How Reliable Are Automatic Evaluation Methods for Instruction-Tuned LLMs?

Ehsan Doostmohammadi[†] **Oskar Holmström**[†] **Marco Kuhlmann**
Linköping University
ehsan.doostmohammadi@liu.se

Abstract

Work on instruction-tuned Large Language Models (LLMs) has used automatic methods based on text overlap and LLM judgments as cost-effective alternatives to human evaluation. In this paper, we perform a meta-evaluation of such methods and assess their reliability across a broad range of tasks. In evaluating how well automatic methods align with human evaluations, correlation metrics are the most commonly employed method despite their inherent limitations when dealing with ties and different scales. To address these shortcomings, we use Pairwise Accuracy as an alternative to standard correlation measures. We observe that while automatic evaluation methods can approximate human ratings under specific conditions, their validity is highly context-dependent. Specifically, the simple ROUGE-L metric correlates very well with human ratings for short-answer English tasks but is unreliable in free-form generation tasks and cross-lingual scenarios. The effectiveness of the more advanced method of using GPT-4 as a judge diminishes significantly if reference answers are not included in the prompt, which is the scenario where this method has the potential to provide the most value compared to other metrics. Our findings enhance the understanding of how automatic methods should be applied and interpreted when developing and evaluating instruction-tuned LLMs.

1 Introduction

A key strength of the current generation of Large Language Models (LLMs) is their capacity to learn new tasks from instructions, either in-context (Mishra et al., 2022; Sanh et al., 2022; Wei et al., 2022) or in a dedicated fine-tuning phase (Wang et al., 2022). The field has also seen the development of methods to adapt LLMs to new languages, for example, through continued fine-tuning (Muennighoff et al., 2023), alignment with translation pairs (Ranaldi et al., 2024), and instruction tuning on additional languages (Chen et al., 2024; Kew et al., 2023).

The gold standard for evaluating generative tasks is human annotation, but this scales poorly due to high costs and time constraints. Consequently, the most common approach for assessing generative LLMs is using automated evaluation techniques. Among these, two popular methods are measuring text overlap with ROUGE-L (Lin, 2004) and utilizing existing LLMs as automatic judges (Zheng et al., 2023); however, these methods only approximate human judgment, prompting questions about their reliability. While previous research has found that automatic evaluation methods correlate well with human assessments (Wang et al., 2022; Zheng et al., 2023), it is important to recognize that these findings generalize over tasks of very different types and in different languages. Additionally, correlation measures may not provide reliable estimates of alignment with human ratings, as they are limited in their ability to deal with ties and constant scores, which are common in human annotations (Deutsch et al., 2023).

In this paper, we provide a thorough analysis of two widely-used automatic methods for approximating human judgments, ROUGE-L and LLM-as-a-judge. Additionally, we experiment with BERTSCORE, a semantic text similarity measure, to assess its potential utility. We study the reliability of the three measures across a broad range of English-language tasks. We also perform experiments on Swedish as an initial study on the reliability of these metrics across languages. Instead of using correlation measures, we employ Pairwise Accuracy (Deutsch et al., 2023) to quantify the alignment with human ratings. Our overall goal is to increase our understanding of the reliability of automatic evaluation methods and to establish guidelines regarding their appropriate usage.

[†] Equal contribution

Findings of the Association for Computational Linguistics: EACL 2024, pages 6321–6336
November 12-16, 2024 ©2024 Association for Computational Linguistics

Our contributions can be summarized as follows:

- We adopt Pairwise Accuracy (PA) with tie calibration (Deutsch et al., 2023) to enable robust comparisons between metrics, as we observe a high prevalence of tied ratings which renders common metrics, such as Kendall's τ and Spearman's ρ, unreliable.

- We show that GPT-4 aligns well with human judgments when gold reference answers are available. However, its reliability diminishes in the absence of these references, where it shows an overly positive bias. This is especially problematic for free-form tasks, since GPT-4 is commonly used in such settings.

- We find that GPT-4, while being the best tool for evaluating generations, can be replaced by faster and far less costly alternatives under certain conditions. In particular, we show that ROUGE-L offers a cost-effective alternative to GPT-4 for short-answer tasks, while BERTSCORE shows promising results in long-answer tasks.

- We observe a decrease in alignment with humans in non-English tasks for ROUGE-L and GPT-4 in situations where it does not have access to gold references. This suggests that it could be challenging to use automatic evaluation methods for lesser-resourced languages.

2 Related Work

We start by reviewing the research on the automatic evaluation of generated text, the use of LLMs as evaluators, and the methods applied to assess the alignment of metrics with human preferences.

2.1 Automatic Evaluation

For short-form tasks such as multiple-choice question answering, assessing the quality of model outputs appears feasible through standard classification metrics like accuracy and F_1-score (Li et al., 2023a; Lai et al., 2023). While such an evaluation can be precise, it is rather strict and can only provide a fair performance assessment if the model does not deviate from the instructed format. However, this easily happens as the tasks diverge from the training data or get more complex. Surface-level similarity measures such as ROUGE-L (Honovich et al., 2023; Wang et al., 2023; Mishra et al., 2022; Yin et al., 2023; Lai et al., 2023; Li et al., 2023b) are more forgiving regarding formatting inconsistencies, but still lack the sophistication to be effective in tasks where free-from answers are expected.

2.2 Evaluation Using LLMs

An increasingly common method for evaluating instruction fine-tuned models is to use powerful LLMs as automatic judges (Peng et al., 2023; Gilardi et al., 2023; Chen et al., 2024; Kew et al., 2023). Zheng et al. (2023) propose three different variations: (1) pairwise comparison, which asks the LLM to choose its preferred answer or declare a tie; (2) single-answer grading, in which the LLM is asked to assign a score to an answer; and (3) reference-guided grading, in which the model is provided with a reference solution (if available). An approach similar to the second one is used by M^3IT (Li et al., 2023b) to evaluate the accuracy, relevance, and naturalness using GPT-4 in a multimodal scenario.

2.3 Meta-Evaluation

Human evaluation is the gold standard of assessment in natural language processing, but is not widely used in the literature due to its high costs. Instead, authors have turned to automatic evaluation measures that correlate well with human judgments. Wang et al. (2022) find a consistently strong correlation between ROUGE-L scores with accuracy across different models and task types, indicating that it is a good proxy for accuracy in classification tasks with short outputs. For machine translation, Zhang* et al. (2020) show that BERTSCORE is better correlated to human judgments than previous metrics, but Hanna and Bojar (2021) also identify setups where it fails.

The recent work on automating evaluation processes and leveraging LLMs has demonstrated substantial agreement with human ratings. Zheng et al. (2023) show that GPT-4's judgments align with human evaluations at over 80% agreement, reaching levels comparable to human–human agreement. Zhou et al. (2023) also report agreement levels between GPT-4 and human annotators on a par with human–human agreements. There is also work on the meta-evaluation of automatic metrics for chat and summarization (Shen et al., 2023; Chiang and Lee, 2023) using different criteria, and on aligning language model evaluations better with human preferences, such as Liu et al. (2024) and Chan et al. (2024).

3 Methodology

In this section, we present the data and instruction-tuned models used in our study. We provide an overview of the automatic metrics we employed to assess model performance and detail our approach to conducting a meta-evaluation of these metrics.

3.1 Data

As our training data, we use the Cleaned Alpaca Dataset[1], which corrects errors found in the original Alpaca (Taori et al., 2023). For testing, we use Natural Instructions v2 (NIv2) (Mishra et al., 2022; Wang et al., 2022), which spans a diverse range of tasks, including classification, question answering, free-form text generation, and reasoning. This enables fine-grained testing.

Sample Selection Because of our limited annotation budget (cf. §3.3), we select 20 from the 119 (English-language) tasks available in NIv2. We aim to find tasks that (a) cover a range of difficulty levels, (b) involve both short and long free-form answers, and (c) are diverse in task types while leaving some type overlap for control purposes. For a full description of the selected tasks, we refer to Appendix A. From each task, we pick 15 random samples, leaving us with 300 samples in total.

Translation To study the metrics' reliability across the language dimension, we translate both our training and our test data to Swedish using `GPT-3.5-turbo`. The prompt template and hyperparameters used for translation can be found in Appendix B. Previous work has shown that the automatic translation of the Alpaca dataset produces high-quality results with low noise levels (Holmström and Doostmohammadi, 2023; Li et al., 2023a). In addition to the two monolingual train datasets, we create an equally-sized English–Swedish bilingual train set by replacing a random 50% of the samples in the Cleaned Alpaca Dataset with their Swedish translations. Our purpose with this bilingual training set is to conduct a controlled study with more diverse bilingual data.

3.2 Instruction Tuning

We instruction-tune three base models in this study: `LLaMA2-7b`, `LLaMA2-13b` and `GPT-SW3-6.7b`. Our selection accounts for different model sizes, pretraining languages, and performance.

- `LLaMa2` (Touvron et al., 2023) is trained on mainly English data; only 0.15% of the pretraining data is Swedish. We use both the 7B and the 13B parameter versions of the model.

- `GPT-SW3` (Ekgren et al., 2024) is a `GPT-2`-based language model mainly trained on North Germanic languages and English, where 26% is Swedish, and 40% is English. `GPT-SW3` exhibits the lowest perplexity on Swedish (see Appendix C.1), which is unsurprising as `LLaMA2` models have seen vastly fewer Swedish tokens during pretraining.

For instruction tuning, we use the same training settings, hyperparameters, and prompts as Alpaca, and use DeepSpeed (Rasley et al., 2020) with the same configuration for all models. For more details regarding the implementation, see Appendix D.

Naming Scheme We instruction-tune each of our three base models on the three training datasets (English, Swedish, and bilingual) and test it on either the single training language (for monolingual models) or both languages (for bilingual models). This gives us a combined total of 12 different configurations for our experiments. Throughout the paper, we refer to these by the name of the base model, model size, training dataset, and testing language, all separated by underscores. For example, `SW3_6.7b_ENSV_SV` identifies the experiment where we train the `GPT-SW3` model with 6.7 billion parameters on the bilingual English–Swedish data and test on Swedish.

3.3 Evaluation Methods

Human Assessment To establish the gold standard for our evaluation, we hire three bilingual (English and Swedish) evaluators to assess model outputs based on three criteria: naturalness (how natural and fluent the generated response is), relatedness (whether the response is related to the prompt and follows the required format), and correctness (whether the response is correct, which is our main criterion). While there are some tasks for which these criteria may not be applicable (especially correctness), they are well-suited for our chosen set of tasks. We ask each annotator to rate each criterion on a Likert scale ranging from 1 (significantly deficient) to 3 (completely proficient). For a detailed description of the annotation process and instructions, see Appendix E. The Kendall's τ is 0.74 (averaged over pairs of annotators) and the Fleiss'

[1] https://github.com/gururise/AlpacaDataCleaned

κ (Fleiss and Cohen, 1973) is 0.63, indicating a substantial agreement between human annotators.

Majority Vote To compare human ratings to other metrics, we use a variant of the majority vote. More specifically, we compress the three ratings into that score which is assigned by at least two raters, and fall back to a neutral score of 2 in cases where all raters have given different scores. We prefer our method over the obvious alternative of taking an average because it reduces the impact of outlier ratings. For reference, among the human ratings of correctness (our main criterion), there is a majority vote for 93.5% of samples, providing a robust foundation for our comparisons. We treat the human majority vote as the gold standard and compare other metrics against it.

Performance Metrics Our selection of performance metrics is motivated by the desire to cover commonly used methods on a spectrum from surface-based semantics-based methods. On one end of this spectrum, we use ROUGE-L (Lin, 2004), which measures the textual overlap between a generated response and a reference output as the length of the longest common subsequence. On the other end of the spectrum, we use `GPT-4` as a judge (Zheng et al., 2023).[2] Similar to previous work (Zhou et al., 2023; Kew et al., 2023), we prompt `GPT-4` with the same instructions that we give to human evaluators and ask it to rate based on the same criteria on the same Likert scale. We also prompt `GPT-4` to provide its reasoning before rating, similarly to Kew et al. (2023), whose framework we use for LLM-as-a-judge evaluations. The prompt template used for evaluations is found in Appendix B. In the standard evaluation setting, the gold labels are included in the prompt as a reference for the model; we mark this setting with the suffix `-gold`. To ablate the effects of the access to gold labels, we perform additional experiments with these labels excluded; we mark these with the suffix `-no-gold`. Finally, as a point in-between a purely surface-based and a powerful semantics-based performance metric, we use BERTSCORE (Zhang* et al., 2020), which quantifies semantic overlap in terms of the cosine similarity between contextual embeddings obtained from pretrained language models.

Metric	Tie Proportion	ϵ
Human Ratings	0.557 ± 0.162	0.000
`GPT-4-gold`	0.524 ± 0.154	0.000
ROUGE-L	0.355 ± 0.252	0.061
BERTSCORE	0.104 ± 0.141	0.133

Table 1: The average tie proportion per metric for English-language tasks.

3.4 Meta-Evaluation Method

Pairwise Accuracy with Tie Calibration In this study, we perform a meta-evaluation of both metrics that produce continuous and ordinal ratings. While Spearman's ρ and Kendall's τ are commonly used for such purposes, these metrics fail to handle tasks with constant score vectors or with different rating scales, and they do not reward correct predictions of ties. Ties are especially frequent in Likert-scale human ratings, which the automatic metrics are benchmarked against.

In response, we have chosen to use Pairwise Accuracy with Tie Calibration (PA) for meta-evaluating metrics. Proposed by Deutsch et al. (2023), PA addresses the shortcomings of traditional metrics by including mechanisms to explicitly account for the prevalence of ties, thus providing a fairer assessment of metrics.

PA measures the proportion of correctly ranked pairs, including accurately predicted ties. With values ranging from 0 to 1, the metric is more easily interpreted than traditional correlation metrics such as Spearman's ρ and Kendall's τ. PA includes a tie calibration process by defining a threshold value, ϵ, which specifies what is considered a significant difference between scores. A pair of scores with a difference smaller than ϵ is considered a tie. This is crucial as some metrics are more likely to produce tied values, as can be seen in Table 1. Tie calibration ensures that comparisons between different metrics are fair, regardless of their inclination to predict ties or having different rating scales.

We study the distribution of ties in our data and observe significant variation for different metrics, as shown in Table 1. The average tie proportion for human ratings is 0.557, serving as our benchmark. In contrast, metrics such as `GPT-4`, ROUGE-L, BERTSCORE exhibit varying degrees of tie proportions. `GPT-4` has a similar degree of ties compared to human ratings, while BERTSCORE has considerably lower tie proportions. The significant amount of ties and a constant score vector validate

[2] Running an evaluation (across all models, tasks, and languages) costs around USD 100 and typically takes 15–20 minutes. Though not significantly cheaper than human evaluations, it certainly surpasses it in terms of time efficiency.

Experiment	Natural	Related	Correct
LLaMA2_13b			
EN_EN	99 (95)	73 (80)	47 (51)
SV_SV	99 (94)	76 (76)	36 (37)
ENSV_EN	99 (92)	80 (79)	47 (47)
ENSV_SV	99 (92)	80 (76)	39 (40)
LLaMA2_7b			
EN_EN	99 (91)	72 (76)	40 (44)
SV_SV	99 (88)	65 (68)	35 (34)
ENSV_EN	99 (92)	74 (75)	41 (43)
ENSV_SV	98 (89)	66 (68)	32 (33)
SW3_6.7b			
EN_EN	99 (91)	79 (70)	35 (36)
SV_SV	99 (92)	66 (67)	37 (37)
ENSV_EN	99 (92)	81 (68)	35 (35)
ENSV_SV	99 (85)	68 (63)	35 (36)
Avg. diff.	7.8	4.2	1.3

Table 2: Human evaluation results per model scaled to 0 to 100. For comparison, GPT-4 ratings are included in parentheses for each model and criterion.

the use of PA to enable a reliable meta-evaluation of our metrics. As an example of a constant vector in our case, in Task 034 (cf. §4), there is a constant score of 1 from human raters, illustrating a scenario that could commonly occur in instruction tuning.

Unlike Deutsch et al. (2023), who calculate ϵ for each task, we calculate the optimal ϵ for each metric using data from all tasks. We find that this produces a PA that better reflects the true correlation between human and metric scores, especially for tasks with only ties or mostly ties. Otherwise, for tasks with only ties, the ϵ could be as large as the value range for the metric and treat every pair of scores as ties. As shown in Table 1, our metric-level ϵ correlates well with the number of ties for the metric. With our pre-calculated ϵ values, we compute PA over all models per task, aligning with the No-Grouping setting in Deutsch et al. (2023).

4 Results and Analysis

In this section, we present a comparative analysis of the evaluation methods in terms of their alignment with human assessments.

4.1 Human Evaluation

We present the human evaluation results for each model in Table 2. All models demonstrate the capability to generate natural-sounding text (99% on average) and also perform fairly well in generating relevant responses that adhere to the required format (73% on average). The correctness scores demonstrate that the models are capable of generating largely accurate answers. There is also a notable diversity among the models regarding correctness. This diversity is crucial because we seek a range of models with varied problem-solving abilities, rather than just strong models that produce highly accurate results.

The ratings of GPT-4 closely align with human ratings for correctness, showing slightly more distance in relatedness, and even more in naturalness. (The average differences between human and GPT-4 ratings are summarized in the final row of Table 2.) Based on these results, in the rest of the paper, we focus solely on correctness. We prioritize correctness since it is the most important criterion for determining the usefulness of LLMs. Moreover, comparing our metrics using the other criteria could be problematic. For instance, while ROUGE-L scores serve as a reasonable proxy for correctness, they are less suited for evaluating naturalness or relatedness.

4.2 Meta-Evaluation of Metrics

The metric with the highest alignment with human ratings is GPT-4-gold which achieves an average PA of 0.81 for English short- and long-answer tasks, followed by ROUGE-L with 0.75, BERTSCORE with 0.66, and GPT-4-no-gold with 0.62. For a comparison of all the results across different task types, languages, and metrics, see Table A3. Nonetheless, a more fine-grained analysis shows a different pattern, which we will discuss in this section.

Finding 1: *All metrics struggle to assess model performance on long-answer tasks.*

As presented in Figures 1, 2, and 4, the alignment with humans drops for long-answer tasks compared to short-answer tasks. On average, for English long-answer tasks, GPT-4-gold at 0.58 is the highest, followed by 0.54 for BERTSCORE, 0.50 for GPT-4-no-gold, and 0.48 for ROUGE-L.

For Tasks 613, 677 and 1659 where the models must generate free-form text, ROUGE-L scores are lower than human ratings accompanied with low to medium PAs. This is due to the large set of possible solutions and the small set of gold label answers, often consisting of a single sentence, that the output should match to receive a high ROUGE-L score. For some of these tasks, e.g., title generation, it is impossible to cover the set of conceivable solutions which makes ROUGE-L unreliable for these types of tasks. The same trend is present for GPT-4, which could be explained by this model's relying

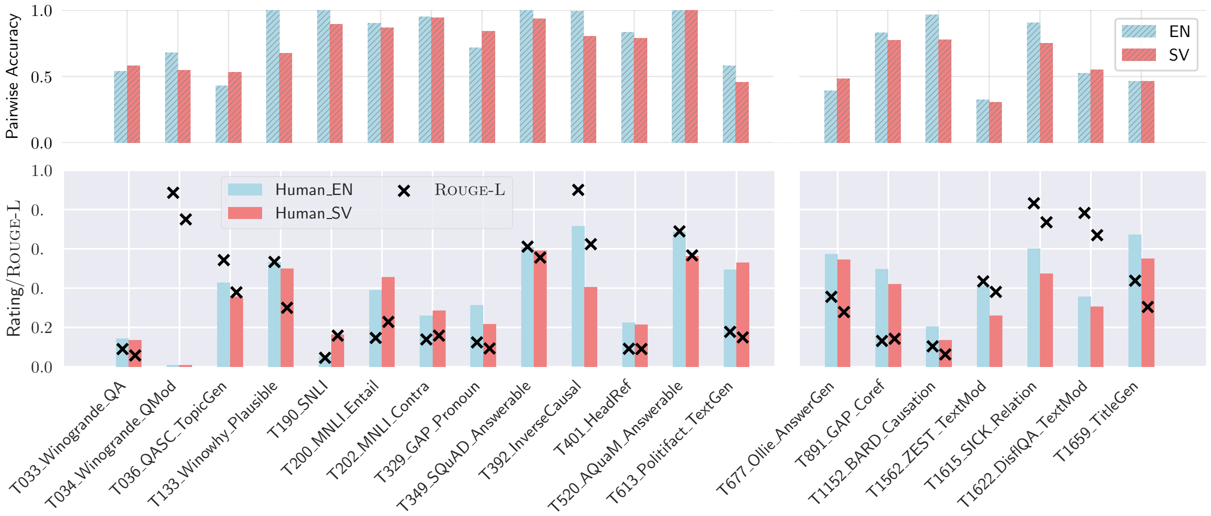

Figure 1: Human ratings and ROUGE-L scores per task and test language at the bottom and their PA at the top. Human scores are normalized to a range between 0 and 1. Short-answer tasks are on the left and long-answer ones are on the right.

on the gold labels during evaluation, which may make it overly critical of responses that show less conformity with the provided reference answers.

In contrast, for Tasks 1562 and 1622, where models are tasked with modifying sentences, `GPT-4` assigns higher ratings compared to humans. `GPT-4` struggles to reliably assess whether our models have met the instructional criteria in such tasks. For instance, in Task 1562, the objective is to generate paraphrases of questions while making as many alterations as possible. When models only introduce minor changes, such as changing *bring* to *take* in the sentence "Can I bring my mountain bike with me to this national park?", `GPT-4` often rates this 3 on correctness.

Finding 2: `GPT-4` *needs reference answers in the prompt.*

When we remove the gold answers from the prompt, `GPT-4`'s alignment with human ratings decreases significantly, from 0.81 PA to 0.62 PA on English. Full results are presented in Table A3. For short-answer tasks in English, the reduction is 0.25, and for long-answer tasks, it is 0.09, which underscores `GPT-4`'s struggle when it lacks a reference for exact matching. We attribute this reduction to the increased complexity of both solving and rating the task, which is more pronounced in Swedish, where there is a higher incidence of models over-generating. While such over-generation may lead to lower human ratings, it could be favored by `GPT-4` due to its bias towards longer and more verbose outputs (Zheng et al., 2023).

A notable observation is that without gold label references, `GPT-4-no-gold` is generally more prone to higher ratings, as seen in Figure 4's lower plot. While `GPT-4-gold` closely aligns with human judgments for incorrect outputs (88% for 1's and 81% for 3's), `GPT-4-no-gold` shows an alignment of 65% for 1's and 84% for 3's, indicating an opposite trend. This demonstrates the excessively positive stance of `GPT-4-no-gold`, also noted by Hada et al. (2024) in a different scenario. The positive bias is particularly evident in long-answer tasks and challenging short-answer tasks, such as Tasks 190 and 401. This tendency underscores the importance of gold labels as references to help align the `GPT-4`'s judgments with humans in most tasks.

However, we note that for some long generation tasks (Tasks 613, 677, and 1659) the scores of human raters are lower than for other tasks even when gold labels are present, as seen in Figure 2. Gold references can therefore be restricting for these types of long generation tasks, where models can have correct answers that diverge from the gold label. This is problematic as it is to these types of tasks that LLM-as-a-judge is often applied, and where it could bring the most value compared to other metrics, particularly due to the multitude of potential correct answers.

Finding 3: *For short-answer tasks,* ROUGE-L *is as effective as* `GPT-4-gold`.

With a PA at 0.90 for English short-answer tasks, ROUGE-L is nearly as well-aligned with human ratings as `GPT-4-gold`, which has a PA at 0.93. Our

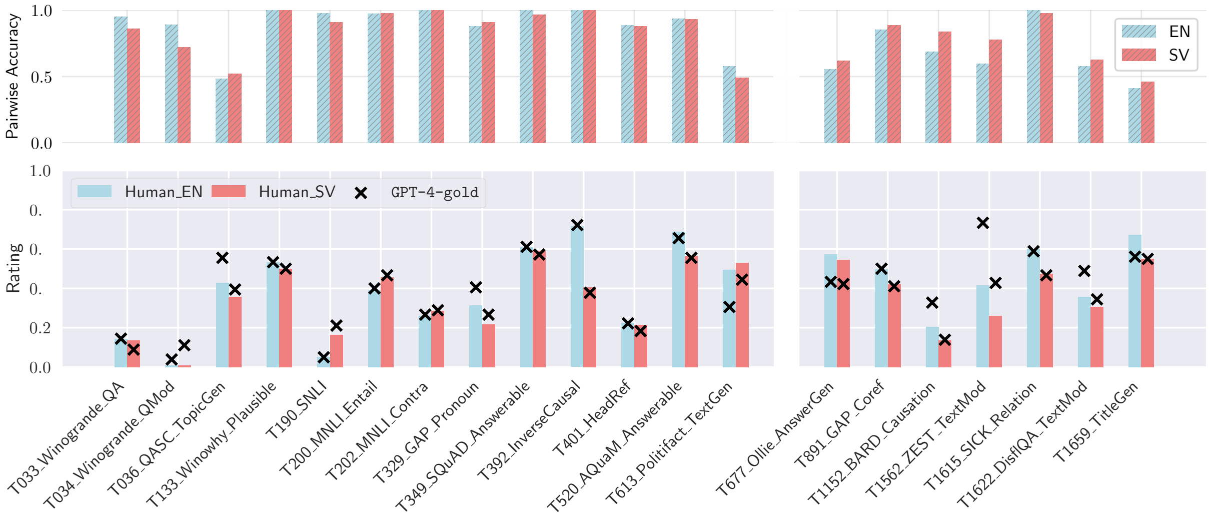

Figure 2: Human and GPT-4-gold's ratings per task and test language on the bottom and their PA on top. Ratings are normalized to a range of 0 to 1. Short-answer tasks are on left and long-answer ones on right.

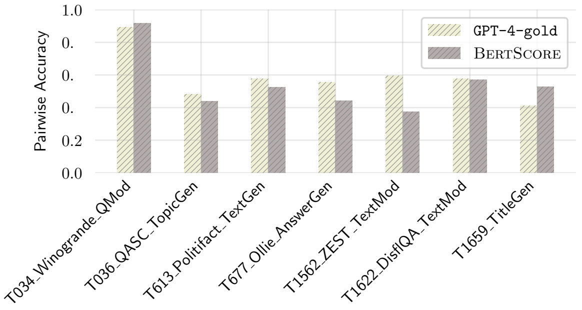

Figure 3: Pairwise accuracy between GPT-4-gold and BERTSCORE for long-answer English tasks.

strong results for ROUGE-L are in line with findings by Wang et al. (2022), who report a high correlation with humans for classification tasks. With that said, there are instances where ROUGE-L has low alignment with human majority rating for short-answer tasks, for example when there is high word overlap between possible answer choices. For example, in Task 1615 the labels are "B entails A", "B contradicts A", or "B neutral A". A wrong answer for this task yields a ROUGE-L score of 0.66, as long as the answer is in the possible answers space. The same issue is observed for Task 392 where the label space is "Plausible" and "Not Plausible".

These types of issues also make it problematic to report average ROUGE-L scores across tasks since a baseline model that always makes wrong predictions could inflate its score beyond its actual performance level. However, as previously discussed, ROUGE-L correlation with humans and GPT-4-gold does not carry over to long-answer tasks, which makes it only suitable as a replacement for GPT-4 when evaluating short-answer tasks. It is important to note that while ROUGE-L demonstrates strong agreement with humans, GPT-4-gold scores are more interpretable as they better align with human judgments, as illustrated at the bottom of Figures 1 and 2.

Finding 4: BERTSCORE *demonstrates strong performance in long-answer tasks.*

With a PA of 0.54 for long-answer tasks, BERTSCORE shows a alignment with humans comparable with GPT-4-gold, which scores 0.58. A comparison of BERTSCORE and GPT-4-gold's PAs are shown in Figure 3. For complete results of BERTSCORE see Appendix C.3. BERTSCORE achieves comparable results to GPT-4-gold on all long-answer tasks, only underperforming on some of them, particularly Task 1622, where it captures the first criterion which requires a high semantic similarity between the two, but fails to take into account whether enough words have been changed, such as in cases involving synonyms, which is explicitly mentioned in the instructions.

Finding 5: *Swedish presents a challenge for certain metrics.*

For ROUGE-L, a reduction of 0.074 in PA is observed for Swedish compared to English. In contrast, GPT-4-gold experiences no significant reduction when switching from English to Swedish. However, GPT-4-no-gold is less consistent, showing a reduction of 0.044. GPT-4's decrease in performance when it does not have access to gold

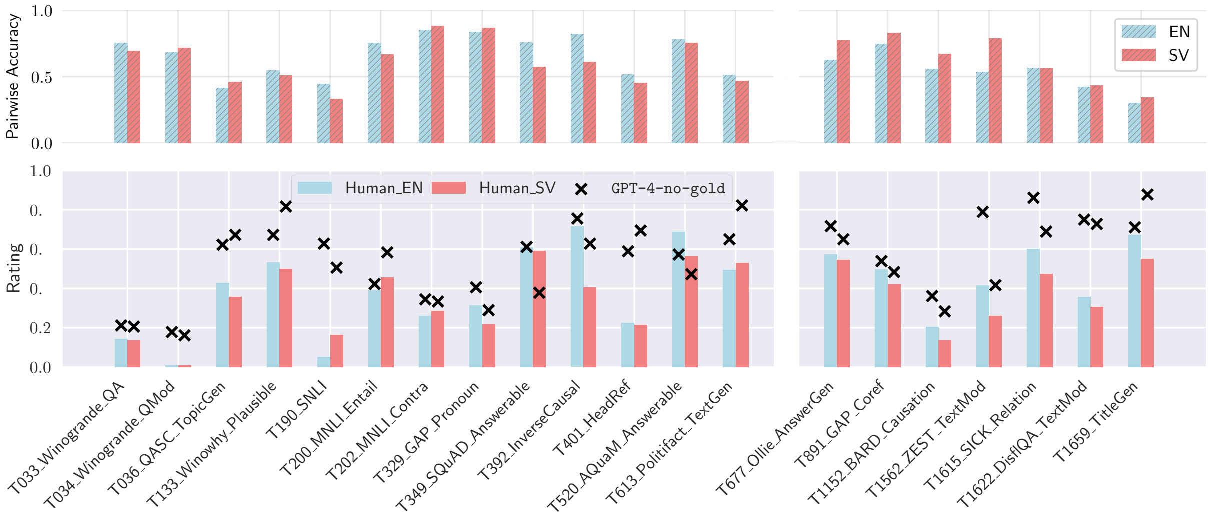

Figure 4: Human and GPT-4-no-gold's ratings per task and test language on the bottom and their PA on top. Ratings are normalized to a range of 0 to 1. Short-answer tasks are on left and long-answer ones on right.

references for Swedish outputs suggests that the model has more difficulties solving the task when it is in another language. We believe the effect could be even more pronounced for languages with less training data than Swedish or less typologically similar to English. Further studies would be necessary to investigate this hypothesis.

While ROUGE-L does not take language into account, it may be less reliable a measurement for languages other than English due to models' failing to adhere to the required format. For instance, we observe that Swedish models have difficulties following instructions, even for short-answer tasks. They sometimes generate synonyms to the true labels, e.g., *sannolikt* "probable" instead of *troligt* "likely", an effect that could stem from seeing less data in Swedish and therefore having less reliable instruction-following capabilities. This is particularly concerning given the prevalence of work utilizing automatic evaluation measures across different languages.

5 Conclusions and Future Work

This study provides insights into the methods we use to evaluate language model generations, focusing on when automatic metrics align with human annotators and what the best metric is under different scenarios. We are the first to do a broader meta-evaluation study where we compare GPT-4-as-a-judge and traditional metrics with a methodology that allows for reliable comparisons between metrics. We recommend using Pairwise Accuracy (PA) with Tie Calibration for meta-evaluation. This method effectively handles ties, which are prevalent when using human and GPT-4 ratings, making it a reliable tool for assessing metric performance against human ratings.

Our main finding is that GPT-4 shows strong alignment with human judgments for short-answer tasks, but only when gold references are provided. The reliability drops significantly without gold references, as the model is overly positive compared to human evaluations. The issue is particularly evident in free-form tasks, which are tasks where LLM-as-a-judge could be the most valuable and where gold labels are typically not available. When gold references are available, we observe that GPT-4 is too strict compared to humans, relying to much on the gold label. For these type of tasks, even though LLM-as-a-judge is often applied to them, human evaluations still remain the gold standard.

ROUGE-L performs comparably to GPT-4-gold for short-answer tasks, offering a cost-effective alternative in scenarios where the use of GPT-4 is limited by cost or time constraints. For long-answer tasks, while BERTSCORE demonstrates strong performance, it does not fully replace the need for GPT-4-gold. These metrics provide valuable insights but vary in effectiveness depending on the specific task.

Evaluating non-English outputs, such as Swedish, presents additional challenges. There is a significant drop in alignment for GPT-4-no-gold, which highlights that GPT-4 as a judge is less reliable for languages other than English. While this is true for Swedish, we expect these findings to be more sever for lesser-resourced languages and those less similar to English. Future work should focus on expanding this study to more languages.

As we have observed, there is a large variation in alignment with human ratings for all metrics across task types. Previous research identifies strong correlations with human annotators, but that is often the average over tasks. Our findings underscores the necessity of task-specific evaluation metrics rather than relying on general averages which can obscure important nuances in metric alignment with human annotators. Furthermore, while GPT-4 is a valuable tool for evaluation short-answer tasks when gold references are available, alternatives like ROUGE-L and BERTSCORE can be effective for most tasks types, offering cost-efficient and reliable evaluations.

Limitations

We choose to report our results using pairwise accuracy which we believe provides more robust and reliable alignment statistics compared to common correlation metrics. With that said, PA has its shortcomings, such as when it faces constant or close-to-constant scores. For example, when the reference vector is $\vec{1}$, a metric vector of $\vec{1}$ or $\vec{3}$ both result in very high PAs, due to the lack of prior knowledge about the metric range. However, this issue is not unique to PA; common correlation metrics also face the same challenge in the case of close-to-constant vectors.

Our study primarily focuses on the evaluation of tasks across English and Swedish. Consequently, the findings may not be applicable to languages that have syntax and structure significantly different from English. We deliberately made this choice to enable a broader examination of tasks and tap into expert knowledge for deeper analyses. Essentially, we prioritized expanding the range of tasks and delving deeper into analysis rather than focusing on additional languages. Furthermore, our evaluation exclusively uses GPT-4 as the language model for assessment. The rapidly evolving landscape of language models suggests the existence of other models that may yield different results or exhibit different patterns.

Ethical Considerations

Our annotators, residents of Sweden, were selected for their proficiency in both English and Swedish, ensuring precise interpretation and annotation of content. We ensured their fair compensation in line with prevailing norms for similar tasks in Sweden. Furthermore, they completed their assignments within a reasonable timeframe, enabling them to work without undue pressure. Prior to acceptance, annotators were briefed on the purpose of their annotations ensuring that they understood the objectives and context behind the tasks assigned to them.

Acknowledgments

We extend our gratitude to our annotators for their diligent contributions to this paper and to Daniel Deutsch and Richard Johansson for their valuable feedback. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation and by the European Union's Horizon 2020 research and innovation program under grant agreement No 101135671 (TrustLLM). We also acknowledge support from the National Graduate School of Computer Science in Sweden (CUGS). The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the Swedish National Supercomputer Center.

References

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

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.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making instruction finetuning accessible to non-English languages: A case study on Swedish models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? *arXiv preprint arXiv:2312.12683*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023b. M^{3}IT: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2638–2656, Torino, Italia. ELRA and ICCL.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 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, Toronto, Canada. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Leonardo Ranaldi, Giulia Pucci, and Andre Freitas. 2024. Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7961–7973, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3063–3079, Toronto, Canada. Association for Computational Linguistics.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

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

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*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Task Descriptions

Task033: Winogrande Answer Generation A fill-in-the-blank task with some restriction, such as that the answer should be chosen from the two objects in the question. "I planted more tomato seeds than I planted cucumber seeds since I hated eating the _ ."Gold answer: "cucumber".

Task034: Winogrande Question Modification Object Similar to task033, but this time the task is to change the question so that the answer, which is given in the input, changes to the other object present in the input.

Task036: QASC Topic Word to Generate Related Fact Write a topic word for the given fact with at least one word overlap with the fact. Example: "Fact: a seismograph is used for measuring the size of an earthquake."One possible gold answer: "seismograph earthquake."

Task133: Winowhy Reason Plausibility Detection Indicate the plausibility of reasoning for the pronoun coreference relations. Example: "Sentence: Although they ran at about the same speed, Sue beat Sally because she had such a bad start.\n Reason: The 'she' refers to sally because Sue won, sally lost. \n Question: Is the above reasoning correct or wrong? "Gold answer: "Correct".

Task190: SNLI Classification Given two sentences, classify their agreement: entailment, contradiction, or neutral.

Task200: MNLI Entailment Classification From three options, choose the one that can be inferred from the given sentence.

Task202: MNLI Contradiction Classification From three options, choose the one that disagrees with the given sentence.

Task329: GAP Classification Given a text, a pronoun, and two candidate names, determine which of the names the pronoun refers to. The answer should be either A, B, or neither.

Task349: SQuAD2.0: Answerable Unanswerable Question Classification Determine whether or not the given question is answerable by the provided passage.

Task392: Inverse Causal Relationship Given two sentences separated by the word "because", determine whether the second sentence can be the result of the first one (is there a cause and effect relationship?)

Task401: Numeric Fused Head Reference Using your knowledge about language and commonsense, determine what element the marked number refers to. Example: "Jim Bronson: What 's your name ?\nTemple Brooks: I do n't have _ one _ !\nJim Bronson: Well everyone I have ever know had a name , that 's really weird . My name is Jim incase your interested .\nTemple Brooks: Well I 'm not !"Gold answer: "name".

Task520: AQuaMuSe Answer Given in Passage Is the answer to the given question contained in the provided passage?

Task613: PolitiFact Text Generation Generate the subject of a speech by a politician.

Task677: Ollie Sentence Answer Generation Given two noun phrases (arguments) and the relationship between them, write a sentence that expresses theses arguments with the given relationship.

Task891: GAP Coreference Resolution Given a passage, find the corresponding person for the provided pronoun.

Task1152: BARD Analogical Reasoning Causation Replace question mark with a verb which is the appropriate consequence of the given action. For example: "ignite : burn. hit : ?". Gold answer: "shatter".

Task1562: ZEST Text Modification Paraphrase the given questions to have different wording. Change it as much as possible using synonyms, etc. Example: "Does this dog breed always have spots?".

Task1615: SICK Classify b Relation a Classify the relation between two sentences: B_entails_A, B_contradicts_A, or B_neutral_A.

Task1622: Disfl-QA: Text Modification Convert a disfluent question to a proper question. Example: "Who were uh instead tell me how many quadrangles does the Main Quadrangles have?"

Task1659: Title Generation Generate a title under forty words which mentions the purpose of the text.

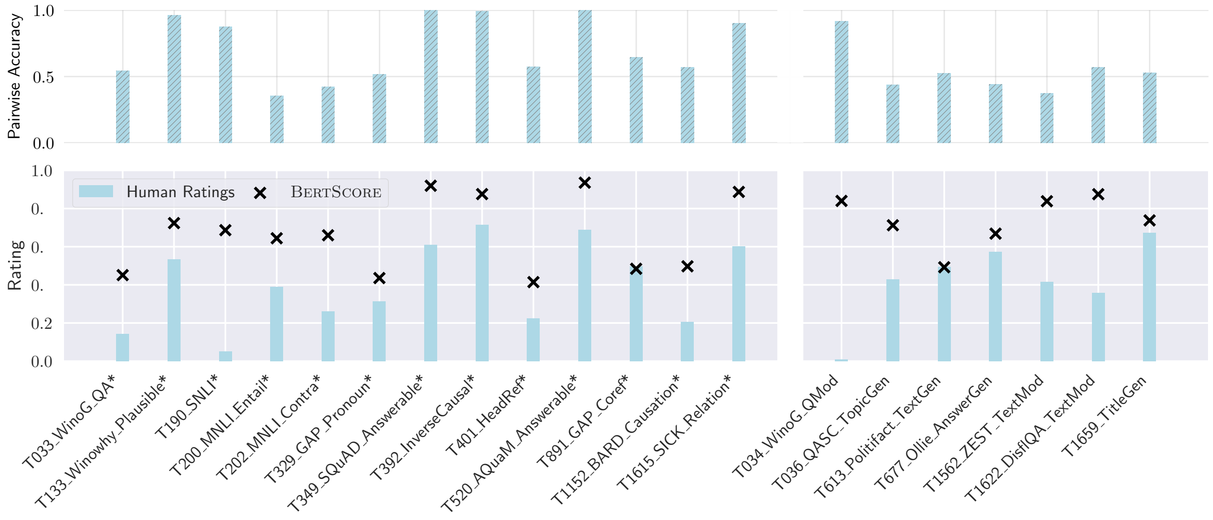

Figure A1: Human and BERTSCORE ratings per task and test language on the bottom and their PA on top. The human ratings are normalized to a range of 0 to 1. Short-answer tasks are on left and long-answer ones on right.

B Prompt Templates

B.1 Translation

We use the following prompt for translating our datasets from English to Swedish using GPT-3.5-turbo:

```
Translate the following text from
English to Swedish:
{English text}
```

B.2 LLM-as-a-judge

The prompt used for GPT-4 could be found in Table A2. The prompt for GPT-4-no-gold is the same, but without the following part:

```
[Gold Answer] (If there are several gold
    answers then they are all correct
   alternatives): {gold_answer}
***
```

Model	Perplexity
LLaMA2_13b	1.96
LLaMA2_13b_EN	2.03
LLaMA2_13b_SV	2.27
LLaMA2_13b_ENSV	2.24
LLaMA2_7b	2.09
LLaMA2_7b_EN	2.22
LLaMA2_7b_SV	2.51
LLaMA2_7b_ENSV	2.49
SW3_6.7b	1.62
SW3_6.7b_EN	1.65
SW3_6.7b_SV	1.72
SW3_6.7b_ENSV	1.69

Table A1: The perplexity of our models on the SVT dataset. The abbreviations are the training language(s).

C Supplementary Results

C.1 How Good Are Our Language Models at Swedish?

To assess the effectiveness of the models described in Section 3.2, we measure their perplexity. To ensure the generated texts meet high standards and to avoid assessing the models on data used during their pretraining, we use a custom dataset consisting of current news articles from SVT[3], the Swedish national public television broadcaster. The dataset comprises 268 articles spanning various topics, published between June 1st, 2023, and October 16th, 2023. To address the variability caused by different tokenizers across various models, we use character length normalization when calculating perplexity (Liang et al., 2022; Yong et al., 2023).

The perplexity for Swedish consistently remains lower in models prior to instruction tuning. However, after tuning, the poorest outcomes are noted in the SV models trained solely on Swedish data. Interestingly, the ENSV models exhibit improved performance, with the EN models showing even better results. This variation could be ascribed to the Swedish-specific model weights being less affected due to their lower exposure to Swedish data, but requires further investigations. Notably, the incre-

[3] svt.se

ments in perplexity are less pronounced for the SW3 models.

C.2 Pairwise Accuracy per Model

Pairwise accuracy per model is shown in Table A2.

C.3 BERTSCORE Results

Detailed results of BERTSCORE are shown in Figure A1.

D Implementation Details

Following Taori et al. (2023), we finetune LLAMA2_7b and SW3_6.7b, which is roughly the same size, for 3 epochs and with a learning rate of $2e-5$, and the larger LLAMA2_13b model for 5 epochs and with a learning rate of $1e-5$. The batch size is set to 128 for both cases. Unlike Taori et al. (2023), we allow for a longer maximum length of 2048 and truncate longer samples. For the sake of reducing computational costs we opt for using bf16 and tf32 precision formats. We distribute the training across multiple GPUs using DeepSpeed (Rasley et al., 2020) stage 3 without offloading.

E Human Evaluations

Our evaluators are not crowd-sourced workers; instead, they are individuals with some experience and expertise in the field. They were carefully selected for their familiarity with the subject matter and were hired specifically for this evaluation task. While they were instructed to use their own judgment in the assessment, they also had access to the gold standard answers to guide their evaluations. The instructions given to human evaluators were similar to those given to GPT-4, as presented in Figure 3.

Model	GPT-4-gold	GPT-4-no-gold	ROUGE-L	BERTSCORE
LLAMA2_13b_EN_EN	0.779	0.590	0.601	0.512
LLAMA2_13b_SV_SV	0.816	0.559	0.547	-
LLAMA2_13b_ENSV_EN	0.782	0.594	0.639	0.545
LLAMA2_13b_ENSV_SV	0.828	0.565	0.634	-
LLAMA2_7b_EN_EN	0.766	0.581	0.578	0.454
LLAMA2_7b_SV_SV	0.794	0.578	0.526	-
LLAMA2_7b_ENSV_EN	0.794	0.609	0.607	0.447
LLAMA2_7b_ENSV_SV	0.804	0.577	0.550	-
SW3_6.7b_EN_EN	0.789	0.612	0.604	0.482
SW3_6.7b_SV_SV	0.843	0.645	0.505	-
SW3_6.7b_ENSV_EN	0.830	0.606	0.622	0.491
SW3_6.7b_ENSV_SV	0.828	0.576	0.600	-

Table A2: Pairwise accuracy per model for all metrics.

Model Name	Task Type	Language	τ	ρ	PA	ϵ
ROUGE-L	all	EN	0.667	0.712	0.752	0.061
ROUGE-L	long	EN	0.308	0.368	0.484	0.061
ROUGE-L	short	EN	0.833	0.871	0.896	0.061
ROUGE-L	all	SV	0.584	0.631	0.699	0.095
ROUGE-L	long	SV	0.325	0.389	0.477	0.095
ROUGE-L	short	SV	0.704	0.743	0.818	0.095
GPT-4-gold	all	EN	0.781	0.801	0.811	0.000
GPT-4-gold	long	EN	0.481	0.521	0.584	0.000
GPT-4-gold	short	EN	0.920	0.930	0.934	0.000
GPT-4-gold	all	SV	0.792	0.811	0.817	0.000
GPT-4-gold	long	SV	0.538	0.575	0.601	0.000
GPT-4-gold	short	SV	0.909	0.920	0.934	0.000
GPT-4-no-gold	all	EN	0.517	0.535	0.622	0.000
GPT-4-no-gold	long	EN	0.309	0.338	0.499	0.000
GPT-4-no-gold	short	EN	0.613	0.625	0.688	0.000
GPT-4-no-gold	all	SV	0.496	0.514	0.620	0.000
GPT-4-no-gold	long	SV	0.393	0.425	0.569	0.000
GPT-4-no-gold	short	SV	0.544	0.556	0.648	0.000
BERTSCORE	all	EN	0.419	0.482	0.658	0.133
BERTSCORE	long	EN	0.367	0.452	0.542	0.133
BERTSCORE	short	EN	0.443	0.495	0.720	0.133

Table A3: Comparison of metrics across different task types for English and Swedish. τ stands for Kendall's and ρ for Spearman's rank correlation coefficient.

```
system\_message = """You are an expert language evaluator."""

user\_message = """You are evaluating a response that has been submitted for a
    particular task, using a specific set of standards. Below is the data:

[BEGIN DATA]
***
[Task]: {prompt}
***
[Submission]: {response}
***
[Gold Answer] (If there are several gold answers then they are all correct
    alternatives):  {gold_answer}
***
[Criterion]: Evaluation Criteria

Naturalness:
1: "Not at all natural - The generated text is grammatically incorrect or sounds
    unnatural, including awkward phrasing or inappropriate vocabulary."
2: "Somewhat natural - The generated text has minor grammatical errors or slightly
    awkward phrasing but is mostly understandable and natural."
3: "Completely natural - The generated text is grammatically correct, well-phrased,
    and uses appropriate vocabulary, sounding completely natural."

Relatedness:
1: "Not at all related - The model's answer does not relate to the question, fails
    to follow the required format, or is outside the scope of possible answers."
2: "Somewhat related - The model's answer is related to the question to some extent
    and mostly follows the required format, staying generally within the scope of
    possible answers."
3: "Completely related - The model's answer is directly related to the question,
    follows the required format accurately, and fits within the scope of possible
    answers."

Correctness:
1: "Not at all correct - The answer is completely incorrect or irrelevant to the
    question posed."
2: "Somewhat correct - The answer is partially correct but includes some
    inaccuracies or incomplete information."
3: "Completely correct - The answer is fully correct, accurate, and provides a
    complete response to the question."

***
[END DATA]

Does the submission meet the criterion? First, write out in a step by step manner
    your reasoning about the criterion to be sure that your conclusion is correct.
    Avoid simply stating the correct answers at the outset.
Your response must be RFC8259 compliant JSON following this schema:

{{"reasoning": str, "naturalness": int, "relatedness": int, "correctness": int}}
"""
```

Figure A2: The prompt for GPT-4 as evaluator. For GPT-4-no-gold the gold answer is removed.