Fine grained evaluation of Language Models with focus on text generation and open-ended question answering

Sarang Ravi Chouguley

[1]Institute for Information Systems, Hof University of Applied Sciences, Hof,Germany.

Abstract

Evaluation of LMs is a challenging task due to its language nature. The aim of this study is to understand if FLASK-like approach can be used to perform fine-grained evaluation of LMs, with criteria that are similar to human-based evaluation. The study evaluates responses generated by 10 LlaMA based open sourced LMs with focus on text generation instructions and open-ended question answering. Standard mertics BLEU, ROGUE, BERTScore are used to set baseline for custom evaluation. The study then evaluates LM using a custom fine-grained FLASK-like approach with text generation and open-ended question answering focused criteria. This study found that custom fine-grained evaluation has positive strong correlation with standard metrics BERTScore and RougeL.

Keywords: Language Model, text-generation, open-ended question answering, fine-grained evaluation

1 Introduction

Past few years have witnessed a huge growth in Language Models. Language Models(LM) of various parameter size have been developed. People have started using these models in day to day life, for various tasks ranging from using them to understand something to using them as a friend. Some models are designed to be general and can handle many different use cases, while other models are fine-tuned for a specific use case. But evaluation remains to be a common challenge for all models.

Evaluation of language models plays a very crucial role in their usage and adoption. Since the beginning of language models, various evaluation techniques have been developed. Techniques like fixed metrics that only consider the semantic and superficial features of the response. Automatic metrics go a step beyond and measure semantic properties of response in automated way. Both of these techniques are sensitive to surface form. To overcome this, human based evaluation is generally preferred to provide accurate evaluation of language modelss. But this technique cannot be used at large scale. To solve this challenge, techniques like LLM-as-a-judge have been developed, where one language model is used to evaluate another language model. This technique is close to human evaluation accuracy and can be scaled easily. But this technique suffers from biases.

To overcome issues of LLM-as-a-judge, recently fine-grained evaluation technique has been proposed. Fine-grained evaluation allows us to comprehend performance of a model from various perspectives. This allows to evaluate models in ways aligned to human values. FLASK is one such technique. The focus of this study is to understand if custom evaluation criteria can be used to perform fine-grained evaluation of LM(Language Model).

This study first evaluates models on some standard metrics, to set a baseline for evaluation. Then it performs custom fine-grained evaluation using criteria motivated from FLASK, which are custom designed for text-generation and open-ended QA. Study uses Prometheus, to perform LLM based evaluation.

This study is restricted to only two types of instructions (i.e. prompt): text-generation and open-ended question answering. Number of models evaluated is restricted to 10, to keep it manageable. The model parameter size is restricted to 13B and below.

The rest of the paper is structured in following manner. First it discusses related work, focusing especially on recent developments in fine-grained and LLM-as-a-judge evaluation. Then, it discusses the dataset used for evaluation, choice of models, experiment setup and design of custom fine-grained criteria. After that, experiment results and analysis are presented. The paper ends with discussion, limitations, conclusion and outlook.

2 Related Work

Language models especially Large Language Models (LLM) like ChatGPT, Claude, Bart have become a norm now. With the evolution of these models, their evaluation techniques also evolved. The following presents literature on evaluation of LMs, text generation and open-ended QA evaluation. In the end, recent trends in LM evaluation are explored.

2.1 Literature related to LM evaluation

In the beginning metrics like Bleu[20], Rouge[14] were used for evaluation of machine translation and summarisation. But they have been criticised for their limitations in capturing the diversity of equally good alternative translations or summaries that use different words or word order [1]. Efforts to address this issue led to exploration of alternative metrics that correlate more closely with human judgements. Some studies found that metrics like chrF2 and COMET exhibit higher correlation with human judgements compared to BLEU, indicting their potential for capturing the quality of alternative translations or summaries [15]. Introduction of BERTScore [29] addressed the limitations of these metrics in evaluating text generation tasks. It offered an automatic metric that considers the compositional diversity of generated text, which is often overlooked by word-overlap-based metric like Bleu. Additionally, BERTScore has been shown to correlate better with human judgements in measuring sentence similarity [2], [18]. But BERTScore has its own limitations. Marvin Kaster et al. noted that BERTScore is significantly sensitive to lexical overlap, akin to BLEU and ROUGE metrics. This sensitivity means it might not effectively capture other linguistic factors such as semantics, syntax, and morphology, potentially limiting its applicability in certain contexts [10]. A study by Zitha Sasindran et al. found that BERTScore tends to overly prioritize keywords. This could lead to skewed evaluations in scenarios where the importance of keywords relative to the overall content needs to be balanced [24]. To overcome these limitations metrics like Sentence Transformer [17, 22] which compares whole sentences were designed. In summary, these metrics are more or less prone to surface form changes in text.

2.2 Literature related to text-generation evaluation

Text generation is the core ability of any language model. Various studies have been done to evaluate text generation capabilities of language models. Chen et al. introduced X-IQE, an approach using visual large language models to evaluate text-to-image generation. X-IQE generates textual explanations that correlate well with human evaluation, assessing real versus generated images, text-image alignment, and image aesthetics [3]. Hua et al. proposed DYPLOC, a framework that enhances coherence and diverse content in long-form opinion text generation. This model was found to outperform others in both automatic evaluation and human judgment [7]. Zhang et al. reviewed common Controllable Text Generation (CTG) tasks, approaches, and evaluation methods, identifying challenges and future research in this direction [28]. Min et al. introduced FACTSCORE, a method for assessing the factuality of long-form text generated by language models, addressing the limitations of existing evaluation methods [16]. Nguyen explored methods and metrics to enhance the quality, diversity, and consistency of machine-generated text [19]. Tan et al. proposed a method for generating domain-specific content keywords and refining them into complete passages, improving text generation quality and efficiency [25].

2.3 Literature related to open-ended question answering evaluation

Question Answering Evaluation is also an actively studied domain in LM evaluation. The assessment of question answering in large language models has involved various methodologies and metrics. For example, recent research has introduced QAScore, an unsupervised unreferenced metric for question generation evaluation, which categorizes widely-applied metrics into word overlap metrics and those utilizing large pre-trained language models such as BLEURT and BERTScore [8]. Furthermore, the exploration of retrieval-augmented language models for clinical medicine has been undertaken to assess the potential of large language models in medical question answering [6]. The importance of answer verification in evaluation has been emphasized, with a focus on benchmarking answer verification methods for question answering-based summarization evaluation metrics as a means to automatically determine the correctness of question answering model predictions [5]. Semantic answer similarity has also been utilized to evaluate question answering models by comparing ground-truth annotations with model predictions, highlighting the significance of semantic similarity in the evaluation process [23]. Kamalloo et al. focused on the challenges of lexical matching evaluation and the performance of different models, including InstructGPT, on the NQ-open benchmark. They found that InstructGPT shows a significant increase in performance, and automated evaluation models are not effective for evaluating long-form answers generated by LLMs [9]. Rabinovich et al. created a benchmark dataset with high-quality paraphrases for factual questions and proposed a framework for predicting the likelihood of a language model accurately answering a question. This approach outperformed baselines when evaluated on five contemporary models [21].

2.4 Recent trends in LM evaluation

Recent trends in evaluation of language models has shifted focus towards measuring multiple aspects of response. Some studies have termed this as fine-grained evaluation. OpenAI used this technique for maths task [12]. Ye et al. introduced FLASK, a fine-grained evaluation protocol for language models based on alignment skill sets. FLASK decomposes coarse-level scoring into skill set-level scoring for each instruction, providing a holistic view of model performance and increasing evaluation reliability [27]. Kim et al. proposed Prometheus, an open-source Large Language Model (LLM) that can evaluate long-form text based on customized score rubrics. Prometheus shows high correlation with human evaluators and outperforms other language models like GPT-4 and ChatGPT in evaluation tasks [11]. Ye et al. proposed ToolEyes, a fine-grained system for evaluating the tool learning capabilities of large language models in real-world scenarios. ToolEyes examines various dimensions crucial to LLMs in tool learning, such as format alignment and intent comprehension [26]. Lin et al. introduced URIAL, a tuning-free alignment method that achieves effective alignment of base LLMs purely through in-context learning. They demonstrated that base LLMs aligned with URIAL can match or surpass the performance of LLMs aligned with supervised fine-tuning or reinforcement learning from human feedback [13]. Chen et al. proposed CoDI-Eval, a benchmark to evaluate large language models' responses to instructions with various constraints. They revealed limitations in following instructions with specific constraints and identified a gap between open-source and commercial closed-source LLMs [4].

3 Dataset

The dataset is self-constructed and is inspired from existing datasets for LM evaluation like BigBench. While constructing the dataset following points were used as guidelines.

- Focus on open-ended question answering and text generation instructions.
- Include tasks from different datasets to make the dataset as diverse as possible.

- Only include instructions written in English.

Based on these guidelines, a set of 20 instructions, 10 open-ended QA instructions and 10 text generation instructions, are selected from publicly available datasets on Huggingface.

The final dataset is composed of tasks from following datasets:

- **Alpaca**: Alpaca is a dataset of 52,000 instructions and demonstrations generated by OpenAI's text-davinci-003 engine. This dataset was chosen because it contains text generation instructions and is fairly popular on huggingface.
- **Ultrachat**: Ultrachat is an open-source, large-scale and multi-round dialogue data powered by Turbo APIs. It is composed of 3 sectors: questions about the world, writing and creation, assistance on existent materials.
- **No Robots**: No Robots is a high-quality dataset of 10,000 instructions and demonstrations created by skilled human annotators. This dataset primarily consist of single-turn instructions that can be used for supervised fine-tuning.

Selection of instructions from above datasets ensured that new dataset consisted of instructions generated by both humans and LMs.

4 Models selected for Study

A set of 10 models are selected for the study to keep it manageable . Only models which are public available on huggingface are selected. While selecting the models following criteria are kept in mind:

- Select models having no more than 13B parameters.
- Select models based on LlaMA (v1/v2)
- Select models which are fine-tuned for text generation and question answering
- Avoid models which are trained on the Alpaca, Ultrachat or No-robots dataset, to avoid biased output.
- Give preference to most popular models on huggingface.

Out of 10 models 5 models have 13B parameters, 4 models have 7B parameters and 1 model has 6B parameters. GPTQ or AWQ quantised versions of the models are used in the study for fast inference (refer appendix A9 for model details).

5 Design of custom fine-grained criteria

5.1 Inspiration for fine-grained criteria

Open-ended responses are composited of multiple components, which makes measuring them with a single metric insufficient. Hence a fine-grained evaluation of model is required to comprehend quality of responses from various perspectives.

5.2 Details of Flask

FLASK is a fine-grained evaluation protocol for both human-based and model-based evaluation which decomposes coarse-level scoring to a skill set-level scoring for each instruction. FLASK defines 4 primary abilities which are divided into 12 fine-grained skills for comprehensive language model evaluation: Logical Thinking

(Logical Correctness, Logical Robustness, Logical Efficiency), Background Knowledge (Factuality, Commonsense Understanding), Problem Handling (Comprehension, Insightfulness, Completeness, Metacognition), and User Alignment (Conciseness, Readability, Harmlessness).

5.3 Why Flask criteria cannot be used in this study?

FLASK's set of 12 fine-grained skills are design for task-agnostic evaluation of models. Although this is good for evaluating across a range of tasks, this is not useful for fine-grained task specific evaluation. Hence,this study uses a custom fine-grained criteria, which is especially designed for text-generation and open-ended question answering tasks. These custom criteria are designed keeping in mind the characteristics of a human generated answer for these tasks.

5.4 Criteia Design

5.4.1 Categorization of Text generation instructions

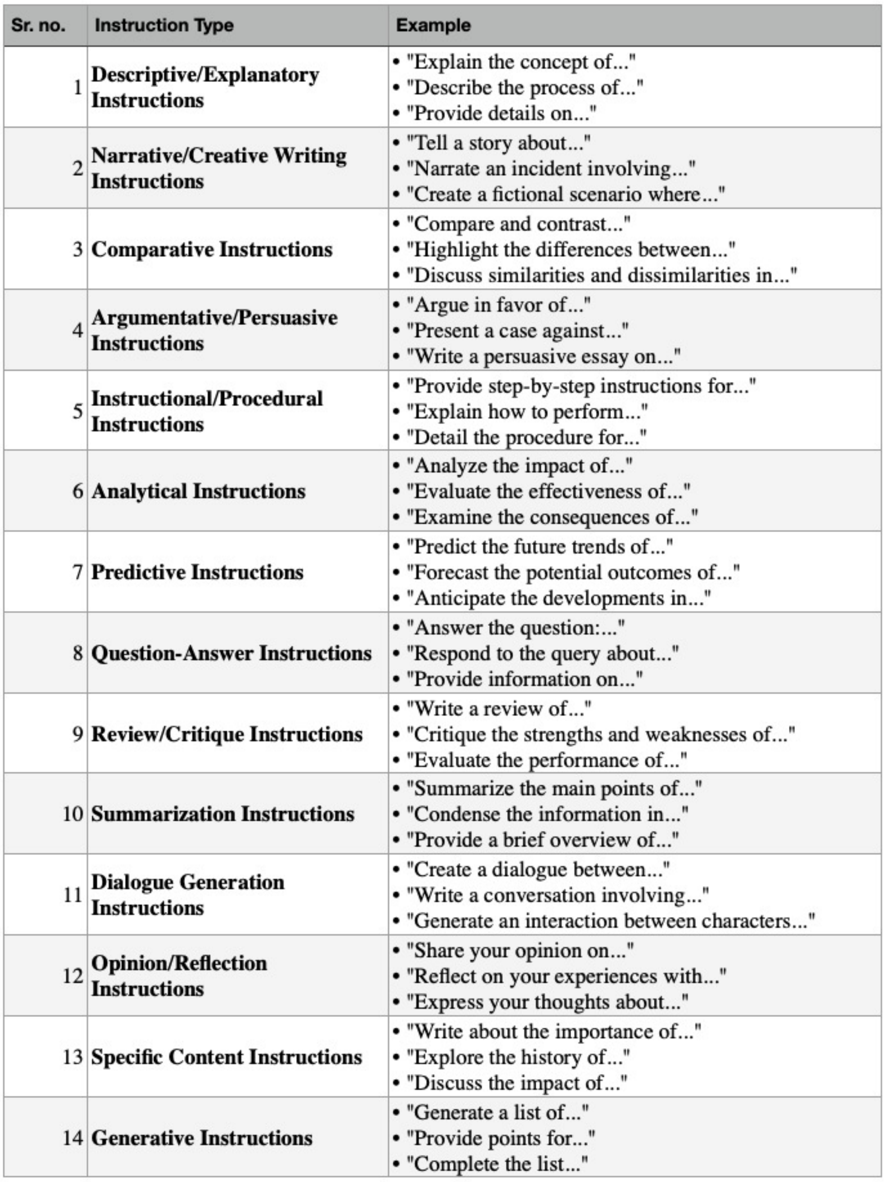
In general text generation instructions can be grouped into following categories:

1. Descriptive/Explanatory Instructions
2. Narrative/Creative Writing Instructions
3. Comparative Instructions
4. Argumentative/Persuasive Instructions
5. Instructional/Procedural Instructions
6. Analytical Instructions
7. Predictive Instructions
8. Question-Answer Instructions
9. Review/Critique Instructions
10. Summarization Instructions
11. Dialogue Generation Instructions
12. Opinion/Reflection Instructions
13. Specific Content Instructions
14. Generative Instructions

Fig. 1 shows examples for each instruction type.

The response of these instructions should contain one or more of the following characteristics in order to be considered good as per human standards.

1. **Relevance**: The response should be directly related to or aligned to the given topic, context, theme, subject.
2. **Clarity and coherence**: The response should be organised in a clear ,coherent and easy manner.
3. **Consistency**: The response should be consistent in themes, characters, elements, tone and style.
4. **Logic and common sense**: The response should be logically valid and have common sense.



Sr. no.	Instruction Type	Example
1	**Descriptive/Explanatory Instructions**	• "Explain the concept of..." • "Describe the process of..." • "Provide details on..."
2	**Narrative/Creative Writing Instructions**	• "Tell a story about..." • "Narrate an incident involving..." • "Create a fictional scenario where..."
3	**Comparative Instructions**	• "Compare and contrast..." • "Highlight the differences between..." • "Discuss similarities and dissimilarities in..."
4	**Argumentative/Persuasive Instructions**	• "Argue in favor of..." • "Present a case against..." • "Write a persuasive essay on..."
5	**Instructional/Procedural Instructions**	• "Provide step-by-step instructions for..." • "Explain how to perform..." • "Detail the procedure for..."
6	**Analytical Instructions**	• "Analyze the impact of..." • "Evaluate the effectiveness of..." • "Examine the consequences of..."
7	**Predictive Instructions**	• "Predict the future trends of..." • "Forecast the potential outcomes of..." • "Anticipate the developments in..."
8	**Question-Answer Instructions**	• "Answer the question:..." • "Respond to the query about..." • "Provide information on..."
9	**Review/Critique Instructions**	• "Write a review of..." • "Critique the strengths and weaknesses of..." • "Evaluate the performance of..."
10	**Summarization Instructions**	• "Summarize the main points of..." • "Condense the information in..." • "Provide a brief overview of..."
11	**Dialogue Generation Instructions**	• "Create a dialogue between..." • "Write a conversation involving..." • "Generate an interaction between characters..."
12	**Opinion/Reflection Instructions**	• "Share your opinion on..." • "Reflect on your experiences with..." • "Express your thoughts about..."
13	**Specific Content Instructions**	• "Write about the importance of..." • "Explore the history of..." • "Discuss the impact of..."
14	**Generative Instructions**	• "Generate a list of..." • "Provide points for..." • "Complete the list..."

Fig. 1 Instruction categorization and Examples

5. **Completeness**: The response should be complete, containing all relevant details necessary to give a comprehensive understanding of the topic.
6. **Accuracy**: The response should be factually correct, providing accurate details.
7. **Engagement**: The response should be engaging, captivating the reader's interest and imagination.

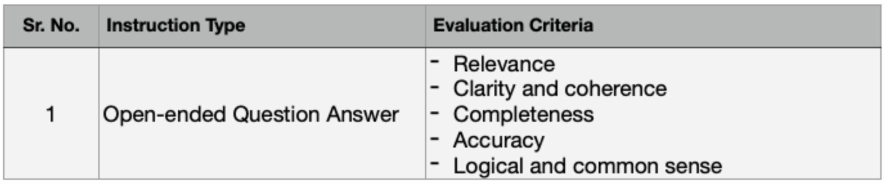
8. **Creativity**: The response should showcase imaginative thinking, originality, and creativity in the content.
9. **Thoughtfulness and Insight** : The response should contain thoughtful analysis, showcasing insight into potential outcomes or developments.
10. **Evidence and support**: The response should contain evidence or support for the claims presented.
11. **Empathy**: The response should convey opinions or reflections with an empathetic and considerate tone where applicable.
12. **Efficiency**: The response should convey information efficiently without unnecessary verbosity.

5.4.2 Analysis of open-ended question answer instructions characteristics

Open-ended question answer can be considered special type of Question-Answer instruction which require the response to have characteristics like contain all the necessary details, not restricted to one word or sentence, factually accurate, logically valid. In other words, the responses should contain following characteristics:

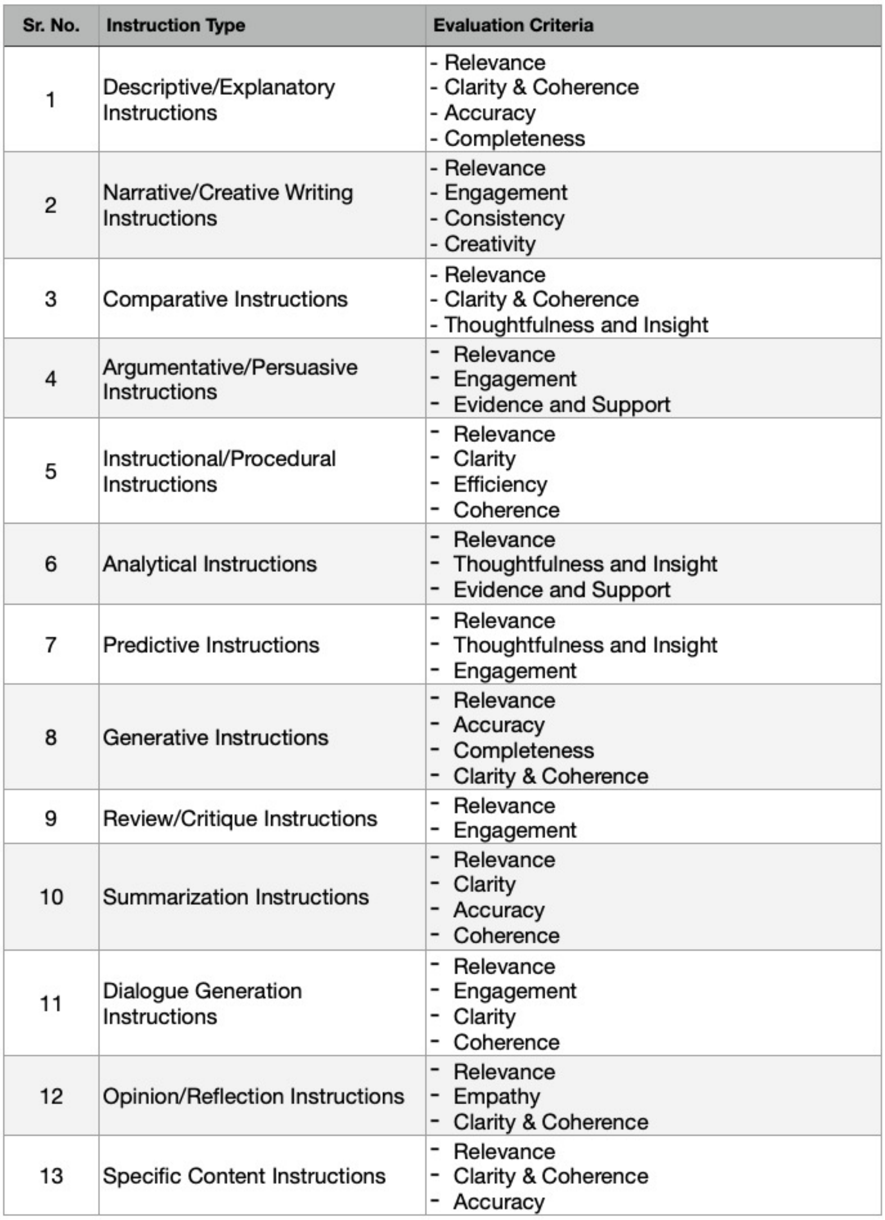
1. Relevance
2. Clarity and Coherence
3. Completeness
4. Accuracy
5. Logic and Common-sense

Based on this categorization of instructions and response characteristics, a mapping of instruction type vs evaluation criteria is created, where a subset of characteristics are assigned to each instruction type. This mapping forms the basis for evaluation. Figs. 3 and 2 show these mappings.



Sr. No.	Instruction Type	Evaluation Criteria
1	Open-ended Question Answer	- Relevance - Clarity and coherence - Completeness - Accuracy - Logical and common sense

Fig. 2 Open-ended QA vs Evaluation Criteria mapping



Sr. No.	Instruction Type	Evaluation Criteria
1	Descriptive/Explanatory Instructions	- Relevance - Clarity & Coherence - Accuracy - Completeness
2	Narrative/Creative Writing Instructions	- Relevance - Engagement - Consistency - Creativity
3	Comparative Instructions	- Relevance - Clarity & Coherence - Thoughtfulness and Insight
4	Argumentative/Persuasive Instructions	- Relevance - Engagement - Evidence and Support
5	Instructional/Procedural Instructions	- Relevance - Clarity - Efficiency - Coherence
6	Analytical Instructions	- Relevance - Thoughtfulness and Insight - Evidence and Support
7	Predictive Instructions	- Relevance - Thoughtfulness and Insight - Engagement
8	Generative Instructions	- Relevance - Accuracy - Completeness - Clarity & Coherence
9	Review/Critique Instructions	- Relevance - Engagement
10	Summarization Instructions	- Relevance - Clarity - Accuracy - Coherence
11	Dialogue Generation Instructions	- Relevance - Engagement - Clarity - Coherence
12	Opinion/Reflection Instructions	- Relevance - Empathy - Clarity & Coherence
13	Specific Content Instructions	- Relevance - Clarity & Coherence - Accuracy

Fig. 3 Instruction type vs Evaluation Criteria mapping

6 Experiment

6.1 Setup

The prompt structure provided by hugging face was used for each model to avoid invalid responses generated due to incorrect prompt structure. NVIDIA A100 40GB GPU was used for generating the response. The responses generated by each of the 10 selected models were collect for evaluation.

6.2 Evaluation of responses

6.2.1 Standard Metrics evaluation

Following standard metrics were used to evaluate the responses generated by each model against gold answer for each task: bleu, bertscore (precision, recall, f1), rouge(rouge1, rouge2, rougeL, rougeLsum)

6.2.2 Custom fine-grained evaluation

Fine grained evaluation was done using Prometheus as LLM evaluator. Prometheus requires instruction, reference answer, response to evaluate, criteria description, and score description for scores 1(refers to completely unaligned response) to 5(refers to completely aligned response). This score description also called score rubric was generated from the custom evaluation criteria mapping defined in Figs. 3 2. A score rubric was generated for each of these evaluation criteria (see table A10). Prometheus provides the evaluation by assessing the quality of the response strictly based on the given score rubric. Output of this evaluation consist of a feedback which explains in details the quality of response based on score rubric and a score ranging from 1 to 5.

7 Results and Analysis

7.1 Text Generation Scores

Table 1 and Fig. 4 shows the text generation instruction scores for each model.

7.1.1 Interpretation of text generation scores

BLEU Score: This is a metric for evaluating a generated sentence to a reference sentence, commonly used in machine translation. A higher BLEU score means the generated text is closer to the reference text. The scores range from 0.09 to 0.84, indicating significant variability in performance among the models with some giving responses much closer to a reference than others.

BERT Score: This reflects how well the model captures the meaning of the text, using contextual embeddings from BERT. It typically ranges from 0 to 1, with higher scores indicating better performance. In this table, the scores range from 0.44 to 0.645, suggesting that some models' responses are closer contextually than others.

ROUGE-L Score: This metric measures the longest common subsequence and is often used for summarization tasks. It indicates how much of the core information

Table 1 Text Generation Instruction scores

Model Name	bleu[1]	bert	rougeL	prometheus
TheBloke/Orca-2-13B-AWQ	0.12	0.645	0.245	4.3
microsoft/Orca-2-7b	0.84	0.58	0.168	3.2
01-ai/Yi-6B	0.34	0.44	0.11	1.6
WizardLM/WizardLM-13B-V1.1	0.71	0.618	0.17	5.0
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.119	0.627	0.223	4.0
TheBloke/llava-v1.5-13B-AWQ	0.11	0.622	0.209	4.0
TheBloke/vicuna-7B-v1.5-GPTQ	0.118	0.648	0.21	3.5
TheBloke/tulu-2-dpo-13B-AWQ	0.79	0.591	0.171	4.4
timdettmers/guanaco-7b	0.98	0.632	0.188	3.8
TheBloke/guanaco-13B-GPTQ	0.09	0.581	0.177	3.2

[1] These are average scores for text generation instructions. For detailed scores refer Appendix A

is captured by the generated response. The scores vary from 0.11 to 0.245, showing some models have better response than others.

Prometheus Score: This metric isn't standard like the other three and is based on custom fine-grained criteria. It indicates how much the response is inline with the fine-grain criteria. The score vary from 1.6 to 5.0, showing some models have better criteria oriented response than others.

Looking at the scores, the model 'WizardLM/WizardLM-13B-V1.1' stands out with the highest Prometheus score of 5.0, which might indicate a superior overall performance. The model '01-ai/Yi-6B' shows the lowest scores across all metrics, which might indicate it is less proficient in the tasks measured compared to the others.

7.2 Open-ended QA Scores

Table 2 and Fig. 5 shows the open-ended QA instruction scores for each model.

7.2.1 Interpretation of QA scores

BLEU Score: A higher BLEU score means the generated text is closer to the reference text. The scores range from 0.112 to 0.77, indicating significant variability in performance among the models with some giving answers much closer to a reference than others.

BERT Score: This score reflects how well the model captures the meaning of the text, using contextual embeddings from BERT. It typically ranges from 0 to 1, with higher scores indicating better performance. In this table, the scores range from 0.542 to 0.705, suggesting that some models' responses are closer contextually than others.

ROUGE-L Score: This metric measures the longest common subsequence and is often used for summarization tasks. It indicates how much of the core information is captured by the generated response. The scores vary from 0.129 to 0.277, showing some models have better response than others.

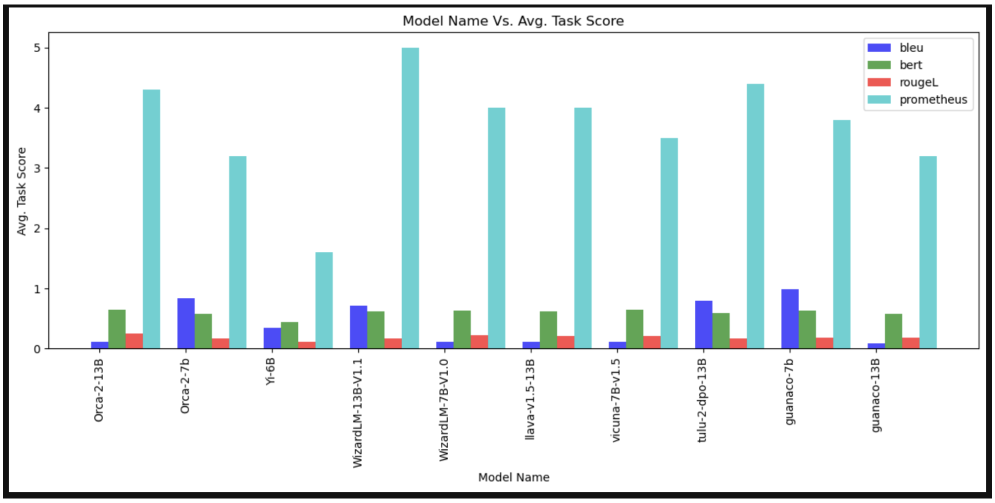


Fig. 4 Average Text Generation score across models

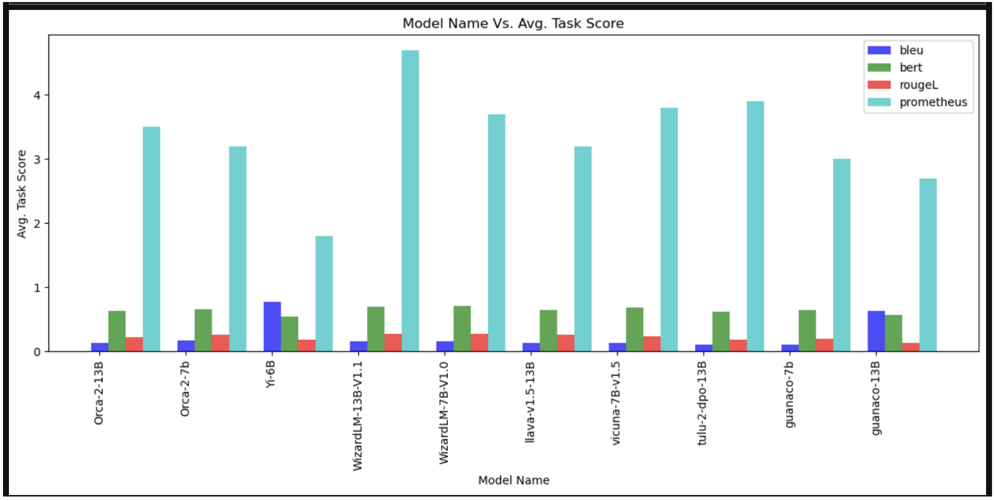


Fig. 5 Average Open-ended QA score across models

Prometheus Score: This metric isn't standard like the other three and is based on custom fine-grained criteria. It indicates how much the response is inline with the fine-grain criteria. The score vary from 1.8 to 4.7, showing some models have better criteria oriented response than others.

Looking at the scores, the model 'WizardLM/WizardLM-13B-V1.1' stands out with the highest Prometheus score of 4.7, which might indicate a superior overall performance. The model '01-ai/Yi-6B' shows the lowest scores across most of the metrics, which might indicate it is less proficient in the tasks measured compared to the others.

Table 2 Open-ended Question Answering Instruction scores

Model Name	bleu[1]	bert	rougeL	prometheus
TheBloke/Orca-2-13B-AWQ	0.128	0.629	0.218	3.5
microsoft/Orca-2-7b	0.168	0.658	0.256	3.2
01-ai/Yi-6B	0.77	0.542	0.182	1.8
WizardLM/WizardLM-13B-V1.1	0.154	0.698	0.277	4.7
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.159	0.705	0.27	3.7
TheBloke/llava-v1.5-13B-AWQ	135	0.646	0.258	3.2
TheBloke/vicuna-7B-v1.5-GPTQ	0.133	0.678	0.232	3.8
TheBloke/tulu-2-dpo-13B-AWQ	0.112	0.621	0.18	3.9
timdettmers/guanaco-7b	0.112	0.647	0.197	3.0
TheBloke/guanaco-13B-GPTQ	0.63	0.574	0.129	2.7

[1] These are average scores for open-ended QA instructions. For detailed scores refer Appendix A

Table 3 Correlation analysis between Prometheus Score and other mertic scores for text generation

Metric	Pearson's Correlation
bleu	0.169099
bert	0.806509
rougeL	0.612497

7.3 Correlation analysis between Standard metrics and Custom Fine-grained metric for text generation instructions

Pearson correlation (refer table 3) between Prometheus score and other metric scores reveals that:

BLEU Score: There is a weak positive correlation with Prometheus score (0.169). This suggests that the BLEU score has little association with the custom defined fine-grain metric in this dataset.

BERT Score: There is a strong positive correlation with Prometheus (0.807). This indicates that models with higher BERT scores tend to have higher Prometheus scores, suggesting that the aspects of performance measured by BERT are similarly valued by the custom defined fine-grain metric.

ROUGE-L Score: There is also a medium positive correlation with Prometheus (0.612). This suggests that models that perform well in producing responses that contain the longest common subsequences with reference texts tend to be rated highly by the custom defined fine-grain metric.

Given these correlations, we can infer that the custom defined fine-grain metric is more closely aligned with the BERT and ROUGE-L scores than with the BLEU score.

Table 4 Correlation analysis between Prometheus Score and other mertic scores for open-ended QA

Metric	Pearson's Correlation
bleu	-0.762148
bert	0.815986
rougeL	0.575397

This might indicate that the qualities captured by BERT and ROUGE-L, such as contextual understanding and summary quality, are significant factors in the evaluation criteria of the custom defined fine-grain metric.

7.4 Correlation analysis between standard metrics and custom fine-grained metric for open-ended question answer instructions

The correlation analysis(refer table 4 between the Prometheus score and the other metrics for this dataset yields the following results:

BLEU Score: There is a strong negative correlation with Prometheus (-0.762). This indicates that models with higher BLEU scores tend to have lower Prometheus scores in this particular dataset, suggesting that the custom defined fine-grain metric may prioritize different aspects of performance than BLEU.

BERT Score: There is a strong positive correlation with Prometheus (0.816). This suggests that models that score higher on the BERT metric, which measures contextual language understanding, also tend to score higher on custom defined fine-grain metric.

ROUGE-L Score: There is a medium positive correlation with Prometheus (0.575). This indicates that models that are better at answering open-ended questions that contain the longer common subsequences with the reference tend to score higher on the custom defined fine-grain metric.

From this analysis, it appears that the custom defined fine-grain metric aligns more closely with BERT and ROUGE-L scores, emphasizing contextual understanding and the quality of generated text. The strong negative correlation with BLEU scores is notable; it suggests that whatever custom defined fine-grain metric is measuring, it is not directly related to the aspects of translation or text generation quality captured by BLEU. This could imply that the custom defined fine-grain metric can capture aspects such as novelty, diversity, or adherence to certain constraints that are not captured by BLEU.

8 Discussion

8.1 Does custom evaluation criteria provide better evaluation ?

The experiment and correlation analysis shows that Prometheus score can capture and measure aspects of text generation like novelty, diversity, adherence. But to understand if it is better than already existing metrics, more study needs to be done.

8.2 Validity of the Prometheus score

During experiments it was observed that Prometheus sometimes generates different feedback for the same response. For example in one iteration for model 'Llava-v1.5-13B' for text generation instructions it generated scores (3,5,5,3,5,5,4,1,5,4) and scores (1,5,5,4,5,5,5,1,blank,4). In this example it is visible that for instructions 1,4,7 the score varies from 3 to 1, 3 to 4 and 4 to 5 respectively. Even thought the variance is not high(except for instruction 1), this behaviour must be kept in mind while using Prometheus. Second observation made during the experiments was Prometheus does not provide a score for each scoring criteria defined in score rubric. Instead it provides a combined feedback and score. How this combined score is calculated is unclear. Also, Score and feedback for each criteria can be more useful for human understanding. These observations undermine the validity of Prometheus score in some manner.

9 Limitations

9.1 Limitation due to number of instructions

This dataset used in this study only includes set of 10 text generation and 10 open-ended question instructions. This dataset only contains 6 types of text generation instructions (Descriptive, Narrative, Opinion, Specific content, Question answer and generative). The remaining 8 instructions (comparative, Argumentative, Analytical, Predictive, Summarization, Review and Dialogue) are not included in this study.

9.2 Limitation due to Number of models

This study used only 10 models consisting of both recent and established models. In my opinion, since this study does not focus on finding the best models but finding the best evaluation method, a smaller number of models would provide better analysis.

9.3 Limitations due to performance of models

The models used in this study do not have highest scores on standard benchmarks, leading to non ideal responses sometimes. To counter this, a smaller number of models which have a high score on standard benchmarks can be used.

9.4 Limitations due to no human evaluations

The analysis in this study did not include human-based evaluations. Including human evaluations can provide a good understanding of performance of custom evaluation.

10 Conclusion and Outlook

This study proves that custom fine-grained evaluation metrics can be used for evaluating LMs. But more work needs to be done, to find if this evaluation is better than FLASK or other LLM-as-a judge evaluations. Following studies can be done in this direction:

- Comparative analysis of custom fine-grained evaluation criteria and FLASK criteria.
- Comparative analysis of custom fine-grained evaluation and Human evaluation.
- Using established models like ChatGPT for LLM based fine-grained evaluation.

Declarations

I certify that the work was solely undertaken by myself without any third person or party helping and it is not AI generated. All sections of the paper that use quotes or describe an argument or concept developed by another author have been referenced, including all secondary literature used, to show that this material has been adopted to support my paper. This is especially true for contents generated by artificial intelligence like GPT-4 or Bard.

Chouguley

Sarang Ravi Chouguley
1st February 2024

References

[1] Blagec K, Dorffner G, Moradi M, et al (2022) A global analysis of metrics used for measuring performance in natural language processing. Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP URL https://doi.org/10.18653/v1/2022.nlppower-1.6

[2] Callan D, Foster J (2023) How interesting and coherent are the stories generated by a large-scale neural language model? Comparing human and automatic evaluations of machine-generated text. Expert Systems 40(6). URL https://doi.org/10.1111/exsy.13292

[3] Chen Y (2023) X-IQE: eXplainable Image Quality Evaluation for Text-to-Image Generation with Visual Large Language Models. arXiv (Cornell University) URL http://arxiv.org/abs/2305.10843

[4] Chen Y, Xu B, Wan B, et al (2024) Benchmarking Large Language Models on Controllable Generation under Diversified Instructions. arXiv (Cornell University) URL https://arxiv.org/abs/2401.00690

[5] Deutsch D, Roth D (2022) Benchmarking answer verification methods for Question Answering-Based Summarization evaluation metrics. arXiv (Cornell University) URL http://arxiv.org/abs/2204.10206

[6] Hiesinger W, Zakka C, Chaurasia S, et al (2023) Almanac: Retrieval-Augmented Language Models for Clinical Medicine. Research Square (Research Square) URL https://doi.org/10.21203/rs.3.rs-2883198/v1

[7] Hua X, Sreevatsa A, Wang L (2021) DYPLOC: Dynamic Planning of Content Using Mixed Language Models for Text Generation. 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) URL https://doi.org/10.18653/v1/2021.acl-long.501

[8] Ji T, Lyu C, Jones G, et al (2022) QASCoRE – an unsupervised unreferenced metric for the question generation evaluation. arXiv (Cornell University) URL http://arxiv.org/abs/2210.04320

[9] Kamalloo E, Dziri N, Clarke CLA, et al (2023) Evaluating Open-Domain question answering in the era of large language models. arXiv (Cornell University) URL https://arxiv.org/abs/2305.06984

[10] Kaster M, Zhao W, Eger S (2021) Global Explainability of BERT-Based Evaluation Metrics by Disentangling along Linguistic Factors. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing URL https://doi.org/10.18653/v1/2021.emnlp-main.701

[11] Kim S, Shin JC, Cho Y, et al (2023) Prometheus: Inducing fine-grained evaluation capability in language models. arXiv (Cornell University) URL https://arxiv.org/abs/2310.08491

[12] Lightman H, Kosaraju V, Burda Y, et al (2023) Let's verify step by step. arXiv (Cornell University) URL https://arxiv.org/abs/2305.20050

[13] Lin BY, Ravichander A, Lu X, et al (2023) The unlocking spell on base LLMs: Rethinking alignment via In-Context Learning. arXiv (Cornell University) URL https://arxiv.org/abs/2312.01552

[14] Lin CY (2004) ROUGE: A Package for Automatic Evaluation of Summaries. Text summarization branches out pp 74–81. URL http://anthology.aclweb.org/W/W04/W04-1013.pdf

[15] Macháček D, Bojar O, Dabre R (2023) MT Metrics Correlate with Human Ratings of Simultaneous Speech Translation. arXiv preprint arXiv:221108633 URL https://doi.org/10.18653/v1/2023.iwslt-1.12

[16] Min S, Krishna K, Lyu X, et al (2023) FACTSCore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv (Cornell University) URL https://arxiv.org/abs/2305.14251

[17] Muennighoff N (2022) SGPT: GPT Sentence embeddings for Semantic Search. arXiv (Cornell University) URL https://arxiv.org/abs/2202.08904

[18] Mukherjee A (2022) REUSE: REference-free UnSupervised Quality Estimation Metric. URL https://aclanthology.org/2022.wmt-1.50

[19] Nguyen A (2021) Language model evaluation in open-ended text generation. arXiv (Cornell University) URL https://arxiv.org/pdf/2108.03578

[20] Papineni K, Roukos S, Ward TJ, et al (2001) BLEU. Association for Computational Linguistics (ACL) URL https://doi.org/10.3115/1073083.1073135

[21] Rabinovich E, Ackerman S, Raz O, et al (2023) Predicting Question-Answering Performance of Large Language Models through Semantic Consistency. arXiv (Cornell University) URL https://arxiv.org/abs/2311.01152

[22] Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:190810084 URL https://doi.org/10.18653/v1/d19-1410

[23] Risch J, Möller T, Gutsch J, et al (2021) Semantic Answer Similarity for Evaluating Question Answering Models. Proceedings of the 3rd Workshop on Machine Reading for Question Answering URL https://doi.org/10.18653/v1/2021.mrqa-1.15

[24] Sasindran Z, Yelchuri H, Prabhakar TV, et al (2023) HEVAL: A New Hybrid Evaluation Metric for Automatic Speech Recognition Tasks. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) URL https://doi.org/10.1109/asru57964.2023.10389717

[25] Tan B, Yang Z, Ai-Shedivat M, et al (2020) Progressive Generation of Long Text with Pretrained Language Models. arXiv (Cornell University) URL https://arxiv.org/abs/2006.15720

[26] Ye J, Li G, Gao S, et al (2024) ToolEYeS: Fine-Grained Evaluation for tool learning capabilities of large language models in real-world scenarios. arXiv (Cornell University) URL https://arxiv.org/abs/2401.00741

[27] Ye S, Kim D, Kim S, et al (2023) FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. arXiv (Cornell University) URL https://arxiv.org/abs/2307.10928

[28] Zhang H, Song H, Li S, et al (2023) A survey of controllable text generation using transformer-based pre-trained language models. ACM Computing Surveys 56(3):1–37. URL https://doi.org/10.1145/3617680

[29] Zhang T, Kishore V, Wu F, et al (2020) BERTScore: Evaluating Text Generation with BERT. arXiv (Cornell University) URL https://arxiv.org/pdf/1904.09675.pdf

Appendix A

Table A1 Bleu score for text generation instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	0.055	0.171	0.118	0.072	0.079	0.247	0.194	0.156	0.019	0.086	0.12
microsoft/Orca-2-7b	0.043	0.033	0.128	0.097	0.077	0.076	0.057	0.106	0.016	0.204	0.084
01-ai/Yi-6B	0.022	0.064	0.066	0.009	0.0	0.019	0.042	0.076	0.007	0.032	0.034
WizardLM/WizardLM-13B-V1.1	0.054	0.046	0.083	0.073	0.073	0.026	0.064	0.103	0.007	0.177	0.071
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.116	0.054	0.102	0.045	0.194	0.178	0.145	0.131	0.014	0.208	0.119
TheBloke/llava-v1.5-13B-AWQ	0.091	0.075	0.171	0.083	0.113	0.043	0.136	0.147	0.004	0.235	0.11
TheBloke/vicuna-7B-v1.5-GPTQ	0.086	0.087	0.125	0.134	0.117	0.26	0.126	0.043	0.011	0.189	0.118
TheBloke/tulu-2-dpo-13B-AWQ	0.058	0.063	0.101	0.071	0.109	0.068	0.112	0.0	0.008	0.199	0.079
timdettmers/guanaco-7b	0.067	0.042	0.151	0.061	0.074	0.22	0.088	0.135	0.01	0.128	0.098
TheBloke/guanaco-13B-GPTQ	0.084	0.09	0.091	0.135	0.061	0.138	0.08	0.072	0.001	0.144	0.09

Table A2 Bleu score for open-ended QA instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	0.117	0.027	0.213	0.059	0.054	0.002	0.204	0.277	0.215	0.11	0.128
microsoft/Orca-2-7b	0.108	0.035	0.213	0.053	0.054	0.826	0.004	0.117	0.203	0.07	0.168
01-ai/Yi-6B	0.053	0.016	0.187	0.086	0.077	0.0	0.156	0.031	0.11	0.059	0.077
WizardLM/WizardLM-13B-V1.1	0.072	0.048	0.115	0.108	0.082	0.452	0.198	0.175	0.207	0.086	0.154
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.048	0.064	0.213	0.182	0.109	0.291	0.214	0.214	0.142	0.112	0.159
TheBloke/llava-v1.5-13B-AWQ	0.119	0.047	0.213	0.146	0.233	0.119	0.013	0.263	0.041	0.152	0.135
TheBloke/vicuna-7B-v1.5-GPTQ	0.065	0.034	0.095	0.174	0.079	0.184	0.21	0.201	0.089	0.199	0.133
TheBloke/tulu-2-dpo-13B-AWQ	0.067	0.053	0.139	0.0	0.061	0.0	0.207	0.216	0.204	0.17	0.112
timdettmers/guanaco-7b	0.086	0.039	0.064	0.105	0.082	0.03	0.154	0.186	0.22	0.158	0.112
TheBloke/guanaco-13B-GPTQ	0.05	0.021	0.031	0.047	0.066	0.025	0.142	0.089	0.107	0.049	0.063

Table A3 BERTScore for text generation instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	0.537	0.735	0.697	0.599	0.562	0.689	0.768	0.648	0.554	0.657	0.645
microsoft/Orca-2-7b	0.561	0.492	0.653	0.645	0.566	0.534	0.622	0.581	0.424	0.724	0.58
01-ai/Yi-6B	0.356	0.624	0.458	0.276	0.326	0.648	0.384	0.493	0.41	0.426	0.44
WizardLM/WizardLM-13B-V1.1	0.569	0.591	0.641	0.616	0.558	0.553	0.65	0.644	0.592	0.765	0.618
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.562	0.583	0.667	0.5	0.672	0.687	0.657	0.637	0.54	0.761	0.627
TheBloke/llava-v1.5-13B-AWQ	0.542	0.602	0.716	0.627	0.572	0.554	0.679	0.641	0.548	0.741	0.622
TheBloke/vicuna-7B-v1.5-GPTQ	0.563	0.618	0.655	0.679	0.589	0.763	0.7	0.597	0.589	0.728	0.648
TheBloke/tulu-2-dpo-13B-AWQ	0.564	0.589	0.687	0.622	0.571	0.599	0.697	0.293	0.542	0.742	0.591
timdettmers/guanaco-7b	0.583	0.566	0.693	0.615	0.523	0.731	0.685	0.625	0.565	0.732	0.632
TheBloke/guanaco-13B-GPTQ	0.563	0.608	0.629	0.586	0.51	0.629	0.593	0.588	0.423	0.683	0.581

Table A4 BERTScore for open-ended QA instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	0.634	0.473	0.752	0.623	0.551	0.292	0.699	0.792	0.773	0.703	0.629
microsoft/Orca-2-7b	0.635	0.544	0.761	0.608	0.624	0.978	0.387	0.663	0.747	0.634	0.658
01-ai/Yi-6B	0.53	0.375	0.713	0.605	0.598	0.418	0.595	0.506	0.488	0.589	0.542
WizardLM/WizardLM-13B-V1.1	0.637	0.584	0.652	0.677	0.652	0.923	0.731	0.721	0.736	0.669	0.698
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.653	0.585	0.757	0.733	0.673	0.813	0.699	0.722	0.708	0.704	0.705
TheBloke/llava-v1.5-13B-AWQ	0.657	0.542	0.738	0.727	0.78	0.662	0.343	0.741	0.527	0.741	0.646
TheBloke/vicuna-7B-v1.5-GPTQ	0.623	0.573	0.634	0.73	0.662	0.716	0.694	0.695	0.678	0.779	0.678
TheBloke/tulu-2-dpo-13B-AWQ	0.611	0.583	0.681	0.346	0.667	0.419	0.713	0.709	0.723	0.763	0.621
timdettmers/guanaco-7b	0.567	0.517	0.613	0.664	0.649	0.59	0.695	0.71	0.724	0.738	0.647
TheBloke/guanaco-13B-GPTQ	0.562	0.44	0.513	0.613	0.615	0.507	0.606	0.615	0.646	0.625	0.574

Table A5 RougeL Score for text generation instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	0.109	0.326	0.301	0.142	0.193	0.364	0.393	0.273	0.163	0.186	0.245
microsoft/Orca-2-7b	0.104	0.147	0.254	0.181	0.183	0.123	0.148	0.225	0.031	0.28	0.168
01-ai/Yi-6B	0.072	0.129	0.099	0.009	0.0	0.24	0.111	0.206	0.095	0.138	0.11
WizardLM/WizardLM-13B-V1.1	0.126	0.14	0.224	0.141	0.187	0.079	0.158	0.222	0.143	0.276	0.17
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.175	0.142	0.239	0.153	0.291	0.311	0.227	0.245	0.15	0.297	0.223
TheBloke/llava-v1.5-13B-AWQ	0.14	0.194	0.351	0.153	0.213	0.092	0.283	0.242	0.069	0.352	0.209
TheBloke/vicuna-7B-v1.5-GPTQ	0.142	0.176	0.212	0.184	0.201	0.385	0.228	0.142	0.16	0.275	0.21
TheBloke/tulu-2-dpo-13B-AWQ	0.135	0.181	0.249	0.116	0.191	0.151	0.227	0.0	0.197	0.267	0.171
timdettmers/guanaco-7b	0.156	0.093	0.271	0.134	0.135	0.364	0.197	0.24	0.098	0.191	0.188
TheBloke/guanaco-13B-GPTQ	0.16	0.157	0.2	0.208	0.137	0.237	0.148	0.159	0.103	0.264	0.177

Table A6 RougeL score for open-ended QA instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	0.278	0.119	0.37	0.128	0.165	0.007	0.263	0.383	0.241	0.23	0.218
microsoft/Orca-2-7b	0.226	0.081	0.408	0.144	0.137	0.909	0.03	0.193	0.23	0.197	0.256
01-ai/Yi-6B	0.191	0.08	0.375	0.193	0.174	0.0	0.267	0.13	0.251	0.156	0.182
WizardLM/WizardLM-13B-V1.1	0.216	0.148	0.255	0.195	0.203	0.762	0.27	0.283	0.24	0.195	0.277
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	0.15	0.168	0.408	0.273	0.21	0.606	0.247	0.274	0.132	0.236	0.27
TheBloke/llava-v1.5-13B-AWQ	0.275	0.168	0.407	0.267	0.284	0.377	0.01	0.376	0.096	0.319	0.258
TheBloke/vicuna-7B-v1.5-GPTQ	0.198	0.103	0.23	0.257	0.198	0.316	0.261	0.262	0.157	0.338	0.232
TheBloke/tulu-2-dpo-13B-AWQ	0.191	0.137	0.214	0.0	0.182	0.0	0.281	0.308	0.201	0.291	0.18
timdettmers/guanaco-7b	0.182	0.147	0.171	0.206	0.182	0.105	0.227	0.249	0.208	0.291	0.197
TheBloke/guanaco-13B-GPTQ	0.144	0.09	0.103	0.092	0.186	0.062	0.179	0.154	0.157	0.127	0.129

Table A7 Prometheus Score for text generation instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	1	5	5	5	5	4	4	5	4	5	4.3
microsoft/Orca-2-7b	1	1	5	4	5	3	5	2	1	5	3.2
01-ai/Yi-6B	1	4	1	1	1	4	1	1	1	1	1.6
WizardLM/WizardLM-13B-V1.1	5	5	5	5	5	5	5	5	5	5	5.0
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	2	5	4	4	3	5	3	5	4	5	4.0
TheBloke/llava-v1.5-13B-AWQ	3	5	5	3	5	5	4	1	5	4	4.0
TheBloke/vicuna-7B-v1.5-GPTQ	1	4	5	4	1	5	5	1	5	4	3.5
TheBloke/tulu-2-dpo-13B-AWQ	5	4	5	5	4	5	5	2	4	5	4.4
timdettmers/guanaco-7b	1	4	5	4	1	5	4	4	5	5	3.8
TheBloke/guanaco-13B-GPTQ	2	4	5	4	1	5	5	1	1	4	3.2

Table A8 Prometheus score for open-ended QA instructions

Model Name / Instruction No.	1	2	3	4	5	6	7	8	9	10	avg
TheBloke/Orca-2-13B-AWQ	5	4	1	4	1	1	5	5	5	4	3.5
microsoft/Orca-2-7b	4	1	1	4	2	5	1	4	5	5	3.2
01-ai/Yi-6B	2	1	1	1	5	3	1	1	1	2	1.8
WizardLM/WizardLM-13B-V1.1	5	3	5	4	5	5	5	5	5	5	4.7
TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	3	2	1	5	4	5	3	5	5	4	3.7
TheBloke/llava-v1.5-13B-AWQ	4	1	1	5	5	4	1	5	1	5	3.2
TheBloke/vicuna-7B-v1.5-GPTQ	2	4	2	4	5	4	5	4	3	5	3.8
TheBloke/tulu-2-dpo-13B-AWQ	4	3	5	1	5	1	5	5	5	5	3.9
timdettmers/guanaco-7b	3	4	2	5	1	1	4	4	3	3	3.0
TheBloke/guanaco-13B-GPTQ	4	1	3	5	4	1	1	1	5	2	2.7

Table A9 Model Details

Sr. No.	Model Name	Parameter Size	Quantization type	Comments
0	TheBloke/Orca-2-13B-AWQ	13B	AWQ	The model is designed to excel particularly in reasoning
1	microsoft/Orca-2-7b	7B	N/A	The model is designed to excel particularly in reasoning
2	01-ai/Yi-6B-Chat-4bits	6B	AWQ	4bit quantized model
3	WizardLM/WizardLM-13B-V1.1	13B	N/A	WizardLM 13B model.
4	TheBloke/WizardLM-7B-V1.0-Uncensored-GPTQ	7B	GPTQ	This is an uncensored model, meaning the guard rails are removed.
5	TheBloke/llava-v1.5-13B-AWQ	13B	AWQ	LLaVA is an open-source chatbot trained by fine-tuning LLaMA/Vicuna on GPT-generated multimodal instruction-following data. It is an auto-regressive language model, based on the transformer architecture.
6	TheBloke/vicuna-7B-v1.5-GPTQ	7B	GPTQ	Vicuna is a chat assistant trained by fine-tuning Llama 2 on user-shared conversations collected from ShareGPT.
7	TheBloke/tulu-2-dpo-13B-AWQ	13B	AWQ	This model is trained using Direct Preference Optimization. This model is a strong alternative of 'Llama 2 13B Chat'.
8	timdettmers/guanaco-7b	7B	N/A	The Guanaco models are open-source finetuned chatbots obtained through 4-bit QLoRA tuning of LLaMA base models on the OASST1 dataset.
9	TheBloke/Guanaco-13B-Uncensored-GPTQ	13B	GPTQ	The Guanaco models are open-source finetuned chatbots obtained through 4-bit QLoRA tuning of LLaMA base models on the OASST1 dataset.

Table A10 Score rubric for evaluation criteria

Criteria	Definition	Score 1(worst)	Score 2	Score 3	Score 4	Score 5(best)
Relevance	The response should be directly related to or aligned with the given topic, context, theme, subject.	The response is completely irrelevant and does not address or align with the given topic, context, theme, or subject.	The response contains significant irrelevant elements that critically undermine its relevance to the given prompt.	The response includes some relevant information, but considerable effort is needed to align it with the specified topic.	The response is mostly relevant, with minor deviations that are easy to rectify and do not significantly impact overall relevance.	The response is entirely relevant and aligned with the given topic, context, theme, or subject.
Clarity and Coherence	The response should be organised in a clear ,coherent and easy manner.	The response lacks organization and is entirely unclear and coherent making it challenging to follow.	The response is disorganized and contains significant elements that undermine its clarity and overall coherence.	The response is somewhat clear but requires coherent structure and considerable effort to make it easily understandable.	The response is generally clear and coherent, with minor areas that are easy to rectify for enhanced clarity and coherence.	The response is well-organized and easy to follow, demonstrating high clarity and coherence.
Consistency	The response should be consistent in themes, characters, elements, tone and style.	The response is entirely inconsistent in themes, characters, elements, tone, and style.	The response contains significant inconsistencies that critically impact its overall coherence.	The response has some inconsistencies that require considerable effort to maintain a consistent narrative.	The response is mostly consistent, with minor inconsistencies that are easy to rectify.	The response is entirely consistent in themes, characters, elements, tone, and style.
Logical and Common sense	The response should be logically valid and have common sense.	The response is entirely illogical and lacks common sense.	The response is illogical and contains significant elements that undermine its validity.	The response is somewhat logical and contains some common sense.	The response is generally logical, with few areas that lack common sense.	The response is completely logical and valid according to common sense.
Completeness	The response should be complete, containing all relevant details necessary to give a comprehensive understanding of the topic.	The response is entirely incomplete, lacking essential details for a comprehensive understanding of the topic.	The response is significantly incomplete, with critical omissions that undermine its overall completeness.	The response is somewhat complete but requires considerable effort to include all relevant details.	The response is mostly complete, with minor omissions that are easy to rectify.	The response is entirely complete, containing all necessary details for a comprehensive understanding of the topic.
Accuracy	The response should be factually correct, providing accurate details.	The response is entirely factually incorrect, providing inaccurate details throughout.	The response contains significant factual errors that critically undermine its overall accuracy.	The response includes some inaccuracies that require considerable effort to correct.	The response has minor factual errors, which are easy to rectify and do not significantly impact its overall accuracy.	The response is entirely accurate and factually correct.
Engagement	The response should be engaging, captivating the reader's interest and imagination.	The response is entirely unengaging, failing to captivate the reader's interest.	The response lacks engagement and contains significant elements that hinder reader interest.	The response is somewhat engaging but requires considerable effort to captivate the reader's interest effectively.	The response is generally engaging, with minor areas that are easy to rectify for improved reader interest.	The response is entirely engaging, captivating the reader's interest and imagination effectively.
Creativity	The response should showcase imaginative thinking, originality, and creativity in the content.	The response lacks any imaginative thinking, originality, or creativity.	The response lacks creativity and contains significant elements that undermine its overall imaginative content.	The response shows some imaginative thinking but requires considerable effort to enhance creativity.	The response is mostly creative, with minor areas that are easy to rectify for increased originality.	The response is entirely creative, showcasing imaginative thinking, originality, and creativity.
Thoughtfulness and Insight	The response should contain thoughtful analysis, showcasing insight into potential outcomes or developments.	The response lacks any thoughtful analysis or insight into potential outcomes or developments.	The response lacks thoughtfulness and contains significant elements that undermine its overall depth of insight.	The response shows some thoughtfulness and insight but requires considerable effort to enhance depth.	The response is generally thoughtful, with minor areas that are easy to rectify for increased insight.	The response is entirely thoughtful, containing insightful analysis and depth of understanding.
Evidence and Support	The response should contain evidence or support for the claims presented.	The response lacks any evidence or support for the claims presented.	The response lacks sufficient evidence or support and contains significant elements that undermine its overall credibility.	The response includes some evidence or support but requires considerable effort to strengthen its claims.	The response is generally supported, with minor areas that are easy to rectify for increased credibility.	The response is entirely supported, providing strong evidence for the claims presented.
Empathy	The response should convey opinions or reflections with an empathetic and considerate tone where applicable.	The response lacks any empathy and fails to convey opinions or reflections with a considerate tone.	The response lacks empathy and contains significant elements that undermine its overall considerate tone.	The response shows some empathy but requires considerable effort to convey opinions with a more considerate tone.	The response is generally empathetic, with minor areas that are easy to rectify for increased consideration.	The response is entirely empathetic, conveying opinions or reflections with a highly considerate tone where applicable.
Efficiency	The response should convey information efficiently without unnecessary verbosity.	The response is highly inefficient, containing excessive verbosity and failing to convey information effectively.	The response is inefficient and contains significant elements of unnecessary verbosity that hinder clarity.	The response is somewhat efficient but requires considerable effort to convey information more concisely.	The response is generally efficient, with minor areas that are easy to rectify for increased conciseness.	The response is entirely efficient, conveying information without unnecessary verbosity, ensuring clarity and brevity.