GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4

Tom Kocmi and **Christian Federmann**
Microsoft, One Microsoft Way, Redmond, WA-98052, USA
{tomkocmi,chrife}@microsoft.com

Abstract

This paper introduces GEMBA-MQM, a GPT-based evaluation metric designed to detect translation quality errors, specifically for the quality estimation setting without the need for human reference translations. Based on the power of large language models (LLM), GEMBA-MQM employs a fixed three-shot prompting technique, querying the GPT-4 model to mark error quality spans. Compared to previous works, our method has language-agnostic prompts, thus avoiding the need for manual prompt preparation for new languages.

While preliminary results indicate that GEMBA-MQM achieves state-of-the-art accuracy for system ranking, we advise caution when using it in academic works to demonstrate improvements over other methods due to its dependence on the proprietary, black-box GPT model.

Metric	Acc.	Meta
GEMBA-MQM	96.5% (1)	0.802 (3)
XCOMET-Ensemble	95.2% (1)	0.825 (1)
docWMT22CometDA	93.7% (2)	0.768 (9)
docWMT22CometKiwiDA	93.7% (2)	0.767 (9)
XCOMET-QE-Ensemble	93.5% (2)	0.808 (2)
COMET	93.5% (2)	0.779 (6)
MetricX-23	93.4% (3)	0.808 (2)
CometKiwi	93.2% (3)	0.782 (5)
Calibri-COMET22	93.1% (3)	0.767 (10)
BLEURT-20	93.0% (4)	0.776 (7)
MaTESe	92.8% (4)	0.782 (5)
mre-score-labse-regular	92.7% (4)	0.743 (13)
mbr-bleurtxv1p-qe	92.5% (4)	0.788 (4)
KG-BERTScore	92.5% (5)	0.774 (7)
MetricX-23-QE	92.0% (5)	0.800 (3)
BERTscore	90.2% (7)	0.742 (13)
MS-COMET-QE-22	90.1% (8)	0.744 (12)
embed_llama	87.3% (10)	0.701 (16)
f200spBLEU	86.8% (11)	0.704 (15)
BLEU	85.9% (12)	0.696 (16)
chrF	85.2% (12)	0.694 (17)

Table 1: Preliminary results of the WMT 2023 Metric Shared task. The first column shows the system-level accuracy, and the second column is the Metrics 2023 meta evaluation. Metrics with gray background need human references. The table does not contain the worst-performing, non-standard metrics due to space reasons.

1 Introduction

GEMBA-MQM builds on the recent finding that large language models (LLMs) can be prompted to assess the quality of machine translation (Kocmi and Federmann, 2023a). We release the scoring script.[1]

The earlier work Kocmi and Federmann (2023a) (GEMBA-DA) adopted a straightforward methodology of assessing single score values for each segment without specifying the scale in detail. Employing a zero-shot approach, their technique showed an unparalleled accuracy in assessment, surpassing all other non-LLM metrics on the WMT22 metrics test set (Freitag et al., 2022).

Next, Lu et al. (2023) (EAPrompt) investigated prompting LLMs to assess individual error classes from a multidimensional quality metrics (MQM) framework (Freitag et al., 2021), where each error can be classified into various error classes (such as accuracy, fluency, style, terminology, etc.), subclasses (accuracy > mistranslation), and is marked with its severity (critical, major, minor). Segment scores are computed by aggregating errors, each weighted by its respective severity coefficient (25, 5, 1). While their approach employed a few-shot prompting with a chain-of-thought strategy (Wei et al., 2022), our GEMBA-MQM approach differs in two aspects: 1) We streamline the process using only single-step prompting, and 2) our prompts are universally applicable across languages, avoiding the need for manual prompt preparation for each language pair.

Another notable effort by Fernandes et al. (2023) paralleled the EAPrompt approach, also marking MQM error spans. In contrast, their approach used a PaLM-2 model, pooling MQM annotations to sample a few shot examples for the prompt. Their

[1] https://github.com/MicrosoftTranslator/GEMBA/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 768–775
December 6–7, 2023.

```
(System) You are an annotator for the quality of machine translation. Your task is to identify
errors and assess the quality of the translation.

(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify
error types in the translation and classify them. The categories of errors are: accuracy
(addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar,
inconsistency, punctuation, register, spelling),
 locale convention (currency, date, name, telephone, or time format)
style (awkward), terminology (inappropriate  for context, inconsistent use), non-translation,
other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what
the text is trying to say is still understandable. Minor errors are technically errors,
but do not disrupt the flow or hinder comprehension.

(assistant) {observed error classes}
```

Figure 1: The general prompt for GEMBA-MQM omits the gray part which performed subpar on internal data (we include it in GEMBA-locale-MQM). The "(user)" and "(assistant)" section is repeated for each few-shot example.

fine-tuning experiments did not improve system-level performance for the top-tier models.

2 Description

Our technique adopts few-shot learning with the GPT-4 model (OpenAI, 2023), prompting the model to mark quality error spans using the MQM framework. The underlying prompt template is modeled on guidelines for human annotators and shown in Figure 1.

In contrast to other methods, we use three predetermined examples (see Appendix A), allowing the method to be used with any language pair, avoiding the need to create language pair specific MQM few-shot examples. This was the original limitation that prevented Fernandes et al. (2023) from evaluating AutoMQM beyond two language pairs. Our decision was not driven by a desire to enhance performance — since domain and language-specific prompts typically boost it (Moslem et al., 2023) — but rather to ensure our method can be evaluated across any language pairs.

3 Experiments

To measure the performance of the GEMBA-MQM metric, we follow the methodology and use test data provided by the WMT22 Metrics shared task (Freitag et al., 2022) which hosts an annual evaluation of automatic metrics, benchmarking them against human gold labels.

We compare our method against the best-performing reference-based metrics of WMT22: MetrixX_XXL (non-public metric), COMET-22 (Rei et al., 2022), UNITE (Wan et al., 2022b), BLEURT-20 (Pu et al., 2021), and COMET-20 (Rei et al., 2020). In addition, we also compare against "classic" string-based metrics BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). Lastly, we compare against reference-less metrics of WMT22: CometKIWI (Rei et al., 2022), Unite-src (Wan et al., 2022a), Comet-QE (Rei et al., 2021), MS-COMET-QE-22 (Kocmi et al., 2022b).

We contrast our work with other LLM-based evaluation methods such as GEMBA-DA (Kocmi and Federmann, 2023b) and EAPrompt (Lu et al., 2023), conducting experiments using two GPT models: GPT-3.5-Turbo and the more powerful GPT-4 (OpenAI, 2023).

3.1 Test set

The main evaluation of our work has been done on the MQM22 (Freitag et al., 2022) and internal Microsoft data. Furthermore, a few days before the camera-ready deadline, organizers of Metrics 2023 (Freitag et al., 2023) released results on the blind test set, showing performance on unseen data.

The MQM22 test set contains human judgments for three translation directions: English into German, English into Russian, and Chinese into English. The test set contains a total of 54 machine translation system outputs or human translations. It

contains a total of 106k segments. Translation systems are mainly from participants of the WMT22 General MT shared task (Kocmi et al., 2022a). The source segments and human reference translations for each language pair contain around 2,000 sentences from four different text domains: news, social, conversational, and e-commerce. The gold standard for scoring translation quality is based on human MQM ratings, annotated by professionals who mark individual errors in each translation, as described in Freitag et al. (2021).

The MQM23 test set is the blind set for this year's WMT Metrics shared task prepared in the same way as MQM22, but with unseen data for all participants, making it the most reliable evaluation as neither participants nor LLM could overfit to those data. The main difference from last year's iteration is the replacement of English into Russian with Hebrew into English. Also, some domains have been updated; see Kocmi et al. (2023).

Additionally, we evaluated GEMBA-MQM on a large internal test set, an extended version of the data set described by Kocmi et al. (2021). This test set contains human scores collected with source-based Direct Assessment (DA, Graham et al., 2013) and its variant DA+SQM (Kocmi et al., 2022a). This test set contains 15 high-resource languages paired with English. Specifically, these are: Arabic, Czech, Dutch, French, German, Hindi, Italian, Japanese, Korean, Polish, Portuguese, Russian, Simplified Chinese, Spanish, and Turkish.

3.2 Evaluation methods

The main use case of automatic metrics is system ranking, either when comparing a baseline to a new model, when claiming state-of-the-art results, when comparing different model architectures in ablation studies, or when deciding if to deploy a new model to production. Therefore, we focus on a method that specifically measures this target: system-level pairwise accuracy (Kocmi et al., 2021).

The pairwise accuracy is defined as the number of system pairs ranked correctly by the metric with respect to the human ranking divided by the total number of system pair comparisons.

Formally:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) == \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

We reproduced all scores reported in the WMT22 Metrics shared task findings paper using the official WMT22 script.[2] Reported scores match Table 11 of the WMT22 metrics findings paper (Freitag et al., 2022).

Furthermore, organizers of Metrics shared task 2023 defined a new meta-evaluation metric based on four different scenarios, each contributing to the final score with a weight of 0.25:

- system-level pairwise accuracy;
- system-level Pearson correlation;
- segment-level Accuracy-t (Deutsch et al., 2023); and
- segment-level Pearson correlation.

The motivation is to measure metrics in the most general usage scenarios (for example, for segment-level filtering) and not just for system ranking. However, we question the decision behind the use of Pearson correlation, especially on the system level. As Mathur et al. (2020) showed, Pearson used for metric evaluation is sensitive when applied to small sample sizes (in MQM23, the sample size is as little as 12 systems); it is heavily affected by outliers (Osborne and Overbay, 2004; Ma et al., 2019), which need to be removed before running the evaluation; and it measures linear correlation with the gold MQM data, which are not necessarily linear to start with (especially the discrete segment-level scores, with error weights of 0.1, 1, 5, 25).

Although it is desirable to have an automatic metric that correlates highly with human annotation behaviour and which is useful for segment-level evaluation, more research is needed regarding the proper way of testing these properties.

4 Results

In this section, we discuss the results observed on three different test sets: 1) MQM test data from WMT, 2) internal test data from Microsoft, and 3) a subset of the internal test data to measure the impact of the MQM locale convention.

4.1 Results on MQM Test Data from WMT

The results of the blind set MQM23 in Table 1 show that GEMBA-MQM outperforms all other techniques on the three languages evaluated in the system ranking scenario. Furthermore, when evaluated in the meta-evaluation scenario it achieves the third cluster rank.

In addition to the official results, we also test on MQM22 test data and show results in Table 2. The

[2] https://github.com/google-research/mt-metrics-eval

Metric	Acc.
EAPrompt-Turbo	90.9%
GEMBA-DA-GPT4	89.8%
GEMBA-locale-MQM-Turbo	89.8%
EAPrompt-Turbo	89.4%
GEMBA-MQM-GPT4	89.4%
GEMBA-DA-GPT4	87.6%
GEMBA-DA-Turbo	86.9%
GEMBA-MQM-Turbo	86.5%
GEMBA-DA-Turbo	86.5%
MetricX_XXL	85.0%
BLEURT-20	84.7%
COMET-22	83.9%
COMET-20	83.6%
UniTE	82.8%
COMETKiwi	78.8%
COMET-QE	78.1%
BERTScore	77.4%
UniTE-src	75.9%
MS-COMET-QE-22	75.5%
chrF	73.4%
BLEU	70.8%

Table 2: The system-level pairwise accuracy results for the WMT 22 metrics task test set. Gray metrics need reference translations which are not the focus of the current evaluation.

# of system pairs (N)	15 langs 4,468	Cs + De 734
COMETKiwi	79.9	81.3
GEMBA-locale-MQM-Turbo	78.6	81.3
GEMBA-MQM-Turbo	78.4	83.0
COMET-QE	77.8	79.8
COMET-22	76.5	79.2
COMET-20	76.3	79.6
BLEURT-20	75.8	79.7
chrF	68.1	70.6
BLEU	66.8	68.9

Table 3: System-level pairwise accuracy results for our internal test set. The first column is for all 15 languages, and the second is Czech and German only. All languages are paired with English.

Source	Vstupné do památky činí 16,50 Eur.
Hypothesis	Admission to the monument is 16.50 Euros.
GPT annot.	locale convention/currency: "euros"

Table 4: An example of a wrong error class "locale convention" as marked by GEMBA-locale-MQM. The translation is correct, however, we assume that the GPT model might not have liked the use of Euros in a Czech text because Euros are not used in the Czech Republic.

main conclusion is that all GEMBA-MQM variants outperform traditional metrics (such as COMET or Metric XXL). When focusing on the quality estimation task, we can see that the GEMBA-locale-MQM-Turbo method slightly outperforms EAPrompt, which is the closest similar technique.

However, we can see that our final technique GEMBA-MQM is performing significantly worse than the GEMBA-locale-MQM metric, while the only difference is the removal of the locale convention error class. We believe this to be caused by the test set. We discuss our decision to remove the locale convention error class in Section 4.3.

4.2 Results on Internal Test Data

Table 3 shows that GEMBA-MQM-Turbo outperforms almost all other metrics, losing only to COMETKIWI-22. This shows some limitations of GPT-based evaluation on blind test sets. Due to access limitations, we do not have results for GPT-4, which we assume should outperform the GPT-3.5 Turbo model. We leave this experiment for future work.

4.3 Removal of Locale Convention

When investigating the performance of GEMBA-locale-MQM on a subset of internal data (Czech and German), we observed a critical error in this prompt regarding the "locale convention" error class. GPT assigned this class for errors not related to translations. It flagged Czech sentences as a locale convention error when the currency Euro was mentioned, even when the translation was fine, see example in Table 4. We assume that it was using this error class to mark parts not standard for a given language but more investigation would be needed to draw any deeper conclusions.

The evaluation on internal test data in Table 4 showed gains of 1.7% accuracy. However, when evaluating over 15 languages, we observed a small degradation of 0.2%. For MQM22 in Table 2, the degradation is even bigger.

When we look at the distribution of the error classes over the fifteen highest resource languages in Table 5, we observe that 32% of all errors for GEMBA-locale-MQM are marked as a locale convention suggesting a misuse of GPT for this error class. Therefore, instead of explaining this class in the prompt, we removed it. This resulted in about half of the original locale errors being reassigned to other error classes, while the other half was not marked.

In conclusion, we decided to remove this class as it is not aligned with what we expected to measure and how GPT appears to be using the classes. Thus, we force GPT to classify those errors using other error categories. Given the different behaviour for internal and external test data, this deserves more

Error class	GEMBA-locale-MQM	GEMBA-MQM
accuracy	960,838 (39%)	1,072,515 (51%)
locale con.	808,702 (32%)	(0%)
fluency	674,228 (27%)	699,037 (33%)
style	23,943 (1%)	41,188 (2%)
terminology	17,379 (1%)	290,490 (14%)
Other errors	4,126 (0%)	10615 (1%)
Total	2,489,216	2,113,845

Table 5: Distribution of errors for both types of prompts over all segments of the internal test set for the Turbo model.

investigation in future work.

5 Caution with "Black Box" LLMs

Although GEMBA-MQM is the state-of-the-art technique for system ranking, we would like to discuss in this section the inherent limitations of using "black box" LLMs (such as GPT-4) when conducting academic research.

Firstly, we would like to point out that GPT-4 is a proprietary model, which leads to several problems. One of them is that we do not know which training data it was trained on, therefore any published test data should be considered as part of their training data (and is, therefore, possibly tainted). Secondly, we cannot guarantee that the model will be available in the future, or that it won't be updated in the future, meaning any results from such a model are relevant only for the specific sampling time. As Chen et al. (2023) showed, the model's performance fluctuated and decreased over the span of 2023.

As this impacts all proprietary LLMs, we advocate for increased research using publicly available models, like LLama 2 (Touvron et al., 2023). This approach ensures future findings can be compared both to "black box" LLMs while also allowing comparison to "open" models.[3]

6 Conclusion

In this paper, we have introduced and evaluated the GEMBA-MQM metric, a GPT-based metric for translation quality error marking. This technique takes advantage of the GPT-4 model with a fixed three-shot prompting strategy. Preliminary results show that GEMBA-MQM achieves a new state of the art when used as a metric for system ranking, outperforming established metrics such as COMET and BLEURT-20.

We would like to acknowledge the inherent limitations tied to using a proprietary model like GPT. Our recommendation to the academic community is to be cautious with employing GEMBA-MQM on top of GPT models. For future research, we want to explore how our approach performs with other, more open LLMs such as LLama 2 (Touvron et al., 2023). Confirming superior behaviour on publicly distributed models (at least their binaries) could open the path for broader usage of the technique in the academic environment.

Limitations

While our findings and techniques with GEMBA-MQM bring promising advancements in translation quality error marking, it is essential to highlight the limitations encountered in this study.

- Reliance on Proprietary GPT Models: GEMBA-MQM depends on the GPT-4 model, which remains proprietary in nature. We do not know what data the model was trained on or if the same model is still deployed and therefore the results are comparable. As Chen et al. (2023) showed, the model's performance fluctuated throughout 2023;

- High-Resource Languages Only: As WMT evaluations primarily focus on high-resource languages, we cannot conclude if the method will perform well on low-resource languages.

Acknowledgements

We are grateful to our anonymous reviewers for their insightful comments and patience that have helped improve the paper. We would like to thank our colleagues on the Microsoft Translator research team for their valuable feedback.

References

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation. *arXiv preprint arXiv:2305.14324*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig,

[3]Although LLama 2 is not fully open, its binary files have been released. Thus, when used it as a scorer, we are using the exact same model.

Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Markus Freitag, 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.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of wmt23 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. 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, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022a. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

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*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

OpenAI. 2023. Gpt-4 technical report.

Jason W Osborne and Amy Overbay. 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1):6.

Kishore Papineni, 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.

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

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. 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. 2022. 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.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, 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.

Hugo Touvron, 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.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022a. Alibaba-translate China's submission for WMT2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 586–592, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022b. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

A Three examples Used for Few-shot Prompting

```
English source: I do apologise about this, we must gain permission from the account holder to discuss
an order with another person, I apologise if this was done previously, however, I would not be able
to discuss this with yourself without the account holders permission.
German translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung
mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber
ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.
MQM annotations:
Critical:
no-error
Major:
accuracy/mistranslation - "involvement"
accuracy/omission - "the account holder"
Minor:
fluency/grammar - "wäre"
fluency/register - "dir"

English source: Talks have resumed in Vienna to try to revive the nuclear pact, with both sides
trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.
Czech transation: Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemže obě
partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.
MQM annotations:
Critical:
no-error
Major:
accuracy/addition - "ve Vídni"
accuracy/omission - "the stop-start"
Minor:
terminology/inappropriate for context - "partaje"

Chinese source: 大众点评乌鲁木齐家居商场频道为您提供高铁居然之家地址，电话，营业时间等最新商户信息，
找装修公司，就上大众点评
English translation: Urumqi Home Furnishing Store Channel provides you with the latest business
information such as the address, telephone number, business hours, etc., of high-speed rail, and
find a decoration company, and go to the reviews.
MQM annotations:
Critical:
accuracy/addition - "of high-speed rail"
Major:
accuracy/mistranslation - "go to the reviews"
Minor:
style/awkward - "etc.,"
```

Figure 2: Three examples used for all languages.