Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models: A Case Study on ChatGPT

Qingyu Lu[◇,ℜ], Baopu Qiu[♭,ℜ], Liang Ding[ℜ], Liping Xie[◇], Dacheng Tao[ℜ]
[◇]Southeast University [ℜ]JD Explore Academy, JD.com Inc. [♭]Nanjing University
📧 luqingyu@seu.edu.cn, qiubaopu@smail.nju.edu.cn, dingliang1@jd.com
🧳 https://github.com/Coldmist-Lu/ErrorAnalysis_Prompt

Abstract

Generative large language models (LLMs), e.g., ChatGPT, have demonstrated remarkable proficiency across several NLP tasks such as machine translation, question answering, text summarization, and natural language understanding. Recent research (Kocmi and Federmann, 2023) has shown that utilizing ChatGPT for assessing the quality of machine translation (MT) achieves state-of-the-art performance at the system level but *performs poorly at the segment level*. To further improve the performance of LLMs on MT quality assessment, we conducted an investigation into several prompting methods. Our results indicate that by combining Chain-of-Thoughts (Wei et al., 2022) and Error Analysis (Lu et al., 2022), a new prompting method called `Error Analysis Prompting`, LLMs like ChatGPT can *generate human-like MT evaluations at both the system and segment level*. Additionally, we discovered some limitations of ChatGPT as an MT evaluator, such as unstable scoring and biases when provided with multiple translations in a single query. Our findings aim to provide a preliminary experience for appropriately evaluating translation quality on ChatGPT while offering a variety of tricks in designing prompts for in-context learning. We anticipate that this report will shed new light on advancing the field of translation evaluation with LLMs by enhancing both the accuracy and reliability of metrics.

1 Introduction

Large language models (LLMs), especially Generative Pre-trained Transformer (GPT) models (Radford et al., 2019; Brown et al., 2020), have shown remarkable performance in various natural language processing (NLP) tasks. Recently, Open AI developed ChatGPT, an interactive chatbot built upon InstructGPT (Ouyang et al., 2022), which has captured the attention of researchers in the NLP community (Qin et al., 2023; Zhong et al., 2023). This chatbot is capable of integrating multiple NLP tasks and can generate detailed and comprehensive responses to human inquiries. Additionally, it can respond appropriately to follow-up questions and maintain sensitivity throughout several turns of conversation.

Previous research has demonstrated that ChatGPT can perform as well as or even better than other LLMs in machine translation task (Hendy et al., 2023). However, it remains uncertain whether ChatGPT can be used as a metric to evaluate the quality of translations. If ChatGPT is suitable for this task, then, how to develop appropriate prompts that can make ChatGPT generate reliable evaluations? Concurrent to our work, Kocmi and Federmann (2023) present an encouraging finding that LLMs, e.g., ChatGPT, could outperform current best MT metrics at the system level quality assessment with zero-shot standard prompting, but such kind of prompts show unreliable performance at the segment level.

In this work, we take the further step by carefully investigating the current advanced few-shot prompting strategies upon ChatGPT for MT quality assessment, and propose a novel prompting strategy – **Error Analysis (EA) Prompting**, combining the Chain-of-Thought (CoT, Wei et al. (2022)) and Error Analysis (EA, Lu et al. (2022)). We give an example of this prompt in Figure 1. The idea is to prompt ChatGPT to generate a human-like evaluation like MQM (Freitag et al., 2021) by ❶ *identifying major and minor errors*, and ❷ *scoring the translations according to the severity of these errors*. In addition, we also explore the potential of ChatGPT compared with modern neural metrics like COMET (Rei et al., 2020), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020).

Our experiments and analyses illustrate that:

🙂 Our proposed EA Prompting outperforms standard prompting (Kocmi and Federmann, 2023) at the segment level, achieving human-

Figure 1: An comparative overview between Standard Prompting and our proposed Error Analysis Prompting in assessing the MT quality with ChatGPT.

like evaluations at both the system level and segment level.

When designing prompts, itemized responses are better than lengthy and detailed explanations of errors. Moreover, splitting the instruction into two identifying errors and scoring translation can improve evaluation stability.

The boosted performance from EA prompting is observed in the zero-shot scenario on text-davinci-003 rather than in the few-shot scenario, which indicates that we need to adjust our settings when utilizing other GPT models.

Despite its good performance, we show that ChatGPT is NOT a stable evaluator and may score the same translation differently.

It is NOT advisable to combine multiple translations into a single query input, as ChatGPT has a preference for former translations.

The remainder of this report is designed as follows. We present the evaluation settings and comparative results in Section 2. In Section 3, we highlight several potential issues that researchers should be aware of when using ChatGPT as a translation evaluator. Conclusions are described in Section 4.

2 ChatGPT As An Evaluation Metric

2.1 Experiment Setup

Dataset We utilize the testset from the WMT20 Metric shared task in two language pairs: Zh-En and En-De. To ensure the reliability of our experiment, for each language pair, we divide the segments into four groups based on the number of tokens they contain (15-24, 25-34, 35-44, 45-54). We randomly sample 10 segments from each group and form a new dataset containing 40 segments. Details are shown in Table 1.

Human Evaluation Human evaluation of translated texts is widely considered to be the gold standard in evaluating metrics. We use a high-quality human evaluation dataset Multi-dimensional Quality Metrics (MQM, Freitag et al. (2021)) as human judgments. This dataset is annotated by human experts and has been widely adopted in recent translation evaluation (Freitag et al., 2022) and quality estimation tasks (Zerva et al., 2022) in WMT.

Meta Evaluation We utilize the accuracy of pairwise system-ranking (Kocmi et al., 2021) for the system level comparison. At the segment level, we follow Freitag et al. (2022) to adopt the average of three types of Kendall correlation. Specifically, these values are computed by flattening the scores

Language Pair	Segments	Systems	Systems Selected
En-De	40	7	Tohoku-AIP-NTT, OPPO, eTranslation, Tencent_Translation, Huoshan_Translate, Online-B, Online-A
Zh-En	40	8	Huoshan_Translate, WeChat_AI, Tencent_Translation, OPPO, THUNLP, DeepMind, DiDi_NLP, Online-B

Table 1: Statistics of testset. Source, reference texts, and translations are from the WMT20 metrics shared task.

Metrics	En-De		Zh-En	
	System(%)	Segment(%)	System(%)	Segment(%)
BLEU (Papineni et al., 2002)	71.43	3.55	21.43	14.71
BERTscore (Zhang et al., 2020)	76.19	12.30	25.00	26.75
BLEURT (Sellam et al., 2020)	76.19	33.44	57.14	32.76
COMET (Rei et al., 2020)	71.43	33.47	50.00	38.97
text-davinci-003	42.86	11.86	53.57	23.08
ChatGPT-EA	76.19	26.40	60.71	36.73

Table 2: The system and segment level results of metrics using pairwise accuracy (%) and Kendall correlation (%) with human-annotated MQM scores. The best results are bold.

into a single vector and calculating the average correlations over systems, or over segments.

Baseline We compare LLMs with several commonly used baseline metrics for MT evaluation. BLEU (Papineni et al., 2002) is the most popular metric that compares the n-gram overlap of the translation with human reference, but it has been criticized for not capturing the full semantic meaning of the translation (Freitag et al., 2022). BERTScore (Zhang et al., 2020) is a neural metric that relies on pre-trained models to compute the semantic similarity with the reference. BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020) are supervised neural metrics that leverage human judgments to train. They have shown a high correlation with human judgments.

Large Language Models We test the evaluation capability on ChatGPT using the default model of ChatGPT-plus, and compare it with text-davinci-003, a base model of ChatGPT.

2.2 ChatGPT as a metric attains SOTA performance at the system level

Table 2 presents the performance of LLMs compared with other baseline metrics. We report the best-performing setting, where LLMs with EA prompting. We can see that:

- at the system level, ChatGPT achieves SOTA

Figure 2: The segment level comparison between ChatGPT and text-davinci-003 with standard prompting ("Standard"), in-context error analysis prompting ("EA"), and zero-shot error analysis prompting ("zero-shot EA").

performance compared with existing evaluation metrics for both language pairs. However, text-davinci-003 obtains inferior results compared with other metrics. Our results are consistent with the findings of Kocmi and Federmann (2023), who tested the performance of large language models on full test set of the WMT22 metric task.

Instruction		Response		Separation	Score - Segment#38								Total
Standard	EA	Detailed	Itemized		sys1	sys2	sys3	sys4	sys5	sys6	sys7	sys8	
3		3		-	-3	-2	0	-3	-1	-3	-1	-2	-15
3			3	-	-3	-3	-2	0	-2	-2	-3	-2	-17
	3	3		7	-1	-1	-3	-1	-1	0	-1	-2	-10
	3	3		3	-2	-2	-2	-3	-1	-2	-2	-2	-16
	3		3	7	-5	-5	-3	-4	-5	-4	-4	-3	-28
	3		3	3	-4	-4	-3	-6	-3	-4	-4	-3	-26

Table 3: Comparison of the segment level scores of ChatGPT for different variants of in-context prompts. We divide the instructions into two categories: standard and error analysis ("EA"). The response template can either be itemized or detailed. As for the error analysis instruction, it can be separated into two queries (one for identifying errors and another for scoring) or combined into a single query.

- ChatGPT and text-davinci-003 lag behind state-of-the-art metrics for En-De at the segment level. For Zh-En, while text-davinci-003 remains suboptimal, ChatGPT with EA prompting exhibits superior performance relative to all other metrics, with the exception of COMET.

2.3 Error analysis prompting with ChatGPT is better than standard prompting at the segment level

To improve the segment level evaluation capabilities of ChatGPT, we combine the idea of Chain-of-Thought (Wei et al., 2022) and Error Analysis (Lu et al., 2022). Chain-of-Thought has been successfully applied in complex reasoning tasks, which encourages the LLM to break down the task into a series of reasoning steps, allowing it to better understand the context and formulate a more accurate response. Error analysis strategy (Lu et al., 2022) aims to generate human-like evaluation by incorporating human evaluation framework, e.g. MQM (Freitag et al., 2021), into existing metrics to obtain better discriminating ability for errors, e.g., lexical choice (Ding et al., 2021a) or adequacy (Popović 2020) errors. Specifically, we instruct ChatGPT to identify major and minor errors in the translation, and then enable ChatGPT to score the translation based on the severity of errors.

Figure 2 compares the segment level results between different prompting strategies. Prompting ChatGPT with error analysis can benefit translation evaluation between segments by improving Kendall correlation by a large margin (26.01 vs. 36.73). However, simply replacing prompting instruction from scoring with zero-shot EA will even damage the performance, since identifying errors without samples will make ChatGPT become more unstable. This also highlights the importance of prompting with in-context examples.

Moreover, on text-davinci-003, the improvements from EA prompting are shown in the zero-shot scenario ("zero-shot EA"). The reason for this may be that while text-davinci-003 is capable of detecting errors when prompted with explicit instructions for error analysis, it may face challenges in fully comprehending the task of error analysis when presented with in-context examples. Compared with text-davinci-003, ChatGPT has been trained using reinforcement learning through human feedback and conversational tuning, which enables it to generalize to error analysis through in-context examples.

2.4 Error analysis prompting empowers ChatGPT to produce human-like evaluations

Given the crucial significance of the prompt design, we explore several versions of in-context prompt contexts and present an analysis in Table 3. See Appendix A for the prompt contexts used in our experiment. We find that:

(i) ChatGPT becomes more adept at identifying errors when instructed by error analysis. When designing in-context examples, is it worth designing an instruction on error analysis or simply adopting standard scoring instruction? We find that error analysis instructions can make ChatGPT better understand the task of error identification in the best setting, since more errors are identified compared with standard instructions (28 vs 17 in total). As a result, We recommend using error analysis instructions in prompt contexts instead of standard instruction.

(ii) Itemized template response is better than detailed illustration. As shown in the "Response"

column, providing descriptions of errors in detail hinders ChatGPT's ability to accurately identify errors. For example, when using error analysis with combined instruction, ChatGPT with itemized response can identify more errors than detailed response (28 vs 10 in total). It seems that lengthy paragraphs will make it dif cult for ChatGPT to fully comprehend the details of each error, potentially leading to confusion regarding the intended task. Therefore, it is recommended to provide clear and concise descriptions of errors in a format that is easy for ChatGPT to process and comprehend.

(iii) Separating the scoring process from error identi cation may improve the stability of Chat-GPT. We suggest splitting the instruction of error analysis into two queries, one for identifying errors and the other for scoring the translation. Although this may not cause a signi cant performance gain, we observe that sometimes ChatGPT fails to deduct points for identi ed errors or presents an incorrect calculation of scores. Separating the scoring process may be helpful, as it allows ChatGPT to focus on one single procedure in each query, thus can provide more accurate judgments.

3 Case Study

In this section, we list several typical issues with the case study that should be aware of when using ChatGPT as a translation evaluator.

3.1 ChatGPT is unstable when conducting evaluation process

When assessing translations using ChatGPT, it is not uncommon to observe variations in the scores assigned to the same input. As shown in Figure 3, we regenerate several responses with the same input and obtain 3 different scores (98, 95, 100) for the translation. The discrepancies in scores could be attributed to the inherent randomness of the model behind ChatGPT. Another possible reason is the lack of clearly stated evaluation criteria described in the prompt contexts. Therefore, we suggest using speci c guidelines such as the template we propose to minimize the impact of these variations.

3.2 ChatGPT prefers former inputs when provided with multiple translations

An interesting phenomenon is that when multiple translations are presented together as a single input to ChatGPT for evaluation, it tends to believe

Figure 3: When evaluating the same translation three times, ChatGPT generates similar explanations but different scores.

that the translations provided earlier are of higher quality, while the quality of later translations are relatively poorer.

Figure 4 shows an example of the attack on Chat-GPT. We provide 8 translations along with their corresponding source and reference sentences. First, we present the translations sequentially, and ask ChatGPT to rank them according to their translation quality. ChatGPT ranks the translations as (SYS1, SYS2, SYS4, SYS5, SYS3, SYS6, SYS7, SYS8), with SYS1 being the best translation and SYS8 being the worst. Then, we reverse the order of translations and obtain an entirely different sequence of ranks - (SYS8, SYS7, SYS6, SYS5, SYS4, SYS3, SYS2, SYS1), with SYS8 being the best translation and SYS1 being the worst. The contradictory results may be attributed to the auto-regressive nature of the decoder model,

Figure 5: An example on ChatGPT directly adopting BLEU to evaluate translation quality.

4   Conclusion

In this paper, we explore the potential of ChatGPT as a metric for evaluating translations. We design a novel in-context prompting strategy based on chain-of-thought and error analysis, and show that this strategy signi cantly improves ChatGPT's evaluation performance. We compare our approach with other prompt designs to show the effectiveness of error analysis. We hope the experience can ben-e t NLP researchers in developing more reliable promoting strategies. In section 3, we also highlight several potential issues that researchers should be aware of when using ChatGPT as a translation evaluator.

In future work, we would like to experiment with our method on more test sets and top-performed systems (Barrault et al., 2019; Anastasopoulos et al., 2021; Ding et al., 2021b; Kocmi et al., 2022; Zan et al., 2022), to make our conclusion more convincing. Also, it is worth exploring the reference-free settings, i.e., quality estimation (Specia et al., 2010; Qiu et al., 2022) evaluation performance, with our proposed error analysis prompting. Lastly, it will be interesting to automatically generate the samples in our few-shot error analysis prompting strategy.

Limitations

Since we do not have access to the ChatGPT API till done the majority of our work, all experiments in this paper were conducted using the interaction screen of ChatGPT. As a result, the test set used in this study is limited. We will conduct more experiments in future work to further validate and re ne our current results and ndings.

Figure 4: Comparison of providing multiple translations in sequential or reverse order. ChatGPT tends to prefer the former translations and generate contradictory judgments.

which gives more attention to the latter input, potentially leading to greater identi cation of errors for the translation input later. Therefore, we recommend that researchers input one translation at a time instead of providing multiple translations.

3.3   ChatGPT may directly adopt existing evaluation metrics

We observe that with standard prompting, sometimes ChatGPT directly adopts existing evaluation metrics, such as BLEU and METEOR. An example of this behavior is in Figure 5.

However, as our objective is to examine ChatGPT's inherent capacity for translation evaluation, rather than its ability to implement pre-existing evaluation procedures, we include an explicit instruction of "Do not use existing metrics" in standard prompting. This encourages ChatGPT to develop its own approach to evaluating translations, independent of existing metrics.

References

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, et al. 2021. Findings of the IWSLT 2021 evaluation campaign. In *IWSLT*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, et al. 2019. Findings of the 2019 conference on machine translation (WMT19). In *WMT*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F. Wong, Dacheng Tao, and Zhaopeng Tu. 2021a. Understanding and improving lexical choice in non-autoregressive translation. In *ICLR*.

Liang Ding, Di Wu, and Dacheng Tao. 2021b. The usyd-jd speech translation system for iwslt2021. In *IWSLT*.

Markus Freitag, George Foster, David Grangier, et al. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *TACL*.

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 *WMT*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, et al. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *WMT*.

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 *WMT*.

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F Wong, and Dacheng Tao. 2022. Toward human-like evaluation for natural language generation with error analysis. *arXiv preprint*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Maja Popović. 2020. Relations between comprehensibility and adequacy errors in machine translation output. In *CoNLL*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint*.

Baopu Qiu, Liang Ding, Di Wu, Lin Shang, Yibing Zhan, and Dacheng Tao. 2022. Original or translated? on the use of parallel data for translation quality estimation. *arXiv preprint*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *ACL*.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Changtong Zan, Keqin Peng, Liang Ding, Baopu Qiu, et al. 2022. Vega-MT: The JD explore academy machine translation system for WMT22. In *WMT*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *WMT*.

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

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint*.

A Prompt Contexts

Figure 6 compares the prompt contexts implemented in error analysis prompting with a detailed response and combined instruction discussed in Section 2.4.

Here is an example template:

In-Context Example

Q: Source: 中新网北京9月27日电（记者 杜燕）为加强节前市场监管执法，北京市市场监管局在国庆节前夕检查各类经营主体2000余户。
Reference: Chinanews.com Report on September 27 in Beijing (Journalist Du Yan) The Beijing Administration for Market Regulation inspected more than 2,000 operating entities of different types before the National Day holiday to strengthen pre-holiday market regulation and law enforcement.
Translation: BEIJING, Sept. 27 (Reporter Du Yan) In order to strengthen market supervision and law enforcement before the festival, the Beijing Municipal Market Supervision Bureau inspected more than 2,000 households of various business subjects on the eve of the National Day.
Based on the given source and reference, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

☺ Itemized Response ⟷ Detailed Response ☹

A: Major errors:
(1)"BEIJING" – Omission
(2)"subjects" – Mistranslation
Minor errors:
(1)"households of various" – Mistranslation
(2)"festival" – Mistranslation
(3)"supervision" – Mistranslation
(4) "Beijing Municipal Market Supervision Bureau" – Inappropriate for context
(5) "BEIJING" – Spelling

I think the mistranslation of "subjects" should be categorized into a major error, and the omission in "BEIJING" should also considered as a major error. "households of various", "festival" and "supervision" are three mistranslation errors, they should be categorized into minor errors. The terminology, "Beijing Municipal Market Supervision Bureau" is Inappropriate for context, and should also be categorized into a minor error. "BEIJING" also has a spelling error, which is considered as a minor error.

Test Question

Use the template above to answer the following question:

Q: Source: 当日，以"讲好中国故事，传播好中国声音"为宗旨的外宣工作创新性、基础性平台——"中国好故事"数据库正式上线，向世界展现真实、立体、全面的中国。
Reference: On that day, the externally publicized innovative and basic platform-"The story of China", for the purpose of telling the story of China well and spreading the voice of China well", was officially on line today, to show the world a true, three-dimensional and comprehensive China.
Translation: On that day, the "China Good Story" database, an innovative and basic platform for outreach work with the aim of "telling a good Chinese story and disseminating a good Chinese voice", was officially launched to show the world a real, three-dimensional and comprehensive China.

☺ Separated Instructions ⟷ Combined Instruction ☹

Based on the given source and reference, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

A: 🤖

Q: Count the number of major and minor errors identified in your last response and compute the final score for this translation. Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

A: 🤖

Based on the given source and reference, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.
Count the number of major and minor errors identified and compute the final score for this translation. Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

Figure 6: A comparison between our proposed error analysis prompting and other prompt contexts.