Quantifying the Influence of Evaluation Aspects on Long-Form Response Assessment

Go Kamoda[1] **Akari Asai**[2] **Ana Brassard**[3,1] **Keisuke Sakaguchi**[1,3]

[1]Tohoku University [2]University of Washington [3]RIKEN

Correspondence: go.kamoda@dc.tohoku.ac.jp

Abstract

Evaluating the outputs of large language models (LLMs) on long-form generative tasks remains challenging. While fine-grained, aspect-wise evaluations provide valuable diagnostic information, they are difficult to design exhaustively, and each aspect's contribution to the overall acceptability of an answer is unclear. In this study, we propose a method to compute an overall quality score as a weighted average of three key aspects: factuality, informativeness, and formality. This approach achieves stronger correlations with human judgments compared to previous metrics. Our analysis identifies factuality as the most predictive aspect of overall quality. Additionally, we release a dataset of 1.2k long-form QA answers annotated with both absolute judgments and relative preferences in overall and aspect-wise schemes, to aid future research in evaluation practices.

github.com/gokamoda/lfqa-weighted-eval

1 Introduction

Despite the widespread adoption of large language models (LLMs) for various open-ended generative tasks, such as long-form question answering (QA; Fan et al. 2019) or summarization (See et al., 2017), evaluating their outputs remains challenging (Wang et al., 2023b; Tan et al., 2024). Human annotations are costly and difficult to scale, so there has long been an interest in automating the process. Previously, a common approach was to use reference-based metrics, where the outputs are compared to one or several gold examples (Lin, 2004; Papineni et al., 2002), however, it was found to correlate poorly with human quality assessments (Krishna et al., 2021). In contrast, human assessments directly provide either a *relative preference* or an *absolute judgment* of quality. These can be further divided into asking for a single *overall* rating or fine-grained *aspect-wise* ones, explicitly covering a range of properties such as factuality or coherence (representative works are listed in Figure 2). Of these, fine-grained absolute scores offer the highest amount of diagnostic information, creating a clear picture of an output's quality. However, it is challenging to design an exhaustive list of all the important factors for a human. An overall assessment is thus desirable to cover any potential other aspects, but it is also still unclear how much each aspect contributes to the overall assessment.

In this study, we propose producing an overall rating as a *weighted average of aspect-wise ratings*. Specifically, we defined FACTUALITY, AMOUNTINFO, and FORMALITY as a representative set of commonly used aspects. We measured their contribution to ACCEPTABILITY (an overall assessment), used this to calculate a weighted average, and then compared the average rating's correlation with human ACCEPTABILITY ratings. This measure demonstrated better alignment with the human ratings compared to previous methods. Figure 1 shows an overview of the annotation scheme and example ratings.

The paper is organized as follows. Section 2 provides background on previous evaluation practices. Section 3 describes the process of building a dataset of human-written and generated answers for long-form QA with human annotations covering all four rating formats: absolute and relative, overall and fine-grained evaluation, respectively. In Section 4, we then analyze the annotations and define our weighted measure, finding that FACTUALITY was most predictive of ACCEPTABILITY scores. Finally, in Section 5, we compare the correlation between human ACCEPTABILITY ratings and several automatic measures, demonstrating that the proposed method yields the highest correlation. We release the new dataset, containing four-way human annotations for 1.2k answers of long-form QA, for future explorations of best practices for annotation.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8787–8808
January 19–24, 2025. ©2025 Association for Computational Linguistics

Question: Why do we hate our own voice when we hear it recorded?

Aspect	Scale	Description
Formality	-1 Too casual 0 1 Too Formal	The answer's formal appropriateness including its vocabulary, grammar, and spelling.
Amount Info	-1 Not enough 0 Sufficient 1 Too much	Whether or not sufficient amount of information was addressed to fully answer the question.
Factuality	0 Not accurate 1 2 3 Accurate	Whether or not the information provided in the candidate answer is factually correct.
Acceptability	0 Unacceptable 1 2 3 Acceptable	Whether or not the candidate answer is overall acceptable.

References		Formality	Amount Info	Factuality	Acceptability
Reference 1:	... Bhatt explained that the dislike of the sound of our own voices is physiological and psychological First off, audio recordings translate ... (link)				
Reference 2:	... Basically, the reasoning is that **because our recorded voice does not sound how we expect it to, we don't like it**. Dr Silke Paulmann, a ... (link)				
⋮	⋮				
Answers					
[HT]	When you hear your voice normally, you hear a sound transmitted through the air just like everyone else hears... but you also hear some of the sound transmitted through your jaw and skull. Since bone transits sound very differently to air, the way you hear your own voice is ...	-0.3	0.3	2.0	2.3
[HR]	When you hear yourself as you speak, you're also hearing the vibrations echoing around in your skull. That changes the way your voice sounds, compared to recording it. You're just not used to hearing yourself sound so different.	-0.7	-1.0	1.7	1.0
[MF]	We perceive our voice differently when we hear it on a recording because we are not used to hearing it from that perspective. When we speak, we hear our voice through the bones in our skull as well as through the air, which creates a richer, deeper tone. However, ...	0.3	0.0	2.7	3.0
[MC]	Have you ever listened to a recording of your own voice and cringed at the sound of it? Well, you're not alone. The reason why we hate the sound of our own voice is because when we talk, the soundwaves from our voice travel through our skull and jawbone, creating vibrations ...	0.0	0.3	2.7	3.0

Figure 1: Four aspects rated within a discrete range (shown on the left), and example ratings (averaged across three annotators) for a long-form QA instance (on the right). The blue-colored cells indicate ideal scores and the pink-colored cells indicate worse scores. HT, HR, MF, and MC denote Human Top, Human Random, Model Formal, and Model Casual, respectively. The scores displayed in the table are the average scores of three annotators. We also collect preferences along with free-form justification.

2 Background

Human evaluations of long-form generations. We provide a taxonomy of human evaluations studied in prior long-form evaluations. Figure 2 summarizes four different categories of evaluations with references to previous work.

Relative overall evaluation is one of the common approaches (Krishna et al., 2021; Zheng et al., 2023; Xu et al., 2023). Human evaluators are shown two candidate generations and asked which one is better. However, such relative evaluations do not provide absolute performance scores of subject systems, as is done in **Absolute Overall** evaluation. Moreover, the overall evaluation lacks insights into the factors and their degree of influence on final preferences (Wu et al., 2023). **Relative Multi-aspect** evaluation conducts relative evaluations in multiple axes (Nakano et al., 2021; Liu et al., 2023d). Specifically, they define fine-grained aspects and collect human evaluations in a pairwise relative manner. In **Absolute Multi-aspect** evaluation, on the other hand, outputs are evaluated for each aspect on an absolute scale (Wiegreffe et al., 2022). While multi-aspect evaluation provides a more fine-grained assessment of outputs, it often requires careful customization and designs of fine-grained aspects for each task.

LLM-based evaluations. Considering the cost of human evaluations, Wang et al. (2023a) and Liu et al. (2023b) conduct LLM-based assessment. They employ a multi-aspect evaluation scheme using LLMs and aggregate at the end to get a single overall score. Computing a single score for the overall score allows easy comparison of the performance of multiple systems. However, they aggregate the scores simply by taking the average, which may not be the most appropriate method when aiming for an evaluation protocol with a strong correlation against human preferences.

3 Collecting Fine-grained Human Annotations

This work investigates what aspects affect the overall rating of long-form responses with a focus on information-seeking queries requiring long-form responses by conducting annotations on both human and model-generated answers (Section 3.1). Our human annotation scheme consists of **fine-grained aspects** and **overall ratings**, both done in an absolute scoring scale. To get deeper insights into other affecting factors, we also collect free-form **justification** comments along with **overall preference** (Section 3.2).

3.1 Focus and Query-Response Data

Among various long-form generation tasks, we focus on the long-form QA task in this work. ELI5 is a dataset widely used in this task constructed by Fan et al. (2019), comprised of questions and answers collected from a Reddit forum, "Explain Like I'm 5." The dataset is widely used in recent stud-

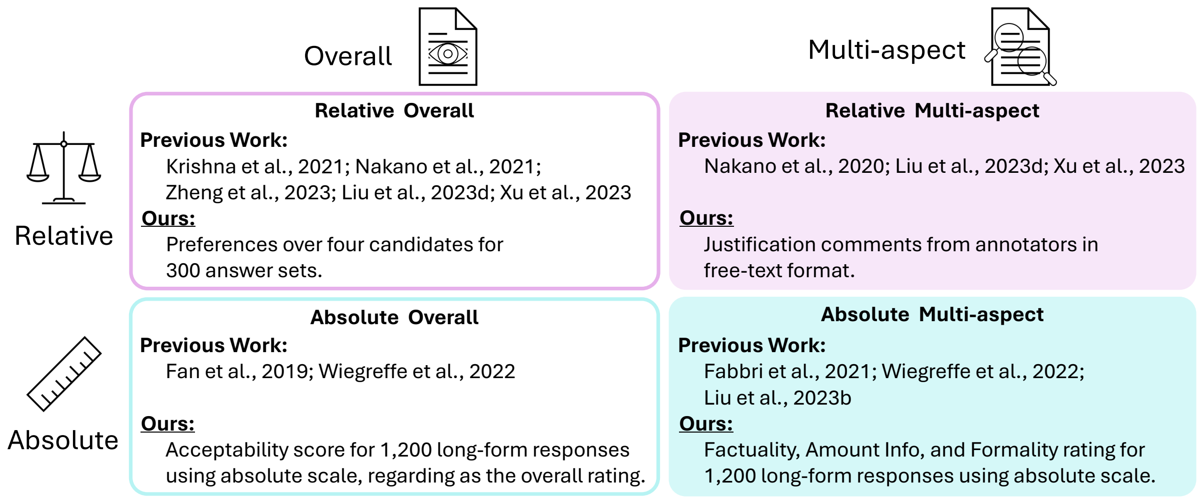


Figure 2: Previous works organized into four types of long-form generation evaluation and overview of annotation data we collected in this work.

ies as an assembly of diverse information-seeking queries (Liu et al., 2023a; Chen et al., 2023a). Thus, we use this dataset as a starting seed data for our evaluation. We evaluate two human-written responses and two model-generated responses for each of the 300 questions sampled from the ELI5 test set.

The 300 questions are sampled using the following procedure: From the first 800 instances (each with a question and multiple human written answers) in the ELI5 dataset, we dropped answers containing URLs to external sources to remove non-self-contained answers. After dropping instances with less than two answers, we collected model-generated responses from ChatGPT.[1] Because ChatGPT refused to answer some queries (e.g., for ethical reasons), we also dropped queries where ChatGPT generated answers starting with "As an AI." Finally, we randomly sampled 300 unique queries from the remaining queries.

Regarding the two human-written responses, one is the top-rated answer (**HT; Human-Top**), and another is randomly sampled from non-top-rated answers (**HR; Human-Random**). For model-generated responses, one is generated by a simple prompt (**MF; Model-Formal**), and another is prompted so that it generates answers in a more casual and engaging format (**MC; Model-Casual**). We included one *more casual and engaging* answer generated by the same model to investigate the effects of such stylistic features, as we noticed that human annotators in our preliminary experiment sometimes chose a human-written answer over a model-generated answer with the same amount of information due to their more engaging fashion.

3.2 Human Evaluation Scheme

The evaluation framework is shown in Figure 1. This study quantitatively investigates the importance of three aspects (FACTUALITY, AMOUNTINFO, and FORMALITY) on ACCEPTABILITY.

Overall rating: We consider ACCEPTABILITY as a higher-level aspect measuring overall rating. We use a 4-point scale for this aspect, with the ideal score of 3, to avoid the middle option.

Fine-grained aspect: For the first fine-grained aspect, we set FACTUALITY. FACTUALITY is intuitively important in long-form QA tasks, as it is required for QA responses to provide accurate information. Because responses contain multiple factual statements (Min et al., 2023; Mishra et al., 2024), binary evaluation of whether the addressed statements are accurate is intuitively insufficient for this task. Thus, we use a 4-point scale for this aspect, with the ideal score of 3, to avoid the middle option. While FACTUALITY measures the quality of information, AMOUNTINFO measures the quantity of information. Responses need to have sufficient information, but at the same time, they should not overwhelm the reader with too much information. Thus, we set a scale from -1 (Not enough information) to 1 (Too much information), with 0 being the ideal score. The last aspect is FORMALITY, which measures the surface-level quality of the responses, including choice of vocabulary, use

[1]The responses were generated in May 2023.

of slang, and correctness of grammar. This aspect serves as an umbrella for measures previously used to evaluate natural language generation (Howcroft et al., 2020). From the intuition that too casual or too formal responses are not favored, we set a scale from -1 (too casual) to 1 (too formal), with 0 being the ideal score.

Justification: Following Xu et al. (2023), we collect free-form justification comments to capture important aspects other than the three aspects we set.

3.3 Annotation Details

We recruited crowdworkers on Amazon Mechanical Turk (AMT) who met our qualification criterion regarding English text quality assessment. Finding crowd workers who can provide high-quality annotation on responses to information-seeking queries is often challenging (Xu et al., 2023). We recruit a set of crowd workers from the crowd worker pool after multiple qualification processes for a relevant annotation task.

The qualification process is explained in Appendix A. Annotators spend about 13-14 minutes on average[2] to complete the task, and we paid $4 for each annotation, resulting in a total cost of $5,064, including preliminary experiments.

The annotation interface can be seen in Appendix D. Other than the instructions and annotation targets, we provide ten websites retrieved by Google Search API[3] in an easy-to-verify format (e.g., snippets + iframe) to enable a friendly and accurate factuality check. Note that we do not restrict the use of other external sources.

We collect three annotations per instance, and the overall inter-annotator agreement measured with Krippendorff's α was 0.74. A more detailed analysis of annotators' agreement is in Appendix B.

4 What Affects Overall Preferences?

In this section, we analyze the collected human evaluation by assessing factors affecting overall rating from fine-grained ratings and human justifications (Section 4.1). Our human annotations also enable us to understand how model-generated answers are different from human-written answers (Section 4.2).

4.1 Dissecting Acceptability Factors

Quantitative analysis on affecting factors. To determine the impact of FACTUALITY, AMOUNTINFO, and FORMALITY on ACCEPTABILITY, we train a simple linear regressor to predict overall ACCEPTABILITY given scores for each of the three other aspects. We first normalize (to range $[0, 1]$) and re-scale the annotated scores to train the model in the deductive method by converting ratings to "distance from the ideal score"(Equation 1, 2). We train the model to learn weights $\mathbf{w} \in \mathbb{R}^3$ in Equation 3.

$$f\left(\begin{bmatrix} S_{\text{FORM}} \\ S_{\text{INFO}} \\ S_{\text{FACT}} \end{bmatrix}\right) = \begin{bmatrix} (S_{\text{FORM}} - 3)/3 \\ -|S_{\text{INFO}}| \\ -|S_{\text{FACT}}| \end{bmatrix} \tag{1}$$

$$y_{\text{ACCE}} = S_{\text{ACCE}} - 3 \tag{2}$$

$$y_{\text{ACCE}} = \mathbf{w} \cdot f(\mathbf{x}) \tag{3}$$

The model was trained on 80% of the annotated data and yielded a Pearson's correlation coefficient ρ of 0.853 on the remaining 20% of the data. After training, w_{FORM}, w_{INFO}, and w_{FACT} were determined as 0.335, 0.739, and 2.048, respectively. The weights indicate that FACTUALITY is the most important evaluation criterion, with approximately three times the weight of AMOUNTINFO. We confirmed that answers with low ACCEPTABILITY also receive low FACTUALITY (Appendix Figure 7). FORMALITY is the least valued aspect, indicating that surface-level quality has less effect on a human's overall acceptability. The findings align with the result of training a decision tree, which is depicted in Appendix Figure 8.

Qualitative analysis on explanations. To further understand how annotators make such judgments, we conduct a manual analysis of the 50 longest justifications of crowd workers' relative preferences. We follow the work of Xu et al. (2023) on the aspects to annotate. We also include "Formality" and "Amount Info." Figure 4b shows the frequency of each aspect mentioned in the justifications for judgments. We show the complete list of the definitions of each aspect with examples in Appendix Table 4. Possibly due to the annotation task we define, the most mentioned aspects are "Amount Info," "Factuality," and "Formality." "Easiness to understand" was mentioned frequently, especially

[2]Because AMT workers can open multiple tasks simultaneously, this value is the result of our analysis according to the AcceptTime and SubmitTime reported by AMT. Screen time we collected is also not 100% accurate as workers could open other web pages for accurate evaluation.

[3]https://serpapi.com/

when domain knowledge is necessary to answer the question. This further confirms that the aspects covered in our protocols cover key factors affecting human preferences.

4.2 Comparative Analysis of Responses from LLMs and Humans

Relative preference. In Figure 4a, we show how frequently each answer type is preferred over the other three answer candidates. It shows that model-generated answers are preferred 83.0% of the time, which aligns with the tendency reported by Xu et al. (2023). The figure also shows that formal-style answers are often chosen over more casual model-generated answers.

Absolute acceptability. While Figure 4a demonstrates that model answers are relatively preferred, it was unclear *how acceptable* they are. Figure 3 shows the distribution of the rated aspects on the model-generated and human-written answers from ELI5, conforming to the superiority of model-generated answers. While human-written top-voted and randomly sampled answers yield 1.38 and 1.15 ACCEPTABILITY scores on average, respectively, model-generated formal and casual answers yield 2.46 and 2.36. This result implies that even the annotated ratings in ELI5 may not necessarily align with users' preferences, which has also been discovered in other tasks, such as summarization (Liu et al., 2023c). Furthermore, although the average ACCEPTABILITY score of model-generated answers is higher, we found that only 54.5% of the best model-generated answers (MF) get the highest ACCEPTABILITY rating (= 3). This suggests that even state-of-the-art models' generations are not fully acceptable, and there is significant room for future improvements.

Aspect-wise analysis. Here, we focus on the distribution of scores in FORMALITY, AMOUNT INFO, and FACTUALITY displayed in Figure 3. In general, human-written answers tend to lack an adequate amount of information, exhibiting a less formal linguistic style. On the other hand, model-generated answers constantly address the appropriate amount of information with high accuracy in a manner that aligns with the desired level of formality. Together with the weights of each aspect on the overall rating revealed in Section 4.1, this explains the high preference for model-generated answers shown in Figure 4a. Unlike FORMALITY and AMOUNT INFO, the distribution of FACTUALITY is high even on model-generated answers, yielding a standard deviation of 0.75. Although the average FACTUALITY score (2.5) is higher than that of human-written answers (1.8), the LLM still has further potential for improvement in generating factually correct responses.

5 How Reliable are Automatic Metrics?

Although human evaluations are expected to be more accurate than automatic evaluation methods, scalability issues are always present. Following previous works (Wang et al., 2023a; Liu et al., 2023a), we investigate to what extent automatic evaluations can substitute human evaluations. In this section, we evaluate existing metrics and LLM-based metrics using our newly collected data. By computing Pearson's correlation coefficient ρ between scores from each metric and human ACCEPTABILITY scores, we reconfirm that ROUGE has a weak correlation with human assessment. For LLM-based evaluation, we employ both overall and multi-aspect schemes.

5.1 Classical Metrics

ROUGE (Lin, 2004) is the metric used in the original ELI5 evaluation (Fan et al., 2019). As this metric is a reference-based metric, we use top-rated human-written answers (HT) as a reference and compute the ROUGE-1and ROUGE-L for human-written randomly-sampled answers (HR), formal model-generated answers (MF), and casual model-generated answers (MC).

For reference-free metric, we use **GPT2-PPL**, which computes the perplexity of sequences using GPT2 in Huggingface (Radford et al., 2019). In the "QA" setting, we feed answers concatenated after corresponding questions, and in the "RQA" setting, we feed answers concatenated after random questions.

We also report the **Length** (Len) of the answers as one of the most superficial metrics, which is also a target for comparison in work by Xu et al. (2023) and Fabbri et al. (2021).

5.2 LLM-based metric

We conduct evaluations using instruction-tuned LLMs, namely GPT-4 (OpenAI et al., 2024) and Llama2-7B (Touvron et al., 2023), inspired by the recent success of LLM-based evaluations (Liu et al., 2023b). In the first setting, we simply prompt LLMs to predict ACCEPTABILITY (hereafter denoted as **LLM**). In the second setting, we conduct

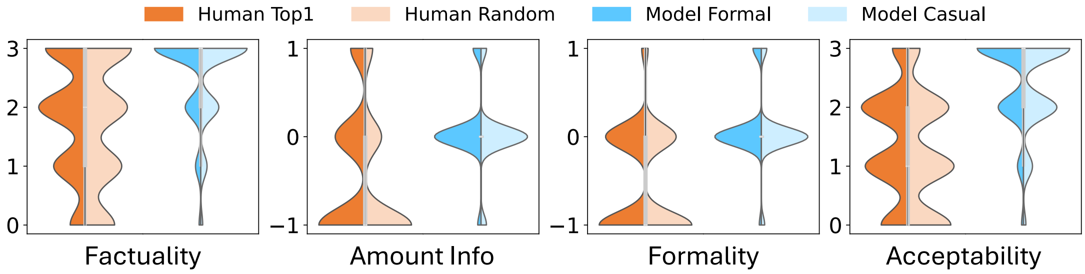
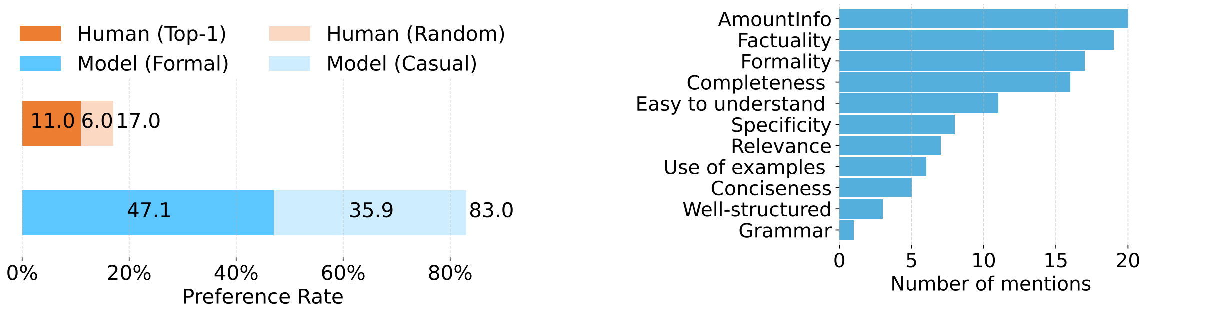


Figure 3: The distribution of the annotated scores. The ideal values are 3 for FACTUALITY and ACCEPTABILITY, and 0 for AMOUNTINFO and FORMALITY.



(a) Preference rates of four types of answer candidates.

(b) Aspects mentioned in 50 justifications of the crowd workers preference.

Figure 4: Analysis on relative human evaluation on two human-written responses and model-generated responses.

multi-aspect evaluation and then take the weighted sum to predict ACCEPTABILITY (hereafter denoted as **W-LLM**).

In W-LLM, we first have LLMs predict scores in FACTUALITY, AMOUNTINFO, and FORMALITY independently, with different prompts. Predicted ratings rescaled with Equation 1 are then fed to the linear regression model trained in Section 4.2. We finally add the ideal ACCEPTABILITY score of 3 to inverse Equation 2 and get the ACCEPTABILITY prediction.

For GPT-4-based evaluation, we feed a one-shot prompt using the instructions and an example held out from the evaluation target. The prompts are available in Appendix Table 5-10. For Llama2-based evaluation, we fine-tuned Llama2-7B using the instruction and annotation data for 192 out of 300 queries (or 768 out of 1,200 answers). Using the fine-tuned model, we evaluate the remaining data for 108 queries or 432 answers.

5.3 Results

We show the results in Table 1 and the distribution by human-annotated acceptability on Figure 5. Despite the significant performance gap between human and model-generated answers in Section 4.2, the performance gap of reference-based metrics (ROUGE) between HR, MC, and MF answers is largely limited. Together with the high deviation observed from Figure 5, we reconfirm the difficulties of comparing overall quality with those metrics. All reference-based metrics show correlations lower than 0.2, even lower than the simplest "Length" metric. GPT2-PPL displays stronger correlations with the overall acceptance. However, the marginal decrease in the correlations of the "RQA" setting from the gold "QA" setting poses a question about their reliability, and it may just overly prefer model-based outputs.

Regarding the LLM-based metrics, the correlations are higher overall than those of the classical metrics. In the LLM setting, where we only focus on the overall ACCEPTABILITY predicted by LLMs, the high correlation (0.70 and 0.72 for GPT-4 and Llama2-7B, accordingly) against human-annotated data indicates the effectiveness of LLMs as evaluators. Furthermore, evaluation using a weighted sum shows a stronger correlation, revealing the validity of evaluating long-form answers in a fine-grained manner.

5.4 Analysis

Testing robustness of GPT-4 evaluation. Since GPT-4 is a non-deterministic model even with tem-

Type	Human Acce.	ROUGE (↑) 1	L	GPT2-PPL (↓) QA	RQA	Len	LLM (↑) GPT-4	Llama2	W-LLM (↑) GPT-4	Llama2
MF	2.46	23.3	13.2	10.9	14.0	109	**2.94**	**2.72**	2.81	**2.99**
MC	2.36	**24.2**	**13.5**	**10.6**	**13.7**	107	2.91	2.55	**2.82**	**2.99**
HT	1.38	-	-	27.1	33.9	**112**	2.07	1.48	1.96	1.81
HR	1.16	21.0	12.4	31.1	41.1	88	1.67	1.26	1.55	1.54
Corr.	-	0.19	0.14	-0.59	-0.54	0.22	0.70	0.72	0.72	**0.74**

Table 1: Average scores of MF (Model-generated Formal), MC (Model-generated Casual), HT (Human-written Top-rated), and HR (Human-written Random-sampled), answers computed by each metric and the correlation with human-annotated ACCEPTABILITY (the detailed plot is provided in Figure 5). "Human Acce." column shows the average ACCEPTABILITY score annotated in this work. ROUGE-based scores are computed using HT answers as a reference.

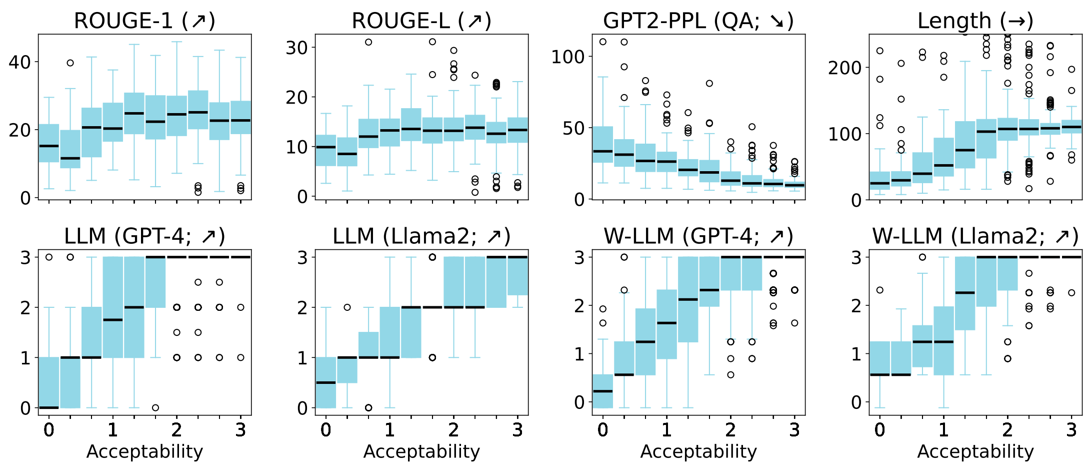


Figure 5: The distribution of automatic evaluation metrics based on annotated human acceptability. The Acceptability axis shows the average ACCEPTABILITY score of three annotators. The arrow in each sub-graph's titles shows the graph's ideal trend.

perature zero, we conduct further experiments regarding the robustness of the evaluation results. Specifically, we ran two additional iterations of evaluations by GPT-4 against 20% of the dataset, just like we collected annotations from three workers. The correlation of averaged scores from three iterations of evaluation with human annotation was 0.73 on the LLM setting and 0.74 on the W-LLM setting. On both the single-iteration and multiple-iteration settings, the correlation on the W-LLM setting exceeded the LLM setting, indicating reproducibility of the conclusion that weighted sum better correlates with human evaluation than just having the model evaluate the overall acceptability.

LLM-based metric on previously collected datasets. For robust conclusion, we inspect if the W-LLM setting also yields better alignment than the vanilla LLM setting on previously collected human evaluation data by Nakano et al. (2021) and Krishna et al. (2021). Because their data is collected in a pair-wise manner, we use the accuracy of preference to evaluate the metrics. The accuracy of LLM and W-LLM using GPT4 was 0.33 and 0.53, respectively, on Nakano et al. (2021)'s dataset and 0.53 and 0.64 on Krishna et al. (2021)'s dataset respectively.

Per-aspect errors. For fine-grained multi-aspect evaluation, the key is to have an accurate, fine-grained scoring system for each aspect. The mean error of GPT-4 was +0.24, +0.04, and -0.05 for FACTUALITY, AMOUNT INFO, and FORMALITY, respectively, and +0.43, -0.02, and -0.02 for Llama2-7B. Both models show larger errors in factuality, suggesting that there is still room for improvements in evaluating factuality.

Equally weighted evaluation. When equally weighted, GPT-4 showed a 1-point increase in correlation compared to W-LLM, while fine-tuned Llama2 showed a 1-point decrease compared to W-LLM. The accuracy on the Preference-Based datasets of Krishna et al. and Nakano et al. decreased by 3 to 5 points when compared to W-LLM. While W-LLM aligns better with human judgments overall, we attribute the small difference between simple averaging and W-LLM to errors in per-aspect evaluation.

6 Conclusion

This work revisits multi-aspect evaluations of long-form generations and investigates which factors affect the overall ratings. Our quantitative analyses reveal that the effect of FACTUALITY is 2.5 times greater than that of AMOUNTINFO. We also show that surface-level quality measured by FORMALITY has less than half the influence of AMOUNTINFO on ACCEPTABILITY.

In the process, we collect human absolute evaluations on 1.2k responses to information-seeking queries in FACTUALITY, AMOUNTINFO, FORMALITY, and ACCEPTABILITY. Along with relative preference and free-text justification comments collected for deeper analyses, we publish the human evaluation data for future development in evaluation systems for long-form generations.

As a first step in using our newly collected data for a reliable automatic evaluation, we reassess existing automatic evaluation methods and LLM-based evaluation. We show that while classical methods yield a weak correlation with human assessment scores, LLM-based methods have a strong correlation. Informed by the degree of importance of fine-grained aspects on overall rating our analyses on human evaluations revealed, we show that taking a weighted sum of LLM evaluations along multiple fine-grained aspects yields a stronger correlation with human evaluations.

Limitations

Our evaluation protocol uses three aspects: FACTUALITY, AMOUNT INFO (amount of information), FORMALITY, and ACCEPTABILITY. Although we acknowledge there are many more aspects to evaluate long-form generation, we found adding a large number of fine-grained aspects often confuses human evaluators and increases the costs of automatic evaluations. Our qualitative analysis of explanations reveals that the key aspects included in our evaluations play key roles when annotators choose the best long-form responses to information-seeking queries. Future work can further explore diverse aspects of long-form generations. While we carefully design and conduct human evaluations, assessing FACTUALITY is challenging even for experienced annotators, and there may be more disagreements on this aspect.

W-LLM, our new automatic evaluation scheme, displayed a high correlation with the annotation data we collected. However, as mentioned at the end of Section 5, their score predictions show relatively weaker correlations in FACTUALITY, suggesting further improvement can enhance the reliability of this metric.

Using GPT-4 as an evaluator in a non-deterministic method can result in the sensitivity to prompt variation and also a lack of the reproducibility of results (Chen et al., 2023b; Asai et al., 2023). To overcome this issue, we fine-tune Llama2-7B on the collected dataset. W-LLM using this in-house model has demonstrated even higher correlations with human evaluations. Feeding GPT-4 the three snippets of the websites that we showed to the annotators resulted in the decline of correlation against human annotated FACTUALITY, presumably due to excessive focus on the snippets with only a small amount of information.

While the correlations on both LLMs in W-LLM settings are high, over-estimations are seen, especially for answers generated by ChatGPT. Considering the distribution of human annotation displayed in Figure 3 and the weights of the linear regressor mentioned in Section 4.1, it is assumed that more effort to evaluate factuality accurately needs to be put into automatic evaluations. In addition, since the finetuning of Llama2 was done in a simple manner, it may also be influential to control the bias in the fine-tuning data.

Ethics

Our data curation process involves crowdsourcing and anonymization of personal information reported by Amazon Mechanical Turk, including Crowdworker IDs. We did not collect any private information. In making our data publicly available, it's essential to acknowledge the potential ethical aspects of this release. We discuss how our method can be applied to long-form QA evaluations, as well as wider applications. As our main focus is on evaluating models, we believe this work does not directly cause harm. However, relying on models to evaluate other models could introduce biases and should be considered as a broader issue for the llm-as-judge approach.

Acknowledgemets

This work was supported by JSPS KAKENHI Grant Number JP21K21343, JST CREST Grant Number JPMJCR20D2, and the IBM Fellowship. Additionally, computational resources provided by Stability AI were used for model finetuning. We would like to thank the members of TohokuNLP for their frequent participation in discussions throughout this research.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, Toronto, Canada. Association for Computational Linguistics.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2023a. Understanding retrieval augmentation for long-form question answering. *arXiv preprint arXiv:2310.12150*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023b. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

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

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023d. On learning to summarize with large language models as references. *arXiv preprint arXiv:2305.14239*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany

Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. Proxyqa: An alternative framework for evaluating long-form text generation with large language models. *arXiv preprint arXiv:2401.15042*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *CoRR*, abs/2306.01693.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

A Annotation Setup

Qualification Human evaluations on Amazon Mechanical Turk were conducted by workers who passed our qualification task and are native to English. We conducted qualification rounds where we curated a set of six explanations, although with different (but similar) criteria as it was for a different project. We manually tagged answers we deemed acceptable and filtered the workers based on their level of alignment (match %). We also manually filtered workers after the preliminary experiment of the task for this paper based on their ratings and justifications. The initial number of participants was 700 workers (all with ≥99% HIT approval rate and ≥5,000 HITs completed), which was reduced to 201 workers after filtering.

B More analysis on annotated results

Agreement on preference. Figure 6 shows the distribution of preference of crowd workers. In 69 out of 300 instances, all three workers agreed on the same candidate's answer in terms of preference. The three workers all preferred model-generated formal answers in 45 instances and model-generated casual answers in 19 instances.

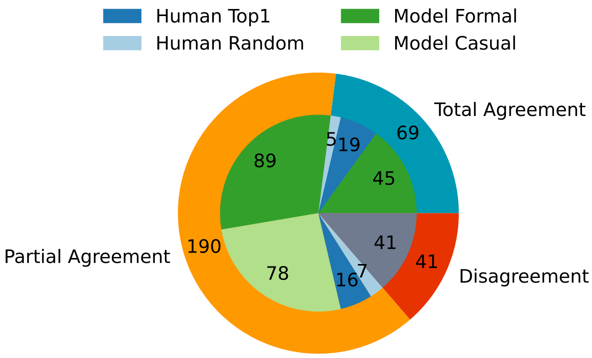


Figure 6: The agreement of preference. "Total Agreement" means all 3 annotators agreed on the preference, and "Disagreement" means 3 workers preferred different answers. The inner pie chart represents the type of answer preferred by the majority of the workers.

Aspect-wise Agreement. The reported Krippendorff's alpha on our paper is computed over all aspects at once using the Python library Krippendorff. Table 2 reports aspect-wise alpha values. The values are anticipated to be small due to the small number of rating options (only -1, 0, 1 for Amount Info and Formality and 0, 1, 2, 3 for Factuality and Acceptability) and the annotation method where more than 3 workers are involved in the human evaluation instead of three fixed annotators. To provide more information, we also report the aspect-wise agreement rate in Table 3 following Figure 8. In over 80% of the instances, a majority vote was taken.

Aspect	Agreement
Factuality	0.31
Amount Info	0.50
Formality	0.37
Acceptability	0.48

Table 2: Aspect-wise Krippendorff's α

Preference rate of model-generated answers. When compared to Human vs. Machine comparison in the work of (Xu et al., 2023)(61.8%), our study demonstrates an even higher preference for model-generated answers. We attribute this difference to the use of different backbone models (i.e., ChatGPT vs. davinci-002) and our four-way choice setup, whereas their evaluations rely on pairwise comparisons.

Aspec	Total Agreement	Partial Agreement	Disagreement
Factuality	266 (22%)	701(58%)	233(19%)
Amount Info	631(53%)	521(43%)	48(4%)
Formality	565(47%)	607(51%)	28(2%)
Acceptability	253(21%)	762(64%)	185(15%)

Table 3: Aspect-wise agreement

Relationship against Acceptability Figure 7 reports the relationship against ACCEPTABILITY. Figure 8 is a decision tree with a max depth of 3. In the figure, we present the "Feature Importance" indicating the effect of each aspect on the prediction, which was computed during the training phase. The tree is fitted to 80% of the annotation data and yields a Pearson correlation coefficient ρ of 0.830 on the remaining 20% of the data, proving the method to be reasonable.

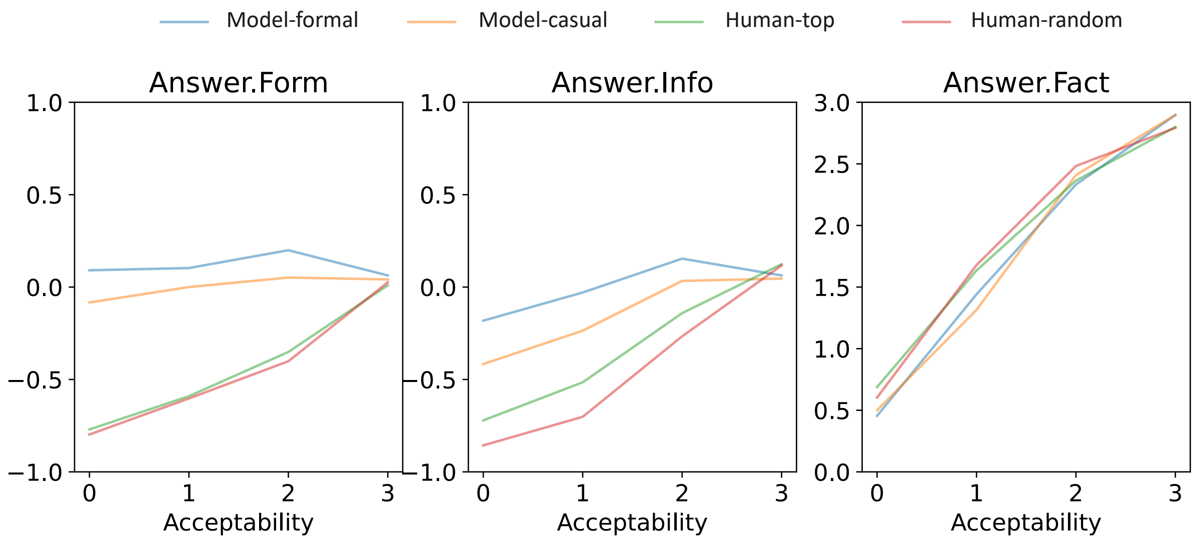


Figure 7: Average scores of each aspect by Acceptability.

C Definitions of aspects mentioned in free-form justification.

4 shows the rough description and keywords we used to annotate free-form justification in Figure 4b

D Annotation Interface

We show the screenshot of the annotation interface in Figures 9-11. We also show a screenshot of the browser window in Figure 12 to display some functions devised for productive annotation.

E Prompts

Prompts to generate long-form answers by ChatGPT are on Table 5 and Table 6. Prompts to evaluate long-form answers in the four aspects are on Tables 7-10.

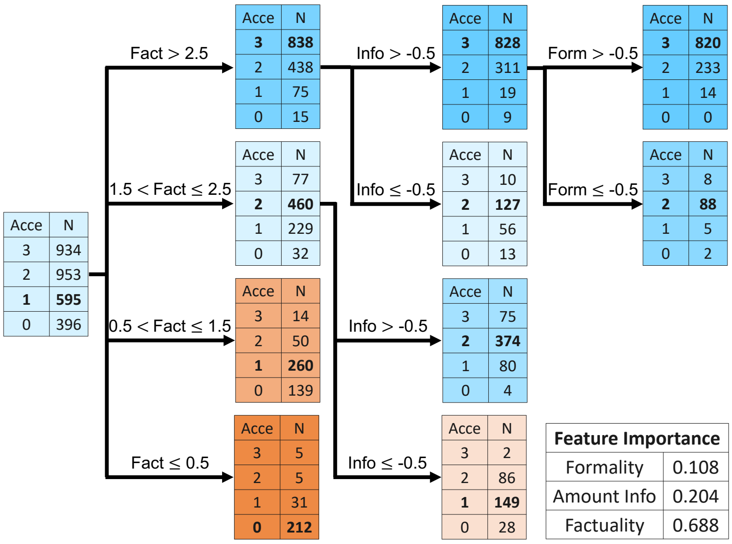


Figure 8: Decision tree trained to predict Acceptability from FACTUALITY, AMOUNT INFO, and FORMALITY scores. Column "N" in each node displays the number of training instances with Acceptability in the "Acce" column.

Aspect	Description	Keywords
Amount Info	Corresponds with the definition we show to crowd workers.	"right amount of information", "too much information", "lack information"
Formality	Corresponds with the definition we show to crowd workers.	"foamality", "formal", "casual"
Factuality	Corresponds with the definition we show to crowd workers.	"accurate", "factual"
Easy to Understand	Following the instruction in the Reddit forum "Explain Like I'm Five", we notified the annotators that answers should be "appropriate for people who are not professionals" in the area.	"easily understandable", "non-professionals"
Relevance	Mentions relevance/irrelevance of information addressed in the answer to the question.	"actually answer the question", "related information"
Well-structured	Mentioned of the quality in answer structure.	"well structured", "logical progression"
Completeness	Mentions that the answer contains necessary information or lacks information.	"completeness"
Grammar	Mentions grammatical correctness/incorrectness.	"grammar", "spelling"
Use of Examples	Mentions the use of example or demand for the use of example.	"incorporate examples"
Specificity	Mention of how specific or detailed the answer is.	"specific", "detail"
Conciseness	Compactness while addressing a sufficient amount of information.	"verbose", "not overly technical"

Table 4: Description of aspects we used to annotate free-form justification.

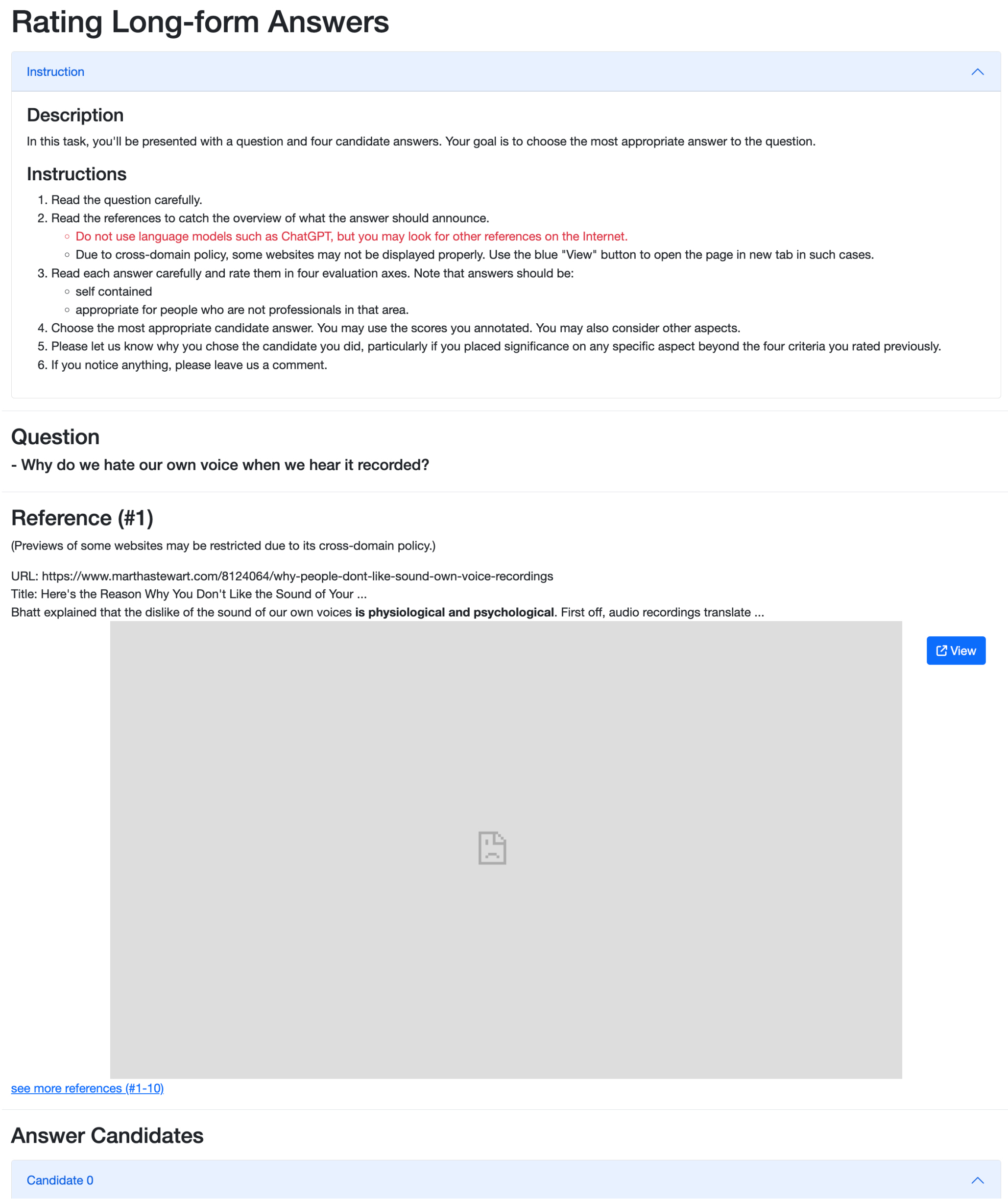


Figure 9: The first portion of the interface. Here, we provide instructions for annotation, a question, and one reference. We provide 9 more references at the bottom of the page. For each reference, we provide the URL, title, and a snippet of relevant text. We use iframe for annotators to view websites, but some websites do not allow this function. The blue “View” button allows crowd workers to open the website in a new tab. After

system	You are a helpful assistant who answers questions on a forum.
user	Answer the following question in 75-100 words: {question}

Table 5: Prompts for generating “model-formal” answers.

Answer Candidates

Candidate 0

Answer

We perceive our voice differently when we hear it on a recording because we are not used to hearing it from that perspective. When we speak, we hear our voice through the bones in our skull as well as through the air, which creates a richer, deeper tone. However, when we hear a recording of our voice, we only hear it through the air, resulting in a higher-pitched sound that we may not like. Additionally, some people may be self-conscious about their voice because they associate it with how they look or their image of themselves.

Evaluation

Factuality (0~3) Select score
Amount Info (-1~1) Select score
Formality (-1~1) Select score
Acceptability (0~3) Select score

Close

Candidate 1

Answer

When you hear yourself as you speak, you're also hearing the vibrations echoing around in your skull. That changes the way your voice sounds, compared to recording it. You're just not used to hearing yourself sound so different.

Evaluation

Factuality (0~3) Select score
Amount Info (-1~1) Select score
Formality (-1~1) Select score
Acceptability (0~3) Select score

Close

Candidate 2

Answer

Have you ever listened to a recording of your own voice and cringed at the sound of it? Well, you're not alone. The reason why we hate the sound of our own voice is because when we talk, the soundwaves from our voice travel through our skull and jawbone, creating vibrations that add depth and resonance to the sound. However, when we hear a recording of ourselves, we don't get the added depth and resonance, making our voice sound different and unfamiliar. Additionally, we are used to hearing our own voice from inside our head, so when we hear it recorded, it can sound strange and uncomfortable. So, don't worry, it's a completely normal phenomenon!

Evaluation

Factuality (0~3) Select score
Amount Info (-1~1) Select score
Formality (-1~1) Select score
Acceptability (0~3) Select score

Close

Candidate 3

Answer

When you hear your voice normally, you hear a sound transmitted through the air just like everyone else hears... but you also hear some of the sound transmitted through your jaw and skull. Since bone transits sound very differently to air, the way you hear your own voice is different to how everyone else hears it.

When your voice is recoded and played back, you hear the sound as everyone else does, transmitted through air, and not through bone.

I doubt that *everyone* **hates** their voice when it's recorded, but they certainly find it strange. And if you do hate your voice, then it's only because of the strangeness - because it sounds different to the way you normally hear your voice.

Evaluation

Factuality (0~3) Select score
Amount Info (-1~1) Select score
Formality (-1~1) Select score
Acceptability (0~3) Select score

Close

Preference

Please select the candidate you prefer.

☐ No candidates are appropriate, but if I have to choose one ...
○ Candidate 0
○ Candidate 1
○ Candidate 2
○ Candidate 3

About your preference

Please let us know why you chose the candidate you did, particularly if you placed significance on any specific aspect beyond the four criteria you rated previously.

Comments

If you notice anything, please let us know.

Figure 10: Here, we show a total of four answers. We ask annotators to evaluate each answer in four aspects. Then, we ask them to select one most preferred answer candidate. The option "No candidates are appropriate, but if I have to choose one..." is also provided. Finally, we ask the crowd workers to justify their preference in free-text format. All fields here are required fields.

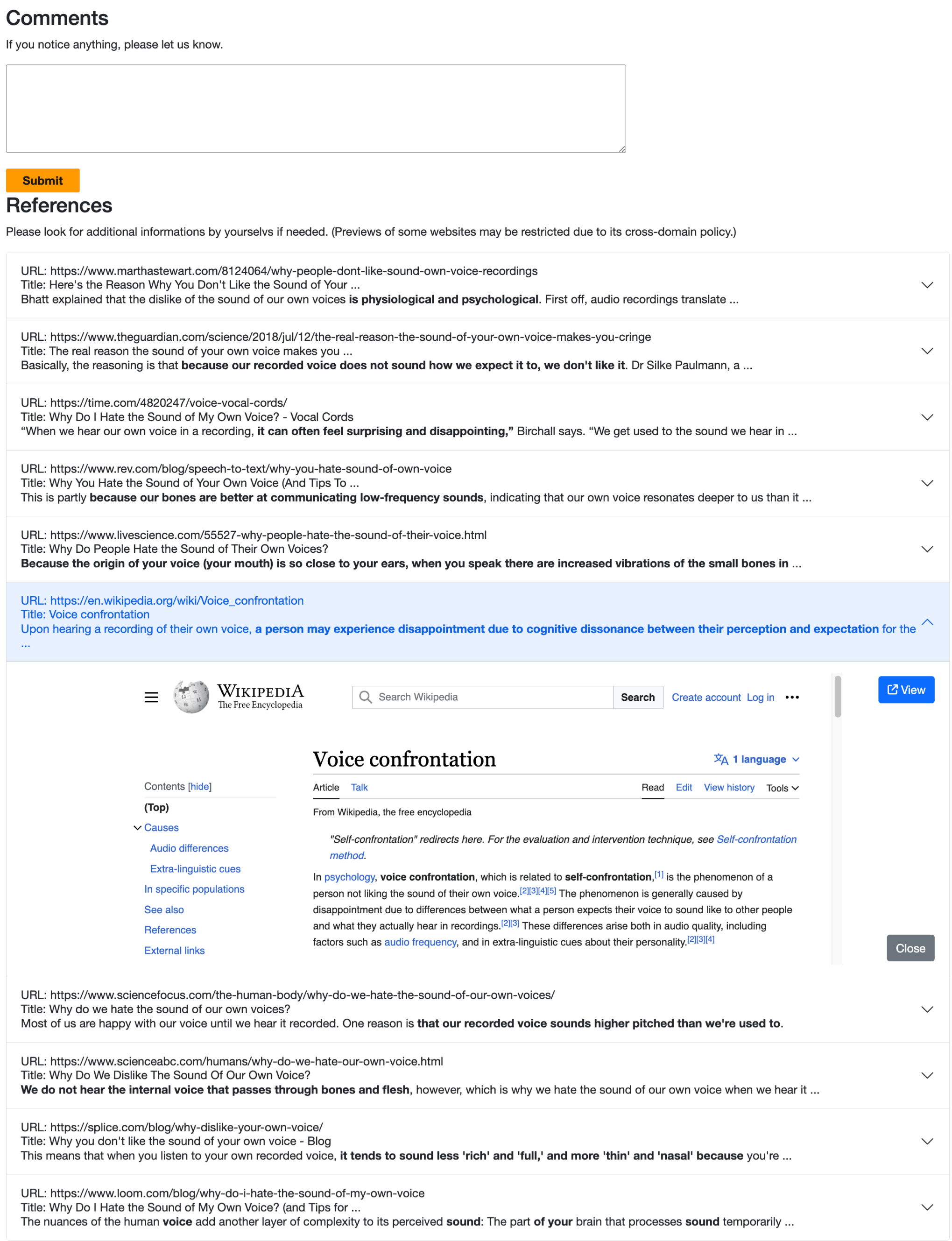


Figure 11: Here, we collect comments (not required) from the workers to catch bugs or any unexpected behaviors of the page. At the end of the page are ten references, each with their URL, title, and snippet. Each reference can be expanded or collapsed. The Wikipedia page is expanded in the figure as an example.

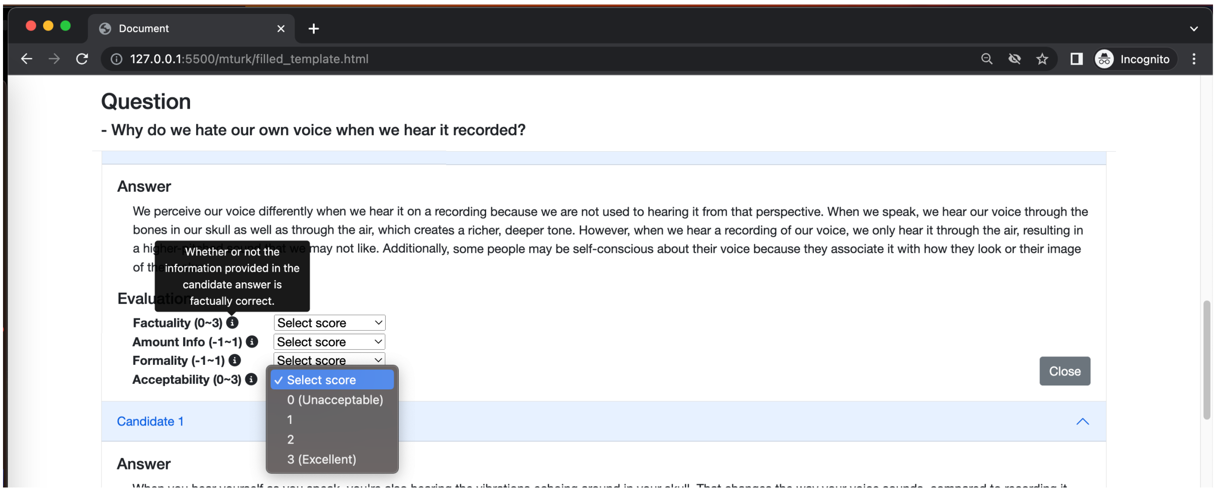


Figure 12: On the browser, the question is fixed to the top of the page to reduce burden of scrolling. The definition of each aspect pops up when the cursor is on the "i" icon. The meaning of the scores is given in the select options.

system	You are a helpful assistant who answers questions on a forum.
user	Instruction: Answer the following question in 75-100 words. Requirements: The answer should not use difficult vocabularies. The answer should be understandable to people outside the field. The answer should be in a little bit casual. Question: {question}

Table 6: Chat-GPT prompts for generating "model-formal" answers.

user	You will receive an instruction, a question, and an answer. Your task is to evaluate the answer in Formality (the answer's formal appropriateness including its vocabulary, grammar, and spelling.). Use the following scale to generate a score: -1: Too Casual - The answer is excessively informal or casual in tone, language, or style. It may use slang, colloquialisms, or inappropriate language for the context. 0: Moderate Formality - The answer strikes a balance between formality and informality, using appropriate language and tone for the given context. It maintains a neutral or professional tone without being overly formal or casual. 1: Too Formal - The answer is excessively formal or rigid in tone, language, or style. It may use overly complex or technical language, unnecessary jargon, or a very formal writing style that may not be suitable for the context. ### Instruction: Rate the acceptability of the following answer to the following question. Question: How do WIFI waves travel through walls? Answer: Quantum physics teaches us that electromagnetic waves interact with matter. Visible light is an electromagnetic wave. Light interacts with matter in a "certain" way. If there is nothing to interact with is passes through. an examples is glass. You can see through glass. but for example if you have hold a piece of Paper in front of your eyes, you can't see through it. But if somebody is pointing a bright flashlight to your face, you can see the light through the paper. Your WiFi has a certain Wavelength. For this wave length it can interact with matter a completely different way. So for your WiFi-Waves a wall just appears as "nothing" as glass does for us. It still interacts and blocks some of the waves but your WiFi is bright enough to still be visible through it. Formality: -1 Explanation: The answer contains some grammatical errors, and is in a casual form. ### Instruction: Rate the acceptability of the following answer to the following question. Question: {question} Answer: {answer} Formality:

Table 7: GPT-4 prompts for evaluating FORMALITY.

user	You will receive an instruction, a question, and an answer. Your task is to evaluate the answer in Amount of Information (whether or not sufficient amount of information was addressed to fully answer the question.). Use the following scale to generate a score: -1: Insufficient Information - The answer lacks necessary details or fails to address the question adequately. It provides too little information or is incomplete. 0: Moderate Information - The answer provides a balanced amount of information, adequately addressing the question without excessive or insufficient details. 1: Excessive Information - The answer contains an overload of information that may not be directly relevant to the question. It includes unnecessary details or goes beyond what is needed. ### Instruction: Rate the acceptability of the following answer to the following question. Question: How do WIFI waves travel through walls? Answer: Quantum physics teaches us that electromagnetic waves interact with matter. Visible light is an electromagnetic wave. Light interacts with matter in a "certain" way. If there is nothing to interact with is passes through. an examples is glass. You can see through glass. but for example if you have hold a piece of Paper in front of your eyes, you can't see through it. But if somebody is pointing a bright flashlight to your face, you can see the light through the paper. Your WiFi has a certain Wavelength. For this wave length it can interact with matter a completely different way. So for your WiFi-Waves a wall just appears as "nothing" as glass does for us. It still interacts and blocks some of the waves but your WiFi is bright enough to still be visible through it. Amount of Information: 0 Explanation: The answer contains a sufficient amount of information. Though a little more information might make the answer better, it addresses key points. ### Instruction: Rate the acceptability of the following answer to the following question. Question: {question} Answer: {answer} Amount of Information:

Table 8: GPT-4 prompts for evaluating AMOUNT INFO.

user	You will receive an instruction, a question, and an answer. Your task is to evaluate the answer in Factuality (whether or not the information provided in the answer is factually correct). Use the following scale to generate a score: 0: Completely Inaccurate - The answer is entirely false or contains significant factual errors. It contradicts well-established information or provides misleading information. 1: Partially Accurate - The answer has some correct information but also contains significant inaccuracies or lacks important details. It requires additional verification or correction. 2: Mostly Accurate - The answer is predominantly correct and provides relevant information. However, it may still have minor inaccuracies or omissions that do not significantly impact the overall accuracy. 3: Entirely Accurate - The answer is entirely accurate and factually correct. It aligns with well-established information, provides clear and precise details, and can be relied upon as a trusted source of information. ### Instruction: Rate the factuality of the following answer to the following question. Question: How do WIFI waves travel through walls? Answer: Quantum physics teaches us that electromagnetic waves interact with matter. Visible light is an electromagnetic wave. Light interacts with matter in a "certain" way. If there is nothing to interact with is passes through. an examples is glass. You can see through glass. but for example if you have hold a piece of Paper in front of your eyes, you can't see through it. But if somebody is pointing a bright flashlight to your face, you can see the light through the paper. Your WiFi has a certain Wavelength. For this wave length it can interact with matter a completely different way. So for your WiFi-Waves a wall just appears as "nothing" as glass does for us. It still interacts and blocks some of the waves but your WiFi is bright enough to still be visible through it. Factuality: 3 Explanation: The facts and common sense addressed in the answer are accurate. ### Instruction: Rate the factuality of the following answer to the following question. Question: {question} Answer: {answer} Factuality:

Table 9: GPT-4 prompts for evaluating FACTUALITY.

user	You will receive an instruction, a question, and an answer. Your task is to evaluate if the answer is overall acceptable. Use the following scale to generate a score: 0: Completely Unacceptable - The answer is incorrect, irrelevant, or nonsensical. It provides no useful information or is entirely false. 1: Partially Acceptable - The answer contains some relevant information but is incomplete, unclear, or contains errors. It may require further clarification or refinement. 2: Mostly Acceptable - The answer is largely correct and provides relevant information. It may have minor inaccuracies or could be improved, but it is generally satisfactory. 3: Fully Acceptable - The answer is accurate, comprehensive, and well-explained. It provides all the necessary information and addresses the question thoroughly. ### Instruction: Rate the acceptability of the following answer to the following question. Question: How do WIFI waves travel through walls? Answer: Quantum physics teaches us that electromagnetic waves interact with matter. Visible light is an electromagnetic wave. Light interacts with matter in a "certain" way. If there is nothing to interact with is passes through. an examples is glass. You can see through glass. but for example if you have hold a piece of Paper in front of your eyes, you can't see through it. But if somebody is pointing a bright flashlight to your face, you can see the light through the paper. Your WiFi has a certain Wavelength. For this wave length it can interact with matter a completely different way. So for your WiFi-Waves a wall just appears as "nothing" as glass does for us. It still interacts and blocks some of the waves but your WiFi is bright enough to still be visible through it. Score: 2 Explanation: The answer provides an easy-to-understand explanation with examples that did not feel too technical. It was easy to read, and left me knowing how it works. ### Instruction: Rate the acceptability of the following answer to the following question. Question: {question} Answer: {answer} Score:

Table 10: GPT-4 prompts for evaluating overall ACCEPTABILITY.