Towards Multiple References Era – Addressing Data Leakage and Limited Reference Diversity in Machine Translation Evaluation

Xianfeng Zeng, Yijin Liu, Fandong Meng* and Jie Zhou
Pattern Recognition Center, WeChat AI, Tencent Inc, China
{xianfzeng, yijinliu, fandongmeng, withtomzhou}@tencent.com

Abstract

Recent research has shown a weak correlation between n-gram-based metrics and human evaluations in machine translation task, particularly when evaluating large language models (LLMs). Additionally, the data leakage risk in LLMs may cause an overestimation problem when evaluating LLMs on downstream tasks. In this work, we identify the limited diversity of references as the primary cause for the inferior performance of n-gram-based metrics and the overestimation problem. To address this issue, we propose to utilize multiple references generated by LLMs, coupled with an effective selection strategy focused on accuracy and diversity, to improve the alignment between automatic metrics and human evaluations. We validate our approach on the WMT22 Metrics benchmark with 4 languages and observe a maximum accuracy gain of 9.5% in F200spBLEU, which makes it on par with computationally expensive neural-based metrics. We also show that using multi-reference with n-gram-based metrics significantly alleviates the overestimation problem when evaluating LLMs with data leakage. Further analysis explores the factors that affect the quality of generated references, offering insights into data synthesis by LLMs. We released the code and generated reference[1] to encourage research on reference quality.

1 Introduction

Due to the inherent diversity and complexity of natural language, human evaluation serves as the gold standard for assessing the quality of machine translation. However, conducting human evaluation is time-consuming and prohibitively expensive in real-world scenarios. Consequently, developing reliable automatic evaluation metrics is crucial for advancing machine translation research and optimizing machine translation systems (Celikyilmaz et al., 2020). The ideal automatic evaluation metrics need to assess the accuracy, fluency, fidelity, and diversity of the generated candidates by the model. It also requires high consistency with human assessment to demonstrate its reliability.

Automatic evaluation metrics for machine translation can be generally classified into two categories: n-gram-based metrics and neural-based metrics. N-gram-based metrics, *e.g.,* BLEU (Papineni et al., 2002) primarily calculate the lexical overlap between the model's outputs and the ground truth. On the other hand, neural-based metrics *e.g.,* BLEURT (Sellam et al., 2020) are trained on a huge amount of text data and score the candidates with a black-box model. Due to the huge amount of training data and parameters, neural-based metrics generally achieve better robustness and generalization. Recent studies (Freitag et al., 2022; Kocmi et al., 2022a) have demonstrated that neural-based metrics exhibit better agreement with human evaluation when compared with n-gram-based metrics.

The emergence of LLMs has further deepened these concerns (Zhao et al., 2023; Hendy et al., 2023) due to the diversity of output results from LLMs. Recent studies (Freitag et al., 2022) have highlighted the challenges in using n-gram-based metrics, such as BLEU, to evaluate LLM-generated hypotheses. Some latest reports (Anil et al., 2023) exclusively rely on neural-based metrics to evaluate the generation quality. Another challenge in evaluating LLMs is the data leakage risk due to their reliance on huge corpora obtained from the web. Recent research (Wei et al., 2023; Zhu et al., 2023) has emphasized that data leakage in LLMs is a prevalent and pressing issue, which underlines the importance of conducting more in-depth research to develop effective methods for addressing this challenge.

In this paper, we examine the performance bottleneck of n-gram-based metrics and identify the

* Corresponding author.
[1] https://github.com/SefaZeng/LLM-Ref

Findings of the Association for Computational Linguistics ACL 2024, pages 11939–11951
August 11-16, 2024

limited diversity of references as the primary cause for the inferior performance of n-gram-based metrics and the overestimation problem. To address this issue, we propose LLM-Ref, a framework that utilizes LLMs to generate multiple references, coupled with an effective selection strategy focused on diversity, to improve the alignment between automatic metrics and human evaluations. Experimental results show that our framework improves the consistency between human evaluation and both kinds of automatic evaluation metrics. Further analysis reveals that multiple references with n-gram-based metrics can effectively mitigate the potential data leakage risk of LLMs while neural-based metrics still face difficulty in overcoming this issue. To our knowledge, this is the first attempt to address the data leakage risk issue in evaluating LLMs.

2 Preliminary Experiments

To verify BLEU's evaluation capability on LLM-generated results and the data leakage issue of LLMs, we conduct preliminary experiments using BLOOMZ-7b on the FLORES-200 test set (Costa-jussà et al., 2022). Previous work (Zhu et al., 2023) has identified data leakage issues with BLOOMZ-7b on FLORES-200 across multiple translation directions.

We conduct supervised fine-tuning (SFT) on the BLOOMz-7b (Scao et al., 2022) with bilingual corpus in Chinese and Japanese from the CCMatrix dataset (Schwenk et al., 2021). To provide a comprehensive comparison, we also evaluate the performance of closed-source large models such as GPT-3.5 and Claude. The baseline machine translation (MT) model is an in-house model with the conventional encoder-decoder architecture.

The evaluation metrics utilized in our experiments include F200spBLEU (Costa-jussà et al., 2022), BLEURT (Sellam et al., 2020), and human evaluation (Human Eval). F200spBLEU is the BLEU (Papineni et al., 2002) score with a multilingual tokenizer trained in 200+ languages. Human Eval involves human experts rating the given translations on a scale ranging from 0 to 5, based on how accurately it reflects the meaning of the original sentence. We acquire 3 scores per sentence from human experts and take the average as the final sentence score. The summarized results are presented in Figure 1. We observe two phenomena from the above experiments:

- F200spBLEU significantly underestimates LLMs like Claude and GPT-3.5-turbo, exhibiting a disparity compared to human evaluations.

- Due to data leakage issues (Zhu et al., 2023), F200spBLEU overestimates BLOOMZ-7b and BLOOMZ-7b-ft, and BLEURT overestimates BLOOMZ-7b-ft, contrasting with lower human evaluation scores.

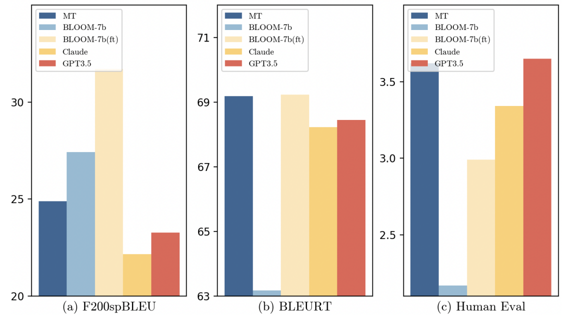

Figure 1: F200spBLEU (a), BLEURT (b), and Human Eval (c) scores of Japanese to Chinese translation task on Flores-200 test set. F200spBLEU underestimates LLMs' results like Claude and GPT-3.5-turbo. Both F200spBLEU and BLEURT overestimate BLOOMz-7b and BLOOMz-7b-ft attributed to the data leakage issue.

ID	Model	DistinctN (n=6) ↑	Unique Token ↑
0	AISP-SJTU	0.7399	7418
1	HuaweiTSC	0.7423	7210
2	M2M100_1.2B-B4	0.7137	7023
3	Online-G	0.7532	7467
4	Ref-A (Ground Truth)	0.7520	7244
5	GPT-3.5-turbo	**0.7857**	**8222**

Table 1: Diversity of model outputs measured by DistinctN and Unique Token number.

The Inadequacy of N-Gram-Based Metrics for Evaluating LLMs. We conjecture that the output distribution of the LLMs differs significantly from that of the MT model, and the n-gram-based nature of BLEU is constrained by the provided reference, resulting in degraded performance when there is a substantial divergence between the output and the reference.

To gain further evidence of the underlying factors contributing to the bias in BLEU's evaluation, we conduct experiments on the Chinese to English dataset from WMT22 (Freitag et al., 2022) to examine the output distribution of different models. We primarily use DistinctN (Li et al., 2015) and the count of unique tokens to analyze the output distribution. We select system outputs and ground truth translations from the task for our analysis. The results are presented in Table 1.

We discover that both DistinctN scores and the count of unique tokens of GPT-3.5-turbo are higher than all MT models, commercial systems, and human-translated references. These findings indicate that the outputs of LLMs exhibit a high degree of diversity, which renders it challenging for n-gram-based metrics to effectively estimate the quality of LLM generations.

LLMs' Data Leakage Risk. The open-source LLMs have gradually become foundational models for most NLP tasks (Touvron et al., 2023; Scao et al., 2022; Zeng et al., 2023b,a). However, due to their reliance on huge corpora obtained from the web, data leakage is an inevitable problem, which may potentially lead to an overestimation of the performance on downstream tasks. Recent study (Wei et al., 2023; Zhu et al., 2023) about LLMs have confirmed that data leakage in some widely used benchmarks is a common issue.

Our hypothesis posits that utilizing multiple references can alleviate overestimation due to data leakage to a certain degree. The tendency of overestimation comes from the model generating almost identical characters as the ground truth, thereby resulting in the performance overestimated by the golden reference. With multiple references provided, the model output close to the golden reference tends to obtain suboptimal performance on other references, which can alleviate the overestimation issue. We will discuss it in detail in Sec 5.4.

3 Methodology

We employ a series of steps to obtain high-quality reference candidates for a more accurate evaluation of generation performance, as outlined below.

3.1 LLMs Reference Candidates Generation

LLMs demonstrate strong capabilities in a variety of natural language processing tasks. We plan to leverage LLMs' powerful generative capabilities to generate diverse reference candidates which further enhance the NLG evaluations. The prompt for invoking LLMs consists of three components.

Rules. The rules mainly contain some characterization (e.g., You are a professional translator...), and rules to follow. These rules aim to facilitate LLMs generating higher quality and more human-preferred reference candidates.

Input. The input mainly consists of a task-specific description (e.g., Please provide 10 high-quality translations of the following...) and the source sentence of the test set.

Ground Truth. Providing manually annotated ground truth enables LLMs to generate more accurate reference candidates and align with human preference.

Observations from the experiments indicate that there are multiple factors (e.g., providing ground truth or not) in the prompt that affect the generation quality. It is worth noting that the quantity of references generated in a single inference by LLMs is the most influential factor. We will delve into a detailed analysis of the impact of prompt variants in Sec 5.2. We list the prompt in Appendix B.

3.2 Quality-Aware Selection

As the number of generated reference candidates increases, the LLMs tend to produce outputs with limited diversity (few word substitutions) and lower accuracy (deviating from the original meaning). These two phenomena can be attributed to the fact that there is a finite number of suitable responses and the hallucination problem of LLMs. Therefore, it is crucial to employ effective strategies for selection from the generated reference candidates.

Diversity Selection. We utilize Self-BLEU (Zhu et al., 2018; Zeng et al., 2021; Meng et al., 2020) as a metric of diversity for the generated reference candidates. Self-BLEU is the BLEU score computed against other reference candidates, which is computed as:

$$\text{Self-BLEU}_i = \text{BLEU}(y_i, \text{Y}) \tag{1}$$

where y_i is the i-th reference candidate, Y is the array of all y except y_i, and each reference will get a score Self-BLEU$_i$.

To detect the anomalous candidates with too large Self-BLEU (limited diversity), we utilize the Inter Quartile Range (IQR) method to decide the threshold for selecting reference candidates, which is a classical method for Outlier Detection. It is computed as:

$$IQR = Q_3 - Q_1 \tag{2}$$

$$upper\_bound = Q3 + 0.5 * IQR \tag{3}$$

where Q_1 is the first quartile and Q_3 is the third quartile for the Self-BLEU array. We filter out references with a score larger than $upper\_bound$, as they have low diversity.

Accuracy Selection. We employ COMET-22 (Rei et al., 2022) as the metric for assessing the accuracy of references, due to its relatively fast computation speed and high accuracy. Subsequently, we utilize the same Interquartile Range (IQR) method for selection.

$$lower\_bound = Q1 - 1.5 * IQR \qquad (4)$$

where Q_1 is the first quartile of references' COMET-22 score. We filter out references with a score lower than $lower\_bound$, as they have low accuracy. We conduct experiments to validate their impact and discuss in Sec 5.6.

3.3 Multi-Reference for Automatic Metrics

N-gram-based metrics (e.g. BLEU) are originally designed to use multiple references so we use the original implementation. In contrast, neural-based metrics assess the model's performance by scoring individual model output in the test set with references and subsequently averaging these scores through the whole test set to get the system-level score for the model. As most neural-based metrics only calculate model output with one reference, we have explored several simple methods(average, top-k, etc.) to combine the scores with multiple references. Empirical results indicate that selecting the maximum value is the best way. Consequently, all subsequent experiments with neural-based metrics are based on the max method, as specified by the following calculation:

$$\text{score} = max([\text{M}(\hat{y}, y_0), \text{M}(\hat{y}, y_i), ...]) \qquad (5)$$

Here $\hat{y}$ is the output from the model, y_i represents the i-th reference candidate generated by LLMs and M is the metric. Every single sentence in the test set gets a score.

4 Experiments

In this section, we describe the benchmarks and the evaluation we used for our experiments.

4.1 Datasets and Evaluation

To evaluate the performance of our proposed framework, we conducted experiments on WMT22 Metrics shared task (Freitag et al., 2022). This task includes human judgments for three translation directions: English to German (EN-DE), English to Russian (EN-RU), and Chinese to English (ZH-EN). The test set for each direction contains around 2000 sentences covering several text domains. The evaluation criteria are based on MQM[2] datasets annotated by domain experts (Freitag et al., 2021).

To evaluate the correlation between automatic metrics and human evaluation, we measured system-level pairwise accuracy, Pearson correlation (ρ), and Kendall-Tau correlation (τ), following the methodology established by the WMT22 Metrics shared task (Kocmi et al., 2021).

We reproduced reported scores in the WMT22 Metrics shared task findings (Freitag et al., 2022) using the official WMT22 script.[3]

4.2 Baseline Metrics

As baseline metrics, we mainly focus on two types of metrics including n-gram-based metrics and neural-based metrics.

N-gram-based metrics. We use the following n-gram-based metrics as baselines, including BLEU (Papineni et al., 2002), F200spBLEU (Costa-jussà et al., 2022), chrF (Popović, 2015). We use the sacrebleu (Post, 2018)[4] toolkit to perform the calculations for the first three metrics.

Neural-based metrics. For the WMT22 Metric Shared Task, we list the results of most neural-based metrics in the task, including BERTscore (Zhang* et al., 2020), MATEESE-QE (Perrella et al., 2022), MS-COMET-QE-22 (Kocmi et al., 2022b), UNITE-src (Wan et al., 2022), COMET-QE (Rei et al., 2022), COMETKiwi (Rei et al., 2022), YiSi-1 (Lo, 2019), MATESE (Perrella et al., 2022), MS-COMET-22 (Kocmi et al., 2022b), UniTE (Wan et al., 2022). We mainly conduct experiments on two most widely used and powerful metrics BLEURT (Sellam et al., 2020) and COMET-20[5] (Rei et al., 2020). For BLEURT[6] and COMET[7], we use the official github example scripts.

We also report the result of GEMBA-GPT4-DA (Kocmi and Federmann, 2023) which uses GPT-4 to directly assess the model outputs.

[2]MQM datasets are composed of multi-dimensional quality scoring by expert translators. Crowdsourced DA datasets are direct assessment scores from crowdsourcing staff.

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

[4]https://github.com/mjpost/sacrebleu

[5]We did not use COMET-22 as its ensemble nature and lack of open-source models.

[6]https://github.com/google-research/bleurt

[7]https://github.com/Unbabel/COMET

ID	Metrics	Accuracy	en-de ρ	en-ru ρ	zh-en ρ
0	GEMBA-GPT4-DA	89.8%	0.899	0.868	0.986
1	**BLEURT-20-LLM-Ref-QAS**	86.4%	0.726	0.962	0.938
2	**COMET-20-LLM-Ref-QAS**	85.4%	0.864	0.946	0.963
3	MetricX XXL	85.0%	0.847	0.949	0.938
4	BLEURT-20	84.7%	0.719	0.959	0.938
5	COMET-22	83.9%	0.771	0.900	0.942
6	COMET-20	83.6%	0.876	0.936	0.970
7	**f200spBLEU-LLM-Ref-QAS**	83.6%	0.451	0.875	0.932
8	UniTE	82.8%	0.624	0.888	0.914
9	MS-COMET-22	82.8%	0.695	0.809	0.909
10	MATESE	81.0%	0.617	0.757	0.856
11	**f200spBLEU-LLM-Ref**	79.6%	0.424	0.851	0.869
12	YiSi-1	79.2%	0.626	0.881	0.935
13	COMETKiwi[noref]	78.8%	0.674	0.763	0.866
14	**chrF-LLM-Ref**	78.4%	0.587	0.878	0.861
15	COMET-QE[noref]	78.1%	0.502	0.468	0.569
16	BERTScore	77.4%	0.428	0.811	0.924
17	**BLEU-LLM-Ref**	76.6%	0.378	0.784	0.783
18	UniTE-src[noref]	75.9%	0.509	0.779	0.874
19	MS-COMET-QE-22[noref]	75.5%	0.539	0.672	0.897
20	MATESE-QE[noref]	74.8%	0.337	0.637	0.767
21	f200spBLEU	74.1%	0.283	0.819	0.728
22	chrF	73.3%	0.346	0.815	0.630
23	BLEU	70.8%	0.179	0.724	0.594

Table 2: System-level Pearson correlation (ρ) and **Accuracy** for WMT22 Metrics Shared Task. The results in the table are ranked by **Accuracy**. The **Bolded** metrics are our results. The metrics with **LLM-Ref** suffix represent that the metrics use multiple references constructed by our method. The metrics with **QAS** suffix denotes Quality-Aware Selection.

4.3 Settings

We use GPT-3.5-turbo for reference generation. When calling the OpenAI API[8], we use the default values for all hyper-parameters. The adjustment of the temperature parameter can generate more diverse candidates, and we leave the exploration of this for future work. We generate 40 references for each sentence on WMT22 Metric Shared Tasks.

4.4 Main Results

We evaluated the performance of LLM-Ref at both the system level and segment level.

Table 2 presents the pairwise accuracy and Pearson correlation at the system level. LLM-Ref demonstrates positive improvements across all metrics, including both neural-based and n-gram-based metrics. The improvement is particularly noticeable in the n-gram-based metrics, with an accuracy increase of up to **+5.5** percentage points when comparing ID 21 to 11. When combined with Quality-Aware selection for reference candidates, there is an improvement of up to **+9.5** percentage points from ID 21 to 7.

For the neural-based metrics, both BLEURT and COMET-20 exhibit improvements of approximately **+1.7** percentage points comparing ID 6 and 4 to 2 and 1. After incorporating LLM-Ref, BLEURT and COMET-20 outperform all metrics except GEMBA-GPT4-DA.

In terms of the Pearson correlation (ρ), the improvement is more pronounced in the n-gram-based metric, with F200spBLEU experiencing an improvement of up to **+19.5** percentage on ZH-EN from ID 21 to 7. The improvement in the neural-based metric is relatively tiny, with a maximum increase of 1 percentage point.

Segment-level experimental results are consistent with system-level. We list the detailed results in the Appendix C.

5 Analysis

In this section, we will first analyze the effects of reference numbers and prompt variants, and then analyze whether multi-reference solves the data leakage issue. We conduct these analyses on the WMT22 Metrics Shared Task.

5.1 Effects of Reference Numbers

Although LLMs are capable of generating a large number of references, it is important to consider to what extent LLMs can scale the reference numbers, as the diversity and accuracy of the generated reference may decrease.

Figure 2 illustrates the effect of reference numbers in system-level Pearson correlation. For n-gram-based metrics, the Pearson correlation tends to increase as the number of references increases. This tendency is more evident for ZH-EN and EN-RU, while the effect is weaker for EN-DE. Furthermore, incorporating Quality-Aware Selection boosts performance in all directions. Notably, the impact of the ground truth on the system-level performance of n-gram-based metrics is consistently negative across all languages, so we only use the generated references.

For neural-based metrics, the Pearson correlation exhibits minor fluctuations. We assume that since BLEURT and COMET-20 already have strong performance (0.97 on ZH-EN and 0.95 on EN-RU), some low-quality reference candidates may be unbeneficial. In addition, neural-based metrics only calculate scores based on a single reference and then perform a basic max operation, which constrains their capability to use multiple references.

Segment-level results are similar to system-level. We list the detailed results in the Appendix C.

[8] https://openai.com/blog/openai-api

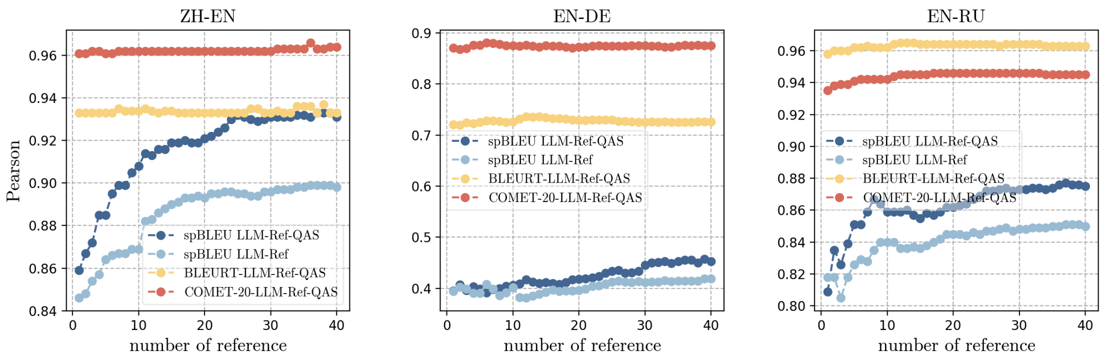

Figure 2: The system-level Pearson correlation (ρ) for each metric along with the increasing reference numbers on the WMT22 Metric task. The advancement of n-gram-based metrics is much larger.

ID	Prompt Variant	(system) ρ	(segment) τ
0	Ground Truth	0.728	0.140
1	MT systems	0.537	0.180
	Reference Candidates Generated By GPT-3.5-turbo		
2	Chinese Prompt	0.780	0.196
3	English Prompt	0.793	0.193
4	3 + No Ground Truth	0.693	0.193
5	3 + No Rules	0.791	0.193

Table 3: The impact of different prompts on the quality of the generated translations. The MT systems used to generate reference candidates are Google and Microsoft translators. The experiments were conducted on the WMT22 Metrics Task ZH-EN test set.

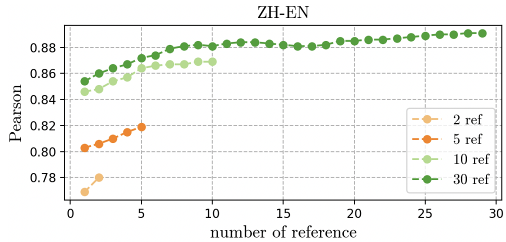

Figure 3: Pearson correlation of F200spBLEU with different reference numbers in a single inference. Larger reference numbers cause higher-quality references.

5.2 Effects of Prompt Variants

We investigate the impact of different prompts on the quality of the generated reference candidates. Commercial translation systems (Google and Microsoft) are also used to generate references. The results are summarized in Table 3. All experiments on prompt variants were conducted on the ZH-EN dataset. Notably, the number of generated reference candidates is constrained to 2 here and we will discuss its impact in Sec 5.3.

References generated by commercial translation systems do significant harm to system-level Pearson correlation (ρ), implying that commercial translation systems may not be able to produce high-quality references to support multiple references. Regarding the choice of language for the prompt, the reference candidates generated by the English prompt slightly outperform the ones generated by the Chinese prompt. Furthermore, omitting the ground truth from the prompt leads to a 10 percentage point decrease in the Pearson correlation (ρ) of the evaluation metric. The rules (characterization and task-specific rules) in the prompt do not appear to have much impact. The Kendall correlation (τ) at the segment level does not vary noticeably across methods.

5.3 Generate As Many References As Possible In One Shot

It is worth noting that the number of generated translations specified in the prompt has a significant effect on the quality of the results. As shown in Figure 3, increasing the number of references from 2 to 30 leads to an 8 percent improvement in the Pearson correlation. We speculate that when LLMs generate more reference candidates at once, they strive to produce reference candidates with finer granularity, enabling more distinct discrimination between high and low-quality candidates. Consequently, the general quality of the reference candidates improves. We believe that this phenomenon can provide valuable insights for enhancing subsequent data synthesis from LLMs.

5.4 Addressing Data Leakage Issues through Multi-Reference

As previously discussed, data leakage can be a concern when using open-source LLMs, as it leads to an overestimation of the downstream task. We aim to alleviate this issue through multi-reference, as models that suffer from data leakage tend to produce nearly identical results to the ground truth, resulting in suboptimal performance on other references. We verify this on the Flores-200 Japanese-Chinese test set (Costa-jussà et al., 2022) as in preliminary experiments. The results are presented in Table 4.

When using a single reference, BLOOMz-7b exceeds all other systems in F200spBLEU including closed-source LLMs with hundreds of billions of parameters even without fine-tuning. The correlation of both F200spBLEU and BLEURT with Human Eval is weak, especially the Kendall score.

After using multiple references, the MT model and other closed-source LLMs in turn outperform BLOOMz-7b-ft. GPT-3.5-turbo also achieves comparable results to the MT model. And the correlation with Human Eval improves a lot for both F200spBLEU and BLEURT. F200spBLEU exhibits better agreements with Human Eval than BLEURT as it accurately measured some hard cases like the MT-ft-test.

Mt-ft-test is the MT model fine-tuned on the test set to simulate the data leakage problem. We observe that F200spBLEU accurately measures the model's performance while BLEURT does not as it overestimates the MT-ft-test too much. We speculate the reason is that neural-based metrics rely on the semantic space to determine the similarity between translations. This inability to discriminate and penalize results in terms of diversity constrains the capability of neural-based metrics to mitigate the overestimation issue.

These observations suggest that using multi-reference with n-gram-based metrics can effectively mitigate the overestimation problem caused by data leakage, which is a pushing issue of open-source LLMs (Wei et al., 2023).

We also apply the direct assessment (DA) scoring method mentioned in GEMBA to evaluate results on the FLORES JA-ZH test set. The performance of GPT-3.5-DA is excellent and unaffected by data leakage from BLOOMz-7b.

5.5 Bias in Generated References towards the Original Model

One critical concern associated with the generated references is whether they exhibit a similar token distribution with the original LLMs and demonstrate any preferential biases when evaluating the original model. We conjecture that as the sentence is transformed into all possible formats, the bias towards the original model diminishes to a significant degree. To verify this assumption, we compute the n-gram overlap for MT model translation, GPT-3.5-turbo, and ground truth with generated references. The results are shown in Figure 4.

Model	Gold-Ref spBLEU	Gold-Ref BLEURT	Multi-Ref spBLEU	Multi-Ref BLEURT	GPT-3.5 DA	Human Eval
Large Language Models						
BLOOMz-7b	27.43	63.17	35.38	67.72	74.80	2.17
BLOOMz-7b-ft	**31.71**	**69.23**	48.53	73.85	83.70	2.99
Claude	22.16	68.23	48.69	74.02	84.55	3.34
GPT-3.5-turbo	23.28	68.45	**53.68**	**74.16**	86.50	3.65
Machine Translation models						
MT	24.89	69.18	52.06	74.41	86.00	3.62
MT-ft-test	33.12	71.21	51.33	75.15	86.50	3.62
Correlation with Human Eval						
Pearson (ρ)	-0.162	0.864	0.970	0.933	0.977	1.000
Kendall (τ)	-0.138	0.276	0.966	0.690	0.928	1.000

Table 4: F200spBLEU, BLEURT and Human Eval score on Flores-200 test set. The **Bolded** scores correspond to the best in LLMs. Using multi-references effectively improves the correlation with Human Eval.

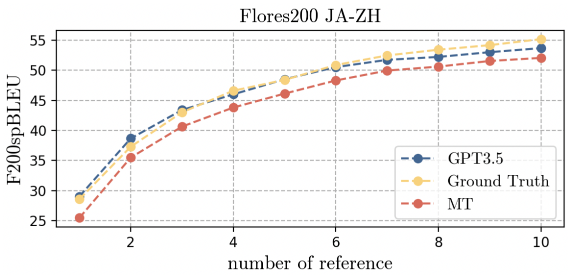

Figure 4: The n-gram overlap of several translation results with generated references.

The scores of GPT-3.5-turbo are larger than the ground truth and MT model when the number of references is relatively small which can be attributed to the similar token distribution. As the reference number increases, the scores of ground truth exceed GPT-3.5-turbo, and the gap between MT and GPT-3.5-turbo becomes closer. This phenomenon validates our assumption that when the reference number increases, the bias gradually decreases.

Given that GPT-3.5-turbo is the model with the highest human evaluation scores, preventing us from observing reference bias, we employ an open-source model, Qwen-7B-chat (Bai et al., 2023), to further investigate this issue. We generate 10 references and list the results in Table 5. When using references generated by Qwen-7B-Chat, the score of itself remains at a reasonable level, without exhibiting excessively high values. Concurrently, the references produced by the open-source LLMs can still enhance the consistency between the metric and human evaluation to a certain extent.

Model	Gold-Ref	Multi-Ref-Qwen	Multi-Ref-GPT	Human Eval
Large Language Models				
BLOOMz-7b	27.43	32.28	35.38	2.17
BLOOMz-7b-ft	**31.71**	41.40	48.53	2.99
Qwen-7B-Chat	19.19	38.36	43.93	2.78
Claude	22.16	39.68	48.69	3.34
GPT-3.5-turbo	23.28	**43.36**	**53.68**	**3.65**
Machine Translation models				
MT	24.89	43.30	52.06	3.62
MT-ft-test	33.12	51.33	42.31	3.62
Correlation with Human Eval				
Pearson (ρ)	0.055	0.936	0.970	1.000
Kendall (τ)	0.097	0.878	0.966	1.000

Table 5: F200spBLEU and Human Eval score on Flores-200 test set with references generated by Qwen-7B-Chat. The **Bolded** scores correspond to the best in LLMs.

Metrics	ZH-EN	EN-RU	EN-DE
spBLEU-LLM-Ref	**0.869**	**0.851**	**0.424**
spBLEU-LLM-Ref w. GT	0.839	0.838	0.364

Table 6: Pearson correlation of spBLEU-LLM-Ref with and without ground truth. GT refers to ground truth.

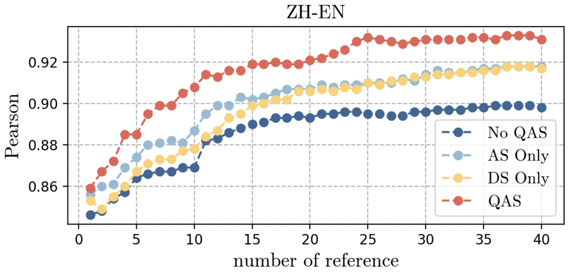

Figure 5: Ablation study on Accuracy Selection (AS) and Diversity Selection (DS) strategies.

Methods	COMET-22	DistinctN	BLEURT
PRISM	67.89	72.42	53.08
LLM-Ref	**82.68**	**75.94**	**73.58**

Table 7: The synthetic reference quality of PRISM and LLM-Ref. We sampled 100 examples from the WMT22 ZH-EN test set and constructed 10 references using both methods.

5.6 Ablation Study

We further conduct some ablation experiments on both Accuracy Selection (AS) and Diversity Selection (DS) of QAS to validate their effectiveness. As shown in Figure 5, using QAS with only the lower bound or upper bound has a positive impact on the Pearson correlation. Combining them can obtain larger improvements, which means these two selection strategies are complementary to each other. The results confirm our hypothesis discussed in Sec 3.2.

5.7 The Effect of Ground Truth

In our experiments, we observed that the presence of ground truth exerts a negative impact on n-gram-based metrics, which aligns with previous studies (Freitag et al., 2020; Bawden et al., 2020). These studies identified that the presence of translationese in ground truth can diminish the correlation between evaluation metrics and human judgments. This phenomenon was consistently manifested across all three directions in our experiments, leading us to exclude ground truth from our experimental setup. The impact of ground truth is demonstrated as shown in Table 6.

6 Related Work

6.1 Automatic Evaluation Metrics

Based on the algorithm used, automated evaluation metrics can be divided into n-gram-based metrics and neural-based metrics. The former computes the accuracy by matching the results generated by the model with the manually generated reference at the n-gram level (Papineni et al., 2002; Popović, 2015). The latter measures the similarity of sentences by learning on a large amount of text data with high-dimensional semantic vectors (Sellam et al., 2020; Rei et al., 2022).

Previous research (Freitag et al., 2020) also investigates the impact of reference quality on automatic evaluation. They collect 2 more references with human annotators while we scale the reference number to 40 with LLMs.

6.2 Paraphrase Reference

Research has also been conducted on the generation of additional references through paraphrasing to enhance the accuracy of automatic evaluation (Kauchak and Barzilay, 2006; Zhou et al., 2006; Bawden et al., 2020). These methods predominantly employ paraphrase models to construct synthetic references. We compared the quality of references generated by PRISM (Bawden et al., 2020) and LLM-Ref, as illustrated in Table 7. The results indicate that the quality of references produced by LLM-Ref is significantly superior to those generated by PRISM. This is because such paraphrasers rely on beam search to generate multiple outcomes, but the beam search curse (Meister et al., 2020) leads to poor results when the beam size exceeds five, making these methods less effective, while LLM is not.

Methods	Pearson	Kendall-tau
Para-Ref	0.813	0.177
LLM-Ref	**0.869**	**0.205**

Table 8: The Pearson and Kendall correlation of Para-Ref and LLM-Ref in WMT22 ZH-EN test set.

6.3 LLMs as Evaluators

The use of LLMs as evaluators have gained popularity in recent studies (Chiang and Lee, 2023; Wang et al., 2023), which leverages instruction-tuned models like ChatGPT to score the results of various NLG tasks including machine translation. It has shown high agreement with human evaluation and has been applied to tasks such as machine translation, text summarization, and image captioning.

The approach closest to ours is one that exploits the paraphrasing ability of LLMs to generate multiple references to augment the consistency of automatic metrics (Tang et al., 2023). We list the results of WMT22 ZH-EN test set of both methods in Table 8. Our work imposes the task information to the generation process (for example the source sentence of translation and the translation rules in Sec 3.1) and proposes effective strategies to select high-quality references based on diversity and accuracy. We further emphasize the multiple references' effect on the overestimation problem caused by data leakage. To our knowledge, this is the first attempt to address the data leakage risk issue in evaluating LLMs.

7 Conclusion

This work presents a study on the limited diversity of references for n-gram-based metrics and the overestimation issue when assessing LLMs that suffer from data leakage. We propose LLM-Ref, a framework that enhances alignments of machine translation evaluation metrics with human evaluation. By generating multiple reference candidates with LLMs and implementing the Quality-Aware selection mechanism, we effectively address these two challenges. Experiments on the WMT22 Metrics shared task including 4 languages demonstrate that LLM-Ref significantly improves the consistency of all automatic evaluation metrics with human assessments. Additionally, our analyses highlight the efficacy of n-gram-based metrics with multiple references in mitigating overestimation issue due to data leakage in LLMs while neural-based metrics struggle to overcome. We explore the factors that affect the quality of generated references by LLMs and provide some insights for future data synthesis from LLMs.

Limitations

Our experiments primarily focused on GPT-3.5-turbo which is a close-source model with hundreds of billion parameters. In future experiments, we plan to investigate the effect of smaller open-source models and generate hundreds of reference candidates.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020. ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 887–894, Online. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. 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.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. 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. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, and Hao Zhou. 2020. WeChat neural machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 239–247, Online. Association for Computational Linguistics.

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.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, October 31 - November 1, 2018*, pages 186–191. 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, 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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, and Furu Wei. 2023. Not all metrics are guilty: Improving nlg evaluation with llm paraphrasing. *arXiv preprint arXiv:2305.15067*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F Wong, Lidia S Chao, Wenqiang Lei, and Jun Xie. 2022. Alibaba-translate china's submission for wmt 2022 metrics shared task. *arXiv preprint arXiv:2210.09683*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023a. Improving machine translation with large language models: A preliminary study with cooperative decoding. *arXiv preprint arXiv:2311.02851*.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023b. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. WeChat neural machine translation systems for WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 243–254, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, Sydney, Australia. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Model	BLEURT	COMET-20
Online-G	0.689	0.456
M2M100$_1.2B-B4$	0.609	0.129
NiuTrans	0.653	0.313
LLM-Ref-1	0.701	0.424
LLM-Ref-2	0.698	0.410

Table 9: BLEURT and COMET-20 score on several models from WMT22 and generated references.

A Accuracy of Generated References

We calculate the BLEURT and COMET scores of the generated references. The generated results are reasonable and exceed the performance of the vast majority of MT systems as shown in Table 9.

B Prompt Templates

Below we show our prompt templates which we use for the experiments to generate multiple reference candidates.

```
You are a professional translator specialized in translating Chinese to English.
During the translation process, please follow the following principles:
1. Strictly adhere to the grammar and vocabulary usage norms of the target
language, and do not add, delete, or adjust the language expression at will;
2. Maintain the meaning and tone of the original text in the translation, and
try not to change the author's original intention;
3. Pay attention to the overall understanding and grasp of the context to ensure
that the translation is complete, coherent, and fluent;
4. Pay attention to language style and localization in the translation, combined
with the cultural background and language habits of the target readers, in order
to make the translated text more natural and fluent.

The following is a difficult-to-translate Chinese sentence:
 {source_sentence}
Its human English translation is as follows:
 {ground_truth}
Please provide 40 high-quality English translations that differ from the above
translation but adhere to the five principles of translation listed above, with
the highest quality at the top and the lowest at the bottom, and one translation
per line.
```

C Segment Level Results of WMT22 Metrics Shared Task

ID	Metrics	Accuracy	en-de τ	en-ru τ	zh-en τ
0	GEMBA-GPT4-DA	89.8%	0.357	0.358	0.382
1	**BLEURT-20-LLM-Ref-QAS**	86.4%	0.342	0.391	0.386
2	**COMET-20-LLM-Ref-QAS**	85.4%	0.323	0.374	0.367
3	MetricX XXL	85.0%	0.360	0.420	0.427
4	BLEURT-20	84.7%	0.344	0.359	0.361
5	COMET-22	83.9%	0.368	0.400	0.428
6	COMET-20	83.6%	0.319	0.330	0.332
7	**f200spBLEU-LLM-Ref-QAS**	83.6%	0.220	0.233	0.224
8	UniTE	82.8%	0.369	0.378	0.357
9	MS-COMET-22	82.8%	0.283	0.351	0.341
10	MATESE	81.0%	0.323	0.279	0.389
11	**f200spBLEU-LLM-Ref**	79.6%	0.231	0.246	0.228
12	YiSi-1	79.2%	0.235	0.227	0.296
13	COMETKiwi[noref]	78.8%	0.290	0.359	0.364
14	**chrF-LLM-Ref**	78.4%	0.246	0.256	0.216
15	COMET-QE[noref]	78.1%	0.281	0.341	0.365
16	BERTScore	77.4%	0.232	0.192	0.316
17	**BLEU-LLM-Ref**	76.6%	0.205	0.214	0.222
18	UniTE-src[noref]	75.9%	0.287	0.342	0.343
19	MS-COMET-QE-22[noref]	75.5%	0.233	0.305	0.287
20	MATESE-QE[noref]	74.8%	0.244	0.229	0.337
21	f200spBLEU	74.1%	0.180	0.153	0.140
22	chrF	73.3%	0.214	0.168	0.147
23	BLEU	70.8%	0.169	0.140	0.145

Table 10: Segment-level Kendall correlation (τ) for WMT22 Metrics Shared Task. The results in the table are ranked by **Accuracy**. The **Bolded** metrics are our results. The metrics with **LLM-Ref** suffix represent that the metrics use multiple references constructed by our method. The metrics with **QAS** suffix denotes Quality-Aware Selection.

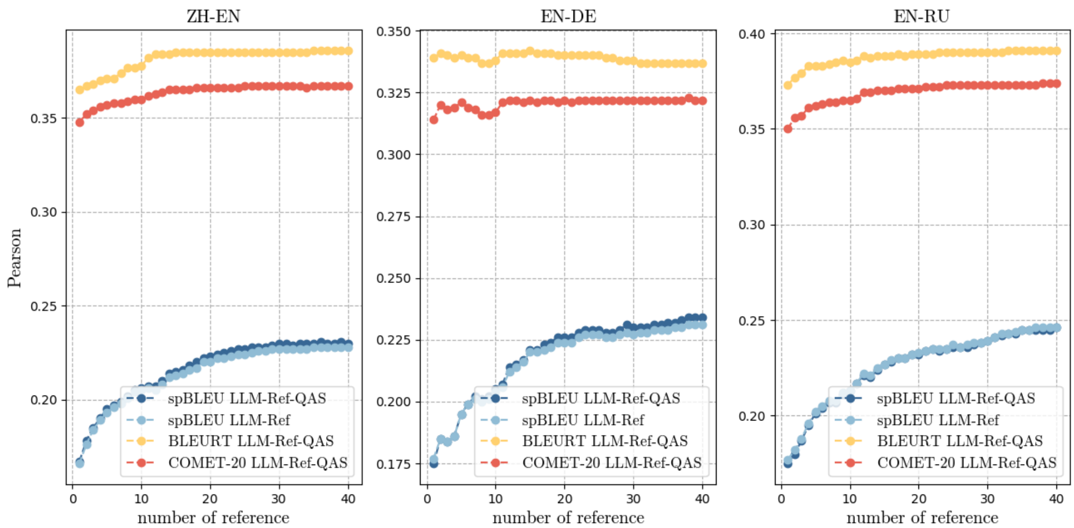

Figure 6: The segment-level Kendall correlation (τ) for each metric along with the increasing reference numbers. Most of the metrics have a stable gain with increasing reference numbers.