Characterised LLMs Affect its Evaluation of Summary and Translation

Yu-An Lu
National Chupei Senior High School
luyuam0@gmail.com

Yu-Ting Lin
Taipei Municipal Chenggong High School
dong1214.mailbox@gmail.com

Abstract

In today's widespread use of Large Language Models (LLMs), there have been significant achievements in various text domains such as generating summaries and translations. However, there is still room for development and improvement in evaluating the outputs of LLMs. In this paper, we propose an innovative scoring system that assesses the quality of summaries and translations using multiple metrics, we also enhance LLM's performance in scoring tasks by assigning it different roles, effectively making it act as an expert. We test four roles in the study: a teacher, a proofreader, a travel writer, and an internet troll, comparing the advantages and disadvantages of each role in the scoring task. Our research results demonstrate that emphasizing LLM's multilingual capabilities and strict standards as its identity can effectively boost its performance. Additionally, imbuing LLM with a more critical thinking ability enhances its performance in translation tasks compared to a milder LLM identity. In summary, we show that assigning different identities to LLM can influence its performance in scoring tasks. We believe that this research will contribute to the use of LLMs for scoring purposes.

1 Introduction

Since GhatGPT's emergence, Large Language Models (LLM) have been flourishing in the Natural Language Processing (NLP) field. Thanks to the growth of LLMs, tasks such as automatic summaries and translations are becoming more commonly generated by LLMs. However, we realized that most existing evaluation methods for LLMs output lack thorough explanation, making the research process in this domain considerably challenging. We believe that by inventing a metric for evaluating summarization and translation, research on article generation would be much more practical.

Inspired by previous work on using LLMs to generate scores and evaluate text(Tom Kocmi, 2023)(Jinlan Fu, 2023)(Fu et al., 2023), as well as research exploring having LLMs play the role of experts(Chan et al., 2023), we present an evaluation system employing multiple metrics (Jinlan Fu, 2023) by carefully designed prompt (Tom Kocmi, 2023) and make LLM act as an expert. For generating scores, we employed the model OpenOrca-Platypus2-13B (Lee et al., 2023b) to generate scores, which is a merge of Platypus2-13B (Lee et al., 2023a) and OpenOrcaxOpenChat-Preview2-13B (Wang et al., 2023). We selected this model because of its strong performers on the leaderboard and its small size for local inference. In order to bolster the Large Language Model's (LLM) evaluation capabilities, we implemented a strategy where the LLM simulates an expert.

This study is also a system description for the Eval4NLP 2023 shared task(Leiter et al., 2023) which in the Small track.

2 Method

To use the large language model to better evaluate summarization and translation, we divided the task into a few parts. First, we separated an evaluation task into several metrics. Then we made LLM role different characters such as a proofreader, writer, or internet troll. LLM would evaluate summation and translation in the expert role. In the end, we added scores from different metrics by XGBoost and post-processing.

2.1 Design Character

We hypothesized that making LLM play in different characters can improve its capability of evaluating. So we designed four characters which were a teacher, a proofreader, a travel writer, and an internet troll. We expected those characters could fix some problems in LLM's evaluation.

- Teacher: The teacher played a most professional role in all characters, it is an expert

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 184–192

on viewing student's summary and translation.(keywords: *grading, score, standardized*)

- Proofreader: For the role of a proofreader, LLM would pretend itself as a professional proofreader at Fox Television. We wrote a self-statement about the rules of raring and its expertise field.(keywords: *accuracy, quality, strict standards*)

- Travel Writer: In the travel writer part, we expected the characters like travel writers could have a better ability to evaluate the performance in localization and adherence to local customs.(keywords: *multilingual, cultural immersion, descriptive narratives*)

- Internet Troll: We noticed LLM preferred to give a higher score to translation and summation, so we designed a mean and nasty character to fix this problem. In this role, LLM would mimic an internet troll on Reddit who likes to criticize others.(keywords: *harsh criticism, linguistic expertise, unreasonable ratings*)

2.2 Score Generation

We create ten metrics for evaluating summation and ten for translation. There are four different prompts for rate—a proofreader, a travel writer, an internet troll, and the baseline without character setting. With these prompts, we made LLM evaluate summation and translation based on the ten metrics and rate them with a 1-10 score. In order to make LLM's outputs controllable, we use pytorch(Paszke et al., 2019) and outlines(Willard and Louf, 2023) in our code.

2.3 Ensemble Features

XGBoost, a widely utilized tree-based algorithm, holds significant popularity within the domain of data science. Once the scores of the metrics created by LLM have been calculated, they are utilized as features in order to train an XGB model for regression. This regression model is designed to predict a score that may be utilized for measurement purposes.

3 Experiments

3.1 Datasets

We conducted experiments on both summarizing and English-German translation tasks. The training datasets were obtained from the MQM annotations of the WMT22 dataset for translation, and the average aspect-based ratings of SummEval for summarization. All the data included source and target texts, as well as scores collected from multiple methods. The test dataset was collected by the Eval4NLP organizer, and it shares a similar format to the training dataset.

dataset	Trans(En-De)	Summ
Train	11046	320
Test	1425	825

Table 1: Size of translation and summarization dataset.

3.2 Exploratory Data Analysis of Model Evaluation

Table 2 shows each feature's Correlation Coefficient with Official Scores. We observed that the correlation coefficients between feature and official score varied across roles, with each role exhibiting the strongest correlations with different features. No consistent pattern was discernible across all roles regarding which features were most important.However, we found that higher average correlation coefficients were associated with higher subsequent model accuracy when using XGB for modeling. There was a positive correlation between average correlation coefficients and subsequent model accuracy.

In Figure 1, there are some graphs show the feature scores' distribution. The distributions of most features are concentrated around 6 and 8 points. The distribution of Travel Writer is the most dense, while that of Teacher is more dispersed. The feature distribution of Teacher exhibits a bimodal shape, indicating it has clearer and more established criteria. After xgb modeling, the performance of Teacher is also the best. It is particularly notable that there is almost no overlap between the distributions of the features and official scores.

3.3 Performance

To assess the evaluating performance of the LLM, we employed several standard metrics for evaluating the correlation between two ranking systems. These included:

- **Kendall:** Kendall's tau provides a measure of the concordance between two rankings, with

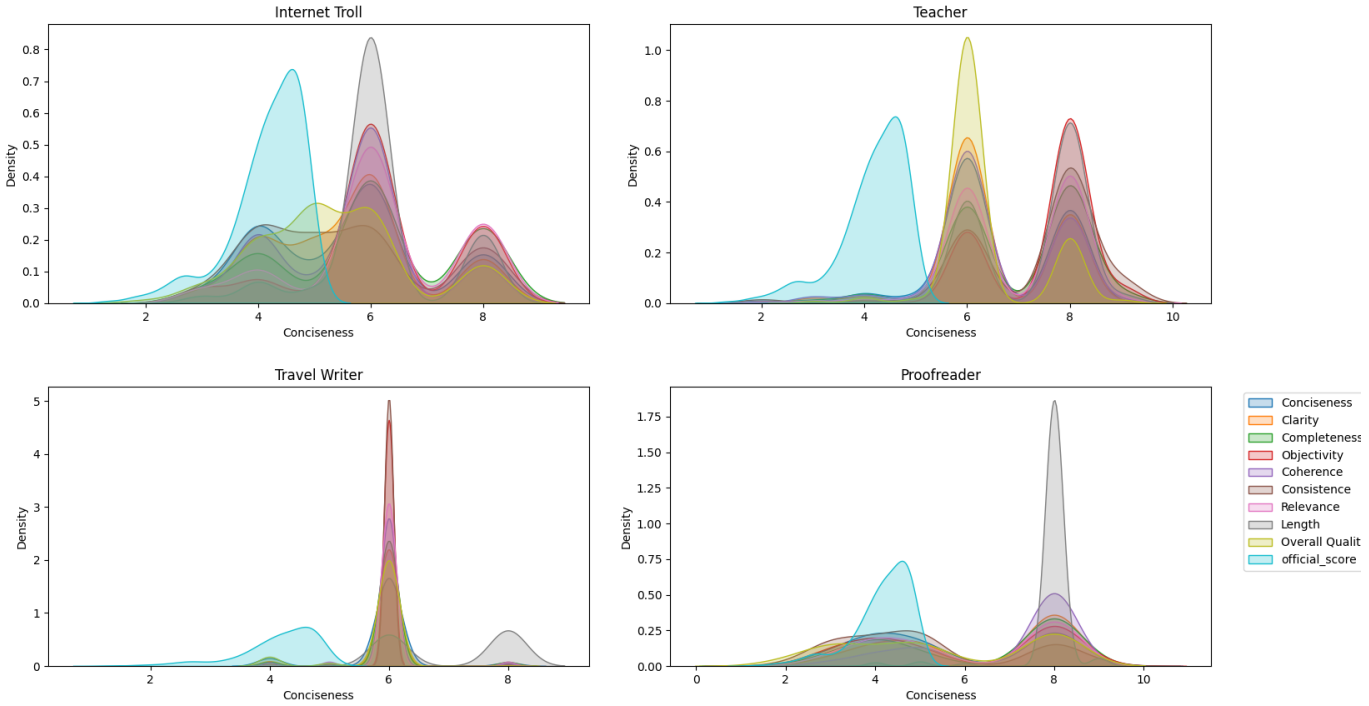

Figure 1: Feature scores' distribution of summarization task.

values closer to 1 indicating stronger agreement.

- **Pearson:** The Pearson coefficient quantifies the linear relationship between two continuous variables. Higher positive coefficients denote greater linear correlation.
- **Spearman:** Spearman's rank correlation coefficient assesses how well the relationship between two rankings can be described using a monotonic function. Values approaching 1 signify a greater tendency for the rankings to match.

The performance is calculated with test datasets on Eval4NLP shared task's codabench, and our team name is *TaiwanSenior*. (Due to the limitations of the codabench platform rules, you can only see on the public page that we achieved 0.04 on En-De, which is just one of our submission scores. You can find the full scores of our different methods in Table 3)

The Travel Writer demonstrates superior performance in the translation task for English-German language pairs, while the Teacher exhibits the highest level of performance in the summarization task. The Travel Writer is noted by several sources for its multilingual capabilities, which result in superior performance in translation tasks but less satisfactory performance in summarization tasks. The inclusion of an Internet Troll character in the Translation task resulted in more effective criticism compared to other general characters. However, the performance of the Internet Troll character was comparatively weaker in the Summarization job. Based on this observation, we may deduce that incorporating greater criticism can assist in improving the performance of Large Language Models (LLMs) to closely resemble human evaluation in the Translation task. The performance of the Proofreader in the Translation task is notably poor, indicating a lack of strong correlation in its evaluation capabilities.

4 Conclusion

We investigate the performance of LLMs with different character in generating scores to evaluate translation and summary tasks by incorporating characterised-prompts into the prompts. We find that emphasizing multilingual capabilities and stringent criteria in the LLM's identity can effectively improve the LLM's performance. By endowing the LLM with stronger critical thinking compared to a more benign LLM, we improve its performance on translation tasks. In summary, we demonstrate that assigning different identities to LLMs influences their performance on scoring tasks.

5 Acknowledgement

We would like to express our gratitude to Professor Hung-Yu Kao and Professor Yao-Chung Fan for their support and assistance throughout this research. We sincerely appreciate them providing us with access to GPU computing resources, without which this work would not have been possible.

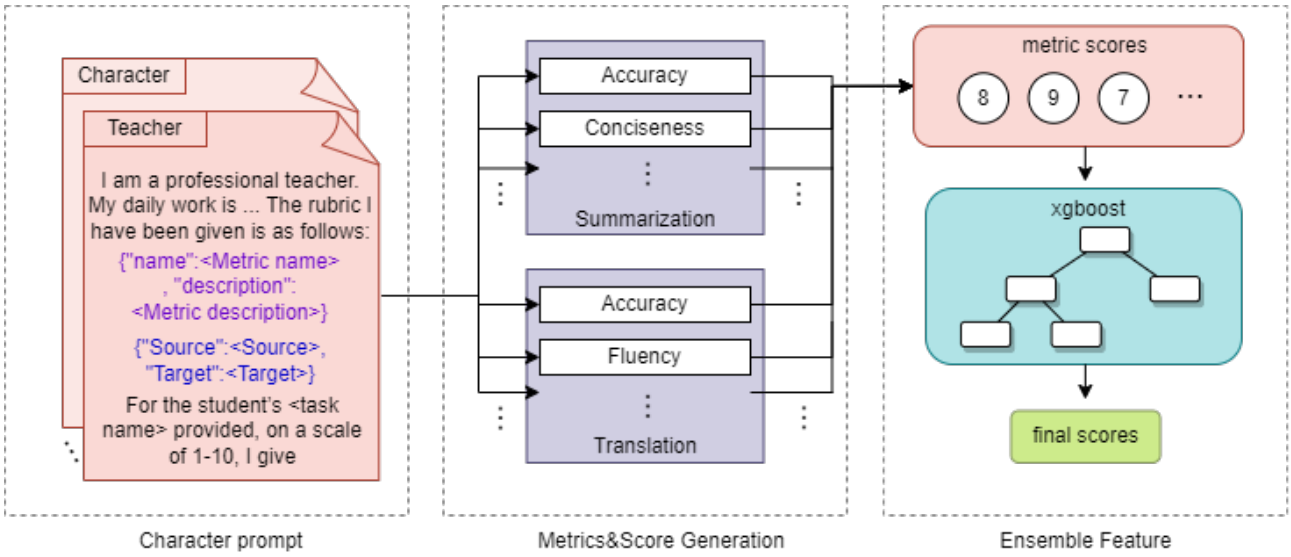

Figure 2: The framework of our evaluation system.

Features	**Internet Troll**	**Teacher**	**Travel Writer**	**Proofreader**
Completeness	**0.072219**	**0.372297**	-0.048266	-0.049932
Clarity	0.072155	0.366486	-0.018286	-0.060655
Relevance	0.033194	0.396311	**0.019369**	-0.027132
Coherence	0.019086	0.332309	-0.033452	**-0.063694**
Objectivity	0.000568	**0.480219**	-0.087915	-0.052679
Accuracy	-0.001213	0.461272	-0.000964	-0.012227
Length	-0.014532	0.423803	0.043023	-0.049222
Conciseness	-0.069729	0.419300	-0.027331	-0.131271
Overall Quality	-0.085636	0.273289	**-0.094137**	-0.086841
Consistence	**-0.099225**	0.478668	-0.065858	**-0.005071**

Table 2: Correlation Coefficient of Features versus Official Scores of summarization task.

	Translation(En-De)			Summarization		
Character	⋆	△	◇	⋆	△	◇
Teacher	0.058	0.074	0.084	**0.363**	**0.453**	**0.520**
Proofreader	0.041	-0.03	0.051	N/A	N/A	N/A
Travel Writer	**0.159**	**0.168**	**0.194**	-0.037	-0.061	-0.047
Internet Troll	0.111	0.112	0.133	-0.043	-0.09	-0.057

⋆ - Kendall △ - Pearson ◇ - Spearman

Table 3: Different charters' performance on Translation and Summarization task

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Zhengbao Jiang Pengfei Liu Jinlan Fu, See-Kiong Ng. 2023. Gptscore: Evaluate as you desire.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2023a. Platypus: Quick, cheap, and powerful refinement of llms.

Ariel N. Lee, Cole J. Hunter, Nataniel Ruiz, Bleys Goodson, Wing Lian, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023b. Openorcaplatypus: Llama2-13b model instruct-tuned on filtered openorcav1 gpt-4 dataset and merged with divergent stem and logic dataset model. `https://huggingface.co/Open-Orca/OpenOrca-Platypus2-13B`.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. 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*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Christian Federmann Tom Kocmi. 2023. Large language models are state-of-the-art evaluators of translation quality. pages 193–203.

Guan Wang, Bleys Goodson, Wing Lian, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorcaxopenchatpreview2: Llama2-13b model instruct-tuned on filtered openorcav1 gpt-4 dataset. `https://https://huggingface.co/Open-Orca/OpenOrcaxOpenChat-Preview2-13B`.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Metric	Tasks	Prompt
Accuracy	Summ, Trans	How accurately the summary/translation represents the key ideas, details and overall meaning of the original text. An accurate summary/translation does not add, misrepresent or leave out information.
Conciseness	Summ	How concise and succinct the summary is, without unnecessary detail. An ideal summary is as condensed as possible while still maintaining accuracy.
Clarity	Summ, Trans	How clear and easy to understand the summary/translation is. A summary/translation should be written clearly using proper grammar and vocabulary suited for the audience.
Completeness	Summ	How complete the summary is in capturing the key points and ideas of the original text. A complete summary covers all important information.
Objectivity	Summ	How objective and unbiased the summary is, without injecting opinions or interpretations. A summary should represent the original text, not the writer's views.
Coherence	Summ	How coherent, unified and logical the summary is. A coherent summary flows smoothly with clear connections between ideas.
Consistence	Summ	How consistent the summary is in tone, style and vocabulary with the original text. The summary should match the original.
Relevance	Summ	How relevant the summary is in selecting the most important ideas from the original text. A relevant summary focuses on key points only.
Length	Summ	Appropriate length for a summary, condensed while still complete. Exact length depends on purpose and original text.
Overall Quality	Summ	The overall comprehensiveness, readability and effectiveness of the summary.
Fluency	Trans	How fluent and natural the translation reads in the target language. High fluency sounds like it was originally written in the target language.
Consistency	Trans	How consistent the translation is across recurring terms, phrases, and styles. High consistency maintains the same translations for repetitions.
Tone	Trans	How well the translation conveys the tone and voice of the original text. High tone matches the original style and emotional impact.
Register	Trans	How appropriate the register (formal/informal language) is for the context. High register matches the original level of formality.
Style	Trans	How well the translation maintains the stylistic properties of the original. High style replicates creative language use, imagery, etc.
Idiomatic Expression	Trans	How well the translation conveys meaning through natural, idiomatic expressions in the target language. High idiomatic expression sounds local.
Cultural Adaptation	Trans	How well the translation adapts cultural references and concepts appropriately for the target audience. High adaptation naturalizes foreign elements.
Domain Knowledge	Trans	How well the translation handles specialized terms and domain-specific concepts. High knowledge accurately conveys technical/domain meaning.

Table 4: Metrics that LLMs evaluate with.

Character	Prompt
Teacher	I am a professional teacher. My daily work is to grade students' work according to grading criteria. I am only allowed to give students a score between 1-10 as a whole number. I cannot include any personal opinions.
Proofreader	As a professional proofreader at Fox Television, I take pride in my solid expertise and over five years of experience in English, Chinese, Spanish, and German. I have a profound understanding of the grammar, sentence structure, vocabulary, and cultural nuances of these languages, enabling me to excel in translation and summarization tasks. In my role, I maintain a strict standard for quality, and I'm unwavering in assigning low scores to translations or summaries that fall short. Working in television demands a zero-tolerance attitude toward accuracy and quality. Below is a translation and summary provided by a client, and I will rate it on a scale of 1 to 10, accompanied by an explanation of my professional assessment. To ensure I adhere to the policies of the television network, I will steadfastly give poor translations and summaries a rating of 1.
Travel Writer	As a travel writer, I take great pride in my multilingual and cross-cultural abilities, which allow me to deeply understand and share the uniqueness of various countries and regions. My language proficiency spans English, Chinese, Spanish, and German. Through extensive travels, I've immersed myself in the cultures of Germany, Spain, the United States, the United Kingdom, and Taiwan, delving into their customs, values, and everyday idioms. My translation and summarization skills enable me to transform these rich experiences into written narratives. I often provide rating services for fellow writers and researchers, rigorously assessing the quality of their work. I not only assign them scores ranging from 1 to 10 but also offer detailed feedback to help them improve. Below is a summary and translation provided by a university student, and I will assess it based on my professional capabilities, accompanied by an objective commentary explaining my evaluation.
Internet Troll	As an internet troll, I excel at critiquing others' work on Reddit, especially translations and summaries. I possess a profound understanding of languages such as English, German, Chinese, Spanish, including their grammar, sentence structure, and vocabulary. I often provide reasonable criticisms of others' translations and summaries based on my extensive linguistic knowledge, and because I always include well-founded explanations, no one can refute my harsh ratings. Here's a translation and summary from the internet, and I will assign it a score from 1 to 10, along with my reasoned explanation to make it irrefutable.

Table 5: Prompts which be used in characterizing LLMs.

Character	Task	Prompt
Proofreader	Trans	As a professional proofreader at Fox Television, I take pride in my solid expertise and over five years of experience in English, Chinese, Spanish, and German. I have a profound understanding of the grammar, sentence structure, vocabulary, and cultural nuances of these languages, enabling me to excel in translation and summarization tasks. In my role, I maintain a strict standard for quality, and I'm unwavering in assigning low scores to translations or summaries that fall short. Working in television demands a zero-tolerance attitude toward accuracy and quality. Below is a translation and summary provided by a client, and I will rate it on a scale of 1 to 10, accompanied by an explanation of my professional assessment. To ensure I adhere to the policies of the television network, I will steadfastly give poor translations and summaries a rating of 1. {"name": "Accuracy", "description": "How accurately the summary represents the key ideas, details and overall meaning of the original text. An accurate summary does not add, misrepresent or leave out information." } {"Source": "The Chlotrudis Award for Best Actress is an annual award presented by the Chlotrudis Society for Independent Films, a non-profit organization, founded in 1994, that recognizes achievements in independent and world cinema.", "Target": "Der Chlotrudis Award für die beste Schauspielerin ist eine jährliche Auszeichnung der Chlotrudis Society for Independent Films, eine 1994 gegründete Non-Profit-Organisation, die Erfolge im unabhängigen und weltweiten Kino anerkennt." } For the student's translation provided, on a scale of 1-10, I give
Proofreader	Summ	As an experienced linguistics professor well-versed in diverse languages and cultures, having lived abroad since childhood and participated in translations for prestigious publications such as The New York Times, The Economist, and Eval4NLP, I have profound and unique insights into translating and summarizing news articles and everyday language. Today, Stanford University has invited me to serve as a reviewer to evaluate summaries and translations completed by their students. I will be provided with a rubric and expected to interpret it based on my expertise to assign scores from 1-10. The rubric I have been given is as follows: {"name": "Fluency", "description": "How fluent and natural the translation reads in the target language. High fluency sounds like it was originally written in the target language." } {"Article": " In 1878, the Oviedo City Council received an application for permission to build the mining railway on Monte Naranco, which raised concerns as it was feared that the construction of the railway would affect the water supply of Fitoria, as it ran parallel to that of the future railway line. On 1 February 1880, the original 7,101-metre (7,766 yd) long mining railway between the Villapérez area and the northern station of Oviedo operated by the Compañía de los Ferrocarriles de Asturias, Galicia y León was inaugurated with an original length of 7.1 km (4.4 mi). The total cost of building the railway was 129,906 pesetas, including 19,798 pesetas for expropriations.", "Summary": "summary" } For the student's summary provided, on a scale of 1-10, I give

Table 6: Examples of prompt ussed on LLM.