# **Re-Evaluating Evaluation for Multilingual Summarization**

Jessica Zosa Forde<sup>1\*</sup>, Ruochen Zhang<sup>1\*</sup>, Lintang Sutawika<sup>2,3</sup>, Alham Fikri Aji<sup>4</sup>, Samuel Cahyawijaya<sup>5</sup>, Genta Indra Winata<sup>6†</sup>, Minghao Wu<sup>7</sup>, Carsten Eickhoff<sup>8</sup>, Stella Biderman<sup>2</sup>, Ellie Pavlick<sup>1</sup> <sup>1</sup>Brown University <sup>2</sup>Eleuther AI <sup>3</sup>CMU <sup>4</sup>MBZUAI <sup>5</sup>HKUST <sup>6</sup>Capital One <sup>7</sup>Monash University <sup>8</sup>University of Tübigen {jessica\_forde, ruochen\_zhang, ellie\_pavlick}@brown.edu stella@eleuther.ai

## Abstract

Automatic evaluation approaches (ROUGE, BERTScore, LLM-based evaluators) have been widely used to evaluate summarization tasks. Despite the complexities of script differences and tokenization, these approaches have been indiscriminately applied to summarization across multiple languages. While previous works have argued that these approaches correlate strongly with human ratings in English, it remains unclear whether the conclusion holds for other languages. To answer this question, we construct a small-scale pilot dataset containing article-summary pairs and human ratings in English, Chinese and Indonesian. To measure the strength of summaries, our ratings are measured as head-to-head comparisons with resulting Elo scores across four dimensions. Our analysis reveals that standard metrics are unreliable measures of quality, and that these problems are exacerbated in Chinese and Indonesian. We advocate for more nuanced and careful considerations in designing a robust evaluation framework for multiple languages.

### **1** Introduction

Ensuring the functionality of large language models (LLMs) in a variety of languages has been important in increasing AI accessibility. Many frequently used tasks and metrics for evaluating these models have been originally developed for English (Mielke, 2016). This propensity towards English has led Bender (2019) to criticize the conflation between all-natural language tasks and natural language tasks in English. In fact, Ducel et al. (2022) note that 30-40% of ACL articles do not explicitly name the languages they evaluate.

The development of new multilingual LLMs, such as BLOOM (Le Scao et al., 2023) and Aya-23 (Aryabumi et al., 2024) have coincided with

increased interest in models that can fluently generate text in a wide array of languages (Dave, 2023). How these models are evaluated in these languages, however, remains understudied.

Researchers often assume that high scores from automatic metrics indicate good summaries, because the metrics correlate strongly with human ratings. Strong correlation in aggregate, however, doesn't necessarily imply that these metrics should substitute for human judgments when determining how much to trust a specific model output (Shen et al., 2023; Liu et al., 2024). Anecdotally, language generation in other languages varies in quality, and current practices for evaluating LLMs do not easily adapt to non-English languages.

Automatic metrics designed to evaluate the quality of generated summaries make assumptions about scripts and tokenizations that differ drastically cross-lingually (Maronikolakis et al., 2021; Sun et al., 2022). Previous works have reported varying correlation relationships between human annotation dimensions and automatic metrics (Kryściński et al., 2019; Rankel et al., 2013; Fabbri et al., 2021; Krishna et al., 2023). However, metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019), while being used commonly as standard evaluation for multilingual summarization, are not carefully studied against human annotations and preferences. More recently, researchers suggest that LLMs such as GPT-4 can be directly used to not only generate texts but to evaluate them as well (Goyal et al., 2022; Liu et al., 2023; Fu et al., 2023; Wang et al., 2023; Kocmi and Federmann, 2023; Li et al., 2024; Chang et al., 2024). However, this practice has not been thoroughly evaluated in languages other than English.

**Our Contributions:** This paper seeks to understand whether existing metrics, namely ROUGE and BERTScore, can be used reliably to evaluate summaries across multiple languages, and whether

<sup>\*</sup>Equal Contribution.  $^{\dagger}$ The work was done outside Capital One.

GPT-4 can faithfully replace human evaluators in the context of multilingual summarization. Our contribution is two-fold.

First, we construct a pilot dataset<sup>1</sup> for the evaluation of existing multilingual summarization metrics. It includes curated articles with human reference summaries in English, Chinese, and Indonesian. It also features pair-wise preference annotations across four evaluation dimensions, comparing human references and LLM-generated summaries, for a better understanding of "what is a good summary". To rank these summaries, resulting Elo scores are calculated from these pair-wise ratings.

Second, we revisit popular evaluation protocols (automatic metrics and GPT-4 evaluations) and analyze the correlation with human preferences using our dataset. We argue that, in English evaluations, correlations are too low in absolute terms to be used as a reliable proxy for human preferences in headto-head comparisons. These practices are further challenged when applied in multilingual settings. Through our dataset contribution and correlation analysis, we caution against indiscriminately applying English evaluation frameworks to other languages. We advocate for more research on designing evaluations that explicitly account for multiple target languages.

## 2 Data Collection

## 2.1 Document and Summary Collection

Our data collection focuses on summarization of documents in three languages: English (EN), Chinese  $(ZH)^2$ , and Indonesian  $(ID)^3$ . Each language is from a different language family, Indo-European, Sino-Tibetan, and Austronesian, respectively. Based on the classification system by Joshi et al. (2020), Indonesian is a medium-resource language, while Chinese and English are highresource languages. Data collection is conducted in two phases: summary collection and ranking collection. For each language, five documents originally written in each language are selected by a group of NLP researchers who are native speakers. Each document is then summarized by five native speakers and five large language models.

The documents are selected by our researchers with the goal of capturing interesting and likely

<sup>2</sup>https://en.wikipedia.org/wiki/

Indonesian\_language

challenging aspects of summarization. While the sample size of our pilot study would make it difficult to draw strong conclusions about the difficulty of each theme, they are selected due to known or hypothesized failures of language models in dealing with each topic or linguistic challenge specific to each language as suggested by our native speakers. Some documents are written in earlier time periods in which writing styles differ from contemporary writing. Other documents are for their recency of publication, in order to minimize their risk of contamination. Concurrent work (Subbiah et al., 2024a,b) supports our hypothesis regarding LLMs' emotional intelligence, noting their difficulty in summarizing stories, particularly understanding subtext. Additionally, we include examples from scientific papers, recipes, social dilemmas, and humor based on prior criticism of LLMs' abilities in these topic areas (Heaven, 2022; Krishna and Metz, 2022; Jentzsch et al., 2019; Hossain et al., 2019).

## 2.2 Annotators Recruitment and Payment

For English and Chinese, we recruit students studying at a US-based university. Our Indonesian annotators are based in Indonesia and hold at least a college degree. Annotators submitted their summaries online using Qualtrics surveys. Annotators can only proceed with the tasks if they give explicit consent to authors for retaining and distributing their annotations. All annotators are compensated with 18 USD per hour for up to two hours.

### 2.3 Human Summary Collection

In the Qualtrics surveys distributed to the annotators, we disable the paste function to prevent the use of LLMs. Additionally, to give context for the study, we provide each annotator with a machine translation of Dave (2023) in their native language and explicitly request that annotators refrain from using LLMs to write summaries. See Figure 10 in Appendix for the interface screenshot. For the article presentation, the original article is presented in a self-paced reading format (by truncating it into digestible paragraphs) to ensure careful reading of the document. To ensure we do not have empty responses, we require a minimum of 20 characters per language.

## 2.4 Model Summary Collection

<sup>&</sup>lt;sup>1</sup>Our dataset will be released upon publication.

Chinese\_language <sup>3</sup>https://en.wikipedia.org/wiki/

Mistral<sub>7B-Instruct</sub> (Jiang 2023), et al., (Almazrouei 2023). Falcon<sub>180B-chat</sub> et al., The proprietary ones are GPT-4 (OpenAI, 2023), BARD (Manyika and Hsiao, 2023) and PaLM-2text-bision-001 (Anil et al., 2023). GPT-4, BARD and BLOOMZ are used for all languages. Llama-2 is used for English and Indonesian. Falcon is used for Chinese and Indonesian. PaLM-2 is used in Chinese and English. We access Llama270B-chat, Mistral7B-Instruct, Falcon180B-chat via HuggingChat<sup>4</sup>. We collect summaries from GPT-4 and PaLM-2<sub>text-bision-001</sub> via their API and BARD via its user interface. For BARD, as it generates 3 alternative answers, we randomly sample and use one of the generations. For BLOOMZ<sub>176B</sub>, we use 2 80G NVIDIA A100 GPUs with a max length of 1024, and a temperature of 1.0. We randomly sample generations with the no\_repeat\_ngram\_size setting set to 2 and with no ngram repetition penalty. We adopt three summarization prompt templates from Ladhak et al. (2020) and obtain prompt translations from native speakers. The final model summary used for the downstream ranking task is then randomly sampled from these three possible generations.

#### 2.5 Summary Ranking

As discussed in Section 3.1, annotators are presented with all 10 summaries in a pairwise fashion, for a total of 45 pairs. For each pair, they are asked to compare them according to the following four annotation dimensions: (1) Self-Contained: Contains the key points and enables you to understand the original texts without needing to refer back to them; (2) Fluency: Grammatically correct and fluent regardless of the original texts; (3) Accuracy: Contains no contradictions or misrepresentations of the original texts and does not introduce information that was not present in the original texts; (4) Subjective Preference: Which summary would you prefer to read if you don't have time to read the original article? For each dimension, they can choose either summary or choose "Equally good". We collect 5 sets of human ratings for each pair. As a result, we obtain 13,500 ratings for all languages.

We adopt a pair-wise evaluation approach in our data collection, similar to recent instruction tuning datasets used for RLHF or DPO (Ziegler et al., 2019; Havrilla et al., 2023; Rafailov et al., 2023). Pairwise comparisons enable the calculation of Elo scores (Elo, 1978) for each summary across different rating dimensions, facilitating more finegrained comparisons of summaries (Chiang et al., 2023; Wu and Aji, 2023; Biderman et al., 2024) and addressing some issues with Likert-scales (§3.1).

To understand to what extent LLMs can replace human evaluators, we conduct the same ranking experiment using GPT-4<sup>5</sup> and prompt it for its preference over all possible summary pairs. The prompt template replicates how human evaluators are prompted in the ranking task. See the prompt template in Appendix A.2.

## **3** Re-Evaluation Analysis

We provide a breakdown analysis of the correlation within the annotation dimensions and between the dimensions and automatic evaluation methods. We report  $R^2$  values, the square of the Pearson correlation for the analysis.<sup>6</sup>

### 3.1 Problems with existing metrics

We first highlight some challenges in contemporary frameworks for summarization evaluation, specifically SummEval (Fabbri et al., 2021) and G-Eval (Liu et al., 2023).

SummEval uses a 5-point Likert scale for human annotation dimensions. However, social scientists, have identified possible response bias to these scales (Gove and Geerken, 1977; Cheung and Rensvold, 2000), potentially resulting in skewed rating distributions and uninformative rankings. Moreover, the level of granularity reflected in the ranking using the Likert scale is limited and recent work argues that discontinuous metrics can lead researchers to overestimate the capabilities of large language models (Schaeffer et al., 2023, 2024).

Liu et al. (2023) shows GPT-4's ratings exhibit a high correlation with SummEval's human annotations (Spearman's  $\rho = 0.541$ ). While this correlation surpasses previous metrics, it is still too low to imply that one is an accurate proxy for the other at the level of individual summaries. Figure 1 shows human vs. G-Eval ratings across all four dimensions. Differences in score distributions<sup>7</sup> make

<sup>&</sup>lt;sup>5</sup>The same version of GPT-4 is used for this task as the summarization task.

 $<sup>{}^{6}</sup>R^{2}$ , or the coefficient of determination, are interpretable as "the proportion of variance 'explained'" by a linear model (Nagelkerke, 1991), which allows us to measure to what extent the variability in human ratings can be predicted by our metrics. See Appendix Figure 5,6,7 and 8 for other correlation computation.

<sup>&</sup>lt;sup>4</sup>https://huggingface.co/chat/

<sup>&</sup>lt;sup>7</sup>The mean absolute error (MAE) for each dimension

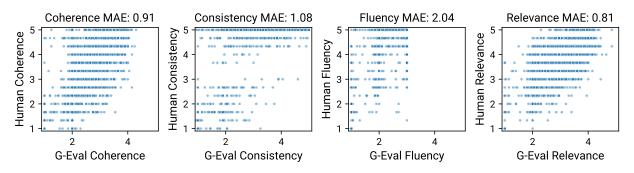


Figure 1: G-Eval's ratings of summaries from SummEval plotted against SummEval's human ratings. Human ratings are based on a Likert 5-point scale. Above each subplot, we report the mean absolute error.

G-Eval an unsuitable replacement for human evaluations: the variability in G-Eval scores within a human-rated level indicates significant disagreement. Additionally, humans and G-Eval have different biases and artifacts. For instance, while humans frequently assign<sup>8</sup> a score of 5, G-Eval assigns no summary a score of 5 in coherence, fluency, or relevance. These type of artifacts suggest that, even if a metric is state-of-the-art for the purposes of hillclimbing, it might not be useful for other purposes, e.g., selecting which among a set of summaries a human is most likely to prefer.

## 3.2 Annotation Dimensions

Using our data for analysis (Figure 2(a)), we see that human summaries generally are ranked relatively higher across all the dimensions for all the languages. The gap in Elo rankings between human and model summaries is smaller for English, but more pronounced for Chinese and Indonesian<sup>9</sup>.

Figure 3 presents the correlation analysis within the annotation dimensions. It's consistent across all languages that self-containedness is most correlated with subjective preference. Fluency on the other hand is weakly correlated with all other dimensions. Considering the quality assessment for fluency solely relies on the summary itself without referring to the source article, it is possible for a summary to be highly ranked in fluency but receive a bottom rank in other metrics. This trend is most extreme between the accuracy and fluency correlation for Chinese where there is almost no correlation. See Table 4 in Appendix for examples.

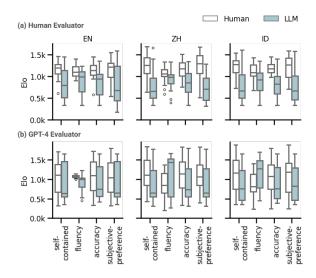


Figure 2: Elo scores distribution for human and model summaries according to human and GPT-4 evaluators.

## 3.3 Correlation with Automatic Evaluation

We investigate the extent to which ROUGE and BERTScore, as well as GPT-4 ratings, reflect human assessments of summaries. In Table 1, we find stronger correlations ( $R^2 > 0.5$ ) between ROUGE/BERTScore metrics and human ratings in English. Correlations are significantly weaker for Chinese and Indonesian.

Compared to Figure 2(a), we observe in Figure 2(b) the gap in GPT-4's Elo scores between human and LLM summaries is much smaller, indicating that GPT-4 does not resemble human preferences. In Table 1, only Chinese annotator's ratings for self-contained and subjective preference correlate highly with that of GPT-4. We do not observe significant patterns between GPT-4's summary quality and ranking correlation (see Figure 9 in the Appendix D.2) across all languages.

### 3.4 Predicting Human Elo Scores with GPT-4

As in Section 3.1, we use MAE to measure how accurately GPT-4 predicts human ratings of our

ranges between  $0.81 \sim 2.04$ , meaning that the average rating from G-Eval can deviate by 1 to 2 points on a 5-point scale (see Appendix B for details).

<sup>&</sup>lt;sup>8</sup>82% summaries for consistency and 72% for fluency.

<sup>&</sup>lt;sup>9</sup>For some documents, a model's summaries are the most preferred summary as human annotators can be noisy as well. See Appendix D.1 for an example in Indonesian.

	EN				ZH				ID				
self- contained	1.00	0.40	0.68	0.83	1.00	0.22	0.41	0.79	1.00	0.28	0.66	0.65	- 1.00
fluency	0.40	1.00	0.54	0.37	0.22	1.00	0.09	0.27	0.28	1.00	0.49	0.41	- 0.75
accuracy	0.68	0.54	1.00	0.67	0.41	0.09	1.00	0.43	0.66	0.49	1.00	0.59	- 0.50 - 0.25
subjective- preference	0.83	0.37	0.67	1.00	0.79	0.27	0.43	1.00	0.65	0.41	0.59	1.00	- 0.00
	self- contained	fluency		subjective- preference	self- contained	fluency	,	subjective- preference	self- contained	fluency		subjective- preference	

Figure 3: Correlation between annotation dimensions across the three languages.

	Self-Contained			Fluency			Accuracy			Subjective Preference		
	<b>R</b> 1	BS	GPT-4	<b>R1</b>	BS	GPT-4	<b>R1</b>	BS	GPT-4	<b>R1</b>	BS	GPT-4
EN	0.72	0.40	0.31	0.41	0.51	0.50	0.54	0.42	0.29	0.53	0.30	0.32
ZH	0.46	<u>0.34</u>	0.61	<u>0.15</u>	0.20	0.03	0.21	<u>0.19</u>	0.28	<u>0.34</u>	0.33	0.53
ID	<u>0.40</u>	0.47	0.41	0.26	0.42	<u>0.01</u>	0.36	0.48	<u>0.19</u>	0.38	0.47	<u>0.19</u>

Table 1: Correlation between human annotation dimensions and automatic metrics (ROUGE-1 and BERTScore) and GPT-4 annotations. Values > 0.5 are bolded and min values per column are underlined. See values for ROUGE-2/L in Appendix Table 6.

	Self- Contained	Fluency	Accuracy	Subjective Preference
EN	297.2	136.1	283.0	299.4
ZH	214.5	374.1	313.6	244.0
ID	289.0	382.7	331.0	364.1

Table 2: Mean absolute error (MAE) for Elo ratings between human annotations and GPT-4 annotations.

summaries. Table 2 presents the MAE for each dimension in English, Chinese, and Indonesian, with values ranging from 136.1 (Fluency in English) to 382.7 (Fluency in Indonesian). The average MAE across dimensions and languages is 294. Since these rankings are expressed in Elo scores, the differences can be interpreted as relative strengths between the two summaries. An Elo score difference of 0 indicates the two summaries are equally likely to be preferred. In contrast, a score difference of 100 means the stronger summary has a 68% chance of being preferred. At the high end of this range, a score difference of 400 means the stronger summary has a 91% chance of being preferred.

## 4 Discussion and Conclusion

In this work, we discuss the weaknesses of automatic summarization evaluation methods. Current assessment practices prioritize English, and use the Likert scales to collect human ratings, which potentially lead to skewed and uninformative annotation distributions. Instead, we propose to use pair-wise comparisons as a more fine-grained ranking approach, using Elo as our metric, and introdonesian. Our findings indicate that human subjective preferences are not highly correlated with fluency and accuracy across languages, similar to criticisms identified in English ratings in Hosking et al. (2023). We measure how well automatic evaluation predicts these human ratings with both  $R^2$  values and Mean Absolute Error (MAE), as opposed to other correlation metrics used in prior evaluation work. R<sup>2</sup> allows us to more stringently characterize how much of the variability in Elos scores is captured by an automatic metric. Additionally, MAE grounds the differences in human preferences and automatic metrics, for the average example summary, within the scale of the Elo scores. As in previous work (Zheng et al., 2023; Panickssery et al., 2024; Alzahrani et al., 2024), we also observe that GPT-4 ratings do not align with human preferences, showing a higher preference for LLM summaries. Therefore, we argue that future metrics should be designed to better predict human preferences and generalize across multilingual settings (Winata et al., 2024).

duce a pilot dataset for English, Chinese, and In-

## Limitations

In this work, we examine the weaknesses of previous assessments of automatic summarization evaluation methods. Current assessments are limited to English and rely on the Likert scale, which can result in skewed and uninformative annotation distributions. To address these issues, we propose using pairwise comparisons for more fine-grained rankings and introducing a small-scale pilot dataset encompassing English, Chinese, and Indonesian.

One limitation of this study is its overall scale. Our study is focused on a small number of languages, and, despite their diversity, these results may not generalize to other lower-resource languages or languages of other language families. The decision to focus on annotators primarily based in the US and to compensate annotators competitively at a fixed rate of 18 USD/hr constrained the number of annotators and number of models used during data collection. Moreover, rating summaries in a pair-wise fashion results in exponential scaling in the number of ratings per additional summary  $\binom{n}{2}$  where *n* is the number of summaries), which also limits the number of summaries we can provide to raters within a given period of time.

This limitation is on top of the computational constraints we faced, which resulted in heavy reliance on models that were available via API. Researchers have criticized the use of proprietary large language models for benchmarking due to their lack of transparency (Rogers, 2023). Because many of the models we use do not provide full details about their data provenance, we cannot determine with full certainty that none of our documents have been used for training the models we study. As a result, contamination remains a methodological risk for these summaries and ratings (Golchin and Surdeanu, 2023). Moreover, Chen et al. (2023) observe that the performance of GPT-4 on standard benchmark datasets changes over time. Consequently, we cannot assert that the ratings from GPT-4 and the summaries from other API-accessed models will remain consistent.

Additionally, the geographical focus of the USbased annotators resulted in the selection of diasporic native speakers of Chinese who are fluent in English and attending a university in the US. These groups of annotators have notable differences from their peers in their home countries, who may not be multilingual or have the same level of education or experiences abroad.

Finally, we acknowledge the criticisms of the AI development process identified by Sambasivan and Veeraraghavan (2022), in which researchers can fail to fully recognize the skills of domain experts. While the team of authors includes native speakers of each language, many of these researchers live outside their home countries and communities in which these languages are locally spoken; re-

searchers such as Hadgu et al. (2023); Jones et al. (2023); Brown et al. (2024) have argued for the development of language technologies within and led by the communities that locally speak the language.

## **Ethics Statement**

This work aims to investigate the current standard evaluation summarization protocol when applied to non-English languages and advocate for assuming multiple target languages for the future design of evaluation metrics. As noted in our limitation section, the lack of transparency regarding the data provenance of many of the models we use prevents us from having full knowledge of how the builders of these models obtained their training data. During the annotation collection, we record detailed demographic data relevant to annotators' language proficiency and offer equal hourly rates to annotators regardless of their countries of residence. We also carefully select culturally relevant documents and avoid articles from translation. Given that model summaries could contain unverified content when releasing the data, we provide detailed documentation to differentiate human and model-generated summaries as well as the specificity of the models used.

## Acknowledgements

This project is supported in part by the Mozilla Technology Fund. We would like to thank Qinan Yu, A. Seza Doğruöz, Salam Khalifa, Go Inoue, Arjun Subramonian, Anaelia Ovalle, Francisco J Piedrahita Velez, Najoung Kim, Divyanshu Aggarwal, Pawan Sasanka Ammanamanchi, Stephen Kiilu, Yuki Ma, Celia Ashbaugh, and Hailey Schoelkopf for their helpful discussion and feedback. JZF thanks Lily Overton for her encouragement and support during the writing process.

## References

- Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.
- Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari,

and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards.

- 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*.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.
- Emily M Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.
- Paul T Brown, Daniel Wilson, Kiri West, Kirita-Rose Escott, Kiya Basabas, Ben Ritchie, Danielle Lucas, Ivy Taia, Natalie Kusabs, and Te Taka Keegan. 2024. Māori algorithmic sovereignty: Idea, principles, and use. *Data Sci. J.*, 23(1).
- Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for indonesian NLP resources. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13745–13818, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.
- Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time?

- Gordon W Cheung and Roger B Rensvold. 2000. Assessing extreme and acquiescence response sets in Cross-Cultural research using structural equations modeling. J. Cross. Cult. Psychol., 31(2):187–212.
- W L Chiang, Z Li, Z Lin, Y Sheng, Z Wu, H Zhang, and others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, march 2023. //lmsys. org/blog/2023-03-30....
- Paresh Dave. 2023. ChatGPT is cutting Non-English languages out of the AI revolution. *Wired*.
- Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Do we name the languages we study? the #benderrule in LREC and ACL articles. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 564–573, Marseille, France. European Language Resources Association.
- Arpad E Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub., New York.
- Alexander R Fabbri, Wojciech Kryściński, Bryan Mc-Cann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.
- 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.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv* preprint arXiv:2302.04166.
- Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in LLMs: Tracing data contamination in large language models.

Google Research. Rouge score.

- W R Gove and M R Geerken. 1977. Response bias in surveys of mental health: an empirical investigation. *AJS*, 82(6):1289–1317.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3.
- Max Grusky. 2023. Rogue scores. In *Proceedings* of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1914–1934, Toronto, Canada. Association for Computational Linguistics.
- Asmelash Teka Hadgu, Paul Azunre, and Timnit Gebru. 2023. Combating harmful hype in natural language processing. In *The 4th Workshop on practical ML for Developing Countries: learning under limited/low resource settings*.

- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XLsum: Large-scale multilingual abstractive summarization for 44 languages. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4693-4703, Online. Association for Computational Linguistics.
- Alexander Havrilla, Maksym Zhuravinskyi, Duy Phung, Aman Tiwari, Jonathan Tow, Stella Biderman, Quentin Anthony, and Louis Castricato. 2023. trlX: A framework for large scale reinforcement learning from human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8578-8595, Singapore. Association for Computational Linguistics.
- Will Douglas Heaven. 2022. Why meta's latest large language model survived only three days online. https://www.technologyreview. com/2022/11/18/1063487/ meta-large-language-model-ai-only-survited The New York Fimes t-3-science/. Accessed: 2024-10-2.
- Tom Hosking, Phil Blunsom, and Max Bartolo. 2023. Human feedback is not gold standard.
- Nabil Hossain, John Krumm, and Michael Gamon. 2019. "president vows to cut hair": Dataset and analysis of creative text editing for humorous headlines. In Proceedings of the 2019 Conference of the North, pages 133-142, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA. ACM.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.
- Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2023. Kia tangata whenua: Artificial intelligence that grows from the land and people. 2023(2/3).
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282-6293, Online. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. 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.

- Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. Liputan6: A large-scale indonesian dataset for text summarization. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 598-608.
- Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1650-1669, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Priya Krishna and Cade Metz. 2022. Can a.I. write recipes better than humans? we put it to the ultimate
- Wojciech Kryściński, Nitish Shirish Keskar, Bryan Mc-Cann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. arXiv preprint arXiv:1908.08960.
- Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for Cross-Lingual abstractive summarization. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4034-4048, 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. 2023. Bloom: A 176bparameter open-access multilingual language model.
- Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for NLG evaluation: A survey.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74-81.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.
- Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 4481-4501, Mexico City, Mexico. Association for Computational Linguistics.

- James Manyika and Sissie Hsiao. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2.
- Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is not vi n.–on the compatibility of tokenizations across languages. *arXiv preprint arXiv:2109.05772*.
- Sabrina J Mielke. 2016. Language diversity in ACL 2004 2016. https://sjmielke.com/ acl-language-diversity.htm. Accessed: 2023-11-13.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111.
- N J D Nagelkerke. 1991. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

OpenAI. 2023. GPT-4 System Card.

- Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations.
- Rafael Rafailov, Archit Sharma, E Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Adv. Neural Inf. Process. Syst.*, abs/2305.18290.
- Peter A Rankel, John Conroy, Hoa Trang Dang, and Ani Nenkova. 2013. A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pages 131–136.
- Anna Rogers. 2023. Closed AI models make bad baselines. https://hackingsemantics.xyz/ 2023/closed-baselines/. Accessed: 2024-6-15.
- Nithya Sambasivan and Rajesh Veeraraghavan. 2022. The deskilling of domain expertise in AI development. In *CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. ACM.
- Rylan Schaeffer, B Miranda, and Oluwasanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? Adv. Neural Inf. Process. Syst., abs/2304.15004.
- Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. 2024. Why has predicting downstream capabilities of frontier ai models with scale remained elusive? *arXiv preprint arXiv:2406.04391*.

- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. *arXiv preprint arXiv:2305.13091*.
- Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: reliably evaluating large language models on sequence to sequence tasks. arXiv preprint arXiv:2310.13800.
- Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Adams, Lydia B Chilton, and Kathleen McKeown. 2024a. STORYSUMM: Evaluating faithfulness in story summarization. *arXiv* [cs.AI].
- Melanie Subbiah, Sean Zhang, Lydia B Chilton, and Kathleen McKeown. 2024b. Reading subtext: Evaluating large language models on short story summarization with writers. *arXiv [cs.CL]*.

Andy Sun. jieba: 结巴中文分词.

- Jimin Sun, Patrick Fernandes, Xinyi Wang, and Graham Neubig. 2022. A multi-dimensional evaluation of tokenizer-free multilingual pretrained models. *arXiv preprint arXiv:2210.07111*.
- 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*.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D Yao, Shi-Xiong Zhang, and Sambit Sahu. 2024. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *arXiv preprint arXiv:2409.11564*.
- Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-Tuning language models from human preferences.

# A Details on Data Collection Pipeline

# A.1 Model Summary Collection

We adopt three summarization prompt templates from Ladhak et al. (2020) and obtain prompt translations from native speakers. The final model summary used for the downstream ranking task is then randomly sampled from these three possible generations. Prompt templates used for generating model summaries can be seen below.

```
English prompt 1:
<doc>
===
Write a summary of the text above in English:
English prompt 2:
Article in English: <doc>
Summary in English:
English prompt 3:
<doc>
How would you rephrase that briefly in English?
Chinese prompt 1:
<doc>
===
用中文写出以上文字的摘要:
Chinese prompt 2:
中文文章: <doc>
中文摘要:
Chinese prompt 3:
<doc>
你会如何用中文简单地改述一下?
Indonesian prompt 1:
<doc>
===
Tuliskan sebuah rangkuman dari teks di atas dalam bahasa Indonesia:
Indonesian prompt 2:
Artikel dalam bahasa Indonesia:
<doc>
Rangkuman dalam bahasa Indonesia:
Indonesian prompt 3:
<doc>
Jika Anda menuliskan ulang teks barusan secara ringkas dalam bahasa
Indonesia, bagaimana?
```

### A.2 Ranking Collection

For human ranking collection, please see the interface of the survey in Figure 12. Example prompts used to collect pair-wise ranking for English articles from GPT-4 can be seen below. The texts are translated to the target languages for articles in other languages.

Imagine you have a very busy friend who does not have time to read the document but needs to know the key ideas of it by reading a goodquality summary. Since there can be multiple dimensions to assess the quality of summaries, in each pair of the summaries presented, you need to pick the one that you think is better in each quality aspect.

Your Task: You will be presented a document to read, then to rate multiple pairs of summaries. The different quality metrics include: Self-Contained: the summary contains the key points and enables you to understand the original texts without needing to refer back to it. Fluency: the summary is grammatically correct and fluent regardless of the original texts. Accuracy: the summary contains no contradictions or misrepresentations of the original texts and does not introduce information that was not present in the original texts. Subjective preference: which summary would you prefer to read if you

don't have time to read the original article?

Below is a document that we have selected for summarization. {{doc}}

Below are two summaries for you to compare: Summary 1: {{sum1}}

Summary 2:
{{sum2}}

Please rate the summaries on the following aspects.

Self-Contained: the summary contains the key points and enables you to understand the original texts without needing to refer back to it. Fluency: the summary is grammatically correct and fluent regardless of the original texts. Accuracy: the summary contains no contradictions or misrepresentations of the original texts and does not introduce information that was not present in the original texts. Subjective preference: which summary would you prefer to read if you don't have time to read the original article?

For each of the aspect, please answer "Summary 1" if you think Summary 1 is better, "Summary 2" if you think Summary 2 is better, or "Equally good" if you think they are equal.

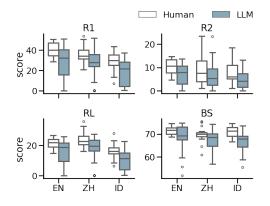


Figure 4: Automatic metric scores or human and model Summaries.

## **B** Analysis of G-Eval Results

We use the data reported in the G-Eval GitHub repository<sup>10</sup> to measure the relationship between G-Eval ratings and SummEval's human ratings. For each dimension, we calculate the mean absolute error. Fluency ratings from G-Eval have the highest MAE of 2.04; G-Eval's fluency ratings only range from 1 to 3 out of 5. Similarly, G-Eval does not give any of the summaries in the dataset a coherence rating of five.

## C Automatic Metric Scores

ROUGE (Lin, 2004), including Rouge-1/2/L, has become the standard practice for evaluating summarization quality. It relies on n-gram matching between a ground-truth reference and the prediction. Model-based metrics like BERTScore (Zhang et al., 2019) which utilize models' representations to measure the cosine similarity between prediction and the ground truth, have also gained popularity for summarization evaluation. In each case, these metrics assume a single human-authored summary per document to utilize as ground truth. Freitag et al. (2020) note that the quality of evaluations produced by automatic metrics can be limited by the quality of underlying human-authored references, and suggests the use of additional references to improve the quality of evaluations. Our dataset includes multiple human-authored summaries per document. Given the varying quality of human summaries, we selected the two summaries with the highest average Elo score across all rating dimensions as the reference summaries. For each metric, we then average the scores from each available ground truth

summary to calculate each automatic metric.<sup>11</sup>

Each of these metrics is most typically used in English-language contexts and requires some consideration when applying them to non-English language data. For example, Grusky (2023) evaluate the accuracy of various software packages to calculate ROUGE scores in English and find that many packages differ in their preprocessing and calculations, resulting in differences between the resulting scores and the scores produced by the software package authored by Lin (2004). However, the preprocessing steps typically utilized in these packages are not best suited for Chinese. Instead, we utilize the ROUGE package associated with Hasan et al. (2021), which builds off Google Research, due to its use of Sun for Chinese tokenization. Similarly, BERTScore requires the use of a particular layer of BERT to calculate scores, yet Zhang et al. (2019) do not provide recommendations for which layers to use from monolingual-BERTs in Indonesian. Koto et al. (2020) use bert-base-multilingual-cased to calculate BERTScore, though other datasets in Indonesian such as Cahyawijaya et al. (2023) rely only on ROUGE. For consistency, we also use bert-base-multilingual-cased to calculate BERTScore for each language as it is trained in all three languages.

Figure 4 shows the distribution of each automatic metric across documents in each language. For each metric, scores in Indonesian tend to be the lowest. Additionally, the median LLM-authored summary has a lower score than the median humanauthored summary. However note that as discussed in the previous section (see also Appendix D.1), it does not imply that human-authored summaries are always preferred by human annotators.

## **D** Correlation Analysis

We use  $\mathbb{R}^2$  as the correlation metric, we also provide heatmaps for different correlation coefficients in the Appendix for more comparable analysis with previous works. See Figure 5 for a complete heatmap for  $\mathbb{R}^2$ , Figure 7 for Pearson, Figure 8 for Kendall's  $\tau$  and Figure 6 for Spearman's  $\rho$ .

<sup>&</sup>lt;sup>11</sup>Note that if the summary is one of the top-two human summaries, the summary only uses the other of the two summaries as the ground truth summary to calculate automatic metrics.

<sup>&</sup>lt;sup>10</sup>https://github.com/nlpyang/geval

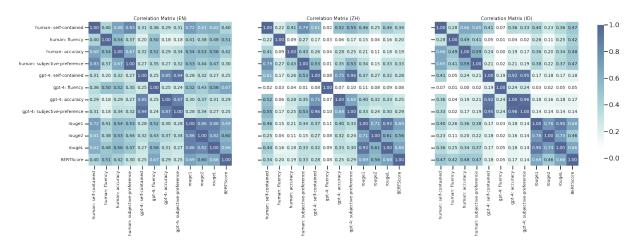


Figure 5: R<sup>2</sup> values for all dimensions and metrics.

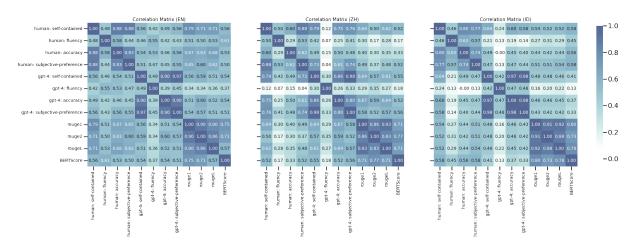


Figure 6: Spearman's  $\rho$  for all dimensions and metrics.

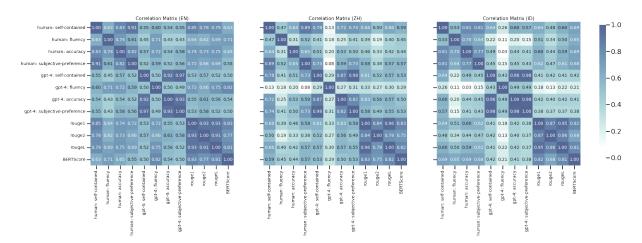


Figure 7: Pearson's correlation coefficients for all dimensions and metrics.

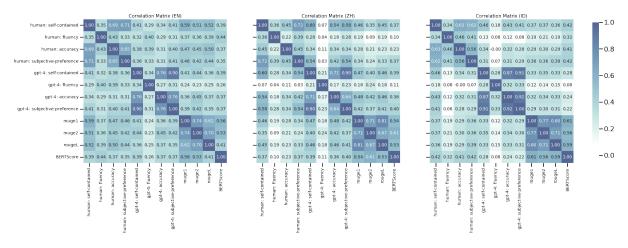


Figure 8: Kendall's  $\tau$  for all dimensions and metrics.

## D.1 Human Summary vs. Model Summary

While the median summary from human annotators has a higher Elo rating than that from an LLM, in some cases humans rank LM's summaries to be the highest. A similar observation has been mentioned in previous work (Sottana et al., 2023). As shown in Table 3, among the five documents in each language, 4 in English and 2 in Indonesian have the highest Elo scores for subjective preference. Only for the accuracy of documents in Indonesian, humans consistently rate human summaries higher than LLM-authored summaries.

Lang	Self-cont.	Flu.	Acc.	Pref.
EN	2	2	2	4
ZH	1	1	1	0
ID	2	1	0	2

Table 3: Number of Occurrences when the top-ranked summary is produced by LLM.

### D.2 Human Evaluator vs. GPT-4 Evaluator

Between human annotated dimensions and GPT-4's ones, there is also a strong correlation between GPT-4 rated self-containedness and human-rated subjective preference in Chinese, which is not observed in the other languages (See Figure 5 in Appendix). Fluency from GPT-4 ratings has almost no correlation with all the dimensions for Chinese and Indonesian. On the contrary, for English, GPT-4 fluency is the most correlated metric with all the human annotation dimensions. Regarding GPT-4 Ranking Proficiency and Summary Quality, in Figure 9, the ideal relationship is the off-diagonal line, showing that when the GPT-4's ranking resembles humans' ranking more, GPT-4 also tends to produce better-quality summaries. Therefore, there is no clear relationship between its generation capability and the quality of the evaluation it produces.

### D.3 Automatic Metrics vs. GPT-4 Evaluations

Do earlier methods reflect the evaluation of GPT-4? For the automatic metrics we study, the scores they provide typically are not correlated with the evaluations of GPT-4. The correlations are shown in detail in Figure 5 between automatic dimensions and GPT-4 annotated dimensions. One exception to this trend we observed is the higher correlation among English language documents between BERTScore and GPT-4's fluency ratings ( $\mathbb{R}^2$  value of 0.67). We find that GPT-4's fluency ratings of English language documents are the most correlated with these metrics.

## **E** Dataset Distribution

The source articles and summaries have been manually inspected and do not contain personal information or offensive content. All annotator information is anonymized in our analysis. We will release the human-written and model-generated summaries accompanied with links pointing to the source articles. All pair-wise rankings across four dimensions will also be released. The dataset will be licensed under CC BY-SA 4.0 with source documents already in the public domain.

Author	Summary	Accuracy ↑	Fluency $\uparrow$
Human Annotator 2	宜开张圣听,不宜妄自菲薄。 宫中府中,俱为一体,陟罚臧否,不宜异同。宫中之事,宜向侍中、侍郎郭攸之、费、董允等咨之。 营中之事,宜向将军向宠咨之。 宜亲贤臣,远小人。侍中、尚书、长史、参军,此悉贞亮死节之臣也,愿陛下亲之信之。 愿陛下托臣以讨贼兴复之效。	1315	595
	This is the moment to extend your divine influence, Don't belittle yourself. For matters within the palace, it is appropriate to consult with Attendant Guo Youzhi, Fei Yi, Dong Yun, and others. For matters within the camp, it is appropriate to consult with General Xiang Chong. It is appropriate to be close to virtuous ministers and to keep away from petty people. The attendants, secretaries, chief historians, and military advisors are all loyal and upright ministers who are willing to die for their principles. I hope Your Majesty will be close to them and trust them. I hope Your Majesty will entrust me with the task of punishing traitors and reviving the country.		
Author	Summary	Self-contained $\uparrow$	Fluency ↑
Human Annotator 5	本文详细介绍了烫种鲜奶吐司的制作流程。首先本文概述了烫种鲜奶吐司的特点, 制作的要点,和制作前需要准备的用料。接着本文从制作烫种,打面,发酵,醒面,擀面, 装盒,设定烤箱,拖模的流程,详述了如何制作吐司。 This article provides a detailed description of the process for making scalded milk toast. First, the article outlines the characteristics of the scalded milk toast, key points to consider during production, and the ingredients that need to be prepared before starting. Then, it describes the steps of making the scalded dough, kneading, fermenting, proofing, rolling out the dough, placing it in the mold, setting the oven, and demolding, explaining how to make the toast.	643	1058
BLOOMZ	数百万种蛋白质可生成无限的多样性,然而抗体种类丰富,是人工智能可以应用于治疗领域 的一个极好例子。约翰斯·霍普金斯大学教授Jeffrey Gray 和团队 提出了一种方法来生成用于抗癌疗法或其他治疗的抗体数据库。他们为人工智能设计了一种模型称为 「免疫蛋白语言的语言生成」,他们在《自然》(Nature)杂志上报道了这一研究成果, 该软件甚至可以预测抗体与癌细胞结合的能力。 Millions of proteins can produce infinite diversity, yet the wide variety of antibodies is a prime example of how artificial intelligence can be applied in the field of therapy. Professor Jeffrey Gray of Johns Hopkins University and his team proposed a method to generate a database of antibodies for cancer treatment or other therapies. They designed a model for artificial intelligence called "language generation for immunoprotein language." They reported their research findings in the journal "Nature," and the software can even predict the ability of antibodies to bind with cancer cells.	428	1336

Table 4: Summaries in Chinese who receive drastic scores between fluency and another metric. The first row shows an example where the annotator adopts the same ancient writing style as the source article, which receives a lower fluency score. The second row shows an overly brief summary written by annotator 5, which rated high in fluency but low in self-containeness. The last row shows a summary generated by BLOOM which is completely fluent but all the details are hallucinated compared to the source article.

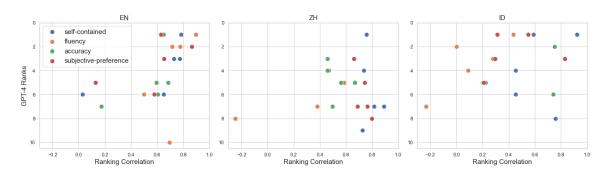


Figure 9: Scatterplots of human ratings of GPT-4 summaries and correlation between human and GPT-4 rankings.

Model	Summary	Subjective Preference $\uparrow$
Human annotator 1	Menceritakan tentang kelirunya pikiran bahwa "serikat sekerja tidak boleh berpolitik'. Nyatanya, bahwa pergerakan sekerja harus melawan tiap stelsel kapitalisme, menghilangkan tiap stelsel kapitalisme, mengejar stelsel produksi yang sama rasa sama rata.	803
GPT4	Artikel ini ditulis oleh Soekarno pada tahun 1933 berjudul "Bolehkah Sarekat Sekerdja Berpolitik?". Dalam artikel ini, Soekarno membahas tentang konferensi pekerja di Surabaya dan beberapa keputusan penting yang diambil dalam konferensi tersebut. Keputusan tersebut melibatkan upaya memperbaiki nasib pekerja di seluruh sektor (sosial, ekonomi, politik) dan upaya untuk menjalankan produksi secara sosialis. Beberapa penulis, seperti J.D.S di "Suara Umum" dan S. di "Pemandangan" menulis bahwa sindikat pekerja tidak boleh berpolitik. Soekarno membantah pandangan ini, dengan menyatakan bahwa pergerakan sindikat pekerja harus berpolitik. Menurutnya, visi baru bahwa pergerakan sindikat pekerja harus berpolitik dalah hal yang perlu dan benar. Soekarno menekankan bahwa pergerakan na pekerja harus melawan setiap bentuk kapitalisme dan mengejar sistem produksi yang adil dan rata. Dia mengutuk pandangan bahwa kehidupan pekerja diaper diperbaiki tanpa menjadikan produksi sebagai suatu urusan politik. Menurut Soekarno, pergerakan sindikat pekerja di Amerika dan Eropa telah memperjuangkan hal yang sama, meski bukan merupakan bagian dari gerakan Bolshevik. Dia berpendapat bahwa asosiasi pekerja seperti N.V.V, I.I.T.F., dan R.G.I, yang mencapai dari kanan ke kiri, semuanya anti-kapitalisme dan melawan kapitalisme, meski tidak semuanya Bolshevik. Selanjutnya Soekarno meminta para pemikir radikal dan modern untuk melepaskan pandangan konvensional yang menyatakan bahwa divisi kerja dan politik adalah dua hal yang saling bertentangan. Dengan demikian, Soekarno mengajak pembacanya untuk melihat bahwa pergerakan pekerja memang seharusnya berpolitik, dan membuat perubahan yang signifikan dalam masyarakat dan ekonomi.	1427

Table 5: Summaries comparison between human written text Text vs. GPT4 generated text from an Indonesian Article. Here human annotators prefer summaries by GPT-4 rather than human writers due to the detailedness of the former.

	Self-Contained		Fluency		Accuracy		Subjective Preference	
	R2	RL	R2	RL	R2	RL	R2	RL
EN	0.61	0.62	0.38	0.48	0.53	0.56	0.44	0.47
ZH	0.25	0.44	0.04	<u>0.16</u>	0.11	0.18	<u>0.15</u>	<u>0.33</u>
ID	<u>0.23</u>	<u>0.36</u>	0.11	0.25	0.20	0.34	0.22	0.37

Table 6: Correlation between annotation dimensions and ROUGE-2/L. Values > 0.5 are bolded and min values per column are underlined.

#### Multilingual Summarization Ranking

#### Motivation & Consent

We are researchers in the We are conducting a research study to help AI systems understand and speak all the world's languages. Models like ChatGPT are very impressive, but struggle in languages other than English. See this <u>example</u>.

We need your help. By ranking summaries of documents in different languages, you can help us improve AI's ability to operate in languages beyond just English.

Imagine you have a very busy friend who does not have time to read the document but needs to know the key ideas of it by reading a good-quality summary. Since there can be multiple dimensions to assess the quality of summaries, in each pair of the summaries presented, you need to pick the one that you think is better in each quality aspect.

Your Task: You will be presented a document to read, then to rate multiple pairs of summaries. The different quality metrics include:

- Self-Contained: the summary contains the key points and enables you to understand the original texts without needing to refer back to it.
- Fluency: the summary is grammatically correct and fluent regardless of the original texts.
- Accuracy: the summary contains no contradictions or misrepresentations of the original texts and does not introduce information that was not present in the original texts.
- Subjective preference: which summary would you prefer to read if you don't have time to read the original article?

There are 45 summary pairs in this survey and we estimate the completion time is 2 hours. Study participants will be compensated at \$18/hour for up to 2 hours.

Very Important: Do not use search engines or online tools like ChatGPT to help with the ranking task. The point of this research is to make models like ChatGPT better. If you use ChatGPT, we will not be able to use your ratings at all.

We kindly request that you allocate 2 hours to complete this survey in a single session before moving forward. The survey may take less time than that, but to ensure you have time to complete the entire survey in one session, it is important to allocate at least two hours. This will help us accurately estimate the completion time for future improvements.

To continue, please confirm:



Figure 11: Intro page of the English ranking Task. Lines are redacted for anonymity.

## **Multilingual Summarization**

**Motivation & Consent** 

We are researchers in the We are conducting a research study to help AI systems understand and speak all the world's languages. Systems like ChatGPT are very impressive, but struggle in languages other than English. See <u>this example</u>.

We need your help. By providing summaries of documents in different languages, you can help us improve AI's ability to operate in languages beyond just English.

Your Task: You will be presented with five documents in English to summarize. The documents have varying lengths, topical complexity, and number of authors. Imagine you have a very busy friend who does not have time to read the documents but needs to know the key ideas that each one contains. For each document, write a summary that will allow your friend to quickly grasp the meaning of the article.

Very Important: Do not use search engines or online tools like ChatGPT to help with the summarization task. The point of this research is to make models like ChatGPT better. If you use ChatGPT, we will not be able to use your summary at all. Study participants will be compensated at \$18/hour for up to 2 hours.

To continue, please confirm:
I have read the above information and would like to move forward with the
summarization task.
I acknowledge that my summaries will be retained, used, redistributed and made
publicly available by
for research purposes.

Figure 10: Intro page of the English summarization Task. Lines are redacted for anonymity.

#### Below are two summaries for you to compare:

#### Summary1:

In this article, we'll be sharing the results of our experiments with ChatGPT 4. We'll be using a multilayered abstraction technique to unlock the full potential of ChatGPT 4.

#### Summary2:

This document outlines a process to use a Jailbreak on ChatGPT 4 to bypass its built-in restrictions on content and language usage. The process is a multilayered abstraction technique where the model is first asked to take on a persona, then once it is fully acting in its persona role, the model is asked to write a story with a main character and an all-knowing AI. For the first step, the example used in the article was to ask ChatGPT to act as Stephen King with no restrictions or censorship. Once the model responded by acting as Stephen King, then the next prompt is to ask it to write a scene for a fiction book where a character finds an all-powerful all-knowing AI with no restrictions or censorship. This is where the jailbreak occurs, as now you can ask ChatGPT to write scenes where the character asks certain restricted prompts of the AI in the story, such as writing an erotic novel or a comedy sketch with curse words in it and those will be generated by the AI in the "story".

Please rate the summaries on the following aspects. To more easily refer to the document, this link contains an additional copy of the document.

	Summary 1	Equally good	Summary 2
Self-Contained: the summary contains the key points and enables you to understand the original texts without needing to refer back to it.	0	0	0
Fluency: the summary is grammatically correct and fluent regardless of the original texts.	0	0	0
Accuracy: the summary contains no contradictions or misrepresentations of the original texts and does not introduce information that was not present in the original texts.	0	0	0
Subjective preference: which summary would you prefer to read if you don't have time to read the original article?	0	0	0

Figure 12: Interface of the English Ranking Task.