LTRC_IIITH's 2023 Submission for Prompting Large Language Models as Explainable Metrics Task

Pavan Baswani*, **Ananya Mukherjee***, **Manish Shrivastava**
Language Technologies Research Center, KCIS, IIIT Hyderabad, India.
{pavan.baswani, ananya.mukherjee}@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

In this report, we share our contribution to the Eval4NLP Shared Task titled "Prompting Large Language Models as Explainable Metrics." We build our prompts with a primary focus on effective *prompting strategies*, *score-aggregation*, and *explainability* for LLM-based metrics. We participated in the track for smaller models by submitting the scores along with their explanations. According to the Kendall correlation scores on the leaderboard, our MT evaluation submission ranks second-best, while our summarization evaluation submission ranks fourth, with only a 0.06 difference from the leading submission. Our code is available at https://github.com/pavanbaswani/Eval4NLP_SharedTask

1 Introduction

With groundbreaking advancements in unsupervised learning and scalable architectures, the possibilities and associated risks, of automatically generating audio, images, videos, and text have become incredibly daunting. Conducting human evaluations of such content is not only costly but often logistically challenging. Consequently, there is a pressing need for automatic metrics that can reliably assess the quality of generation systems and their outputs. Presently, the state-of-the-art metrics for evaluating natural language generation (NLG) systems still fall short of replicating the proficiency of human experts. These metrics primarily rely on neural language models and typically yield a single quality score at the sentence level. This singular score makes it arduous to explain their internal decision-making processes and their resulting assessments (Leiter et al., 2023a).

The introduction of APIs for large language models (LLMs), such as ChatGPT, and the recent open-source availability of LLMs like LLaMA have ignited a surge in NLP research, including the development of LLM-based metrics (Chiang and Lee, 2023). Noteworthy examples include GEMBA (Kocmi and Federmann, 2023a), which delves into using prompts with ChatGPT (OpenAI, 2023a) and GPT4 (OpenAI, 2023b) directly as metrics, and Instructscore (Xu et al., 2023), which takes a different approach by fine-tuning a LLaMA model to provide a detailed error diagnosis of machine-translated content.

It is important to note that current research lacks systematic evaluation of potential prompts and prompting techniques for metric usage. This includes approaches that involve instructing a model or having the model explain a task on its own. Additionally, there is a scarcity of assessments regarding the performance of recent open-source LLMs, despite their critical role in enhancing the reproducibility of metric research compared to closed-source alternatives.

This year's Eval4NLP shared task (Leiter et al., 2023b) addresses these gaps, providing open-source, pre-trained LLMs (Table 1) for assessing machine translations and summaries. The focus is on prompting techniques without LLM fine-tuning, aiming to improve alignment with human evaluations and enhance metric interpretability while identifying promising models for future fine-tuning.

The shared task aims to achieve the following objectives:

- Development of prompting strategies for LLM-based metrics.
- Establishment of a score aggregation method for LLM-based metrics.
- Enhancement of explainability in the context of LLM-based metrics.

Our submission aligns with these objectives. We attain these goals by utilizing the orca_mini_v3_7b

* Authors contributed equally

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 156–163

Model	Language	Params	Seq Length	Size (GB)
Guanaco-65B-GPTQ	multilingual	65B	2048	33.5
Platypus2-70B-Instruct-GPTQ	english	70B	4096	35.3
WizardLM-13B-V1.1-GPTQ	english	13B	2048	7.45
Nous-Hermes-13b	english	13B	2048	26
OpenOrca-Platypus2-13B	english	13B	4096	26.03
orca_mini_v3_7b	english	7B	4096	13.48

Table 1: List of LLMs provided in the Shared Task

(Mathur, 2023) model and crafting prompts through a combination of fine-grained and chain-of-thought prompting strategies. Additionally, we have adapted 4-bit quantization to optimize model loading. We submit reference-free a) segment-level quality scores for all the language pairs (en-de, en-zh, en-es) listed under the MT evaluation task and b) summary-level quality scores for all the documents provided.

2 Background

2.1 LLM-Based Evaluation

Large Language Model (LLM)-based evaluation involves employing sophisticated language models (such as GPT-3 or similar) to evaluate the accuracy and quality of machine-generated text. An example of this is the work by Liu et al., 2023, who introduced G-Eval, a summarization evaluation model built on GPT-4. Notably, G-Eval surpassed all previous baseline models in summarization evaluation performance according to their research. In the recent WMT22 metrics shared task (Freitag et al., 2022), the best-performing MT evaluation metric is METRICX XXL, a massive multi-task metric fine-tuned on LLM model checkpoints. However, Kocmi and Federmann, 2023b shows that GEMBA, a GPT-based metric that works both with a reference translation and without has outperformed all the metrics that participated in the WMT22 shared task.

It's important to note that LLM-based evaluations usually generate a single score but lack the capacity to provide detailed reasoning or explanations behind that score.

2.2 Explainability

Explainability has gained significant importance in AI research in recent years, offering potential benefits for AI system users, designers, and developers (Leiter et al., 2023a). Explainability is particularly desirable for evaluation metrics. Sai et al., 2022 explainable Natural Language Generation (NLG) metrics should prioritize offering comprehensive information beyond a single score. Eval4NLP 2021 (Fomicheva et al., 2021) was the first shared task to emphasize explainability in MT evaluation.

Explainable evaluations are assessment methods that not only provide a numerical score for the quality of machine-generated text but also offer detailed insights or explanations regarding why a particular score was assigned. These metrics aim to make the evaluation process more transparent and interpretable by highlighting specific strengths and weaknesses in the generated text, such as fluency, accuracy, coherence, relevance, or semantic fidelity. They are valuable for both improving NLG systems and enabling users to better understand the quality of text.

2.3 Prompt Engineering

Prompt engineering is a dual-purpose AI engineering technique: it fine-tunes large language models with specific prompts and guides the process of refining inputs for generative AI services to create text or images. In the following, we'll discuss some prompt-engineering techniques.

1. Zero-Shot Prompting: Zero-shot prompting is an AI technique where models respond effectively to prompts they've never seen before during training. It leverages general knowledge to generate context-aware responses, often by providing auxiliary information or examples. This approach enhances the adaptability of AI models in tasks like language understanding and generation. It's particularly valuable in diverse, real-world applications.

2. Few-Shot Prompting: Few-shot prompting is an AI approach where models are trained to perform tasks or generate responses with very limited examples or data, typically fewer than five instances. It relies on techniques like meta-learning and transfer learning to enable models to generalize effectively from minimal training data. This method is essential for applications requiring rapid adaptation to new tasks or domains.

3. Chain of Thought (CoT): Chain of thought prompting is a cognitive technique involving structured, sequential prompts or questions designed to guide systematic thinking and exploration of a topic. Large Language Models (LLMs) have shown enhanced capabilities of solving novel tasks by reasoning step-by-step (Kim et al., 2023).

4. Fine-Grained Analysis: Fine-grained prompting is a method that involves detailed examination and analysis of data or information at a granular level. It is employed to gain a deeper and more comprehensive understanding by breaking them down into smaller, distinct components for in-depth exploration and assessment. Fine-grained prompting is often used in research, data analysis, and various industries to extract valuable insights and make informed decisions.

5. Translational Probability: Translational probability prompting involves assessing the likelihood that a given translation accurately represents the intended meaning of the source text. It's a key factor in evaluating the quality and fidelity of machine-generated translations. This technique helps measure how well an MT system produces translations that align with the expected or reference translations, contributing to the assessment of translation accuracy and effectiveness.

6. Majority Vote: Majority vote prompting is a decision-making approach that relies on aggregating the opinions or votes of multiple individuals or systems to make a final decision. This technique is used to enhance decision-making by leveraging collective wisdom and improving the accuracy or robustness of choices.

7. Self-Refinement: Self-refinement is a process of continuous improvement or self-development. Self-refinement prompting involves providing prompts or questions that prompt reflection and self-assessment. These prompts encourage models to identify areas for improvement and take action to enhance their performance.

Each of these concepts plays a crucial role in various domains, from machine learning and artificial intelligence to cognitive psychology and decision-making processes. Understanding and effectively applying these concepts can lead to more robust and informed solutions in a wide range of applications.

3 System Description

We opted for orca_mini_v3_7b among the provided LLMs due to its smaller size, which accommodated our resource constraints. We encountered challenges when attempting to load other LLMs. We curated prompts using a blend of fine-grained and chain-of-thought prompting strategies. Furthermore, using bitsandbytes[1] we employed 4-bit quantization to enhance model loading efficiency and considered MAX TOKENS as 512 during inference (refer Appendix 7 for computation details).

Our submission includes: a) Summary-level quality scores for all the documents provided in the task. b) Segment-level quality scores for language pairs (en-de, en-zh, en-es) in the MT evaluation task, without relying on references.

The summary-level scores and segment-level scores lies in the range of 0-100, where 0 is the least score that can be awarded to a bad translation/summary and 100 is the highest score that can be assigned to a perfect translation/summary.

3.1 Dataset

Table 2 illustrates the provided test sample statistics. The reported token counts were computed using bert tokenizer[2].

		# Entries	min tokens	max tokens	average tokens
summarization	*source (en)*	825	144	818	279.413
	target (en)		9	402	51.697
en_de	*source (en)*	1425	18	137	37.935
	target (de)		17	156	41.297
en_es	*source (en)*	1834	15	137	37.472
	target (es)		19	149	41.683
en_zh	*source (en)*	1297	18	137	37.856
	target (zh)		21	212	51.436

Table 2: Test Data Statistics

3.2 Our Prompting Strategies

We outline our prompting strategies for this shared task as follows.

3.2.1 Approach-1 (Zero-shot W/o explanation)

"Zero-shot prompting without explanation" means prompting the LLM to generate a response without providing any additional information or context to clarify or support the prompt. It relies solely on the initial instruction without further elaboration.

3.2.2 Approach-2 (Zero-shot w/ explanation)

"Zero-shot prompting with explanation" involves providing a prompt or instruction to a system and supplementing it with additional information or context to clarify or support the prompt (refer Table 3 & 4). This approach aims to enhance the

[1] https://huggingface.co/blog/4bit-transformers-bitsandbytes#advanced-usage

[2] https://huggingface.co/bert-base-multilingual-cased

system's understanding of the task or request by offering more details or background information alongside the initial instruction.

3.2.3 Approach-3 (CoT + Fine-grained w/ explanation)

We aim to incorporate a strategic approach to facilitate a deeper understanding, ultimately enhancing the LLM's ability to provide improved responses. Our approach involves a combination of chain of thought (CoT) prompting and fine-grained analysis, specifically focusing on the aspects of Relevance, Consistency, Coherence, and Fluency for Summarization; and emphasizing on Adequacy, Faithfulness, and Fluency for MT

- **Fine-grained Analysis for Summarization:** Firstly, the LLM is instructed to provide individual scores for Relevance, Consistency, Coherence, and Fluency. These individual scores are then used to prompt the model to provide a final overall summary score, ensuring a comprehensive assessment of the summarization quality (refer Table 5). This approach enables a more detailed and nuanced evaluation of the summary's performance in each aspect.

- **Fine-grained Analysis for MT:** Initially, the LLM generates separate scores for Adequacy, Faithfulness, and Fluency. Subsequently, using these scores, the model is prompted to produce a final translation quality score, ensuring a comprehensive evaluation of the translation's performance in each dimension (refer Table 6). This approach enhances our ability to assess translation quality thoroughly.

4 Results & Analysis

Table 7 depicts the summary-level Kendall correlation scores for the summarization evaluation task. We can infer that our submission (LTRC) ranks 4th with a very minute difference of 0.06 when compared to the top submission. We initially used zero-shot prompting which resulted in a correlation of 0.41 in the leaderboard. After employing CoT + Fine-grained prompting, the Kendall correlation improved to 0.44. Hence, it is evident that strategic prompting has shown a positive improvement in the system's performance.

Table 8, 9, and 10 depict segment-level Kendall correlations for MT on en-de, en-zh, and en-es language pairs respectively. We can notice that our submissions have consistently ranked 2nd (in small models track) across the language pairs.

For the en-de language pair, zero-shot prompting resulted in a correlation of 0.11 which drastically improved to 0.19 with CoT + Fine-grained prompting. Conversely, for en-zh, when CoT + Fine-grained prompting was applied, the correlation score dropped to 0.09. Hence for en-zh and en-es, we have made our submission with zero-shot prompting.

An interesting point to observe is that our submissions have surpassed most of the submissions made in the large model track except NLLG for en-de and en-es, and MysteryTest for en-es.

4.1 Error Analysis

We conducted manual analysis on a few English-German MT samples. During this analysis, we identified a minor scoring issue emanating from language compatibility [3]. To illustrate this, we've provided a few examples in Table 11. It's notable that the zero-shot prompting strategy yielded a notably high score, even though it overlooked translation accuracy (in the first case) and generated inaccurate explanations (in both examples). On the other hand, CoT + fine-grained prompting has penalized the first example by awarding a score of 70 but in the explanation, it failed to identify the missing info and rather provided an incorrect assessment of text fluency. This observation underscores the need for a more nuanced evaluation approach that considers not only the final scores but also the accuracy and reliability of the explanations provided by the model.

5 Challenges

- **Resource Constraints:** The process of loading and utilizing large language models demands substantial computational resources. Unfortunately, due to limited available memory, we encountered difficulties loading alternative models. Despite successfully loading the large models, we encountered issues when attempting to perform inference.

- **Language Compatibility:** Using an English-trained (orca_mini_v3_7b) model to evaluate German, Spanish, and Chinese translations may have performance implications.

[3] orca_mini_v3_7b was originally trained on English text

Instruction
The task is to provide the overall score for the given summary with reference to the given article on a continuous scale from 0 to 10
along with explanation in JSON format with "score" and "explanation" keys as follows: {"score": <float-value>, "explanation": <explanation-text>}.
Where a score of 0 means the summary is "irrelevant, factually incorrect and not readable" and score of 10 means "relevant, factually correct, good readability".
You must justify the score that you provided with clear and concise reason within 2 sentences interms of justifying the relevance, readability, factuality metrics.
The article text and summary text is given in triple backticks ‘‘‘ with ### Article: and ### Summary: as prefix respectively.
Note: The generated response must be in json format without any missed braces or incomplete text. Also, it should not provide any additional information other than JSON output.

Article: ‘‘‘{}‘‘‘
Summary: ‘‘‘{}‘‘‘
Response:

Table 3: Zero-shot prompting for evaluating Summary

Instruction:
The task is to score a translated text from {English} to {German} with respect to the source sentence on a continous scale from 0 to 100,
along with explaination in JSON format with "score" and "explanation" keys as follows: {"score": <float-value>, "explanation": <explanation-text>}.
Where a score of zero means "no meaning preserved and poor translation quality" and score of one hundred means "excellant translation quality with perfect meaning and grammar".
You must justify the score that you provided with clear and concise reason within 2 sentences interms of justifying the adequacy, fluency, faithfulness metrics.
The source sentence and target sentence is given in triple backticks with ### source sentence: and ### target sentence: as prefix respectively.
Note: The generated response must be in json format without any missed braces or incomplete text. Also, it should not provide any additional information other than JSON output.

source sentence: ‘‘‘{}‘‘‘
target sentence: ‘‘‘{}‘‘‘
Response:

Table 4: Zero-shot prompting for evaluating MT

Instruction
You will be given one summary written for a news article.

Your task is to assign the single score for the summary on continuous scale from 0 to 10 along with explanation.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing,
and refer to it as needed. You must justify the score that you provided with clear and concise reason within 2 sentences in
terms of justifying the relevance, fluency, coherence and consistency metrics.

The article text and summary text is given in triple backticks ‘‘‘ with "Source Text:" and "Summary:" as prefix respectively.

Evaluation Criteria:
1) Relevance (1-5) - selection of important content from the source. The summary should include only important information
from the source document. Annotators were instructed to penalize summaries which contained redundancies and excess information.
Here, 1 is the lowest and 5 is the highest.
2) Consistency (1-5) - the factual alignment between the summary and the summarized source. A factually consistent summary
contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained
hallucinated facts. Here, 1 is the lowest and 5 is the highest
3) Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and
coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related
information, but should build from sentence to a coherent body of information about a topic.". Here, 1 is the lowest and 5 is the highest.
4) Fluency (1-3): the quality of the summary in terms of grammar, spelling, punctuation, word choice, and sentence structure.
- 1: Poor. The summary has many errors that make it hard to understand or sound unnatural.
- 2: Fair. The summary has some errors that affect the clarity or smoothness of the text, but the main points are still comprehensible.
- 3: Good. The summary has few or no errors and is easy to read and follow.

Evaluation Steps:
1. Read the summary and the source document carefully.
2. Compare the summary to the source document and identify the main points of the article.
3. Assign scores for Relevance, Consistency, Coherence and Fluency based on the Evaluation Criteria.
4. By utilizing the generated scores of Relevance, Readability, Coherence and Fluency, aggregate these scores to assign the single score
for the summary on continuous scale from 0 to 10 along with explanation in JSON format with "score" and "explanation" keys as follows:
{"score": <float-value>, "explanation": <explanation-text>}.

Source Text: ‘‘‘{}‘‘‘
Summary: ‘‘‘{}‘‘‘
Response:

Table 5: CoT + fine-grained prompting for evaluating summaries

Instruction
You will be given one translated sentence in {Spanish} for a source sentence in {English}.

Your task is to assign the single score for the translation on continuous scale from 0 to 100 along with explanation.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing,
and refer to it as needed. For explanation, you must justify the score that you provided with clear and concise reason within
2 sentences interms of justifying the adequacy, fluency and faithfulness metrics.

The source text and translation text is given in triple backticks “‘ with "Source Text:" and "Translation:" as prefix respectively.

Evaluation Criteria:
1) Adequacy (1-5) - the correspondence of the target text to the source text, including the expressive means in translation.
Annotators were instructed to penalize translation which contained misinformation, redundancies and excess information.
Here, 1 is the lowest and 5 is the highest.
2) Faithfulness (1-5) - translation faithfulness to the meaning depends on how the translator interprets the speaker’s intention
and does not imply that one should never or always translate literally. Here, 1 is the lowest and 5 is the highest.
3) Fluency (1-3): the quality of the translation in terms of grammar, spelling, punctuation, word choice, and sentence structure.
- 1: Poor. The translation has many errors that make it hard to understand or sound unnatural.
- 2: Fair. The translation has some errors that affect the clarity or smoothness of the text, but the main points are still comprehensible.
- 3: Good. The translation has few or no errors and is easy to read and follow.

Evaluation Steps:
1. Read the translation and the source document carefully.
2. Compare the translation to the source text.
3. Assign scores for Adequacy, Faithfulness and Fluency based on the Evaluation Criteria.
4. By utilizing the generated scores of Adequacy, Faithfulness and Fluency, aggregate these scores to assign the single score for the
translation on continuous scale from 0 to 100 along with explanation in JSON format with "score" and "explanation" keys as follows:
{"score": <float-value>, "explanation": <explanation-text>}.

Source Text: “‘{}“‘
Translation: “‘{}“‘
Response:

Table 6: CoT + fine-grained prompting for evaluating MT

Track	Team Name	Summ
Small	DSBA	0.5
	iML	0.49
	IUST_NLP_Lab	0.48
	LTRC	0.44
	CompetitionEntrants	0.44
	Beginners	0.38
	ManCity	0.25
Large	NLLG	0.35

Table 7: Summary-level Kendall Correlation for Summarization Task

Track	Team Name	en-de
Small	HIT-MI&T Lab	0.49
	LTRC	0.19
	uOttawa	0.12
	TaiwanSenior	0.04
Large	NLLG	0.24
	MysteryTest	0.17
	Eval4NLP	0

Table 8: Segment-level Kendall Correlation for MT on English-German pairs.

Track	Team Name	en-zh
Small	HIT-MI&T Lab	0.32
	LTRC	0.13
Large	NLLG	0.13
	MysteryTest	0.1
	Eval4NLP	0.01

Table 9: Segment-level Kendall Correlation for MT on English-Chinese pairs.

Track	Team Name	en-es
Small	HIT-MI&T Lab	0.42
	LTRC	0.11
Large	NLLG	0.18
	MysteryTest	0.12
	Eval4NLP	-0.02

Table 10: Segment-level Kendall Correlation for MT on English-Spanish pairs.

6 Conclusions

In this paper, we present our contribution to the Eval4NLP shared task, which focuses on the "Prompting Large Language Models as Explainable Metrics Task." Our submission is specifically tailored to the small model track. Our evaluation involved the use of the "orca_mini_v3_7b" model to assess machine translation (MT) and summarization test data. To generate scores accompanied by explanations, we employed both zero-shot and fine-grained + CoT prompting strategies. In our findings, we provide Kendall correlation scores in comparison to other submissions. We also conducted error analysis by discussing the model's performance on a few samples. Overall, our submission ranked 2nd in the machine translation task and 4th in the summarization task. We believe that further exploration of strategic prompting methods holds the potential to enhance model performance in terms of both scoring and explainability.

References

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. 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.

Seungone Kim, Se Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning.

Tom Kocmi and Christian Federmann. 2023a. 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 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.

		Zero-Shot Prompting		CoT + Fine-grained Prompting		Human Evaluation	
English (en)	**German (de)**	**Score**	**Explanation**	**Score**	**Explanation**	**Score**	**Explanation**
Samson Lawal (born 25 April 2004) is a Nigerian professional footballer who plays as a midfielder for Livingston.	Samson Lawal (* 25. April 2004 in Livingston) ist ein nigerianischer Fußballspieler.	90	The translation is adequate, with the main information about the player and his birthdate preserved. However, the sentence structure is not as fluent as it could be, and the use of the German word "ist" instead of "ist in" could be improved.	70	The translation is adequate as it conveys the main information from the source text, including the birthdate, nationality, and position of the footballer. However, the sentence structure is not as fluent as it could be, and there are some minor grammatical errors.	60	The translation misses some information like \textit{Samson Lawal plays as a midfielder for Livingston} and instead mentions Livingston as his birthplace.
Andrii Ivanovych Dubchak (born 19 October 1976, Kalynivka, Vinnytsia Oblast) is a Ukrainian photographer, videographer, photojournalist, and war correspondent.	Andrii Ivanovych Dubchak (* 19. Oktober 1976 in Kalynivka, Vinnytsia Oblast) ist ein ukrainischer Fotograf, Videograf, Fotojournalist und Kriegskorrespondent.	90	The translation is very fluent and maintains the meaning of the original sentence. However, there are some minor grammatical errors, such as the use of 'ist' instead of 'ist ein' for the first occurrence of 'Andrii Ivanovych Dubchak' and the missing comma after 'Kalynivka, Vinnytsia Oblast).	80	The translation is adequate, as it conveys the main information about Andrii Ivanovych Dubchak's profession and birthplace. The fluency is good, with no major grammatical or spelling errors. However, the faithfulness could be improved, as the sentence structure is not entirely faithful to the original source text.	95	hethetheThe translation is accurate and preserves source meaning. The only minor issue is that letter 'U' should be capitalized in \textit{ukrainischer}.

Table 11: Analysis on en-de MT pairs.

Christoph Leiter, Piyawat Lertvittayakumjorn, M. Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2023a. Towards explainable evaluation metrics for machine translation. *ArXiv*, abs/2306.13041.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023b. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison for NLP systems*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Pankaj Mathur. 2023. orca_mini_v3_7b: An explain tuned llama2-7b model. `https://https://huggingface.co/psmathur/orca_mini_v3_7b`.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4 technical report.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback.

7 Appendices

We used the following computation for all inferences.

1. CPU:
    - **Name**: Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz
    - **Total**: 40
    - **# Cores**: 10
    - **cache size**: 25600 KB
2. GPU:
    - **Name**: NVIDIA GeForce RTX 2080 Ti
    - **Total**: 4
    - **Memory/GPU**: 11GB