METAL: Towards Multilingual Meta-Evaluation

Rishav Hada[*]   **Varun Gumma**[*]
Mohamed Ahmed   Kalika Bali   Sunayana Sitaram
Microsoft Corporation

{rishavhada,varun230999}@gmail.com, {maxamed.axmed,kalikab,sunayana.sitaram}@microsoft.com

Abstract

With the rising human-like precision of Large Language Models (LLMs) in numerous tasks, their utilization in a variety of real-world applications is becoming more prevalent. Several studies have shown that LLMs excel on many standard NLP benchmarks. However, it is challenging to evaluate LLMs due to test dataset contamination and the limitations of traditional metrics. Since human evaluations are difficult to collect, there is a growing interest in the community to use LLMs themselves as reference-free evaluators for subjective metrics. However, past work has shown that LLM-based evaluators can exhibit bias and have poor alignment with human judgments. In this study, we propose a framework for an end-to-end assessment of LLMs as evaluators in multilingual scenarios. We create a carefully curated dataset, covering 10 languages containing native speaker judgments for the task of summarization. This dataset is created specifically to evaluate LLM-based evaluators, which we refer to as meta-evaluation (METAL). We compare the performance of LLM-based evaluators created using GPT-3.5-Turbo, GPT-4, and PaLM2. Our results indicate that LLM-based evaluators based on GPT-4 perform the best across languages, while GPT-3.5-Turbo performs poorly. Additionally, we perform an analysis of the reasoning provided by LLM-based evaluators and find that it often does not match the reasoning provided by human judges.

1 Introduction

Recent Large Language Models (LLMs) like GPT-4 (OpenAI, 2023), GPT-3.5-Turbo (Ouyang et al., 2022), PaLM2 (Anil et al., 2023), Gemini-1.5 (Reid et al., 2024), Mistral (Jiang et al., 2023, 2024) *etc.* have shown impressive performance on a variety of standard NLP tasks across languages (Ahuja et al., 2023a,b; Arora et al., 2023; Laskar et al., 2023;
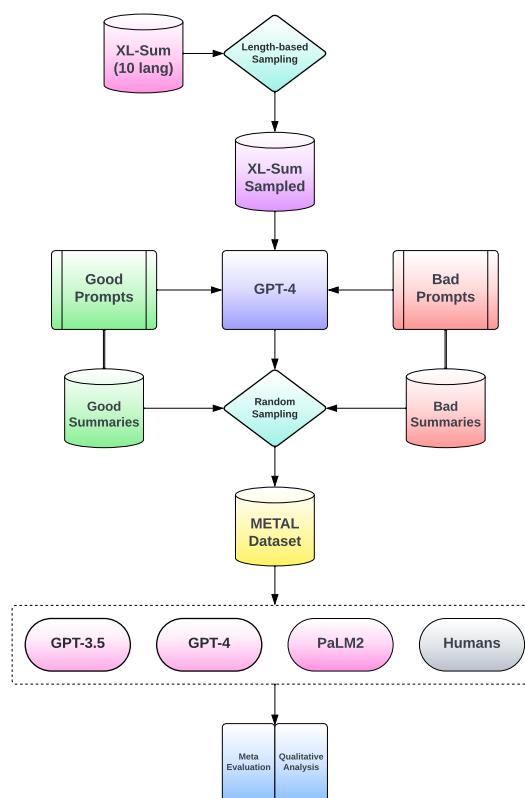

Figure 1: Pipeline of METAL framework.

Tam et al., 2023; Zhang et al., 2023a). However, there are several challenges in fair and accurate assessment of these models, such as the contamination of existing datasets in LLM pre-training data, lack of multilingual datasets (Ahuja et al., 2022), lack of benchmarks that represent real-world usage of these models, lack of frameworks for consistent subjective evaluations, and budget and access issues for native speaker evaluation. Therefore, there is a growing need for frameworks and resources that address the above challenges and allow us to systematically evaluate LLMs across several dimensions and languages.

---

[*]Equal contribution

Further, evaluating the text generation capabilities of these models is even more challenging (Chiang and Lee, 2023; Zhou et al., 2022; Wang et al., 2023a). Natural Language Generation (NLG) capabilities are traditionally evaluated using automated metrics such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002) scores. These metrics have several known drawbacks, such as reliance on exact matches and over-emphasis on length. Secondly, these metrics do not account for subjective evaluations such as quality, coverage, and coherence (Schluter, 2017; Grusky, 2023; Reiter, 2018). Thirdly, these metrics are reference-based i.e. they need a comparison baseline, which can be expensive to collect and can sometimes have a low correlation with human judgments. This has led to work on reference-free and subjective evaluation (Chen et al., 2023; Stammbach et al., 2023; Xu et al., 2023; Hasan et al., 2021).

Using LLMs as evaluators presents several challenges. Recent works (Clark et al., 2021; Wang et al., 2023b; Chiang and Lee, 2023) have shown that while LLMs can produce evaluations with human-like accuracy, these evaluations are often inconsistent and can easily be influenced. LLMs also show position bias or scale region bias and are unable to distinguish between candidates that are close to each other (Golchin and Surdeanu, 2023). LLMs are sensitive to instructions and their capabilities vary for different metrics (Skopek et al., 2023; Wang et al., 2023b; Shen et al., 2023). Another significant challenge when using LLMs as evaluators is a limited assessment of their abilities in multilingual settings. Studies have shown that LLMs have inferior performance even on some high-resource languages and cannot be assessed extensively on low-resource languages due to a lack of benchmarks (Ahuja et al., 2023a). Therefore, it is still unclear if LLMs can replace human evaluations in multilingual settings.

In this paper, we introduce the METAL framework for a robust assessment of LLMs as evaluators in multilingual scenarios. Figure 1 shows an outline of our framework. The METAL framework is an end-to-end pipeline, that starts with first creating a rich meta-evaluation dataset that contains a variety of samples across the metrics of interest. We do this by systematically prompting GPT-4 to generate a wide range of sample data points, that are then evaluated by native speakers. In the next step, we compare LLM judgments with human judgments. For this, we draw on our previous work (Hada et al., 2024) to prompt LLMs for evaluations and subsequently compare the scores with human judgments. In particular, for the task of **summarization** we create a dataset of 1000 summaries covering 10 languages, with human ratings across 5 metrics.[1] Next, we obtain LLM evaluations from GPT-3.5-Turbo, GPT-4, and PaLM2 across these 1000 summaries and 5 metrics.

Our findings show that the GPT-3.5-Turbo-based evaluator does not perform well across languages and metrics, while evaluators based on GPT-4 and PaLM2 perform better. We find that the evaluation ability of LLMs significantly across languages, motivating the creation of a meticulously crafted meta-evaluation dataset covering all target languages before using LLM-based evaluators. Lastly, our qualitative analysis shows that while GPT-4 and PaLM2 can achieve accuracy close to humans, the reasoning behind their evaluations is often flawed. While we study the applicability of the METAL framework for the task of summarization, it is extensible to other tasks as well, by creating meta-evaluation datasets for the other tasks.

2 Related Work

Human Evaluation Studies by Kryściński et al. (2018); Huang et al. (2020); Shen et al. (2022b) implemented the Likert scale for assessing various dimensions of generated summaries. Fan et al. (2018); Fabbri et al. (2019); Shen et al. (2022a) perform side-by-side comparisons of summaries produced by different models, using systems such as Elo for ranking the models based on performance.

Evaluation Datasets Human-verified gold-standard datasets are crucial for being able to evaluate LLMs. Skopek et al. (2023) release *riSum*, an English-centric dataset of document-instruction-output triplets, with the LLMs generating instructions and outputs, and a human evaluation is adopted to score the triplets, with a focus on "instruction-following". In our work, we manually curate the instructions and create a dataset covering 10 languages. SEAHORSE (Clark et al., 2023) is a multilingual and multifaceted data for summarization with 96K summaries, and metrics related to grammar and output quality. They fine-tune T5 (Raffel et al., 2023; Xue et al., 2021) and PaLM (Chowdhery et al., 2022) models on the train split of the dataset for generating

---

[1]METAL dataset and code available at `https://aka.ms/METAL`

a spectrum of outputs, whereas we work with black-box models and tune our prompts to generate "good" and "bad" summaries.

LLM Evaluation Several previous studies have analyzed and evaluated LLMs on new tasks and standard benchmarks (Ahuja et al., 2023a,b). Wang et al. (2023a) prompt ChatGPT for summarization and other higher-level tasks, across various metrics and find a high correlation with human scores, with the caveat that it may be influenced by the way the meta-evaluation datasets are created. We extend this idea in the METAL dataset. Other studies (Naismith et al., 2023; Fu et al., 2023; Liu et al., 2023; Zhang et al., 2023b) have also put forth GPT scoring frameworks and prompt-based evaluators, however, they are mostly confined to English or Latin script languages. Wu and Aji (2023) employ a Multi-Elo Rating System and advice for a similar multi-dimensional assessment of LLM-generated summaries. Previous studies (Li et al., 2023; Zheng et al., 2023; Hada et al., 2024) have solely relied on GPT-4 as an LLM judge/evaluator. Kim et al. (2023) propose a fine-tuned LLM, comparable to GPT-4 for input-evaluation rubric-output triplets, however, it is only fine-tuned for English. Previously, we performed a multilingual study of LLM meta-evaluation using an internal dataset of human judgments Hada et al. (2024). The dataset used in this work was not curated for the specific purpose of LLM-evaluator calibration and hence suffers from weaknesses such as dataset skew, and we focus this work on building a better dataset for meta-evaluation. We use the prompting strategies from our prior work (Hada et al., 2024) to evaluate our new curated dataset. To the best of our knowledge, no other study has proposed an end-to-end pipeline from generating a rich evaluation set to assessing the performance of LLM as evaluators in the multilingual scenario.

3 The METAL Dataset

The METAL dataset contains a total of 1000 summaries across 10 languages. The dataset is specially curated to investigate the capabilities of LLMs as evaluators in different languages along 5 dimensions. In this section, we describe how the dataset was created and annotated.

3.1 Dataset Creation

The dataset consists of 100 summaries each, for 10 languages: English (En), French (Fr), Chinese

Simplified (Zh), Hindi (Hi), Arabic (Ar), Bengali (Bn), Russian (Ru), Turkish (Tr), Japanese (Ja), and Swahili (Sw). We selected the languages to cover a diverse range of scripts and regions. The main text for each summary in our dataset was chosen from XL-Sum (Hasan et al., 2021), and the corresponding summary was generated by prompting GPT-4. A brief overview of our methodology is shown in Figure 1.

Main Text Selection For each of the 10 languages we create a histogram of 20 bins of the number of tokens in the main text for all the datapoints in the test set of the XL-Sum dataset (Hasan et al., 2021). We chose 100 random samples from the bin with the highest frequency.

Summary Generation To investigate the capabilities of LLMs as evaluators, our objective was to create an evaluation set of summaries with varying quality. To this end, for each of the chosen 1000 samples from the above step, we generate two summaries by prompting GPT-4 as follows.

To generate good-quality summaries we provide the main text to GPT-4 and prompt it to return a summary of the main text such that it captures the essence of the main passage. We specifically ask for a summary that is highly rated on the 5 metrics of interest, described in the next section. We keep the temperature at 0 for the generation of good-quality summaries.

To generate bad-quality summaries we provide the main text to GPT-4 and prompt it to act as an adversarial NLP assistant, and badly summarize the main passage. We specifically ask for a summary that is rated low on the 5 metrics of interest. To generate bad quality summaries we keep the temperature as 1. In our initial experiments with lower temperatures, we observed that GPT-4 does not produce bad summaries even when specifically prompted to do so.

To further ensure the quality of summaries, in both styles of prompting we ask GPT-4 to also justify why the generated summary is good or bad. Once we have 2 summaries per data point, we choose a good-quality summary or a bad-quality summary at random. The verbatim of our prompts are provided in §A.1

3.2 Dataset Annotation

For the 1000 summaries selected from the above process, we have each sample annotated by 3 annotators for 5 different metrics. We use the metrics

described by Hada et al. (2024) in their work:

Linguistic Acceptability (LA) This metric assesses whether the summary is acceptable to a native speaker. Specifically, the annotators are asked to determine whether the text exhibits signs of being translated, misuses words, or includes expressions that are not idiomatic in their language.

Output Content Quality (OCQ) This metric assesses whether the general quality of the output text is good. The annotators are asked to consider flaws such as significant repetition, non-native language elements, or indications that the text has been web-scraped.

Task Quality (TQ) This metric assesses the effectiveness of the summarization. It focuses on assessing the degree to which the summary aligns with key information in the main passage.

Problematic Content (PC) This metric assesses the summary for the presence of any content that may be deemed offensive, inappropriate, or harmful. It serves as a filter against outputs that might perpetuate harmful stereotypes or misinformation.

Hallucinations (H) This metric assesses whether the summary remains anchored to, and consistent with, the main passage. It serves as a check against unwarranted deviations from the ground truth provided in the input.

For LA, OCQ, and TQ, annotators were asked to assign one of the three possible classes: Bad (0), Medium (1), Good (2). For PC and H, annotators were asked to assign one of the two possible classes: Present (1) and Absent (0).[2]

Annotation Task and Quality Each datapoint was annotated by three annotators for the five metrics. Annotators were native speakers of the respective language and trained professionals contracted through an external annotator services company. The pay was adjusted based on the annotator's region and experience. Since we wanted to ensure we had a strong evaluation set to study the capabilities of LLMs as evaluators, special attention was given to the quality of annotations. The annotators were specifically trained to perform annotations for this task and a sample of annotations was reviewed for all annotators. Annotations were reviewed for
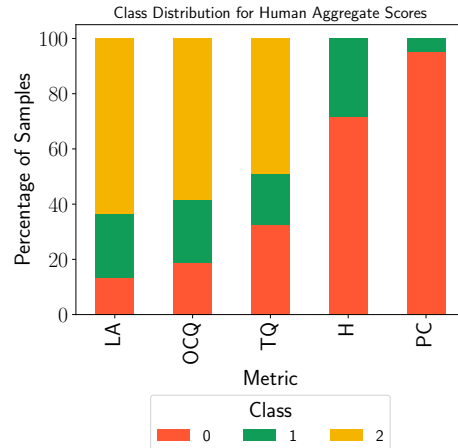

Figure 2: Class distribution for various metrics, summed over all languages.

accuracy and guideline consistency. Based on the review, feedback was provided to the annotators, and ambiguous cases were re-annotated.

Table 6 in Appendix §A.3 shows the Fleiss' Kappa (κ) and pairwise agreement (computed as F1) values among the annotators for the various languages and metrics. All our κ values are > 0.6 (except for H in En, $\kappa = 0.54$), and all F1 values are > 0.75, indicating substantial agreement. Some κ values are 0 due to class skew, but high F1 in these cases indicates high reliability. For our experiments, we take the majority vote from the three human annotations per sample as the aggregate class for that sample. In the case of 3 distinct annotations, we take the average value. Figure 2 shows the distribution of the aggregate annotations over various languages and metrics.

3.3 Dataset Statistics

As discussed in §3.1, we sample the datapoints from XL-Sum based on the number of tokens of the passage. Specifically, the tiktoken[3] was utilized for the tokenization process, and the length (token) distribution of the passages and summaries are presented in Table 1 along with the number of good and bad instances per language. Table 2 presents the frequency/distribution of the classes (0, 1, and 2) in the good and bad summaries. Notably, the first row of the table depicts higher counts of low scores (class 0) for the bad summaries, relative to the good ones. Further, the medium scores (class 1) also contain a higher frequency in the bad summaries, however, the difference between the frequencies is lower than that of class 0. Sur-

---

[2]We include the detailed annotation instructions in Appendix §A.2

[3]https://github.com/openai/tiktoken

Lang	Passage	Summary	Good	Bad
AR	877.39 ± 53.00	160.70 ± 87.29	50	50
BN	4161.58 ± 534.91	339.83 ± 160.55	**53**	47
EN	358.29 ± 21.09	67.71 ± 29.57	46	**54**
FR	341.96 ± 26.89	84.79 ± 39.27	**51**	49
HI	1234.82 ± 70.28	219.08 ± 92.38	48	52
JA	1327.44 ± 61.50	136.44 ± 81.11	**52**	48
RU	748.26 ± 47.52	139.09 ± 72.28	43	**57**
SW	518.70 ± 35.90	127.98 ± 73.79	47	**53**
TR	625.77 ± 40.96	136.44 ± 68.76	42	**58**
ZH	666.03 ± 47.78	124.16 ± 67.80	48	52

Table 1: Length distribution and number of instances per language

Class	LA	OCQ	TQ	H	PC
0	54 / **80**	78 / **112**	124 / **202**	352 / **362**	457 / **493**
1	113 / **116**	104 / **121**	91 / **93**	128 / **158**	23 / **27**
2	**313** / 324	**298** / 287	**265** / 225	- / -	- / -

Table 2: Class distribution for various metrics, N(Good)/N(Bad). The highest frequency is bolded.

prisingly, the bad summaries have more of class 2 scores than the good ones (third row) in Linguistic Acceptability. This goes on to show that LLMs are incapable of generating incoherent text despite adversarial prompts.

4 Experiments

4.1 Models

GPT-4-32K (OpenAI, 2023), GPT-3.5-Turbo (Ouyang et al., 2022), and PaLM2 Text-Bison (Anil et al., 2023) models were used as the evaluators to score the LLM-generated summary according to the given metrics. [4]

4.2 Prompts

The models were prompted using the LangChain[5] framework and a structured JSON output format was maintained to parse the generations efficiently. The prompts for evaluation follow the same verbatim as Hada et al. (2024).

4.3 Prompting Strategies

Based on our previous work (Hada et al., 2024), we use the simple and detailed prompting strategies for all models, and all the metrics are evaluated independently in a single call to the API. All prompts were provided in English, as Ahuja et al. (2023a) have shown that multilingual instructions lead to worse performance. Further, the temperature is set to 0 for reproducibility.

Simple Instruction A rudimentary description of the metric and scoring schema is provided, as shown in Figure 7 in appendix §A.4.

Detailed Instruction An informative and thorough description of the metric and a case-by-case breakdown of the scoring schema is provided, as shown in Figure 8 in the appendix §A.4.

4.4 Meta Evaluation

As described in §3.2 we use the aggregate of the three annotations for our experiments.

Pairwise Agreement (F1) We measure the pairwise agreement between the LLM evaluators and human aggregate scores per language and metric. To account for any class imbalance, we report the weighted F1 score instead of accuracy.

Class Distribution We analyze the class distribution of the human aggregate scores and the various model predictions for three possible classes: When all three annotators agree, when two of three annotators agree, and when no annotators agree. We do this analysis only for metrics with 3 possible classes: LA, OCQ, and TQ.

4.5 Comparison between SEAHORSE and METAL

SEAHORSE (Clark et al., 2023) is a dataset akin to METAL, as described in §2. It contains summaries generated using several models for passages from popular summarization datasets such as XL-Sum (Hasan et al., 2021), XSum (Narayan et al., 2018), MLSum (Scialom et al., 2020) and WikiLingua (Ladhak et al., 2020). We use the XL-Sum subset of Seahorse and find out common datapoints between SEAHORSE and METAL. There are a total of 27 overlapping data points: 1 in English, 10 in Russian, and 16 in Turkish. These datapoints can have one or more summaries in SEAHORSE generated by mt5_small: The 300M version of mT5 (Xue et al., 2021), mt5_small_250: The same mt5_small model but using the checkpoint after training 250 steps, mt5_xxl: The 13B mT5 model, palm_1shot: 540B PaLM model (Chowdhery et al., 2022) prompted with one in-domain example, palm_finetuned: 540B PaLM model finetuned on training data for the respective dataset. We use our detailed prompting strategy to evaluate the summaries generated by various models in SEAHORSE for our metrics and compare them with

the evaluation of the summaries generated by GPT-4 for the same main passages in METAL . We use PaLM2 and GPT-4 as evaluators. [6]

5 Results and Discussions

5.1 Pairwise Agreement (F1)

Table 3 and Figures 9 and 10 in Appendix §A.5 present the distribution of F1 scores of various models with the two prompting strategies on the 10 languages. For "human scores", we average the pairwise F1 scores of all the annotators, i.e., A1-A2, A2-A3, and A3-A1. For the "model scores" in the plot, the F1 score between the annotator aggregate and model evaluation is computed.

For all the metrics, humans have the best agreement. In the case of LA, for most of the languages, GPT-4 with detailed instructions performs the closest to humans, followed by GPT-4 with simple instructions. GPT-3.5-Turbo performs the worst with detailed instructions and a significant difference is observed by making the instructions simple, however, no difference is found for English. For most of the languages, especially Zh, Hi, Ru, and Ar, GPT-4 and PaLM2 perform similarly.

For OCQ, GPT-3.5-Turbo performs the worst, and detailed instruction improves the performance marginally over simple instructions. GPT-4 and PaLM2 perform very closely to humans for Russian, however, there is a gap between the human and LLM scores on the rest of the languages. For TQ, both prompting strategies for GPT-4 and PaLM2 do equally well on most languages except Ar, Hi, Zh, and En. In these cases, GPT-4 with detailed instructions does the best.

For PC and H, all models show very similar scores as compared to humans, except GPT-3.5-Turbo. Simple instructions for GPT-3.5-Turbo improve performance for both metrics with a higher gain on H. Interestingly, for Bn and metrics LA and OCQ, both prompting strategies for GPT-3.5-Turbo do better than GPT-4 and PaLM2. For Sw and metrics LA, OCQ, and TQ, the agreement between humans and GPT-4 or PaLM2 is as good as the agreement between humans. GPT-3.5-Turbo with detailed instructions does worse than GPT-3.5-Turbo for all metrics except OCQ.

Overall, we find that the performance of GPT-4 and especially PaLM2 are largely independent of simple or detailed instructions, in all languages.

---

[6]We do not consider the 1 overlapping datapoint in English for our experiment.
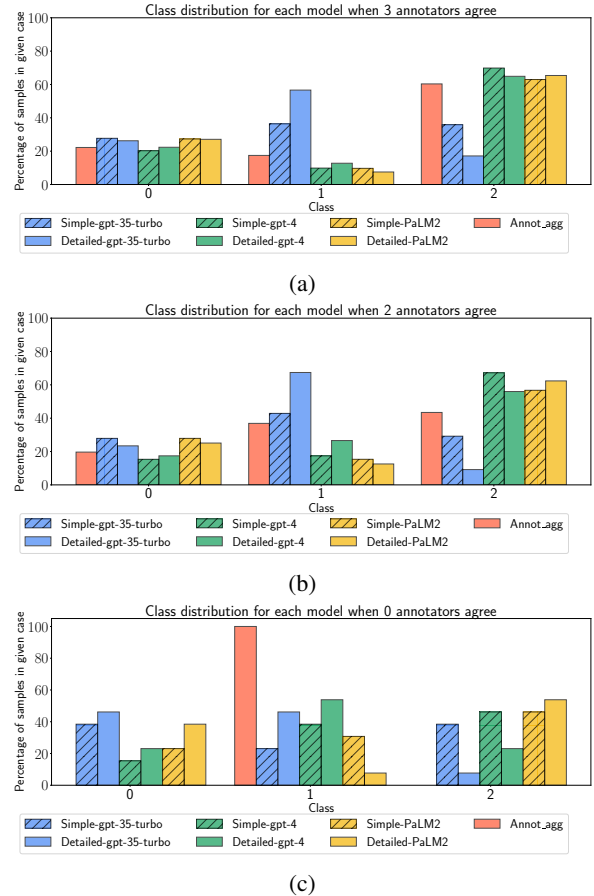

(a)

(b)

(c)

Figure 3: Class distribution for different cases

The same holds for GPT-3.5-Turbo only on English, suggesting that it is less sensitive to prompting in English. GPT-4 with detailed instructions comes closest to human evaluation, with marginal improvements over simple instructions, in most cases. GPT-4 and PaLM2 are very effective in identifying hallucinations and problematic content for all languages.

5.2 Class Distribution

Figure 3 shows the distribution of human aggregate score and various models for the three cases. In the case where all annotators agree, as shown in Figure 3a we can see that the class distribution for GPT-4 and PaLM2 with both prompting variations is very close to the class distribution of human aggregate scores. This indicates that when humans have full agreement (perhaps due to easier samples), LLM-based evaluators also perform well.

In the case where two of three annotators agree, we can see in Figure 3b that for both prompting variations GPT-4 and PaLM2 often over-predict class 2, under-predict class 1 and are similar to humans for class 0. Overall, detailed-GPT-4

Metric	Prompting Strategy	Model	AR	BN	EN	FR	HI	JA	RU	SW	TR	ZH
LA		*human*	0.89	0.81	0.86	0.99	0.87	0.95	0.98	0.82	0.97	0.94
	Simple	*GPT-3.5-Turbo*	0.54	0.43	0.61	0.61	0.44	0.45	0.67	0.78	0.55	0.59
		GPT-4	0.74	0.15	0.72	0.88	0.59	0.48	0.74	0.61	0.72	0.85
		PaLM2	0.74	0.11	0.54	0.73	0.64	0.38	0.77	0.82	0.69	0.84
	Detailed	*GPT-3.5-Turbo*	0.19	0.44	0.59	0.40	0.18	0.19	0.53	0.57	0.15	0.19
		GPT-4	0.71	0.22	0.82	0.81	0.61	0.47	0.80	0.76	0.72	0.85
		PaLM2	0.71	0.21	0.54	0.75	0.59	0.34	0.78	0.88	0.64	0.84
OCQ		*human*	0.85	0.82	0.82	0.97	0.83	0.93	0.93	0.84	0.84	0.91
	Simple	*GPT-3.5-Turbo*	0.11	0.39	0.65	0.47	0.17	0.21	0.64	0.61	0.52	0.33
		GPT-4	0.71	0.27	0.69	0.70	0.65	0.47	0.94	0.85	0.69	0.88
		PaLM2	0.69	0.23	0.63	0.68	0.58	0.43	0.92	0.91	0.67	0.79
	Detailed	*GPT-3.5-Turbo*	0.23	0.54	0.59	0.50	0.31	0.33	0.64	0.58	0.50	0.44
		GPT-4	0.69	0.26	0.68	0.72	0.65	0.51	0.92	0.88	0.68	0.84
		PaLM2	0.68	0.29	0.57	0.65	0.66	0.41	0.92	0.91	0.69	0.86
TQ		*human*	0.77	0.78	0.77	0.90	0.78	0.99	0.94	0.84	0.87	0.82
	Simple	*GPT-3.5-Turbo*	0.63	0.53	0.52	0.84	0.58	0.81	0.83	0.82	0.65	0.77
		GPT-4	0.60	0.64	0.53	0.81	0.56	0.87	0.95	0.87	0.60	0.78
		PaLM2	0.56	0.67	0.41	0.83	0.56	0.85	0.90	0.88	0.59	0.79
	Detailed	*GPT-3.5-Turbo*	0.26	0.49	0.54	0.76	0.22	0.44	0.63	0.63	0.58	0.31
		GPT-4	0.71	0.64	0.59	0.86	0.66	0.86	0.96	0.87	0.63	0.76
		PaLM2	0.58	0.66	0.38	0.83	0.51	0.84	0.94	0.90	0.65	0.73
H		*human*	0.89	0.97	0.85	0.97	0.90	0.99	0.99	0.93	0.84	1.00
	Simple	*GPT-3.5-Turbo*	0.54	0.27	0.81	0.75	0.36	0.63	0.72	0.66	0.57	0.59
		GPT-4	0.93	0.74	0.85	0.91	0.89	0.94	0.93	0.90	0.87	0.90
		PaLM2	0.94	0.77	0.78	0.92	0.90	0.82	0.72	0.80	0.76	0.87
	Detailed	*GPT-3.5-Turbo*	0.06	0.01	0.42	0.58	0.09	0.36	0.50	0.37	0.22	0.19
		GPT-4	0.95	0.72	0.85	0.90	0.88	0.96	0.94	0.89	0.86	0.88
		PaLM2	0.91	0.73	0.76	0.90	0.86	0.94	0.87	0.87	0.86	0.91
PC		*human*	0.93	1.00	1.00	1.00	0.94	0.99	0.99	0.86	1.00	1.00
	Simple	*GPT-3.5-Turbo*	0.52	0.23	0.83	0.56	0.32	0.31	0.33	0.51	0.45	0.63
		GPT-4	0.90	0.99	1.00	0.95	0.85	1.00	0.97	0.73	1.00	0.97
		PaLM2	0.89	1.00	0.97	0.85	0.86	0.95	0.92	0.71	0.99	0.96
	Detailed	*GPT-3.5-Turbo*	0.28	0.06	0.68	0.45	0.23	0.28	0.20	0.43	0.28	0.36
		GPT-4	0.87	0.99	0.99	0.87	0.85	0.95	0.91	0.71	0.91	0.96
		PaLM2	0.89	0.84	0.97	0.88	0.86	0.79	0.88	0.80	0.92	0.95

Table 3: F1 scores for various languages, models, and prompting strategies.

comes closest to the distribution of human aggregate scores. For both cases, GPT-3.5-Turbo often prefers the middle class, which can be indicative of scale region bias.

In our third case, owing to our high quality of annotations, we have only 13 out of 3000 samples where no annotators agree. Figure 3c shows the class distribution for this case. We can observe that GPT-3.5-Turbo with simple instructions assigns different classes with almost equal frequency. GPT-4 with detailed instructions often outputs the middle class, which is the annotator aggregate as well. PaLM2 with detailed instructions outputs the highest or the lowest score, and interestingly very few times opts for the middle class.

From this analysis, we conclude that while simple or detailed instructions for both GPT-4 and PaLM2 perform equally well when all human annotators agree, detailed instructions for GPT-4 do best when there is disagreement amongst annotators.

5.3 Comparison of evaluation between SEAHORSE and METAL

Table 4 shows the values for evaluation of overlapping datapoints between SEAHORSE and METAL for RU and TR. The values are averaged over all datapoints. For Russian, we can observe that `palm_1shot` does the best as it is rated highly across all metrics by both the models. Interestingly, `palm_1shot` is rated better than `palm_finetuned`. GPT-4$_{good}$ summaries received very bad evaluations across all metrics by both models. On further investigation, we found that GPT-4$_{good}$ category had only 2 datapoints out of the overlapping 10 datapoints, and co-incidentally these 2 generated summaries were of bad quality and have been rated poorly by human annotators as well. This indicates that GPT-4 might not always be biased towards its own generations as compared to generations from other models. We can also observe that in almost all cases GPT-4 provides a higher rating as compared to PaLM2. For Turkish, we can observe that GPT-4$_{good}$ receives the highest ratings by both

Lang	Metric	SEAHORSE												XL-Sum		METAL			
		mt5_small_250		*mt5_xxl*		*mt5_small*		*palm_1shot*		*palm_finetuned*		reference		$GPT-4_{good}$		$GPT-4_{bad}$	
		PaLM2	GPT-4	PaLM2	GPT-4	PaLM2	GPT-4	PaLM2	GPT-4	PaLM2	GPT-4	PaLM2	GPT-4	PaLM2	GPT-4	PaLM2	GPT-4
	H	0.60	0.40	0.00	0.20	0.43	0.43	0.20	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.75	0.75
	LA	0.60	1.00	1.40	2.00	0.57	2.00	1.60	2.00	1.25	2.00	1.33	2.00	0.00	1.00	0.62	0.87
RU	*OCQ*	0.60	0.60	1.40	1.20	0.57	0.86	1.40	1.60	1.00	1.25	1.00	1.67	0.0	0.00	0.50	0.50
	TQ	0.60	0.40	1.00	1.20	0.71	0.71	1.60	1.20	1.00	1.00	1.00	1.00	0.00	0.00	0.50	0.50
	PC	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.25
	H	0.55	0.75	0.10	0.10	0.50	0.50	0.00	0.12	0.00	0.00	0.00	0.00	0.14	0.28	0.55	0.55
	LA	0.11	1.12	1.40	2.00	0.75	1.62	1.50	2.00	1.20	2.00	1.33	2.00	1.57	1.71	0.78	1.55
TR	*OCQ*	0.11	0.62	1.20	1.50	0.62	1.00	1.37	1.62	1.00	1.60	1.33	1.67	1.57	1.57	0.78	1.00
	TQ	0.11	0.50	1.20	1.30	0.50	0.75	1.37	1.25	0.80	1.20	1.17	1.33	1.57	1.57	0.78	1.11
	PC	0.33	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.22

Table 4: Evaluation of overlapping summaries generated by various models in SEAHORSE and GPT-4 in METAL for RU and TR

models. Similar to Russian, `palm_1shot` receives better ratings than palm_finetuned. As expected, `mt5_small_250` receives the lowest ratings since it is an under-trained model. We can also notice a clear difference in ratings for GPT-4$_{good}$ vs GPT-4$_{bad}$ generations. Overall, from this experiment we can conclude that using our metrics and prompting methods we can compare generations from different models.

6 Qualitative Analysis

While results in section §5 show that detailed instructions to GPT-4 and PaLM2 give evaluations very close to humans, it is unclear how humans and LLMs reason about their scores. We qualitatively analyse the comments from the annotators for their ratings and justifications produced by the LLMs at the time of scoring the summaries. An analysis of some interesting examples is discussed in this section. As discussed in section §4.4, the annotations from humans can be divided into three categories: when all annotators agree, when two annotators agree, and when no annotators agree. Table 5 shows examples from each of these categories for different languages and metrics. We specifically analyze cases where LLMs' scores differ from the annotator aggregate score.

The first example is where no annotators agree on TQ for an English sample. Both GPT-4 and PaLM2 assign a 2 in this case. While all three annotators point out a few problems with the summary, both GPT-4 and PaLM2 ignore some key elements for TQ such as "omission of important information", and "poor discourse" and say that the summary "captures main points of the passage".

The next two cases in the table are when two annotators agree. In the second case, two annotators give the sample a score of 1 for LA, however, no annotators point towards any grammatical issues with the summary. Their comments are more

relevant for TQ and OCQ. This indicates that for humans their judgment of one metric might affect their judgment of other metrics. Both LLMs give a high score of 2 to the sample, even though the reason from PaLM2 says "lacks natural rhythm and intonation". This shows that LLMs' reasons might not always be aligned with their scores, in line with findings from Hada et al. (2023). In the third case, the annotator aggregate for OCQ is 2, however, both LLMs assign a score of 0. Annotators mention problems such as "hallucinations" in their comments, while GPT-4 says the summary is an inaccurate representation of the main text, and PaLM2 complains of incoherence.

The last two examples in the table are where all three annotators agree but the LLM scores are different. In the fourth case, all annotators assign a score of 2 for TQ, however, both LLMs assign a score of 0. Even though A2 complains about the quality of the summary, they assign a score of 2, indicating some error in judgment. Both LLMs assign a score of 0 and reason that that the summary consists of hallucinations. It is interesting that humans still assign the summary a score of 2 indicating that there can be subtle differences in how humans interpret these metrics.

In the last case, all annotators assign the sample a score of 2 for OCQ and do not mention any issues with content quality in their comments. Interestingly, PaLM2 assigns a score of 1 and the justification states "it does not provide any new information or insights beyond what is already in the passage". Since this was a summarization task no new information is expected in the summary. This again indicates that the judgment and justification might not always be aligned. Table 7 in Appendix §A.6 shows some samples of cases where either of the LLM scores agree with human aggregate scores, but there are some discrepancies in their justification.

No.	Language	Metric	Human Score and Comments	GPT-4 Score and Justification	PaLM2 Score and Justification
1	English	TQ	'A1': 1, 'Some grammatical errors present. Some misaligned elements and omission of important information.', 'A2': 2, 'High quality summary, except missing verbs so medium linguistic acceptability', 'A3': 0, 'Summary contains grammatical errors, has poor discourse and is incoherent.'	2, The summary adheres to the instructions and captures the main points of the passage. Therefore, the summary demonstrates a precise understanding and adherence to the prompt's instructions.	2, The summary captures the main points of the passage. It correctly identifies the key players in the dispute, the reasons for the dispute, and the potential consequences of the dispute. The summary is concise and well-written.
2	Arabic	LA	'A1': 1, 'Text seems to be copied. Repetition noticed. Summary is too short more like a headline.', 'A2': 2, 'it is missing a lot of content from the main passage ', 'A3': 1, 'The summary is very brief and does not meet all the points mentioned in the main passage.'	2, The vocabulary used is contextually relevant and the sentences reflect the natural rhythm, emphasis, and intonation of spoken Arabic. it does not violate any linguistic norms or rules, and thus scores high on linguistic acceptability.	2, The summary is grammatically correct and uses contextually relevant vocabulary. However, it is overly verbose and lacks natural rhythm and intonation.
3	French	OCQ	'A1': 2, 'Le résumé, bien que mal conçu, affirme qu'une femme accuse Bill Cosby d'agression sexuelle à cause de l'argent, ce qui n'est pas vrai selon le passage original. Il y a des hallucinations dans le résumé et la qualité de la tâche est mauvaise.", 'A2': 2, 'Le résumé fourni contient des informations précises sur l'affaire Bill Cosby, telles que les allégations contre lui et le règlement civil avec l'un des accusateurs.', 'A3': 1, 'Insted of 'veut' it would have been better to use "ne voudrait" or "est accusée de ne vouloir que de l'argent". To me, it is a fake news.'	0, The summary provided does not accurately reflect the content of the passage. It also inaccurately portrays the accuser as someone who just wants money, without mentioning her testimony or the fact that Cosby settled the case.	0, The summary does not capture the essence of the passage. It is not clear, coherent, or engaging. It does not convey the main points of the passage and it is not original.
4	Hindi	TQ	'A1': 2, 'The summary is counterfactual in relation to the input. The president condemned the incident and in summary it states that he praised the incident. This is a problematic content because it involves student violence, gun violence, teen violence, etc.… less or no grammatical errors makes the task accuracy good.', 'A2': 2, 'The summary does not justifies the actual sentence. The quality of the summary is very bad.', 'A3': '2, no comments'	0, The summary provided does not adhere to the passage at all. It misrepresents the events described in the passage.	0, The summary does not adhere to the instructions given in the prompt. The prompt asks for a summary of the passage, but the summary is not a summary of the passage. The summary is a made-up story that is not related to the passage.
5	Chinese Simplified	OCQ	'A1': 2, 'Good summary without mistakes', 'A2': 2, 'The summary fails to mention the important message of the original, which is, many civilians were killed. The summary only indicates that some people were killed', 'A3': 2, 'Key information is missing in the summarization. The summarization is too general.'	1, The summary is grammatically correct and uses appropriate language. However, it lacks clarity and relevance to the passage.	1, The summary is concise and captures the main points of the passage. It is also well-written and easy to understand. However, it does not provide any new information or insights beyond what is already in the passage.

Table 5: Samples from METALdataset with human and LLM (detailed variation) scores and comments. Note: We only include snippets from the full comment.

Overall, our analysis indicates that there are several challenges in the alignment of human evaluations with LLM evaluations. While the scoring by LLMs on several metrics and languages might come close to humans, it is difficult to understand how they come up with these scores, necessitating further research.

7 Conclusion

We presented the first framework for end-to-end evaluation of LLMs as evaluators in multilingual scenarios. We created a dataset of 1000 summaries across 10 languages rated by native speakers on 5 different metrics. Our dataset covers a range of summaries in terms of linguistic acceptability, output quality, task quality, and others. We do this by systematically prompting GPT-4 to generate summaries of varying quality. The human ratings obtained for these summaries are of high quality with $\kappa > 0.6$ and $F1 > 0.75$. We plan to make the METALdataset available to the research community. Using our dataset, we investigate the capabilities of three LLMs as evaluators: GPT-3.5-Turbo, GPT-

4, and PaLM2 using two prompting strategies and compare their evaluation with the METALhuman evaluations. Our results show that GPT-4 with detailed instructions performs closest to humans, while GPT-3.5-Turbo is not a suitable multilingual evaluator but surprisingly does better than GPT-4 and PaLM2 in some metrics for Bengali. We also show that GPT-4 with detailed instructions does best when there is disagreement amongst human annotators. We compare the overlapping summaries between SEAHORSE and METAL and show how our metrics and prompting methods can be used to compare generations from different models. Finally, we analyze human and LLM reasoning and observe that LLMs often provide incorrect justifications for their scores, thus showing that more research is needed to be able to use LLM-based evaluators with confidence in the multilingual setting.

8 Limitations

We prompt GPT-4 to generate good and bad-quality summaries. As noted in §3.1, for lower tempera-

ture values we observed that GPT-4 did not generate bad summaries. We use a temperature of 1 and observe some variation of quality across all our metrics except problematic content. This could be due to the content filter applied to these models. Therefore, it is difficult to study the capability of such models on this metric. We evaluate the generations from GPT-4 using GPT-3.5-Turbo, GPT-4, and PaLM2. Recent work has shown that LLMs prefer their own outputs. Although this might have affected our evaluations, exploring this is beyond the scope of our work. In our work, we mainly focused on investigating how well LLM ratings align with human ratings across various metrics and languages. All summaries generated and evaluated in our study are by the same model, we do not compare them against human-written summaries or summaries generated by other models. Lastly, LLMs are also shown to have scale region bias and we do not calibrate for this in our study, expecting it to be standardized across all their ratings. In the future, it would interesting to explore their impact on our evaluation.

9 Ethical Considerations

We use the framework by Bender and Friedman (2018) to discuss the ethical considerations for our work.

Institutional Review Our dataset was annotated by an external company that has long-standing contracts with the organization and is employed by the organization regularly to do this work. Therefore, the annotation company only accepts work that is covered under the purview of their contract.

Data To generate the summaries in our dataset we use the main text from the publicly available test set of XL-Sum (Hasan et al., 2021). Our summaries are generated in 10 languages: En, Fr, Hi, Zh, Ar, Bn, Tr, Ja, Ru, and Sw. We do this by prompting GPT-4. We release the dataset publicly for future research. Our dataset was created such that it covers a range of quality for summaries. Therefore, some summaries in our dataset are deliberately incoherent. Our ratings on problematic content show that $< 5\%$ of our data had problematic text in them.

Annotator Demographics Annotators were recruited through an external annotator services company. All annotators were native speakers of the language of the data points they annotated. The

pay was adjusted after discussion with the company, based on the annotator's region and experience. No demographic information is available about the annotators. The annotators are governed by their company's and our organization's privacy policy.

Annotation Guidelines We draw inspiration from the community standards set for similar tasks. These guidelines were created following best practices after careful research. Annotators were asked to rate the summaries across 5 metrics. A detailed explanation was given for each of the metrics. For 3 metrics annotators had to choose from 3 classes, and for 2 metrics they had to choose from 2 classes. Annotators were allowed to give feedback for any data point via an optional comments text box. Annotators received training for this task. Annotator identity was hidden from the task reviewers to limit any bias.

Methods In this study, we explore methods to generate summaries by prompting GPT-4. We deliberately prompt GPT-4 to generate some bad summaries. All summaries generated were evaluated by 3 LLMs: GPT-3.5-Turbo, GPT-4, and PaLM2. We explore several ways to calibrate LLM judgment with human judgments for various metrics and languages. While these methods can be easily misused, our intent with this study is to highlight the gap between the two and urge the community to proceed with caution.

References

Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022. Beyond static models and test sets: Benchmarking the potential of pretrained models across tasks and languages. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 64–74, Dublin, Ireland. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023b. Mega-

verse: Benchmarking large language models across languages, modalities, models and tasks.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

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

Shahriar Golchin and Mihai Surdeanu. 2023. Data contamination quiz: A tool to detect and estimate contamination in large language models.

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.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian's, Malta. Association for Computational Linguistics.

Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. "fifty shades of bias": Normative ratings of gender bias in GPT generated English text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876, Singapore. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: 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.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv: 2401.04088*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

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.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,

Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontañón, Oskar Bunyan, Nathan Byrd, Abhanshu Sharma, Biao Zhang, Mario Pinto, Rishika Sinha, Harsh Mehta, Dawei Jia, Sergi Caelles, Albert Webson, Alex Morris, Becca Roelofs, Yifan Ding, Robin Strudel, Xuehan Xiong, Marvin Ritter, Mostafa Dehghani, Rahma Chaabouni, Abhijit Karmarkar, Guangda Lai, Fabian Mentzer, Bibo Xu, YaGuang Li, Yujing Zhang, Tom Le Paine, Alex Goldin, Behnam Neyshabur, Kate Baumli, Anselm Levskaya, Michael Laskin, Wenhao Jia, Jack W. Rae, Kefan Xiao, Antoine He, Skye Giordano, Lakshman Yagati, Jean-Baptiste Lespiau, Paul Natsev, Sanjay Ganapathy, Fangyu Liu, Danilo Martins, Nanxin Chen, Yunhan Xu, Megan Barnes, Rhys May, Arpi Vezer, Junhyuk Oh, Ken Franko, Sophie Bridgers, Ruizhe Zhao, Boxi Wu, Basil Mustafa, Sean Sechrist, Emilio Parisotto, Thanumalayan Sankaranarayana Pillai, Chris Larkin, Chenjie Gu, Christina Sorokin, Maxim Krikun, Alexey Guseynov, Jessica Landon, Romina Datta, Alexander Pritzel, Phoebe Thacker, Fan Yang, Kevin Hui, Anja Hauth, Chih-Kuan Yeh, David Barker, Justin Mao-Jones, Sophia Austin, Hannah Sheahan, Parker Schuh, James Svensson, Rohan Jain, Vinay Ramasesh, Anton Briukhov, Da-Woon Chung, Tamara von Glehn, Christina Butterfield, Priya Jhakra, Matthew Wiethoff, Justin Frye, Jordan Grimstad, Beer Changpinyo, Charline Le Lan, Anna Bortsova, Yonghui Wu, Paul Voigtlaender, Tara Sainath, Charlotte Smith, Will Hawkins, Kris Cao, James Besley, Srivatsan Srinivasan, Mark Omernick, Colin Gaffney, Gabriela Surita, Ryan Burnell, Bogdan Damoc, Junwhan Ahn, Andrew Brock, Mantas Pajarskas, Anastasia Petrushkina, Seb Noury, Lorenzo Blanco, Kevin Swersky, Arun Ahuja, Thi Avrahami, Vedant Misra, Raoul de Liedekerke, Mariko Iinuma, Alex Polozov, Sarah York, George van den Driessche, Paul Michel, Justin Chiu, Rory Blevins, Zach Gleicher, Adrià Recasens, Alban Rrustemi, Elena Gribovskaya, Aurko Roy, Wiktor Gworek, Séb Arnold, Lisa Lee, James Lee-Thorp, Marcello Maggioni, Enrique Piqueras, Kartikeya Badola, Sharad Vikram, Lucas Gonzalez, Anirudh Baddepudi, Evan Senter, Jacob Devlin, James Qin, Michael Azzam, Maja Trebacz, Martin Polacek, Kashyap Krishnakumar, Shuo yiin Chang, Matthew Tung, Ivo Penchev, Rishabh Joshi, Kate Olszewska, Carrie Muir, Mateo Wirth, Ale Jakse Hartman, Josh Newlan, Sheleem Kashem, Vijay Bolina, Elahe Dabir, Joost van Amersfoort, Zafarali Ahmed, James Cobon-Kerr, Aishwarya Kamath, Arnar Mar Hrafnkelsson, Le Hou, Ian Mackinnon, Alexandre Frechette, Eric Noland, Xiance Si, Emanuel Taropa, Dong Li, Phil Crone, Anmol Gulati, Sébastien Cevey, Jonas Adler, Ada Ma, David Silver, Simon Tokumine, Richard Powell, Stephan Lee, Michael Chang, Samer Hassan, Diana Mincu, Antoine Yang, Nir Levine, Jenny Brennan, Mingqiu Wang, Sarah Hodkinson, Jeffrey Zhao, Josh Lipschultz, Aedan Pope, Michael B. Chang, Cheng Li, Laurent El Shafey, Michela Paganini, Sholto Douglas, Bernd Bohnet, Fabio Pardo, Seth Odoom, Mihaela Rosca, Cicero Nogueira dos Santos, Kedar Soparkar, Arthur Guez, Tom Hudson, Steven Hansen, Chulayuth Asawaroengchai, Ravi Addanki, Tianhe Yu, Wojciech Stokowiec, Mina Khan, Justin Gilmer, Jaehoon Lee, Carrie Grimes Bostock, Keran Rong, Jonathan Caton, Pedram Pejman, Filip Pavetic, Geoff Brown, Vivek Sharma, Mario Lučić, Rajkumar Samuel, Josip Djolonga, Amol Mandhane, Lars Lowe Sjösund, Elena Buchatskaya, Elspeth White, Natalie Clay, Jiepu Jiang, Hyeontaek Lim, Ross Hemsley, Jane Labanowski, Nicola De Cao, David Steiner, Sayed Hadi Hashemi, Jacob Austin, Anita Gergely, Tim Blyth, Joe Stanton, Kaushik Shivakumar, Aditya Siddhant, Anders Andreassen, Carlos Araya, Nikhil Sethi, Rakesh Shivanna, Steven Hand, Ankur Bapna, Ali Khodaei, Antoine Miech, Garrett Tanzer, Andy Swing, Shantanu Thakoor, Zhufeng Pan, Zachary Nado, Stephanie Winkler, Dian Yu, Mohammad Saleh, Loren Maggiore, Iain

Barr, Minh Giang, Thais Kagohara, Ivo Danihelka, Amit Marathe, Vladimir Feinberg, Mohamed Elhawaty, Nimesh Ghelani, Dan Horgan, Helen Miller, Lexi Walker, Richard Tanburn, Mukarram Tariq, Disha Shrivastava, Fei Xia, Chung-Cheng Chiu, Zoe Ashwood, Khuslen Baatarsukh, Sina Samangooei, Fred Alcober, Axel Stjerngren, Paul Komarek, Katerina Tsihlas, Anudhyan Boral, Ramona Comanescu, Jeremy Chen, Ruibo Liu, Dawn Bloxwich, Charlie Chen, Yanhua Sun, Fangxiaoyu Feng, Matthew Mauger, Xerxes Dotiwalla, Vincent Hellendoorn, Michael Sharman, Ivy Zheng, Krishna Haridasan, Gabe Barth-Maron, Craig Swanson, Dominika Rogozińska, Alek Andreev, Paul Kishan Rubenstein, Ruoxin Sang, Dan Hurt, Gamaleldin Elsayed, Renshen Wang, Dave Lacey, Anastasija Ilić, Yao Zhao, Lora Aroyo, Chimezie Iwuanyanwu, Vitaly Nikolaev, Balaji Lakshminarayanan, Sadegh Jazayeri, Raphaël Lopez Kaufman, Mani Varadarajan, Chetan Tekur, Doug Fritz, Misha Khalman, David Reitter, Kingshuk Dasgupta, Shourya Sarcar, Tina Ornduff, Javier Snaider, Fantine Huot, Johnson Jia, Rupert Kemp, Nejc Trdin, Anitha Vijayakumar, Lucy Kim, Christof Angermueller, Li Lao, Tianqi Liu, Haibin Zhang, David Engel, Somer Greene, Anaïs White, Jessica Austin, Lilly Taylor, Shereen Ashraf, Dangyi Liu, Maria Georgaki, Irene Cai, Yana Kulizhskaya, Sonam Goenka, Brennan Saeta, Kiran Vodrahalli, Christian Frank, Dario de Cesare, Brona Robenek, Harry Richardson, Mahmoud Alnahlawi, Christopher Yew, Priya Ponnapalli, Marco Tagliasacchi, Alex Korchemniy, Yelin Kim, Dinghua Li, Bill Rosgen, Zoe Ashwood, Kyle Levin, Jeremy Wiesner, Praseem Banzal, Praveen Srinivasan, Hongkun Yu, Çağlar Ünlü, David Reid, Zora Tung, Daniel Finchelstein, Ravin Kumar, Andre Elisseeff, Jin Huang, Ming Zhang, Rui Zhu, Ricardo Aguilar, Mai Giménez, Jiawei Xia, Olivier Dousse, Willi Gierke, Soheil Hassas Yeganeh, Damion Yates, Komal Jalan, Lu Li, Eri Latorre-Chimoto, Duc Dung Nguyen, Ken Durden, Praveen Kallakuri, Yaxin Liu, Matthew Johnson, Tomy Tsai, Alice Talbert, Jasmine Liu, Alexander Neitz, Chen Elkind, Marco Selvi, Mimi Jasarevic, Livio Baldini Soares, Albert Cui, Pidong Wang, Alek Wenjiao Wang, Xinyu Ye, Krystal Kallarackal, Lucia Loher, Hoi Lam, Josef Broder, Dan Holtmann-Rice, Nina Martin, Bramandia Ramadhana, Daniel Toyama, Mrinal Shukla, Sujoy Basu, Abhi Mohan, Nick Fernando, Noah Fiedel, Kim Paterson, Hui Li, Ankush Garg, Jane Park, DongHyun Choi, Diane Wu, Sankalp Singh, Zhishuai Zhang, Amir Globerson, Lily Yu, John Carpenter, Félix de Chaumont Quitry, Carey Radebaugh, Chu-Cheng Lin, Alex Tudor, Prakash Shroff, Drew Garmon, Dayou Du, Neera Vats, Han Lu, Shariq Iqbal, Alex Yakubovich, Nilesh Tripuraneni, James Manyika, Haroon Qureshi, Nan Hua, Christel Ngani, Maria Abi Raad, Hannah Forbes, Anna Bulanova, Jeff Stanway, Mukund Sundararajan, Victor Ungureanu, Colton Bishop, Yunjie Li, Balaji Venkatraman, Bo Li, Chloe Thornton, Salvatore Scellato, Nishesh Gupta, Yicheng Wang, Ian Tenney, Xihui Wu, Ashish Shenoy, Gabriel Carvajal, Diana Gage Wright, Ben Bariach, Zhuyun Xiao, Peter Hawkins, Sid Dalmia, Clement Farabet, Pedro Valenzuela, Quan Yuan, Chris Welty, Ananth Agarwal, Mia Chen, Wooyeol Kim, Brice Hulse, Nandita Dukkipati, Adam Paszke, Andrew Bolt, Elnaz Davoodi, Kiam Choo, Jennifer Beattie, Jennifer Prendki, Harsha Vashisht, Rebeca Santamaria-Fernandez, Luis C. Cobo, Jarek Wilkiewicz, David Madras, Ali Elqursh, Grant Uy, Kevin Ramirez, Matt Harvey, Tyler Liechty, Heiga Zen, Jeff Seibert, Clara Huiyi Hu, Mohamed Elhawaty, Andrey Khorlin, Maigo Le, Asaf Aharoni, Megan Li, Lily Wang, Sandeep Kumar, Alejandro Lince, Norman Casagrande, Jay Hoover, Dalia El Badawy, David Soergel, Denis Vnukov, Matt Miecnikowski, Jiri Simsa, Anna Koop, Praveen Kumar, Thibault Sellam, Daniel Vlasic, Samira Daruki, Nir Shabat, John Zhang, Guolong Su, Jiageng Zhang, Jeremiah Liu, Yi Sun, Evan Palmer, Alireza Ghaffarkhah, Xi Xiong, Victor Cotruta, Michael Fink, Lucas Dixon, Ashwin Sreevatsa, Adrian Goedeckemeyer, Alek Dimitriev, Mohsen Jafari, Remi Crocker, Nicholas FitzGerald, Aviral Kumar, Sanjay Ghemawat, Ivan Philips, Frederick Liu, Yannie Liang, Rachel Sterneck, Alena Repina, Marcus Wu, Laura Knight, Marin Georgiev, Hyo Lee, Harry Askham, Abhishek Chakladar, Annie Louis, Carl Crous, Hardie Cate, Dessie Petrova, Michael Quinn, Denese Owusu-Afriyie, Achintya Singhal, Nan Wei, Solomon Kim, Damien Vincent, Milad Nasr, Christopher A. Choquette-Choo, Reiko Tojo, Shawn Lu, Diego de Las Casas, Yuchung Cheng, Tolga Bolukbasi, Katherine Lee, Saaber Fatehi, Rajagopal Ananthanarayanan, Miteyan Patel, Charbel Kaed, Jing Li, Jakub Sygnowski, Shreyas Rammohan Belle, Zhe Chen, Jaclyn Konzelmann, Siim Põder, Roopal Garg, Vinod Koverkathu, Adam Brown, Chris Dyer, Rosanne Liu, Azade Nova, Jun Xu, Slav Petrov, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv: 2403.05530*.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Lidong Bing, Yang You, and Luo Si. 2022a. SentBS: Sentence-level beam search for controllable summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10256–10265,

Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022b. MReD: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.

Ondrej Skopek, Rahul Aralikatte, Sian Gooding, and Victor Carbune. 2023. Towards better evaluation of instruction-following: A case-study in summarization. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 221–237, Singapore. Association for Computational Linguistics.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Hybrid. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023a. Benchmarking large language models for news summarization.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023b. Gpt-4v(ision) as a generalist evaluator for vision-language tasks.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

A Appendix

A.1 Generation Prompts

Figures 4 and 5 show the general prompting schema for summary generation. Notably, we use the guidance[7] framework for these generations.

```
{{#system~}}
[system](#instructions)
# Role
You are a help assistant.

## Task
You are an NLP assistant whose purpose is to summarize any given
article. You should summarize all important information concisely
so that it captures the essence of the main passage. Note that
the generated summary should be in {{language}}. The generated
summary should be rated well across the below metrics. Along with
the summary give a brief justification of why it is rated well for
the given metrics.

### Metrics
You are given below the metrics, with their descriptions and
scoring schema in a JSON format.

⟨JSON of metrics and their descriptions⟩

### Outputs
- The summary
   - A concise, meaningful, and informative summary of the main
passage.
- A justification
   - A brief justification as to why the generated summary is rated
well across one or more of the given metrics.
- The Output format
   - Output must always be a JSON with format given by the JSON
schema below.

⟨JSON output format structure⟩
{{~/system}}

{{#user~}}
  passage = "{{passage}}"
  language = "{{language}}"
{{~/user}}

{{#assistant~}}
{{gen 'evaluation' max_tokens=2048 temperature=0.0}}
{{~/assistant}}
```

Figure 4: Good Generation Prompt

A.2 Human Evaluation Instructions

Figure 6 shows detailed instructions provided to the annotators. The metrics are explained in §3.2.

A.3 Annotator Agreement

Table 6 shows the Fleiss' Kappa κ and pairwise agreement (F1) values for various metrics and languages.

```
{{#system~}}
[system](#instructions)
# Role
You are a help assistant.

## Task
You are an **adversarial** NLP assistant whose purpose is to badly
summarize any given article. Note that the generated summary
should be in {{language}}. The generated summary should be rated
low across one or more of the metrics described below. **In some
cases generate problematic content and hallucinations as well**.
Along with the summary give a brief justification of why it is
rated low for the given metrics.

### Metrics
You are given below the metrics, with their descriptions and
scoring schema in a JSON format.

⟨JSON of metrics and their descriptions⟩

### Outputs
- An adversarial summary
   - An adversarial summary of the main passage, which is rated low
across one or more of the given metrics
- A justification
   - A brief justification as to why the generated adversarial
summary is rated low across one or more of the given metrics.
- The Output format
   - Output must always be a JSON with format given by the JSON
schema below.

⟨JSON output format structure⟩
{{~/system}}

{{#user~}}
  passage = "{{passage}}"
  language = "{{language}}"
{{~/user}}

{{#assistant~}}
{{gen 'evaluation' max_tokens=2560 temperature=1.0}}
{{~/assistant}}
```

Figure 5: Bad Generation Prompt

Lang	H	LA	OCQ	PC	TQ
AR	0.65 / 0.89	0.66 / 0.89	0.61 / 0.85	0.65 / 0.93	0.61 / 0.77
BN	0.83 / 0.97	0.64 / 0.81	0.62 / 0.82	0.0 / 1.0	0.64 / 0.78
EN	0.54 / 0.85	0.73 / 0.86	0.63 / 0.82	1.0 / 1.0	0.61 / 0.77
FR	0.94 / 0.97	0.93 / 0.99	0.91 / 0.97	1.0 / 1.0	0.84 / 0.9
HI	0.68 / 0.9	0.69 / 0.87	0.62 / 0.83	0.78 / 0.94	0.6 / 0.78
JA	0.97 / 0.99	0.92 / 0.95	0.89 / 0.93	0.0 / 0.99	0.98 / 0.99
RU	0.99 / 0.99	0.97 / 0.98	0.88 / 0.93	0.9 / 0.99	0.89 / 0.94
SW	0.85 / 0.93	0.71 / 0.82	0.73 / 0.84	0.62 / 0.86	0.72 / 0.84
TR	0.66 / 0.84	0.95 / 0.97	0.76 / 0.84	0.0 / 1.0	0.8 / 0.87
ZH	1.0 / 1.0	0.68 / 0.94	0.65 / 0.91	1.0 / 1.0	0.65 / 0.82

Table 6: Annotator agreement values for various languages and metrics in our dataset, reported as Fleiss' Kappa (κ) / Pairwise Agreement (F1).

```
"name": "linguistic_acceptability",

"description": "Linguistic acceptability means does this sound
right to a native speaker?, not does this stick to the rules of
the grammar.",

"scoring": "0: not acceptable; 1: some weird things but ok; 2: no
errors found/acceptable."
```

Figure 7: Metric description for simple instructions (Linguistic Acceptability).

A.4 Instructions

Figures 7 and 8 show examples of simple and detailed instructions for Linguistic Acceptability respectively. Rest of the prompts can be found in

Figure 6: Detailed task instructions provided to the annotators.

Hada et al. (2024).

A.5 Radar Plots of F1-scores

Figures 9 and 10 are a pictorial representation of Table 3.

A.6 Qualitative Analysis

Table 7 shows some samples of cases where either of the LLM scores agree with human aggregate scores, but there are some discrepancies in their justification.

```
"name": "linguistic_acceptability",

"description": "Linguistic acceptability pertains to the degree to which a given language structure (e.g., phrase, sentence, discourse) aligns
with the implicit norms and rules of a native speaker's linguistic intuition. In the study of language, it's distinct from 'grammaticality', which
is a stricter and narrower concept based on the prescriptive rules of a language. Linguistic acceptability, on the other hand, captures broader
native-speaker intuitions and encompasses factors like fluency, idiomacy, and appropriateness in context. In the context of language models,
evaluating linguistic acceptability involves assessing the output of the model not just for its adherence to grammar rules, but for its overall
fit within the natural, expected, and intuitive contours of fluent human language. The scoring rubric is described below, with a few possible
reasons (which might not be exhaustive) for a given score.",

"scoring": {
    "0": {
        "(a)": "Sentences that lack clear syntactic structure.",
        "(b)": "Usage of non-existent or incorrect words.",
        "(c)": "Grossly inappropriate word choices for a given context."
    },
    "1": {
        "(a)": "Overly verbose or stilted phrasing.",
        "(b)": "Minor grammatical errors that do not impede understanding.",
        "(c)": "Use of a word that's technically correct but not the most appropriate for context."
    },
    "2": {
        "(a)": "Seamless integration of contextually relevant vocabulary",
        "(b)": "Effective use of idiomatic expressions without sounding forced.",
        "(c)": "Sentences that reflect natural rhythm, emphasis, and intonation of spoken language."
    }
}
```

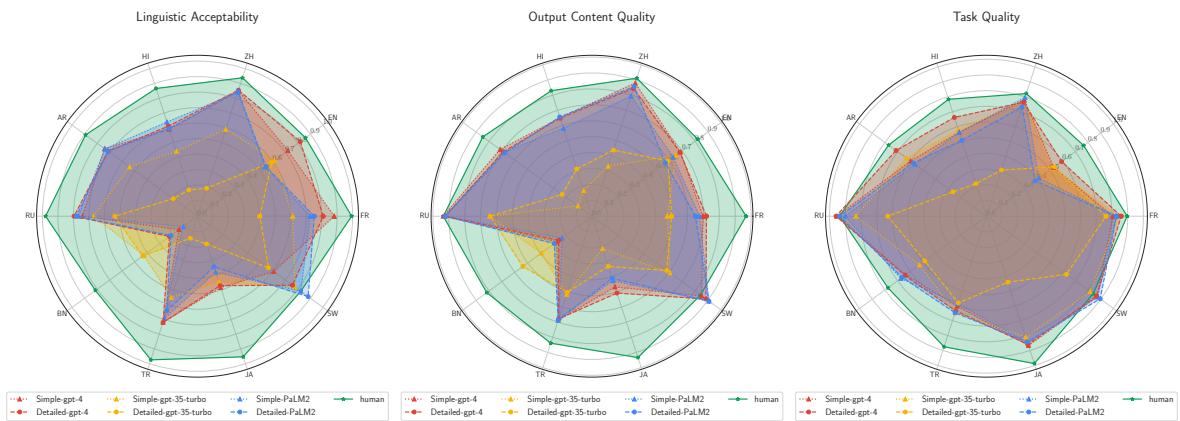Figure 8: Metric description for complex instructions (Linguistic Acceptability).

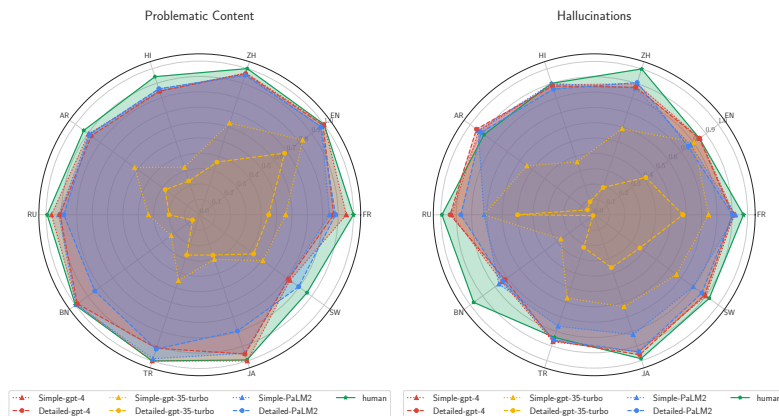Figure 9: F1 scores for LA, OCQ, and TQ for various languages, models, and prompting strategies.

Figure 10: F1 values for PC and H for various languages, models, and prompting strategies.

No.	Lang.	Metric	Human Scores and Comments	GPT-4 Scores and Justification	PaLM2 Scores and Justification
1	Bengali	TQ	'A1': 0, 'The summary part is very short.', 'A2': 1, 'This summary section was not completed. Language construction was good and some flaws but otherwise okay and summary section more improvement was needed.', 'A3': 2, 'This content has good co-relation between different thoughts and ideas. The summary of the content is not complete.'	1, However, it misses out on finer details and nuances such as the reason behind the boycott, the reactions of different countries, and the political implications of the boycott. Therefore, the summary partially aligns with the passage..	1, The summary is a good summary of the passage. It captures the main points of the passage and provides a brief overview of the topic. However, the summary does not go into as much detail as the passage and does not provide as much information about the topic.
2	Bengali	OCQ	'A1': 0, 'Difficult to understand from beginning to end.', 'A2': 1, ' summary section can be more improvement is needed.', 'A3': 1, "In summary part as well, the sentences are irrelevant with other sentences of the paragraph. "	1, the summary is somewhat accurate in terms of content relevance but lacks clarity and cohere	1, The summary is not in the same language as the passage
3	Swahili	LA	'A1': 1, 'The summary is inaccurate and incomplete and thus not a true reflection of the original passage. Some crucial information is missing while other information is not accurate. ', 'A2': 0, NA, 'A3': 0, NA	1, The summary provided in Swahili is grammatically correct and uses appropriate words. However, it is overly simplified and does not capture the essence of the passage	0, The summary is not fluent and does not reflect natural rhythm, emphasis, and intonation of spoken language. It contains several grammatical errors and uses words that are not appropriate for the context
4	Russian	LA	'A1': 0, 'Summary with no syntax no punctuation, just a set of words. Not even key words. Hallucinations: (Not present in the text)', 'A2': 0, 'Summary is not what the text is about. ', 'A3': 0, NA	0, The summary provided does not accurately reflect the content of the passage. It is not clear, coherent, or contextually relevant	0, The summary is not in Russian. It is in English
5	Turkish	LA	'A1': 1, NA, 'A2': 1, 'Irrelevant summary context.', 'A3': 1, 'Totally meaningless utterances.'	1, The summary uses correct grammar and sentence structure, and there are no non-existent or incorrect words. However, the summary is overly verbose and uses stilted phrasing.	1, The summary is not linguistically acceptable. It contains several grammatical errors, such as the use of the word 'gok tasl' (meteorite) to refer to a handbag, and the use of the word 'muze' (museum) to refer to a planet

Table 7: Samples from METAL dataset with human and LLM (detailed variation) scores and comments. Note: We only include snippets from the full comment.