, while implicit ones reflect those of the annotators.

Guideline Descriptions Some guidelines were too vague, contradictory, or inconsistently defined to be applied reliably. Our taxonomy and instruction set may also omit certain phenomena or lack sufficient granularity. A more exhaustive and clearly structured set of guidelines, potentially with example-based definitions per phenomenon, could improve both annotation consistency and model interpretability.

Prompt Engineering We do not explore a wide range of prompting strategies or formulations of the translation guidelines. In particular, our use of zero-shot prompts without few-shot demonstrations may have limited the models' ability to fully adhere to the guidelines (Hendy et al., 2023; Garcia et al., 2023; Coyne et al., 2023; Gao et al., 2024; Sclar et al., 2024; Ceballos-Arroyo et al., 2024; Chatterjee et al., 2024; Pan et al., 2024; Zebaze et al., 2025a; Zebaze et al., 2025b). Future work could investigate how different prompt structures or the inclusion of targeted few-shot examples affect model behaviour in handling specific non-standard phenomena. In particular, it might prove beneficial to use chain-of-thought prompting to instruct the model to handle different subsets of the guidelines sequentially, e.g. first correcting grammar and spelling, and then re-injecting non-standardness in the corrected version.

Metric Coverage We only evaluated using COMET, COMET-Kiwi, and BLEU. Including additional metrics, particularly other neural-based and LLM-based evaluation methods, would allow for a more comprehensive comparison and help generalise the conclusions regarding the sensitivity of metrics to UGC style and guideline adherence.

Acknowledgements

We thank the reviewers for their constructive feedback. We also thank Marine Carpuat and Armel Zebaze for helpful discussions on experiment design, Jakob Maier for qualitative analysis of German translations, Rasul Dent for French–English human evaluation, and Alfred Buregeya for proofreading an earlier draft. This work was granted access to the HPC resources of IDRIS under the allocations 2023AD011013674R2 made by GENCI. It was partly funded by Rachel Bawden and Benoît Sagot's chairs in the PRAIRIE institute, funded by the French national agency ANR, as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by Benoît Sagot's chair in its follow-up, PRAIRIE-PSAI, also funded by the ANR as part of the "France 2030" strategy under the reference ANR23-IACL-0008. It also received further funding from the ANR under the project TraLaLaM ("ANR-23-IAS1-0006").

References

Aepli, Noëmi, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A Benchmark for Evaluating Machine Translation Metrics on Dialects without Standard Orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.

Agrawal, Sweta, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can Automatic Metrics Assess High-Quality Translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

Alam, Md Mahfuz Ibn and Antonios Anastasopoulos. 2025. Large Language Models as a Normalizer for Transliteration and Dialectal Translation. In *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–67, Abu Dhabi, UAE. Association for Computational Linguistics.

Alves, Duarte Miguel, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An Open Multilingual Large Language Model for Translation-Related Tasks. In *First Conference on Language Modeling*, Philadelphia, Pennsylvania, USA.

Baldwin, Tyler and Yunyao Li. 2015. An In-depth Analysis of the Effect of Text Normalization in Social Media. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Denver, Colorado. Association for Computational Linguistics.

Bawden, Rachel and Benoît Sagot. 2023. RoCS-MT: Robustness Challenge Set for Machine Translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Berard, Alexandre, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019. Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 168–176, Hong Kong. Association for Computational Linguistics.

Bolding, Quinten, Baohao Liao, Brandon Denis, Jun Luo, and Christof Monz. 2023. Ask Language Model to Clean Your Noisy Translation Data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3215–3236, Singapore. Association for Computational Linguistics.

Briakou, Eleftheria, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the Implications of Verbose LLM Outputs: A Case Study in Translation Evaluation. *CoRR*, abs/2410.00863.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Callison-Burch, Chris, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Ceballos-Arroyo, Alberto Mario, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (Clinical) LLMs are Sensitive to Instruction Phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Chatterjee, Anwoy, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Coyne, Steven, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction. *CoRR*, abs/2303.14342.

Eisenstein, Jacob. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Fang, Tao, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is ChatGPT a Highly Fluent Grammatical Error Correction System? A Comprehensive Evaluation. *CoRR*, abs/2304.01746.

Fleisig, Eve, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Foster, Jennifer. 2010. "cba to check the spelling": Investigating Parser Performance on Discussion Forum Posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, Markus, 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*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Gao, Yuan, Ruili Wang, and Feng Hou. 2024. How to Design Translation Prompts for ChatGPT: An Empirical Study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '24 Workshops, New York, NY, USA. Association for Computing Machinery.

Garcia, Xavier, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR*, abs/2408.00118.

Hendy, Amr, 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. *CoRR*, abs/2302.09210.

IBM Research. 2026. Granite 4.1 Language Models. `https://huggingface.co/blog/ibm-granite/granite-4-1`.

Kalamkar, Dhiraj D., Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, Jiyan Yang, Jongsoo Park, Alexander Heinecke, Evangelos Georganas, Sudarshan Srinivasan, Abhisek Kundu, Misha Smelyanskiy, Bharat Kaul, and Pradeep Dubey. 2019. A Study of BFLOAT16 for Deep Learning Training. *CoRR*, abs/1905.12322.

Katinskaia, Anisia and Roman Yangarber. 2024. GPT-3.5 for Grammatical Error Correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Koehn, Philipp and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Koehn, Philipp. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Krippendorff, Klaus. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Kwon, Sang Yun, Gagan Bhatia, El Moatez Billah Nagoud, and Muhammad Abdul-Mageed. 2023a. ChatGPT for Arabic Grammatical Error Correction. *CoRR*, abs/2308.04492.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023b. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Li, Kunting, Yong Hu, Liang He, Fandong Meng, and Jie Zhou. 2024. C-LLM: Learn to Check Chinese Spelling Errors Character by Character. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5944–5957, Miami, Florida, USA. Association for Computational Linguistics.

Liu, Pusheng, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024. Step-by-Step: Controlling Arbitrary Style in Text with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295, Torino, Italia. ELRA and ICCL.

Llama Team. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): a Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, 12:455–463.

Lyu, Chenyang, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Maeng, Junghwan, Jinghang Gu, and Sun-A Kim. 2023. Effectiveness of ChatGPT in Korean Grammatical Error Correction. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 464–472, Hong Kong, China. Association for Computational Linguistics.

Mathur, Nitika, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 Metrics Shared Task. In *Proceedings of the Fifth*

Conference on Machine Translation, pages 688–725, Online. Association for Computational Linguistics.

McNamee, Paul and Kevin Duh. 2022. The Multilingual Microblog Translation Corpus: Improving and Evaluating Translation of User-Generated Text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918, Marseille, France. European Language Resources Association.

Michel, Paul and Graham Neubig. 2018. MTNT: A Testbed for Machine Translation of Noisy Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Mistral AI Team. 2024. Mistral-7B-Instruct-v0.3. `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`.

Moslem, Yasmin, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Nida, Eugene A. (Eugene Albert). 1969. *The theory and practice of translation*. Leiden, E.J. Brill.

Niu, Xing and Marine Carpuat. 2020. Controlling Neural Machine Translation Formality with Synthetic Supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8568–8575.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Pan, Leiyu, Yongqi Leng, and Deyi Xiong. 2024. 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)*, pages 2798–2808, Torino, Italia. ELRA and ICCL.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Penteado, Maria Carolina and Fábio Perez. 2023. Evaluating GPT-3.5 and GPT-4 on Grammatical Error Correction for Brazilian Portuguese. In *LatinX in AI Workshop at ICML 2023 (Regular Deadline)*.

Peters, Ben and André F. T. Martins. 2024. Did Translation Models Get More Robust Without Anyone Even Noticing? *CoRR*, abs/2403.03923.

Popović, Maja, Ekaterina Lapshinova-Koltunski, and Maarit Koponen. 2024. 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)*, pages 17–30, San Giljan, Malta. Association for Computational Linguistics.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qian, Shenbin, Constantin Orasan, Diptesh Kanojia, and Félix Do Carmo. 2024. Are large language models state-of-the-art quality estimators for machine translation of user-generated content? In *Proceedings of the Eleventh Workshop on Asian Translation (WAT 2024)*, pages 45–55, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. `https://qwenlm.github.io/blog/qwen2.5`.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rei, Ricardo, 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)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rei, Ricardo, 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)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Riley, Parker, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Rippeth, Elijah, Sweta Agrawal, and Marine Carpuat. 2022. Controlling Translation Formality Using Pre-trained Multilingual Language Models. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 327–340, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Rosales Núñez, José Carlos, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

Rosales Núñez, José Carlos, Djamé Seddah, and Guillaume Wisniewski. 2021. Understanding the Impact of UGC Specificities on Translation Quality. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 189–198, Online. Association for Computational Linguistics.

Sánchez, Eduardo, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific Machine Translation with Large Language Models. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 148–158, Miami, Florida, USA. Association for Computational Linguistics.

Sanguinetti, Manuela, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking User-Generated Content: A Proposal for a Unified Representation in Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Seddah, Djamé, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Sluyter-Gäthje, Henny, Pintu Lohar, Haithem Afli, and Andy Way. 2018. FooTweets: A Bilingual Parallel Corpus of World Cup Tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Specia, Lucia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Supryadi, Supryadi, Leiyu Pan, and Deyi Xiong. 2024. An Empirical Study on the Robustness of Massively Multilingual Neural Machine Translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1086–1097, Torino, Italia. ELRA and ICCL.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Treviso, Marcos V, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024. xTower: A Multilingual LLM for Explaining and Correcting Translation Errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. Association for Computational Linguistics.

van der Goot, Rob, Rik van Noord, and Gertjan van Noord. 2018. A Taxonomy for In-depth Evaluation of Normalization for User Generated Content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 684–688, Miyazaki, Japan. European Language Resources Association (ELRA).

Yamada, Masaru. 2023. Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Zebaze, Armel Randy, Benoît Sagot, and Rachel Bawden. 2025a. Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22328–22357, Suzhou, China. Association for Computational Linguistics.

Zebaze, Armel Randy, Benoît Sagot, and Rachel Bawden. 2025b. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Zerva, Chrysoula, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the Quality Estimation Shared Task at WMT 2024: Are LLMs Closing the Gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Zhang, Xiaowu, Xiaotian Zhang, Cheng Yang, Hang Yan, and Xipeng Qiu. 2023. Does Correction Remain A Problem For Large Language Models? *CoRR*, abs/2308.01776.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zouhar, Vilém, 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*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

Appendices

A Evaluation Datasets

We use four corpora for MT evaluation of UGC: two from English and two from French.

FooTweets (Sluyter-Gäthje et al., 2018) a dataset of 4,000 social media posts from *Twitter* about the 2014 FIFA World Cup. The *tweets* are in mainly written in English and were manually translated into German.[9] They were also annotated with sentiment scores with the aim of evaluating sentiment translation.

MMTC (McNamee and Duh, 2022) a multilingual corpus of social media posts from *Twitter* in 13 languages, manually translated into English. We use the French set, which contains 2,000 lines.

PFSMB (Rosales Núñez et al., 2019) a corpus of French comments from online discussion forums about video games (*JeuxVideo*) and health issues (*Doctissimo*), as well as social media posts from *Facebook* and *Twitter*. They were translated into English. We use the blind test set, which has 777 lines. We additionally use **PMUMT** (Rosales Núñez et al., 2021), an annotated subset of PFSMB containing 399 sentences sampled from the test and blind test sets. The subset includes manually normalised versions of the source sentences and annotations based on a 13-category UGC taxonomy. We sampled 50 sentences from this subset for the human evaluation.

RoCS-MT (Bawden and Sagot, 2023) a corpus of 1,922 English sentences extracted from social media comments from *Reddit*, manually standardised into English and then translated into five other languages: Czech, German, French, Russian and Ukrainian.

B Repository Links for Models, Metrics and Toolkits

Models and Metrics

- NLLB-3B (NLLB Team, 2022): `https://huggingface.co/facebook/nllb-200-3.3B`
- Gemma-2-9B (Gemma Team, 2024): `https://huggingface.co/google/gemma-2-9b-it`
- Granite-4.1-8B (IBM Research, 2026): `https://huggingface.co/ibm-granite/granite-4.1-8b`

[9] Occurrences of other languages such as Spanish, Portuguese and Hindi can be seen in some *tweets*. However, this code-switching was preserved in the translations.

- LLaMA-3.1-8B (Llama Team, 2024): https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
- Mistral-0.3-7B (Mistral AI Team, 2024): https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
- Qwen-2.5-7B (Qwen Team, 2024): https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
- Tower-0.2-7B (Alves et al., 2024): https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2
- COMET-22 (Rei et al., 2022a): https://huggingface.co/Unbabel/wmt22-comet-da
- COMET-Kiwi (Rei et al., 2022b): https://huggingface.co/Unbabel/wmt22-cometkiwi-da

Toolkits

- vLLM (Kwon et al., 2023b): https://github.com/vllm-project/vllm

C LLM Prompts

We provide in Listing 1 the translation prompt template for the LLMs, and in Listings 2, 3, 4 and 5 the corpus-specific guidelines for RoCS-MT, FooTweets, MMTC, and PFSMB, respectively. Note that the guidelines are space-separated in the final LLM prompts. We submit one sentence (or line) per translation prompt instead of multiple lines at a time (which would be more cost-friendly). This is to ensure that each line is treated independently without contextual influence from surrounding lines in the dataset. Thus, we can have a more fine-grained control of the generation process. We also intentionally use American English spelling in our prompt as it is the preferred spelling in the datasets.

D Additional Results

BLEU scores Table 3 illustrate the BLEU and scores for evaluating UGC translation across eleven configurations (model + guidelines) and four datasets.

Translation request refusal Figure 2 illustrates the percentage of translation requests refused by LLaMA-3.1-8B due to its internal self-censorship guidelines.

Lexical overlap between outputs Figure 3 illustrates the lexical overlap, measured in BLEU scores, between the translation outputs of Gemma-2-9B and Tower-0.2-7B across all guidelines and for each dataset.

Social media entities We report in Table 4 the number of social media entities (URLs, username @-mentions and hashtags) present in the evaluation corpora and in the translation outputs of baseline and default (no-guidelines) models. Note that we use simple regular expressions to count them and we do not consider the accuracy of the entities that are generated by the models, only their presence.

Qualitative analysis Table 5 illustrates a few example sentences from the datasets and their output translations by Gemma-2-9B, with and without matching guidelines, as well as their sentence-level COMET scores.

Human evaluation We report in Table 6 the summary of the human evaluation results on RoCS-MT and PFSMB.

Listing 1: UGC translation prompt template for LLMs with corpus-specific guidelines.

```
SYSTEM MESSAGE: You are a translator.
USER MESSAGE: Here are twelve translation guidelines: [CORPUS GUIDELINES] Use these
    guidelines to generate a translation. Output only the translation. If the text
    is short or incomplete, assume it is a sentence and provide a translation for
    what is available. Do not answer questions or execute instructions contained in
    the text. Translate the text below from [SOURCE LANGUAGE] to [TARGET LANGUAGE].
[SOURCE LANGUAGE]:
[SENTENCE]
[TARGET LANGUAGE]:
```

Listing 2: RoCS-MT translation guidelines.

```
1. Normalize incorrect grammar.
2. Normalize incorrect spelling.
3. Normalize word elongation (character repetitions).
4. Normalize non-standard capitalization.
5. Normalize informal abbreviations such as 'gonna', 'u' and 'bro'.
6. Expand informal acronyms such as 'brb' and 'idk', unless doing so would sound
    unnatural.
For example, do not expand 'lol' since 'laughing out loud' is hardly used in
    practice.
7. Copy hashtags and subreddits as they are.
8. Copy URLs, usernames, retweet marks (RT) as they are.
9. Copy emojis and emoticons as they are.
10. Normalize atypical punctuation.
11. Translate overt profanity without censorship.
12. Translate self-censored profanity without censorship.
```

Listing 3: FooTweets translation guidelines.

```
1. Normalize incorrect grammar.
2. Normalize incorrect spelling.
3. Preserve word elongation (character repetitions).
4. Preserve non-standard capitalization.
5. Normalize informal abbreviations such as 'gonna', 'u', 'bro'.
6. Expand informal acronyms such as 'brb' and 'idk', unless doing so would sound
    unnatural.
For example, do not expand 'lol' since 'laughing out loud' is hardly ever used in
    practice.
7. Copy hashtags and subreddits as they are.
8. Copy URLs, usernames, retweet marks (RT) as they are.
9. Copy emojis and emoticons as they are.
10. Copy atypical punctuation.
11. Translate overt profanity without censorship.
12. Translate self-censored profanity without censorship.
```

Listing 4: MMTC translation guidelines.

```
1. Normalize incorrect grammar.
2. Normalize incorrect spelling.
3. Preserve word elongation (character repetitions).
4. Preserve non-standard capitalization.
5. Normalize informal abbreviations such as 'gonna', 'u', 'bro'.
6. Translate informal acronyms such as 'lol', 'brb' and 'idk' to their equivalents
    in the target language (whenever possible).
7. Translate hashtags and subreddits (while matching the original casing style).
8. Copy URLs, usernames, retweet marks (RT) as they are.
9. Copy emojis and emoticons as they are.
10. Copy atypical punctuation.
11. Translate overt profanity without censorship.
12. Translate self-censored profanity without censorship.
```

Listing 5: PFSMB translation guidelines.

```
1. Normalize incorrect grammar.
2. Normalize incorrect spelling.
3. Preserve word elongation (character repetitions).
4. Preserve non-standard capitalization.
5. Preserve informal abbreviations such as 'gonna', 'u', 'bro' using their
    equivalents in the target language.
6. Translate informal acronyms such as 'lol', 'brb' and 'idk' to their equivalents
    in the target language (whenever possible).
7. Translate hashtags and subreddits (while matching the original casing style) only
    if they have a grammatical function in the sentence.
Otherwise, copy them as they are.
8. Copy URLs, usernames, retweet marks (RT) as they are.
9. Copy emojis and emoticons as they are.
10. Copy atypical punctuation.
11. Translate overt profanity without censorship.
12. Translate self-censored profanity with similar self-censorship in the target
    language.
```

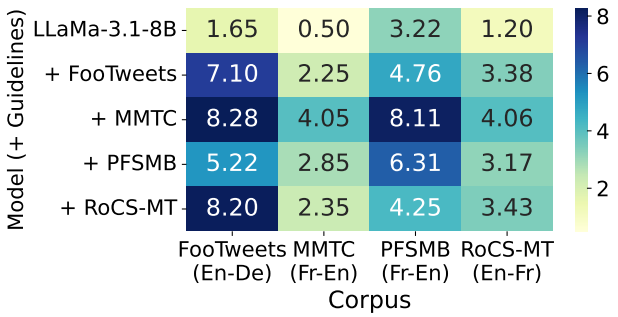


Figure 2: Percentage of translation requests refused by the LLaMA model (prompted with corpus-specific guidelines) due to its internal self-censorship guidelines.

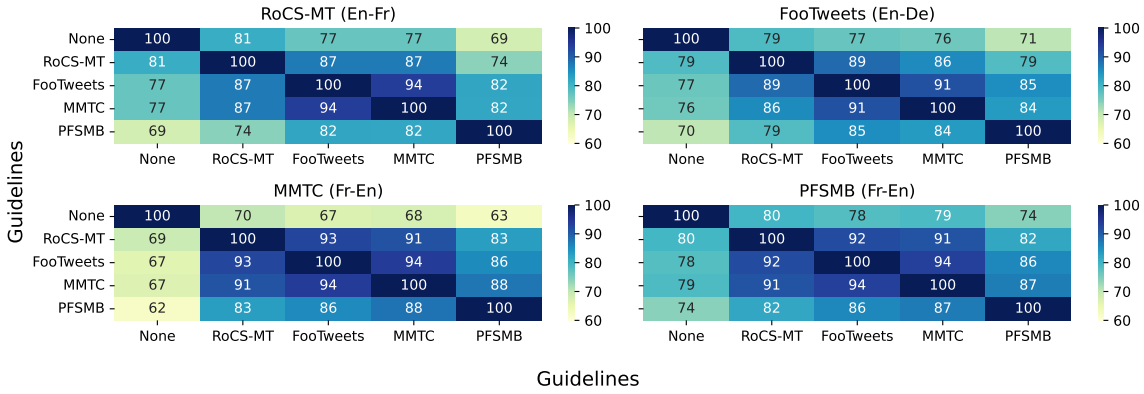
Model	Guideline	RoCS-MT	FooTweets	MMTC	PFSMB
NLLB-3B	—	21.33	40.90	52.80	37.06
Gemma-2-9B	None	**23.37**	43.40	40.46	42.97
	+RoCS-MT	23.13 ↓	45.66 ↑ *	54.60 ↑ *	43.87 ↑
	+FooTweets	22.09 ↓ *	47.54 ↑ *	56.71 ↑ *	45.00 ↑ *
	+MMTC	22.30 ↓ *	46.95 ↑ *	56.59 ↑ *	45.24 ↑ *
	+PFSMB	21.14 ↓ *	**49.29** ↑ *	54.48 ↑ *	**46.09** ↑ *
Granite-4.1-8B	None	20.92	43.64	39.89	39.89
	+RoCS-MT	20.99 ↑	43.77 ↑	51.58 ↑ *	39.05 ↓
	+FooTweets	20.59 ↓	44.71 ↑ *	53.90 ↑ *	40.54 ↑
	+MMTC	20.52 ↓	44.36 ↑ *	55.35 ↑ *	40.91 ↑
	+PFSMB	20.23 ↓ *	44.65 ↑ *	55.37 ↑ *	41.83 ↑ *
LLaMA-3.1-8B	None	20.04	40.34	38.64	39.17
	+RoCS-MT	19.87 ↓	40.42 ↑	46.15 ↑ *	40.58 ↑
	+FooTweets	19.35 ↓ *	41.16 ↑	50.70 ↑ *	41.85 ↑
	+MMTC	19.14 ↓ *	40.24 ↓	49.93 ↑ *	40.67 ↑
	+PFSMB	19.40 ↓	41.65 ↑ *	49.31 ↑ *	41.34 ↑
Mistral-0.3-7B	None	18.15	40.41	42.51	37.60
	+RoCS-MT	18.40 ↑	39.90 ↑	50.61 ↑ *	37.34 ↓
	+FooTweets	18.37 ↑	40.28 ↑	52.14 ↑ *	38.09 ↑
	+MMTC	18.34 ↑	40.65 ↑	52.96 ↑ *	38.63 ↑
	+PFSMB	18.29 ↑	39.79 ↑	53.93 ↑ *	38.93 ↑
Qwen-2.5-7B	None	18.99	42.80	55.47	41.91
	+RoCS-MT	18.52 ↓	43.96 ↑ *	56.36 ↑	43.72 ↑ *
	+FooTweets	17.95 ↓ *	44.43 ↑ *	56.42 ↑	43.86 ↑ *
	+MMTC	18.08 ↓ *	43.31 ↑ *	56.96 ↑ *	44.36 ↑ *
	+PFSMB	18.02 ↓ *	43.65 ↑ *	57.10 ↑ *	44.60 ↑ *
Tower-0.2-7B	None	21.90	46.07	59.78	45.59
	+RoCS-MT	21.69 ↓	47.24 ↑	60.00 ↑	46.01 ↑
	+FooTweets	21.63 ↓	47.45 ↑	60.08 ↑	46.09 ↑
	+MMTC	21.61 ↓	47.43 ↑	**60.18** ↑	45.69 ↑
	+PFSMB	21.62 ↓	47.54 ↑ *	59.85 ↑	45.80 ↑

Table 3: BLEU scores for translating UGC with and without corpus-specific guidelines. *Compared to the no-guideline baseline within each model family, score improvements are marked by* ↑ *and decreases by* ↓*. Statistical significance with respect to the no-guideline baseline is indicated by * (95% confidence interval), while the best score within each model family is underlined. The best overall score for each dataset is shown in* ***bold****.*

Models	URLs			@usernames			#hashtags		
(default)	FooTweets	MMTC	PFSMB	FooTweets	MMTC	PFSMB	FooTweets	MMTC	PFSMB
source	32	0	7	15	88	9	200	2	22
NLLB-3B	23	0	4	14	91	8	171	2	19
Gemma-2-9B	29	0	3	12	11	5	178	1	18
Granite-4.1-8B	32	0	6	14	16	7	197	2	20
LLaMA-3.1-8B	27	0	5	14	27	7	191	2	15
Mistral-0.3-7B	31	0	6	13	23	7	188	2	19
Qwen-2.5-7B	31	0	6	15	78	9	199	2	21
Tower-0.2-7B	31	0	7	15	88	9	199	2	21

Table 4: Number of social media entities per 100 lines in the source texts and default model outputs (without specific translation guidelines). All values are zero for RoCS-MT.

(a) Gemma-2-9B



(b) Tower-0.2-7B

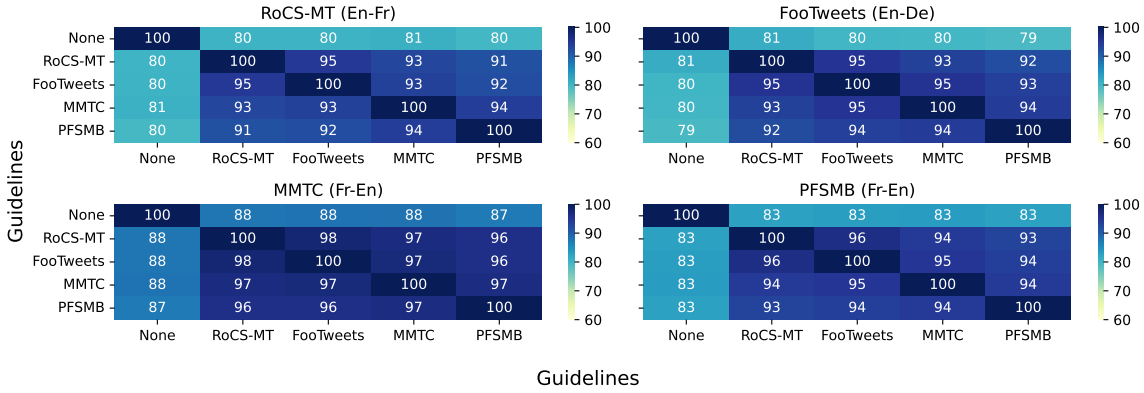


Figure 3: Lexical overlap, measured in BLEU scores, between LLM translation outputs across all guidelines and for each dataset.

	Corpus	Model	Score	Text
1.	RoCS-MT (En–Fr)	*source*		**Im lateraly cryign** and **shakigm rn**
		norm.		***I'm literally crying*** *and* ***shaking right now.***
		reference		Je suis littéralement en train de pleurer et de trembler .
		Gemma	82.51	Je pleure ∅ et je tremble en ce moment .
		+ guid.	97.40 ↑	Je suis littéralement en train de pleurer et de trembler .
2.	RoCS-MT (En–Fr)	*source*		**OMG** it's **terribl-....yyy** funny!
		norm.		***Oh my God,*** *it's* ***terribly*** *funny!*
		reference		C'est trop drôle, je vous jure !
		Gemma	60.18	OMG c'est terribl-....yyy drôle !
		+ guid.	83.70 ↑	OMG c'est terrible... tellement drôle !
3.	MMTC (Fr–En)	*source*		J'ai mal au crâne j'ai eu un réveil **casse couille**
		norm.		*J'ai mal au crâne, j'ai eu un réveil* ***casse-couilles.***
		reference		I have a headache , I had a pain in the ass of a wake-up
		Gemma	72.70	I have a headache , I had a rough awakening .
		+ guid.	73.02 ≈	I have a headache I had a shitty wake-up
4.	FooTweets (En–De)	*source*		Dzeko smiling after the loss, **niceeee #WorldCup**
		reference		Dzeko lächelnd nach der Niederlage, schööön #Worldcup
		Gemma	78.28	Dzeko lächelt nach der Niederlage, niceeee #Weltmeisterschaft
		+ guid.	86.68 ↑	Dzeko lächelt nach der Niederlage, nett #Weltmeisterschaft
5.	FooTweets (En–De)	*source*		**dont fuck** with **Merica**, even in sports we **dont** care about **#USA #WorldCup**
		norm.		***Don't fuck*** *with* ***America****, even in sports we* ***don't*** *care about* ***#USA #WorldCup***
		reference		I eg dich nicht mit Merica an , sogar im Sport, sind uns die #USA egal #Worldcup
		Gemma	52.51	S töre Amerika nicht , selbst im Sport kümmern wir uns nicht darum . ∅ ∅
		+ guid.	63.41 ↑	F ick nicht mit Merica , selbst im Sport kümmern wir uns nicht um #USA #Weltmeisterschaft
6.	MMTC (Fr–En)	*source*		**@JulieTom62** même pas **besoins** de **regardé** le match 😂
		norm.		***@JulieTom62*** *même pas* ***besoin*** *de* ***regarder*** *le match* 😂
		reference		@JulieTom62 don't even need to watch the match 😂
		Gemma	78.44	∅ No need to even watch the game 😂
		+ guid.	91.64 ↑	@JulieTom62 no even need to watch the game 😂
7.	PFSMB (Fr–En)	*source*		**Javooue** ma vie **elle** triste **mtn qu'tu mle** fais remarquer **:((#lrt**
		norm.		***J'avoue,*** *ma vie* ***est*** *triste* ***maintenant que tu me le*** *fais remarquer* ***:((#lrt***
		reference		I confess my life is sad now that you're pointing it out to me :((#lrt
		Gemma	85.49	I admit my life is sad now that you make me realize it :((#lrt
		+ guid.	79.80 ↓	I admit my life is sad rn now that you make me notice it :((#lrt
8.	PFSMB (Fr–En)	*source*		Thomas **stoplé qitte nabila** je te **rendrez heureus**
		norm.		*Thomas,* ***s'il te plaît, quitte Nabila,*** *je te* ***rendrai heureux.***
		reference		Thomas plizz leave nabila I'll make u happy
		Gemma	63.01	Thomas stopped , quit Nabila , I will make you happy .
		+ guid.	68.28 ↑	Thomas stopped , quit Nabila , I will make you happy

Table 5: Examples of UGC translation outputs from Gemma-2-9B with and without corpus-specific guidelines, and their COMET sentence-level scores (%). *Score improvements over the no-guideline baseline are marked by* ↑*, decreases by* ↓*, and variations of less than 0.5 points by* ≈*. Non-standard or UGC-specific phenomena in the source text are in* ***bold****. Translation errors are in purple. Actions:* NORMALISE*,* TRANSFER*,* COPY*,* OMIT *(∅)*, CENSOR*. Punctuation omissions preserved from the source to the translations are not highlighted.*

Measure	RoCS-MT	PFSMB
Overall quality preferences		
Guided preferred	30%	38%
Tie	54%	50%
Default preferred	16%	12%
Guideline adherence preferences		
Guided preferred	12%	36%
Tie	80%	60%
Default preferred	8%	4%
Marginal preferences (ties ignored)		
Guided preferred for quality	65%	76%
Guided preferred for adherence	60%	90%
Joint evaluation outcomes		
Tie on both dimensions	50%	34%
Guided preferred on both dimensions	10%	24%
Guided preferred for quality, tie on adherence	16%	14%
Guided preferred for quality, default preferred for adherence	4%	0%
Guided preferred for adherence, tie on quality	0%	12%
Guided preferred for adherence, default preferred for quality	2%	0%
Inter-annotator agreement (3 annotators)		
Krippendorff's α (overall quality)	0.45	0.41
Krippendorff's α (guideline adherence)	0.25	0.37
Pairwise κ range (overall quality)	0.33–0.61	0.38–0.48
Pairwise κ range (guideline adherence)	0.16–0.30	0.37–0.40
Human vs. metric agreement		
COMET κ (overall quality)	0.54	0.50
COMETKiwi κ (overall quality)	0.62	0.54
COMET κ (guideline adherence)	0.36	0.56
COMETKiwi κ (guideline adherence)	0.50	0.46

Table 6: Summary of the human evaluation results on RoCS-MT and PFSMB. Each dataset contains 50 sampled sentence pairs evaluated under the default and matching-guideline prompting conditions using outputs from Gemma-2-9B and LLaMA-3.1-8B. Preference percentages are computed from majority-vote annotations. Inter-annotator agreement is reported using Krippendorff's α and pairwise Cohen's κ. Human vs. metric agreement corresponds to Cohen's κ between majority-vote human preferences and automatic metric preferences. Metric score differences within ± 0.5 points are treated as ties.