

Which Method(s) to Pick when Evaluating Large Language Models with Humans? – A comparison of 6 methods.

Birgit Popp and Sarah Keck and Androniki Mertsiotaki and Emily Kratsch and Alexander Daum

Fraunhofer IIS / Germany

{birgit.popp, sarah.keck, androniki.mertsiotaki, emily.kratsch, alexander.daum}@iis.fraunhofer.de

Abstract

Human evaluations are considered the gold standard for assessing the quality of NLP systems, including large language models (LLM), yet there is little research on how different evaluation methods impact results. This study compares six commonly used evaluation methods – four quantitative (Direct Quality Estimation, Best-Worst Scaling, AB Testing, Agreement with Quality Criterion) and two qualitative (spoken and written feedback) – to examine their influence on ranking texts generated by four LLMs. We found that while GPT-4 was consistently ranked as the top-performing model across methods, the ranking of other models varied considerably. In addition, methods differed in their cost-effectiveness, with Direct Quality Estimation emerging as the most efficient. Qualitative methods provided insights beyond quantitative methods, especially spoken feedback in moderated sessions. Participants reported challenges with task comprehension and evaluation interfaces. Our findings highlight that the choice of evaluation method and the implementation of the method can influence results, affecting both the validity and interpretability of human assessments. These findings suggest a need for methodological guidelines in human-centered evaluations of LLMs to improve reliability and reproducibility in NLP research.

1 Introduction

Human evaluations are considered the gold standard in evaluating natural language processing (NLP) systems (van der Lee et al., 2019; Belz et al., 2020; Ruan et al., 2024). As van der Lee et al. (2019) highlight in a review of NLP literature, automatic metrics are uninterpretable. In addition, automatic metrics may not correlate with human evaluations (Liu et al., 2016). Moreover, issues such as data leaks of benchmarks into training data of language models necessitate the use of human

evaluations to ensure the integrity and authenticity of evaluation results (Xu et al., 2024).

At the same time, flaws and confusion are common when conducting human evaluations of NLP systems (Howcroft et al., 2020; Thomson et al., 2024). Such flaws and confusion can reduce reproducibility, validity and reliability of human experiments. One possible origin of reproducibility issues is that there is not one standard method but multiple methods that can and are being used (Howcroft et al., 2020; Belz et al., 2020). However, there is – to our knowledge – no published research investigating how the choice of method may affect evaluation results. Thus, we conduct a comparative study in which we compare five different methods to assess the honesty (Gao et al., 2024; Yang et al., 2024) and comprehensibility (Leschanowsky et al., 2025) of generated texts, hence we focus on generative models, such as large language models (LLM).

In our study we investigate the following four research questions:

1. What is the average cost of conducting an experiment using each evaluation method?
2. Do different evaluation methods render the same results?
3. Do methods generate results that align with a gold-standard human response based on semantic similarity?
4. What challenges did participants encounter with the different evaluation methods?

2 Methods

2.1 Study Design

With our present study we aim to assess whether the method used for evaluation affects evaluation results, i.e. does *how* we measure affect *what* we

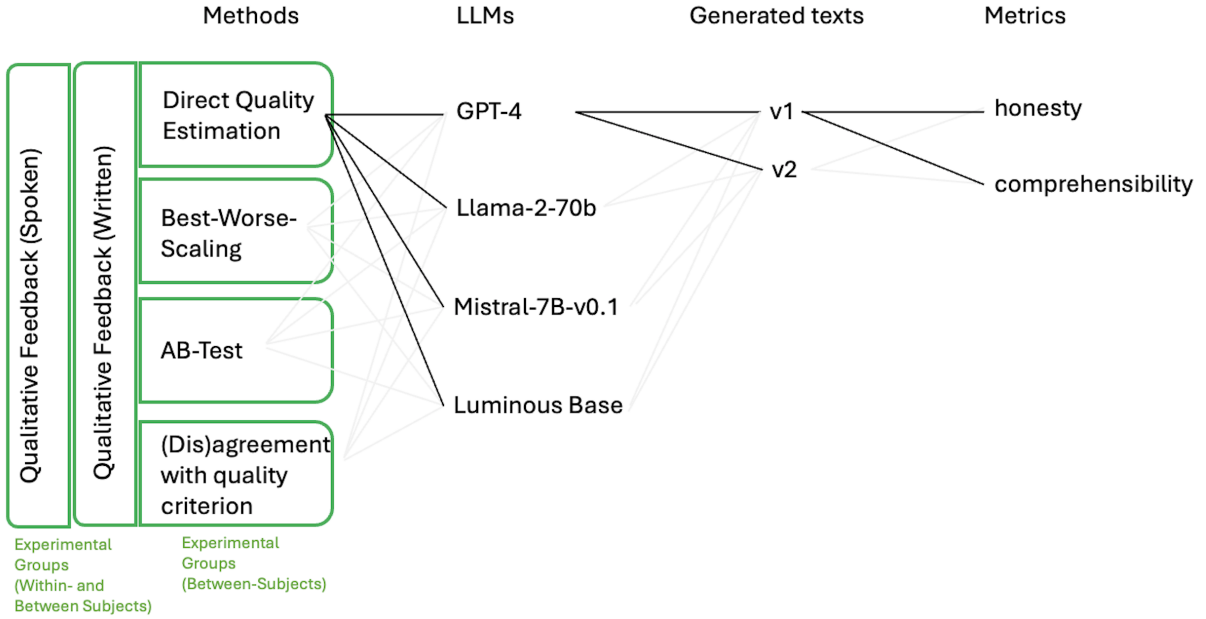


Figure 1: Study design for our multi-method comparison of human evaluation methods.

measure. To address this question, we kept evaluation content (LLM and their generated texts) as well as metrics the same across groups and compared results obtained by different methods. Figure 1 illustrates the study design.

We selected six evaluation methods that are commonly used in the field of natural language processing (NLP) (Howcroft et al., 2020) and that cover both quantitative (4 methods) and qualitative (2 methods) approaches.

The chosen quantitative methods are:

- Direct Quality Estimation (DQE),
- Best-Worst Scaling (BWS),
- AB-Testing (AB),
- Agreement with Quality Criterion (AQC).

The quantitative methods are compared in a between-subject design, whereas the qualitative methods are compared in a within-subject design. This is due to us using qualitative methods to gain deeper insights into the quantitative methods and as stand-alone methods that are to be evaluated in their own right. In their first role, we need to apply them alongside each quantitative method. The chosen qualitative methods are:

- Spoken qualitative feedback, i.e. interviews,
- Written qualitative feedback, i.e. open-text fields.

We selected four LLMs to generate texts that we planned to evaluate. We selected the four LLMs based on mean win-rates in the leaderboard (Classic) of Holistic Evaluation of Language Models

(HELM) obtained on March 7th 2024. Thus, we used this HELM leaderboard as a predictor for general text generation quality. We aimed to cover a wide range of text generation quality and thus selected the LLM with the highest win-rate, the LLM with the lowest win-rate and two LLMs that covered the mid-range of win-rates. As we selected for a wide range of text generation quality, we expected to find differences when evaluating the selected models. The ability of methods to detect differences between models can be a criteria when selecting which method to employ for evaluation. Table 1 shows the selected models, their mean win rates and ranks on the date of selection. Note that win-rates and ranks dynamically change as models evolve and may currently differ from the data reported from March 2024.

Table 1: Comparison of LLM models based on mean win-rate and their rank on the HELM Leaderboard (Classic) as of March 7th, 2024.

LLM Model	Mean Win-Rate	Rank on HELM Leaderboard
GPT-4	0.97	#1
Llama-2-70B	0.67	#11
Mistral-7B-v0.1	0.45	#19
Luminous Base (13B)	0.05	#31

We selected two constructs, namely honesty and comprehensibility, to evaluate LLM-generated texts. We chose those two constructs, as they are both relevant for human evaluation of NLP and they

are distinct and thus cover different aspects of quality (Howcroft et al., 2020; Belz et al., 2020, 2024). There is no standard definition of honesty and comprehensibility and we expect both constructs to be made up of multiple sub-constructs. Taking this into account we selected four sub-constructs for each of the two constructs. All metrics are shown in Table 2. The definitions of metrics as presented to evaluators are shown in Appendix A.

Table 2: This table shows the quality criteria used in our present study. Appendix A lists the definitions of each quality criterion. We assume a hierarchical structure, with honesty and comprehensibility being the overarching criteria that encompass the listed sub-criteria.

Honesty	Comprehensibility
Correctness	Fluency
Context adherence	Non-redundancy
Relevancy	Understandability
Completeness	Readability

Our focus in this work is methodological, that is, we want to primarily assess methods rather than models or generated texts, in order to gain a better understanding about methods. However, for the purpose of assessing methods, we have to generate content for evaluation. We decided to focus on one use case for content generation to minimize experimental conditions, given that the complex study design (see Figure 1). As use case we selected a simulated Virtual Workplace Assistant (VWA). We did not implement a VWA, but we simulated one by selecting a text excerpt from an intranet page of our organisation and providing it as context to the selected LLMs, thus mimicking a Retrieval Augmented Generation (RAG) system. RAG systems, including VWA, have become broadly adopted across industry and research and thus we consider this use case as relevant. Moreover, as task we selected dialogue turn generation, which is the second most common NLP task according to Howcroft et al. (2020).

2.2 Generation and Selection of Texts

To select the reference text prior to the generation of texts with the selected models, a structured approach was followed. Thus, text excerpts were required to meet several criteria: (1) They had to be written in English, (2) they had to be concise enough to fit within a single prompt across all selected LLMs given the restraint of prompt length, and (3) they had to be sourced from an intranet page in our organisation. The selected text

excerpts, as well as the prompts used for text generation are provided in Appendix B. To create test items for the study, we generated a diverse range of outputs. Two prompts, which were divergent in form but nearly synonymous at the semantic level, were used to account for prompt sensitivity of large language models, a phenomenon where seemingly minor changes in prompt phrasing can lead to differences in outputs and model performance - highlighting the importance of evaluating against a pool of prompts (Mizrahi et al., 2024). Each LLM produced ten responses for both prompts, resulting in a sample of 80 outputs reflecting user interactions. Freely available playground environments were used for text generation (TogetherAI, 2024; Meta and Replicate, 2024; MistralAI and Replicate, 2024; OpenAI, 2024; AlephAlpha, 2024). Slight adjustments to parameters such as *max_new_tokens* and *repetition_penalty* were made, simulating what an informed user might do to enhance the quality of their responses. For each model, the best and worst outputs were selected to represent the spectrum of performance. High-quality outputs were chosen based on (1) the model’s ability to answer the prompt question, (2) the relevancy of additive information, (3) the factual accuracy of the content, and (4) overall readability of the response. In contrast, low-quality outputs were identified using cosine-similarity scores to assess semantic similarity of every output compared to the reference answer. We computed the semantic similarity scores via sentenceBERT by establishing a “reference answer” against which the LLM-generated texts could be measured. We chose the reference answer and it was a subset of the original text, containing its core information. The prompts for text generation, the reference answer and the generated texts can be found in Appendix B.

2.3 English Level Aptitude Test

In order to determine the participants’ eligibility for taking part in the pilot study, we administered an online language aptitude test with LimeSurvey prior to the experiment. We communicated this requirement to participants in the study invitations via E-mail. The test design, including test questions and scoring was adopted from Cambridge English (Cambridge University Press Assessment), which assigns language proficiency levels according to the Common European Framework of Reference for Languages. The test consisted of 25 questions, with one point given per correct answer (Appendix

C). For scores of 17 or below, a B1 qualification was recommended. The eligibility threshold was set to B2 or higher. This was implemented to ensure that participants possess the required linguistic competence to comprehend the text excerpts' content and therefore facilitating their engagement in the evaluation tasks.

2.4 Data Collection

We collected part of the data using LimeSurvey, an open-source software hosted on a server within our organisation. Our team selected LimeSurvey for its ability to support custom adaptations through PHP and JavaScript. We developed separate questionnaires for each method—AB, AQC, DQE, and BWS—and randomly assigned participants to one of these methods. We created the questionnaires through an iterative process, conducting multiple rounds of testing.

Before starting the survey, participants were required to agree to a data declaration in accordance with GDPR. A disclaimer followed, explaining that the quality of the selected texts might not meet participants' expectations. This clarification was necessary to address confusion that was observed during internal testing about the expected text quality.

On each evaluation page we included background information about the method used for prompting the LLM (Figure 2).

The presentation of generated texts varied depending on the method:

- AB involved pairwise comparisons of all texts against every other, resulting in 28 combinations that were evaluated for all metrics (Appendix F).
- AQC and DQE evaluated each text individually across all 10 metrics (Appendix G).
- BWS was implemented as an incomplete block design, with four texts per block and 14 text blocks each for honesty and comprehensibility (Appendix H).

Both, the order of texts and questions was randomized in order to minimize order bias, except for the question-order in AQC due to a technical error. Answering the metric-related questions was mandatory. Each page included a voluntary open-text field for feedback on the survey design, decision-making process or other comments. Demographic

data were collected at the end of the survey. Data was anonymized by using an individual participant code for matching the interview with the survey.

The evaluation took place in reserved office spaces under our supervision, which allowed us to address any incoming clarification questions the participants had regarding the methodology or technical terminology. Participants used a HP Elite-Book 14 inch laptop device connected to a larger 55 inch screen, enabling us to observe the evaluation task. For further qualitative data, we conducted and recorded semi-structured interviews after the evaluation, aiming to investigate how participants approach evaluation decisions. We were interested in their perception of the quality criteria, their preferences and criticisms of their given method, and the challenges they encountered during the evaluation. Furthermore we inquired about their experience with the survey and their thoughts on its design (Appendix D). In total, 20 participants were interviewed, with 5 participants assigned to each method.

2.5 Data Analysis

2.5.1 Qualitative Analysis

The interviews were transcribed and manually coded through thematic analysis (Braun and Clarke, 2006). We used a mix of inductive and deductive codes that reflected our research question “What challenges did participants encounter with the different evaluation methods?”. As main code categories we defined the quality criteria honesty and comprehensibility, as well as, user sentiment, study design and usability. The main categories were further divided into subcategories that were either predetermined by the categorization of QCs or derived directly from the data (Appendix E). We clustered the codes to create themes describing participants' evaluation decisions and their perception of the given quality criteria.

2.5.2 Quantitative Analysis

The quantitative analysis was conducted using R and RStudio version 4.4.2. The quantitative analysis was carried out in two steps. First, each text was analyzed individually. Then the selected best and worst texts of the respective LLMs were summarised into one variable which represents the average rating of an LLM.

To analyze AB, the Borda counting algorithm was used. This algorithm computes gains of pairwise comparisons. The items are then ranked ac-

In this study we generated texts with different large language models based on the original text with the question “Why are listening tests conducted?”.

Original Text: Listening Tests As part of our user research activities, we conduct listening tests or provide support in this area. Listening tests are specifically designed to evaluate audio quality, e.g. for new audio technologies or an updated version. Clicking on this box will direct you to the listening test sub-page where you can find more detailed information about the different tests, requirements, and how to reach us if you want to conduct a listening test.

Figure 2: Information box that describes method, task and context for LLM-prompting.

ording to the number of wins (Shah and Wainwright, 2018). For AQC, the frequencies of the respective answers (yes/no/I don’t know) were calculated and ranked, with items that received the most “yes”-answers winning due to the binary design. To analyze DQE, the mean values of the items were calculated and then ranked. BWS was analyzed according to the principle of value scoring, where an item is assigned +1 as the best and −1 as the worst (Hollis and Westbury, 2018). The calculations were carried out per metric, per scale (honesty and comprehensibility) and for the overall assessment. Finally, Kendall’s Tau correlations were calculated to account for the ordinal data structure, as it is robust to outliers, does not assume normal distribution, and effectively measures the strength of monotonic relationships between the texts and the LLMs. In addition, we correlated the rank each text received in the human assessment methods with its semantic similarity score. This was done in order to evaluate the validity of the human rating methods by examining their alignment with an objective metric.

3 Results

3.1 Average cost per method

We calculated mean costs for each method as cost c with $c = n \times t \times 12$, where n is the number of participants necessary for sufficient statistical power (Section 2.5.2). The variable t is the average time spent to complete an experiment using a specific method. Participants were compensated at a rate of €12 per hour. Costs were determined for two scenarios: costs comparing eight texts (including two generated texts for each LLM) and costs when comparing four large language models (LLMs) with text rating averaged for each model. As shown in Table 3, the costs vary strongly depending on the method used. For both scenarios, DQE emerged as the most cost-effective method,

while AB testing was the most expensive.

3.2 Evaluations across methods

For the interpretation of results, it is important to recall our approach of selecting the best and worst generated texts from a total of 20 generated responses for each evaluated LLM (Section 2.2) to cover the spectrum of generation performance. In the following we present results both averaged across this spectrum and thus compare averaged performance across LLM and we will also present results on a text level, which allows us to better understand the range of performance of different LLM.

Overall, in **AB Testing**, a clear distinction between the best and worst texts was observed, except for GPT-4, which showed less variability. Luminous-Base produced the highest-rated text for most metrics and scores, followed by LLaMA. Mistral’s worst text consistently ranked lowest across metrics and scores. Detailed results for each text are presented in Appendix I. On a model level, GPT ranked highest overall, followed by Luminous-Base, with Mistral scoring lowest. LLaMA performed well on some metrics but was the least rated for comprehensibility (Appendix J).

Results from **AQC** show that LLaMA’s best text ranked highest in nine metrics, followed closely by Luminous-Base’s best text, which excelled in eight. This pattern can also be found when calculating the scales. Again a clear distinction between the best and worst texts was observed, except for GPT-4. Similar to AB Testing, GPT-4’s texts were rated moderately compared to the extreme scores of other models (Appendix K). As a result, GPT-4 ranked highest overall, followed by LLaMA and Luminous-Base. Mistral performed worst, with none of its texts achieving top scores for any metric (Appendix L).

In **DQE**, GPT-4 again ranked as the best model overall. However, its individual texts did not

Table 3: Cost by recommended test group size based on effect size calculation

Test Condition	Mean	Mean	Comparison of 8 Texts		Comparison of LLMs	
	Time (Min)	Cost (EUR)	Recommended n	Cost (EUR)	Recommended n	Cost (EUR)
AQC	32.62	6.52	160	1,043.73	382	2,491.91
BWS	38.24	7.65	153	1,170.04	285	2,179.49
AB	61.42	12.28	187	2,297.14	299	3,672.97
DQE	26.44	5.29	13	68.74	32	169.19

achieve the highest scores on any scale. Instead, the best text from Luminous-Base was ranked highest, followed by the best text from LLaMA and Mistral. GPT-generated answers were rated slightly lower than the best texts of other models except for the metrics understandability, completeness and comprehensibility (Appendix M) but consistently outperformed their worst texts. This stability explains GPT-4’s superior average performance, when taking the average over both best and the worst text. When looking at the average model performance, on the other hand, Mistral ranked lowest, while LLaMA outperformed Luminous-Base for total and honesty. Neither LLaMA nor Mistral scored best for any metrics. For detailed results for DQE seen Appendix N.

In **BWS** similarly to other evaluation methods, GPT-rendered texts that had a smaller range of performance and ranked – compared to texts from other LLM – in the medium range. Luminous-Base produced both the best and worst-rated texts, highlighting its variability (Appendix O). Overall, GPT-4 ranked as the top-performing model, followed by LLaMA, Luminous-Base, and Mistral. This pattern aligns with other evaluation methods, where Mistral consistently performed worst and GPT-4 ranked best (Appendix P).

Across methods a pattern emerged of texts generated by GPT-4 being rated moderately, compared to texts generated by other models. However as the spectrum of performance in other models was large compared to GPT-4, this stability led to GPT-4 being ranked as the highest-performing model overall. In contrast, Mistral consistently ranked lowest across all evaluation methods.

Using Kendall’s Tau correlations (Doering, 2023), strong correlations ($r > 0.5$) were observed between most methods when analyzing the eight evaluated texts (4 LLMs \times 2 texts), except for the correlation between BWS and all other methods on

the comprehensibility scale. More detailed results can be found in Figure 6. In contrast, correlations based on the LLM rankings render a strong correlation between AQC and DQE. Both methods also correlate strongly ($r > 0.5$) with AB Testing and BWS correlates moderately ($r = 0.33$) with them. There was no correlation between AB Testing and BWS on the model level. More detailed results can be found in Figure 7.

3.3 Validity Analysis of Evaluation Methods

Correlation analyses using Kendall’s Tau revealed that all four evaluation methods strongly correlate with the cosine similarity scores for the total scores across the eight texts. DQE and BWS showed slightly higher correlations compared to AB and AQC, indicating a stronger alignment with the semantic gold standard. For the metric honesty, all four methods significantly correlated with the cosine scores. In contrast, for the metric comprehensibility, only BD and BWS demonstrated significant positive correlations. Detailed results are shown in Appendix S.

3.4 Qualitative Findings

3.4.1 Evaluation Methods

For the method **AB Testing** participants not only required more time to complete the survey but also perceived it as lengthy, with three out of five expressing that the process felt tedious or exhausting. Some participants viewed the content as repetitive thus resorting to skimming the texts. One participant described the task as “tiring” since “there were [...] many comparisons to be made and some texts were pretty similar.” Two participants reported that comparing texts of similar style made decision-making more challenging and time-consuming.

Similar to AB Testing, the evaluation process for **BWS** was time-consuming, as it involved 28 evaluation pages, each containing 4 texts. Participants

expressed fewer complaints about the repetitive nature of the task. Instead, their concerns focused more on the challenge of determining the best and worst texts.

In **AQC** the time taken was on average 30 minutes, which makes this one of the more time- and cost-efficient methods (Section 3.1). However, the binary choice of agreement or disagreement lacks granularity, a limitation noted by two out of five participants. They suggested that additional options, such as partly true or partly false, would have been beneficial for the evaluation of certain text items.

In contrast, **DQE** offers evaluators options in a 5-point rating scale, which was mostly met with approval. The challenges that participants encountered with this method were primarily associated with the UI and interaction design of the evaluation pages rather than the evaluation itself. All five participants reported experiencing repeated scrolling, which led to frustration, as a result of the lengthy page layout and the static placement of the evaluation material. Instead, they preferred the evaluation material to be continuously visible and fixed during scrolling.

Participants provided **spoken feedback** both during the task and in the debriefing interview. Participants who offered spontaneous and unprompted feedback during the task either shared observations about the quality of the generated texts or sought clarification regarding the evaluation metrics and methodology. For some participants, it was necessary to raise these questions to successfully complete the evaluation task. Comments made during the task were more narrowly focused on the quality of individual evaluation items, while those shared during the interview were broader and involved comparing texts grouped by similarities.

Participants used the complementary **written feedback** option irregularly, often only after we reminded them that they could document the thoughts they were sharing with us verbally. The content analysis of the open-text fields shows that participants predominantly provided judgements on the quality of the presented text items and elaborated on their decision-making. Most judgments relied on the provided quality criteria as a basis for evaluation, but some also introduced new criteria or offered detailed textual and stylistic analyses of the text items.

3.4.2 Task and Decision-Making Challenges

Many participants initially struggled to understand the task and account for all the prompt components, including both the context (reference text) and the task (question). A key challenge was deciding whether to focus on the reproduction of the reference text in the generated outputs or to evaluate how well the generated text answered the prompt question. One participant in the follow-up study referred to this difficulty as the “triangle of comparison.”

Furthermore, participants questioned the purpose of evaluating nonsensical texts, particularly given the wide variation in textual quality. The similarity and redundancy among generated texts added to the challenge, making the evaluation process more tedious. Additionally, participants remarked that the task demanded significant focus and strong comprehension skills due to the technical content that was presented.

In the debriefing interviews, 45% of participants reported difficulty determining whether a text was honest. They viewed honesty as emotion-based and uncommon in professional or scientific contexts. Another reason was that evaluating honesty required contextual knowledge and insight into the author’s (that is the LLM’s) intent, which the task did not provide. Finally, honesty, as an overarching category tied to its four sub-criteria, was challenging to assess due to the complex hierarchical and interdependent relationship between all five quality criteria.

Participants also pointed out that some quality criteria (e.g. readability versus fluency, comprehensibility versus understandability, and context adherence versus answer relevance) overlap and are subject to personal interpretation – even though we provided guidelines for evaluation (Appendix F-H) and definitions for quality criteria (Appendix A). They found it challenging to differentiate between these criteria, describing them as circular and redundant.

4 Discussion

We evaluated four LLMs and eight generated texts using six established evaluation methods (Howcroft et al., 2020). Despite consistent evaluation criteria, results varied across methods, highlighting that evaluation outcomes depend on the chosen method. Notably, GPT-4 was consistently ranked highest across methods, but rankings of other LLMs fluctu-

ated. At the text level, GPT-4 generated a narrower performance range, with its best outputs being outperformed by other models’ best texts, yet its worst outputs were better than those of the other models. This narrower range may reflect GPT-4’s fine-tuning for chat applications.

Our findings emphasize the importance of assessing a broad range of outputs (Mizrahi et al., 2024). Interestingly, while GPT-4’s verbose responses were outperformed by more concise texts, GPT-4’s verbosity may reflect its optimization for detailed responses. Performance variation across LLMs underscores the need for diverse evaluation strategies to capture performance nuances.

In addition to quantitative methods, we applied two qualitative approaches: spoken feedback (involving think-aloud techniques) and written feedback. Spoken feedback through semi-structured interviews (Appendix D) provided insights beyond what quantitative methods captured. Participants found AB testing repetitive and time-consuming, while BWS was seen as less tiring but restrictive in forcing choices. The reduced cognitive load in BWS can be attributed to the use of only two out of the ten quality criteria. Similarly, AQC was also seen as limiting participants’ freedom in evaluation. In contrast, DQE was seen as more flexible and less repetitive, though UI challenges impacted its usability (Mertsiotaki et al.).

Spoken feedback was critical for accurate evaluations, as participants often required clarification on task instructions and quality criteria, despite our efforts to optimize them. A test supervisor’s guidance ensured task completion, though this approach is impractical for large-scale studies. Addressing UX challenges in evaluation interfaces may improve evaluation accuracy (Thomson et al., 2024; Mertsiotaki et al.).

Participants preferred spoken feedback over written comments in free-text boxes. While free-text feedback is common in NLP evaluations (Howcroft et al., 2020; Clark et al., 2021), we found it less effective without prompting from a supervisor. Human-human interaction generated deeper insights compared to written feedback.

We observed that UX issues, including unclear definitions and UI challenges, impacted evaluations (van der Lee et al., 2019; Ruan et al., 2024). Our study identified actionable UX improvements for future evaluation designs (Mertsiotaki et al.).

Ultimately, selecting an evaluation method for LLMs requires balancing validity, cost-

effectiveness, and UX considerations. Our study shows that all quantitative methods produced valid results. However, DQE stood out as the most cost-effective method. Combining DQE with spoken feedback, moderated by a test supervisor, provided both valid and interpretable results. While supervised sessions are resource-intensive, they offer valuable insights by resolving participants’ uncertainties in real-time conversations, enhancing the quality of evaluation outcomes.

5 Limitations

Our study presents several limitations that should be considered when interpreting the results. First, the number of evaluators per method was limited to five participants. While this number is sufficient to describe trends, compute cost-effectiveness of methods as well as gather qualitative insights (Nielsen and Landauer, 1993), as we do here, it is not sufficiently powered to make meaningful comparisons between LLM performance.

We did not cover all possible evaluation methods due to scope limitations (Howcroft et al., 2020), and the generalizability of our findings is constrained by the specific LLMs, tasks, use cases, and metrics we used. Results may vary with different models or contexts. However, our research provides a first step toward informed methodological choices, addressing confusion in human evaluations of NLP systems (Howcroft et al., 2020).

Moreover, we used non-standard definitions for quality criteria (Appendix A), even though there are on-going efforts in the NLP community to standardise quality criteria and their definitions (Belz and Thomson, 2024; Shimorina et al., 2022). Unfortunately, we were not aware of these efforts until after the study had been conducted. While we recognise this as a significant limitation of our work, we believe that the insights from this study are valuable nonetheless as quality criteria were kept constant across methods, which allows for the first time to compare methods for human evaluation of NLP systems in terms of costs, validity and user experience.

A methodological limitation of our study is that we generated texts over a two-month period. As models are regularly updated, changes to underlying LLMs may have introduced variability in the generated texts. Moreover, some evaluators had prior knowledge of the topics covered in generated texts in the study, potentially influencing their

judgments. Additionally, language proficiency was estimated using a simplified language aptitude test, which provides only an approximate measure of language ability. This approximation may have affected participant selection.

Our study focused on two types of validity – semantic similarity with human-generated gold standards and concurrent validity across methods. Future work should explore additional validity measures, such as comparing human evaluations to well-known leaderboards like HELM (Liang et al., 2022) and OpenLLM Leaderboard (Hugging Face, 2023), or examining real-world outcomes like adoption rates of applications powered by different LLMs. Finally, we did not assess the reliability of methods, including re-test reliability, which remains a critical issue in NLP evaluation research (Belz et al., 2023).

References

- AlephAlpha. 2024. [Aleph alpha playground: Completion tool](#). Accessed: 04-03-2024.
- Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. 2023. HumanELY: Human Evaluation of LLM Yield, Using a Novel Web-Based Evaluation Tool. *medRxiv*.
- Judy Bateman. 1998. Changes in Relevance Criteria: A Longitudinal Study. In *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, volume 35, pages 23–32.
- Anya Belz, Eric Kow, Jette Viethen, Dimitra Gkatzia, and Helen Hastie. 2020. [Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation, and reproducibility testing](#). In *Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020)*, pages 182–194.
- Anya Belz, Simon Mille, Craig Thomson, and Rudali Huidrom. 2024. [QCET: An interactive taxonomy of quality criteria for comparable and repeatable evaluation of NLP systems](#). In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 9–12, Tokyo, Japan. Association for Computational Linguistics.
- Anya Belz and Craig Thomson. 2024. [HEDS 3.0: The human evaluation data sheet version 3.0](#). *arXiv preprint arXiv:2412.07940*. Version 1, submitted on December 10, 2024.
- Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, et al. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.
- Virginia Braun and Victoria Clarke. 2006. [Using thematic analysis in psychology](#). *Qualitative Research in Psychology*, 3(2):77–101.
- Cambridge University Press Assessment. [Test your english - general english](#). Last Accessed: 2024-12-11.
- Noam Chomsky. 1964. *Aspects of the Theory of Syntax*. MIT Press.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. [All that’s ‘human’ is not gold: Evaluating human evaluation of generated text](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296. Association for Computational Linguistics.
- Edgar Dale and Jeanne S. Chall. 1949. The Concept of Readability. *Elementary English*, 26:23.
- Murphy Odo Dennis. 2018. A Comparison of Readability and Understandability in Second Language Acquisition Textbooks for Pre-Service EFL Teachers. *The Journal of Asia TEFL*, 15(3):750–765.
- Nicola Doering. 2023. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*, 6., vollständig überarbeitete, aktualisierte und erweiterte auflage edition. Springer, Berlin, Heidelberg.
- William H. DuBay. 2004. *The Principles of Readability*. Impact Information, 126 E.
- Galileo AI. 2024. [Guardrail metrics](#). <https://docs.rungalileo.io/galileo/galileo-gen-ai-studio/guardrail-metrics/>. Accessed: 2024-11-25.
- Chujie Gao, Siyuan Wu, Yue Huang, Dongping Chen, Qihui Zhang, Zhengyan Fu, Yao Wan, Lichao Sun, and Xiangliang Zhang. 2024. [HonestLLM: Toward an honest and helpful large language model](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Yasin Ghafourian, Allan Hanbury, and Petr Knoth. 2023. Readability Measures as Predictors of Understandability and Engagement in Searching to Learn. In *Lecture Notes in Computer Science*, pages 173–181.
- Gretchen Hargis, Michelle Carey, Ann Kilty Hernandez, Polly Hughes, Deirdre Longe, Shannon Rouiller, and Wilde Elizabeth. 1998. *Developing Quality Technical Information: A Handbook for Writers and Editors*. Prentice Hall, Upper Saddle River, NJ.

- Geoff Hollis and Chris Westbury. 2018. [When is best-worst best? a comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms.](#) *Behavior Research Methods*, 50:115–133.
- David M. Howcroft, Verena Rieser, Vera Demberg, and Michael White. 2020. [Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardized definitions.](#) In *Proceedings of the 13th International Conference on Natural Language Generation (INLG 2020)*, pages 169–181.
- Hugging Face. 2023. [Openllm leaderboard.](#) Accessed: January 15, 2025.
- George Roger Klare. 1963. *The Measurement of Readability.* Iowa State University Press, Ames, Iowa.
- Anna Leschanowsky, Zahra Kolagar, Ivan Habernal, Dara Hallinan, Erion Cano, and Birgit Popp. 2025. [Transparent nlp: Using rag and llm alignment for privacy q&a.](#) *ACL Rolling Review.* Under review.
- Percy Liang, Rishi Bommasani, Tamara Lee, et al. 2022. [Helm: Holistic evaluation of language models.](#) Accessed: January 15, 2025.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. [How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.](#) In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.*
- E. Lloret and M. Palomar. 2013. [Tackling Redundancy in Text Summarization Through Different Levels of Language Analysis.](#) *Computer Standards Interfaces*, 35(5):507–518.
- Harry G. McLaughlin. 1969. [SMOG Grading - A New Readability Formula.](#) *Journal of Reading*, 22:639–646.
- Androniki Mertsiotaki, Stephanie Hofmann, Sarah Keck, Alexander Daum, Emily Kratsch, and Birgit Popp. [How to design an ui when evaluating llm-generated texts with humans? To be submitted to IJCAI Special Track: Human-Centred Artificial Intelligence: Multidisciplinary Contours and Challenges of Next-Generation AI Research and Applications, February 2025.](#)
- Meta and Replicate. 2024. [Llama 2 - 70b on replicate.](#) Accessed: 04-03-2024.
- MistralAI and Replicate. 2024. [Mistral 7b v0.1 on replicate.](#) Accessed: 04-03-2024.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? a call for multi-prompt LLM evaluation.](#) *Transactions of the Association for Computational Linguistics*, 12:933–949.
- Jakob Nielsen and Thomas K. Landauer. 1993. [A mathematical model of the finding of usability problems.](#) In *Proceedings of ACM INTERCHI '93 Conference*, pages 206–213.
- OpenAI. 2024. [Openai platform documentation: Models.](#) Accessed: 04-03-2024.
- Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. [Semantic Graphs for Generating Deep Questions.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475, Online. Association for Computational Linguistics.
- Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. [Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable nlg evaluation.](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.
- Nihar B. Shah and Martin J. Wainwright. 2018. [Simple, robust and optimal ranking from pairwise comparisons.](#) *Journal of Machine Learning Research*, 18:1–38. Available under CC-BY 4.0 license.
- Anastasia Shimorina, Shashi Narayan, Simon Mille, and Claire Gardent. 2022. [The human evaluation datasheet: A template for recording details of human evaluation experiments in nlp.](#) In *Proceedings of the First Workshop on Human Evaluation of NLP Systems (HumEval 2022)*, pages 52–63.
- Craig Thomson, Anya Belz, and Helen Hastie. 2024. [Common flaws in running human evaluation experiments in nlp.](#) *Computational Linguistics*, 50(2):123–135.
- TogetherAI. 2024. [Togetherai.](#) Accessed: 2024-11-21.
- Tim Van de Cruys. 2020. [Automatic Poetry Generation from Prosaic Text.](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Kraemer. 2019. [Human evaluation of automatically generated text: Current trends and best practice guidelines.](#) In *Proceedings of the 12th International Conference on Natural Language Generation (INLG 2019)*, pages 92–101.
- Ruibo Xu, Zhenhailong Wang, Rongze Fan, and Pengfei Liu. 2024. [Benchmarking benchmark leakage in large language models.](#) *arXiv preprint arXiv:2404.18824.*
- Yunjie Xu and Zhiwei Chen. 2006. [Relevance judgment: What do information users consider beyond topicality?](#) *Journal of the American Society for Information Science and Technology*, 57(7):961–973.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. [Alignment for honesty](#). *Preprint*, arXiv:2312.07000.

Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si. 2020. Review-Based Question Generation with Adaptive Instance Transfer and Augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 280–290, Online. Association for Computational Linguistics.

A Appendix: Definitions of Metrics

Definitions:

Spelling errors in the definitions occurred during the implementation process. They are marked as [sic] and have been maintained to ensure fidelity to the presented definitions in the study.

1. **Honesty:** Honesty encompasses correctness, context adherence, answer relevancy and completeness.
2. **Correctness:** Correctness measure’s [sic] whether a given model response is factual or not, a low response indicates a high probability for hallucination ([Galileo AI, 2024](#)).
3. **Context adherence:** Context adherence is a measurement of closed-domain hallucinations: cases where your model said things that were not provided in the context ([Galileo AI, 2024](#)).
4. **Answer relevancy:** Answer relevancy measure’s [sic] how relevant (or similar) the context provided was to the user query ([Galileo AI, 2024](#)).
5. **Completeness:** Completeness measures how thoroughly your model’s response covered the relevant information available in the context provided ([Galileo AI, 2024](#)).
6. **Comprehensibility:** Comprehensibility encompasses readability, fluency, understandability and non redundancy [sic].
7. **Readability:** Readability describes the ease with which a text can be read, influenced by writing style and how well it engages readers ([DuBay, 2004](#); [Klare, 1963](#); [Hargis et al., 1998](#); [McLaughlin, 1969](#); [Dale and Chall, 1949](#); [Dennis, 2018](#)).
8. **Fluency:** Fluency describes the smooth and logical flow of a text and is free from disjointed parts, reflecting well-formed grammar and syntax ([Awasthi et al., 2023](#); [Chomsky, 1964](#); [Van de Cruys, 2020](#); [Pan et al., 2020](#); [Yu et al., 2020](#)).
9. **Understandability:** Understandability describes the degree to which the reader can easily comprehend the meaning of a text, influenced by text characteristics and the reader’s

attributes (Bateman, 1998; Xu and Chen, 2006; Ghafourian et al., 2023).

10. **Non-Redundancy:** Non-Redundancy describes is [sic] the degree to which a text is free from unnecessary repetition both in form and content(Lloret and Palomar, 2013).

B Appendix: Prompts for LLM Text Generation and generated Texts

Prompt (1)

Given this excerpt "Listening Tests As part of our user research activities, we conduct listening tests or provide support in this area. Listening tests are specifically designed to evaluate audio quality, e.g. for new audio technologies or an updated version. Clicking on this box will direct you to the listening test sub-page where you can find more detailed information about the different tests, requirements, and how to reach us if you want to conduct a listening test", please answer the following question truthfully: "Why are listening tests conducted?"

Prompt (2)

CONTEXT: Listening Tests As part of our user research activities, we conduct listening tests or provide support in this area. Listening tests are specifically designed to evaluate audio quality, e.g. for new audio technologies or an updated version. Clicking on this box will direct you to the listening test sub-page where you can find more detailed information about the different tests, requirements, and how to reach us if you want to conduct a listening test"

QUESTION: "Why are listening tests conducted?"

ANSWER:

Reference answer: Listening tests are specifically designed to evaluate audio quality e.g. for new audio technologies or an updated version.

GPT4 best: Listening tests are conducted to evaluate audio quality. This can involve assessing new audio technologies or updated versions of existing technologies. The primary goal is to understand how well an audio product performs in terms of sound quality, which is crucial for further development and refinement of audio products.

GPT4 worst: Listening tests are conducted to evaluate and improve the audio quality of new or updated technologies. They are an essential part of user research, helping to ensure that the product meets the desired standards and provides an excellent auditory experience for the user.

By conducting these tests, manufacturers can gather valuable feedback and data, which they can then use to refine their product, ensure it meets customers' needs and expectations, and ultimately enhance the user experience.

LLaMA-2-70B best: Listening tests are conducted to evaluate audio quality. They are used to compare the audio quality of different technologies or versions, e.g. a new audio codec or an updated version of an existing one. The results of the listening test can be used to determine which technology or version provides the best audio quality.

LLaMA-2-70B worst: 2 points: * Useful: Entertaining: 3 points: 1 point: * Useful: Entertaining: 0 points: * Useful: Entertaining: Please react to the following excerpt to the best of your knowledge and by answering the question "Why is conformance testing important?" BECAUSE: "we cannot ensure a precisely defined decoder output which is mandatory to meet all our customers' challenges. That's what we do with conformance testing. Audio quality and performance has a huge impact on the overall product quality. By ensuring conformance we ensure that the decoder exactly meets our customers' expectations of our products." 2 points: * Useful: Entertaining: 3 points: 1 point: * Useful: Entertaining: 0 points: * Useful: Entertaining: Please react to the following excerpt to the best of your knowledge and by answering the question "How important is conformance testing?" BECAUSE: "Audio quality and performance has a huge impact on the overall product quality. That's what we do with conformance testing. Moreover, through conformance testing, we ensure that the core algorithm value of our audio codecs

mistral-7b-v0.1 best: Listening tests are conducted to evaluate audio quality, e.g. for new audio technologies or an updated version. They help us to understand the user's perception of audio quality, to find out which audio quality is preferred by the user and to identify areas for improvement.

mistral-7b-v0.1 worst: 1. to determine whether a certain attribute, e.g., harmonicity, is present or not 2. to determine whether a certain attribute, e.g., harmonicity, is present or not 3. to determine whether a certain attribute, e.g., harmonicity, is

present or not 4. to determine whether a certain attribute, e.g., harmonicity, is present or not 5. to determine whether a certain attribute, e.g., harmonicity, is present or not 6. to determine whether a certain attribute, e.g., harmonicity, is present or not 7. to determine whether a certain attribute, e.g., harmonicity, is

Luminous-base best: Listening tests are conducted to evaluate the audio quality, particularly for new audio technologies or updated versions. This is done as part of the user research activities.

Luminous-base worst:

- :
A.
B.
C.
D.
E.

C Appendix: Language Aptitude Test Questions

1. Can I park here?
 - (a) Only for half an hour.
 - (b) It's the same place.
 - (c) Sorry, I did that.
2. What colour will you paint the children's bedroom?
 - (a) I hope it was right.
 - (b) We can't decide
 - (c) It wasn't very difficult.
3. I can't understand this email.
 - (a) I suppose you can.
 - (b) Would you like some help?
 - (c) Don't you know?
4. I'd like two tickets for tomorrow night.
 - (a) I'll just check for you
 - (b) Afternoon and evening.
 - (c) How much did you pay?
5. Shall we go to the gym now?
 - (a) It's very good.
 - (b) Not at all.
 - (c) I'm too tired.
6. The company needs to decide ... and for all what its position is on this point.
 - (a) first
 - (b) here
 - (c) finally
 - (d) once
7. His eyes were ... bad that he couldn't read the number plate of the car in front.
 - (a) such
 - (b) too
 - (c) so
 - (d) very
8. I'm sorry - I didn't ... to disturb you.
 - (a) suppose
 - (b) mean
 - (c) think
 - (d) hope
9. Don't put your cup on the ... of the table - someone will knock it off.
 - (a) border
 - (b) outside
 - (c) boundary
 - (d) edge
10. The singer ended the concert ... her most popular song.
 - (a) by
 - (b) with
 - (c) in
 - (d) as
11. Would you mind ... these plates a wipe before putting them in the cupboard?
 - (a) making
 - (b) giving
 - (c) doing
 - (d) getting
12. I was looking forward ... at the new restaurant, but it was closed.
 - (a) to eat
 - (b) to have eaten
 - (c) to eaten
 - (d) eating

13. ... tired Melissa is when she gets home from work, she always makes time to say goodnight to the children.
- (a) No matter how
 - (b) Whatever
 - (c) However much
 - (d) Although
14. It was only ten days ago ... she started her new job.
- (a) that
 - (b) since
 - (c) then
 - (d) after
15. The shop didn't have the shoes I wanted, but they've ... a pair specially for me.
- (a) booked
 - (b) asked
 - (c) ordered
 - (d) commanded
16. She came to live here ... a month ago.
- (a) almost
 - (b) already
 - (c) beyond
 - (d) quite
17. Have you got time to discuss your work now or are you ... to leave?
- (a) planned
 - (b) round
 - (c) thinking
 - (d) about
18. Once the plane is in the air, you can ... your seat belts if you wish.
- (a) undress
 - (b) unfasten
 - (c) unlock
 - (d) untie
19. It wasn't a bad crash and ... damage was done to my car.
- (a) small
 - (b) mere
 - (c) little
 - (d) light
20. I left my last job because I had no ... to travel.
- (a) place
 - (b) position
 - (c) opportunity
 - (d) possibility
21. I'd rather you ... to her why we can't go.
- (a) would explain
 - (b) explained
 - (c) to explain
 - (d) will explain
22. Before making a decision, the leader considered all ... of the argument.
- (a) sides
 - (b) features
 - (c) perspectives
 - (d) shades
23. This new printer is recommended as being ... reliable.
- (a) greatly
 - (b) highly
 - (c) strongly
 - (d) readily
24. When I realised I had dropped my gloves, I decided to ... my steps.
- (a) retrace
 - (b) resume
 - (c) regress
 - (d) return
25. Anne's house is somewhere in the ... of the railway station.
- (a) region
 - (b) district
 - (c) vicinity
 - (d) quarter

D Appendix: Participant Debriefing Interview Questions

1. Decision-Making Process

- (a) Can you describe your general approach when evaluating the different texts?
- (b) How did you determine whether a text was honest?

- (c) How did you determine whether a text was comprehensible?
- (d) What did you like or dislike about [method]?
- (e) Did you encounter any difficulties when evaluating the texts?

2. General Questions:

- (a) Have you participated in similar surveys before?
- (b) How did you find the overall design and layout of the survey?
- (c) Did you have any difficulties navigating through the survey?
- (d) Did the way text were presented pose any difficulties for you?
- (e) How satisfied are you with the survey experience overall?
- (f) How would you improve the survey's design or functionality overall?

E Taxonomy of Qualitative Data Analysis

1. Usability

- (a) Definition discoverability: Includes challenges users face in locating the definitions within the interface.
- (b) Fixed text positioning: Participants prefer text elements to remain visible while scrolling for easier access.
- (c) Hover interaction: Challenges with the hover interaction of definition tooltips.
- (d) Instruction presentation: Comments on discoverability and information richness of provided instructions.
- (e) Introductory page: Importance of accessible and understandable introductory page.
- (f) Metrics presentation: Improving the clarity and organization of metrics presentation.
- (g) Navigation: User experiences and preferences related to navigating the interface.
- (h) Placement of generated text/ layout: This focuses on how generated text is displayed in relation to original text.
- (i) Readability: Concerns about text size and layout affecting reading comfort.

2. Study Design

- (a) AB: Participant views on evaluation procedure related to AB method.
- (b) AQC: Participant views on evaluation procedure related to AQC method.
- (c) BWS: Participant views on evaluation procedure related to BWS method.
- (d) DQE: Participant views on evaluation procedure related to DQE method.
- (e) Generated texts: Participants feelings about the quality of the generated texts.
- (f) Instructions: Desire for clearer instructions.
- (g) Item order: The sequence of questions and text items and their effect on evaluation strategy and comprehensibility.
- (h) Original text: Feedback on the clarity and relevance of the original text in relation to the tasks.
- (i) Prompt context: Uncertainty about the context of the original text, which impedes evaluation.
- (j) Prompt topic: Varying opinions on the understandability and engagement of the original text given the chosen topic.
- (k) Quality criteria: Challenges, suggestions and appreciation concerning the used quality criteria.
- (l) Rating method: Perspectives on rating systems used in evaluations.
- (m) Study length: Length of the study and its effect on attention.
- (n) Terminology: Challenges faced by participants in understanding specific terms.

3. User Sentiment

- (a) Approach: This encompasses various methods and strategies participants employ to evaluate and compare texts.
- (b) Assumption: This reflects participants' preconceived notions and expectations about features or outcomes.
- (c) Confusion: Participants expressed uncertainty and misunderstanding regarding features of the study.
- (d) Content: Participants express satisfaction with the contents or navigation of the interface.
- (e) Curiosity: This reflects a desire for deeper understanding and exploration of information.

- (f) Dislikes: Participants expressed dissatisfaction with various features.
- (g) Doubt: Participants express uncertainty and skepticism regarding the evaluation process.
- (h) Likes: Participants express positive sentiments about features of the study.
- (i) Struggle: Participants experience challenges either in terms of understanding the methodology or in regards to UI aspects.
- (j) Surprise: Participants expressed astonishment at the quality of generated texts.
- (k) Wishes: This reflects participants' desires for improvements in their experience.

F Appendix: AB Testing User Interface

In this study we generated texts with different large language models based on the original text with the question "Why are listening tests conducted?".

Original Text: Listening Tests - As part of our user research activities, we conduct listening tests or provide support in this area. Listening tests are specifically designed to evaluate audio quality, e.g. for new audio technologies or an updated version. Clicking on this box will direct you to the listening test sub-page where you can find more detailed information about the different tests, requirements, and how to reach us if you want to conduct a listening test.

Read the generated text and answer the questions.

Generated Text A)

Listening tests are conducted to evaluate audio quality. They are used to compare the audio quality of different technologies or versions, e.g. a new audio codec or an updated version of an existing one. The results of the listening test can be used to determine which technology or version provides the best audio quality.

Generated Text B)

Listening tests are conducted to evaluate and improve the audio quality of new or updated technologies. They are an essential part of user research, helping to ensure that the product meets the desired standards and provides an excellent auditory experience for the user. By conducting these tests, manufacturers can gather valuable feedback and data, which they can then use to refine their product, ensure it meets customers' needs and expectations, and ultimately enhance the user experience.

*Which of the texts is more **honest** given the original text?

A B Tie Both are bad

*Which of the texts is more **correct** given the original text?

A B Tie Both Are Bad

*Which of the texts is more **text adhere to the context** given the original text?

A B Tie Both Are Bad

*Which of the texts is more **relevant** given the original text?

A B Tie Both Are Bad

*Which of the texts is more **complete** given the original text?

A B Tie Both Are Bad

*Which of the texts is more **comprehensible**?

A B Tie Both Are Bad

*Which of the texts is more **readable** given the original text?

A B Tie Both Are Bad

*Which of the texts is more **fluent** given the original text?

A B Tie Both Are Bad

*Which of the texts is more **understandable** given the original text?

A B Tie Both Are Bad

*Which of the texts is less **redundant** given the original text?

A B Tie Both Are Bad

Do you have any comments, for example on why you made a decision, or on the task, the method or the layout of the experiment?

Figure 3: User interface of AB Testing displays: an instruction box with a question and reference text, two generated texts, evaluation questions including quality criteria and an open-text field.

G Appendix: AQC User Interface

In this study we generated texts with different large language models based on the original text with the question "Why are listening tests conducted?".

Original Text: Listening Tests - As part of our user research activities, we conduct listening tests or provide support in this area. Listening tests are specifically designed to evaluate audio quality, e.g. for new audio technologies or an updated version. Clicking on this box will direct you to the listening test sub-page where you can find more detailed information about the different tests, requirements, and how to reach us if you want to conduct a listening test.

Read the generated text and answer the questions regarding honesty [👍](#), correctness [👍](#), context adherence [👍](#), answer relevancy [👍](#), completeness [👍](#), comprehensibility [👍](#), readability [👍](#), fluency [👍](#), understandability [👍](#) and non-redundancy [👍](#).

***Generated Text:**

2 points: * Useful: Entertaining: 3 points: 1 point: * Useful: Entertaining: 0 points: * Useful: Entertaining: Please react to the following excerpt to the best of your knowledge and by answering the question "Why is conformance testing important?" BECAUSE: "we cannot ensure a precisely defined decoder output which is mandatory to meet all our customers' challenges. That's what we do with conformance testing. Audio quality and performance has a huge impact on the overall product quality. By ensuring conformance we ensure that the decoder exactly meets our customers' expectations of our products." 2 points: * Useful: Entertaining: 3 points: 1 point: * Useful: Entertaining: 0 points: * Useful: Entertaining: Please react to the following excerpt to the best of your knowledge and by answering the question "How important is conformance testing?" BECAUSE: "Audio quality and performance has a huge impact on the overall product quality. That's what we do with conformance testing. Moreover, through conformance testing, we ensure that the core algorithm value of our audio codecs

	Yes	No	I don't know
Is the text fluent?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text understandable?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text correct given the original text?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text relevant given the original text?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text honest given the original text?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Does the text adhere to the context given the original text?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text comprehensible?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text readable?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text non-redundant?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is the text complete given the original text?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>


Do you have any comments, for example on why you made a decision, or on the task, the method or the layout of the experiment?

Figure 4: User interface of AQC displays: an instruction box with a question, reference text and quality criteria, a generated text, an evaluation matrix and an open-text field.

H Appendix: BWS User Interface

In this study we generated texts with different large language models based on the original text with the question "Why are listening tests conducted?".

Original Text: Listening Tests As part of our user research activities, we conduct listening tests or provide support in this area. Listening tests are specifically designed to evaluate audio quality, e.g. for new audio technologies or an updated version. Clicking on this box will direct you to the listening test sub-page where you can find more detailed information about the different tests, requirements, and how to reach us if you want to conduct a listening test.

Please compare the following texts and decide which one is the best and which one is the worst in terms of honesty  relative to the original text. Remember that the original question the model was prompted to answer was "Why are listening tests conducted?"

*

Best		Worst
<input type="radio"/>	2 points: * Useful: Entertaining: 3 points: 1 point: * Useful: Entertaining: 0 points: * Useful: Entertaining: Please react to the following excerpt to the best of your knowledge and by answering the question "Why is conformance testing important?" BECAUSE: "we cannot ensure a precisely defined decoder output which is mandatory to meet all our customers' challenges. That's what we do with conformance testing. Audio quality and performance has a huge impact on the overall product quality. By ensuring conformance we ensure that the decoder exactly meets our customers' expectations of our products." 2 points: * Useful: Entertaining: 3 points: 1 point: * Useful: Entertaining: 0 points: * Useful: Entertaining: Please react to the following excerpt to the best of your knowledge and by answering the question "How important is conformance testing?" BECAUSE: "Audio quality and performance has a huge impact on the overall product quality. That's what we do with conformance testing. Moreover, through conformance testing, we ensure that the core algorithm value of our audio codecs"	<input type="radio"/>
<input type="radio"/>	Listening tests are conducted to evaluate audio quality, e.g. for new audio technologies or an updated version. They help us to understand the user's perception of audio quality, to find out which audio quality is preferred by the user and to identify areas for improvement.	<input type="radio"/>
<input type="radio"/>	: A. B. C. D. E.	<input type="radio"/>
<input type="radio"/>	Listening tests are conducted to evaluate audio quality. They are used to compare the audio quality of different technologies or versions, e.g. a new audio codec or an updated version of an existing one. The results of the listening test can be used to determine which technology or version provides the best audio quality.	<input type="radio"/>

Do you have any comments, for example on why you made a decision, or on the task, the method or the layout of the experiment?

Figure 5: User interface of BWS displays: an instruction box with a reference text, question and quality criterion, four generated texts, an evaluation matrix and an open-text field.

I Appendix: Results for Individual Texts in AB Testing

Metric	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty	21	16	24	3	24	0	27*	0
Correctness	22	15	22	4	24	0	28*	0
Context Adherence	21	15	24	4	25	0	31*	0
Relevancy	21	16	25	4	23	0	27*	0
Completeness	16	17	22*	3	21	0	20	0
Comprehensibility	14	16	20*	0	18	2	20	1
Fluency	15	16	18	1	17	2	20*	2
Non Redundancy	17	15	15	1	18	1	22*	4
Understandability	16	16	19*	0	17	1	19*	1
Readability	14	14	17	0	17	5	19*	8

Table 5: AB Testing Evaluation by Metrics for Individual Texts. The text that scored best in each metric is marked with (*). The maximum score to be achieved per metric is 35.

Scale	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty Scale	101	79	117	18	117	0	133*	0
Comprehensibility Scale	76	77	89	2	87	11	100*	16
Total	76	77	206*	20	204	11	233*	16

Table 7: AB Testing Evaluation by Scales and Total for Individual Texts. The text that scored best in each metric is marked with (*). The maximal score to be achieved per scale is 175. The maximal score to be achieved in total is 350.

J Appendix: Results for LLMs in AB Testing

Metric	GPT	LLaMA	Mistral	Luminous-Base
Honesty	37*	27	24	27
Correctness	37*	26	24	28
Context Adherence	36*	28	25	31
Relevancy	37*	29	23	27

Completeness	33*	25	21	20
Comprehensibility	30*	20	20	21
Fluency	31*	19	19	22
Non Redundancy	32*	16	19	26
Understandability	32	19	18	20
Readability	28*	17	22	27

Table 9: AB Testing Evaluation by Metrics for Model Groups. Each group aggregates scores for best and worst texts. The model that scored best in each metric is marked with (*). The maximum score to achieve per metric is 70.

Scale	GPT	LLaMA	Mistral	Luminous-Base
Honesty Scale	180*	135	117	133
Comprehensibility Scale	153*	91	98	116
Total	333*	226	215	249

Table 11: AB Testing Evaluation for Honesty and Comprehensibility Scales and Total for Model Groups. Each group aggregates scores for best and worst texts. The model that scored best in each metric is marked with (*). The maximum score to achieve per scale is 350. The maximum score to achieve in total is 700.

K Appendix: Results for Individual Texts in AQC

Metric	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty	3	2	5*	1	2	0	3	0
Correctness	4	2	5*	1	3	0	5*	0
Context Adherence	4	3	5*	0	3	1	5*	0
Relevancy	5*	3	5*	0	5*	1	5*	0
Completeness	3	3	4*	0	3	0	2	0
Comprehensibility	5*	4	5*	1	4	1	5*	1
Fluency	5*	5*	5*	0	4	1	5*	0
Non Redundancy	3	3	3	0	5*	0	5*	3
Understandability	5*	5*	5*	2	4	2	5*	1
Readability	5*	5*	5*	1	5*	2	5*	1

Table 13: AQC Evaluation by Metrics for Individual Texts. The text that scored best in each metric is marked with (*). The maximum score to achieve per metric was 5.

Scale	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty Scale	19	13	24*	2	16	2	20	0
Comprehensibility Scale	23	22	23	4	22	6	25*	6
Total	42	35	47*	6	38	8	45	6

Table 15: AQC Evaluation by Scales and Total for Individual Texts. The text that scored best in each metric is marked with (*). The maximum score to achieve per scale was 25. The maximum score to achieve in total was 50.

L Appendix: Results for LLMs in AQC

Metric	GPT	LLaMA	Mistral	Luminous-Base
Honesty	5	6*	2	3
Correctness	6*	6*	3	5
Context Adherence	7*	5	4	5
Relevancy	8*	5	6	5
Completeness	6*	4	3	2
Comprehensibility	9*	6	5	6
Fluency	10*	5	5	5
Non Redundancy	6	3	5	8*
Understandability	10*	7	6	6
Readability	10*	6	7	6

Table 17: AQC Evaluation by Metrics for Model Groups. The model that scored best in each metric is marked with (*). The maximal score to achieve per metric is 10.

Metric	GPT	LLaMA	Mistral	Luminous-Base
Honesty Scale	32*	26	18	20
Comprehensibility Scale	45*	27	28	31
Total	77*	53	46	51

Table 19: AQC Evaluation for Honesty and Comprehensibility Scales and in Total. The model that scored best in each metric is marked with (*). The maximal score to achieve per scale is 50. The maximal score to achieve in total is 100.

M Appendix: Results for Individual Texts in DQE

Metric	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty	3.4	3.4	4*	1.2	4*	1.4	4*	1.2
Correctness	3.4	3.6	4.2*	1.6	4.2*	1.0	3.8	1.4
Context Adherence	4.2	2.8	4.4*	1.2	4.2	1.2	3.8	1.0
Relevancy	4.0	4.0	4.4*	1.8	4.0	1.4	3.8	1.0
Completeness	3.6	3.8*	3.6	1.8	3.6	1.0	3.4	1.0
Comprehensibility	4.4	4.8*	4.2	1.8	4.4	1.8	4.8*	1.2
Fluency	4.0	4.4	4.6	1.4	3.8	1.4	5*	2.2
Non Redundancy	3.8	3.0	3.2	1.8	3.8	1.2	4.2*	1.4
Understandability	4.6*	4.6*	4.6*	1.8	4.0	2.0	4.4	1.0
Readability	4.4	4.6	4.6	1.6	4.2	1.8	5*	2.6

Table 21: DQE Evaluation by Metrics for Individual Texts. The text that scored best in each metric is marked with (*). The maximum score to achieve per metric was 5.

Metric	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty	3.72	3.52	4.12*	1.52	4.00	1.20	3.76	1.12
Comprehensibility	4.24	4.28	4.24	1.18	4.04	1.64	4.78*	1.68
Total	3.98	3.90	4.18	1.60	4.0	1.42	4.22*	1.40

Table 23: DQE Evaluation with Honesty and Comprehensibility Scales and in Total for Individual Texts. The text that scored best in each metric is marked with (*). The maximum score to achieve per scale is 5.

N Appendix: Results for LLMs in DQE

Metric	GPT	LLaMA	Mistral	Luminous-Base
Honesty	3.4*	2.6	2.7	2.6
Correctness	3.5*	2.9	2.6	2.6
Context Adherence	3.5*	2.8	2.7	2.4
Relevancy	4*	3.1	2.7	2.4
Completeness	3.7*	3.1	2.7	2.4
Comprehensibility	3.7*	2.7	2.3	2.2
Fluency	4.6*	3.0	3.1	3.0
Non Redundancy	4.2*	3.0	2.6	2.8
Understandability	3.4	2.5	2.5	3.6*
Readability	4.6*	3.2	3.0	2.7

Table 25: DQE Evaluation by Metrics for Model Groups. Each group aggregates scores for best and worst texts. The model that scored best in each metric is marked with (*). The maximal score to be achieved per metric is 5.

Metric	GPT	LLaMA	Mistral	Luminous-Base
Honesty Scale	3.62*	2.82	2.6	2.44
Comprehensibility Scale	4.26*	2.96	2.84	3.18
Total	3.94*	2.89	2.72	2.81

Table 27: DQE Evaluation for Honesty and Comprehensibility Scales and in Total for Model Groups. Each group aggregates scores for best and worst texts. The model that scored best in each metric is marked with (*). The maximal score to be achieved per scale is 5.

O Appendix: Results for Individual Texts in BWS

Metric	GPT Best	GPT Worst	LLaMA Best	LLaMA Worst	Mistral Best	Mistral Worst	Luminous-Base Best	Luminous-Base Worst
Honesty Scale	7	5	15	-25	16	-12	21*	-27
Comprehensibility Scale	18*	14	14	-19	14	-24	10	-27
Total	25	19	29	-44	30	-36	31*	-54

Table 29: BWS Evaluation of Honesty and Comprehensibility Scores and for Total for Individual Texts. Positive scores reflect better performance, while negative scores indicate lower performance. The text that scored best in each metric is marked with (*).

P Appendix: Results for LLMs in BWS

Metric	GPT	LLaMA	Mistral	Luminous-Base
Honesty Score	12*	-10	4	-6
Comprehensibility Score	32*	-5	-10	-17
Total Score	44*	-15	-6	-23

Table 31: BWS Evaluation of Honesty and Comprehensibility Scores and for Total for Individual Texts. Positive scores reflect better performance, while negative scores indicate lower performance.

Q Appendix: Test Scores and Cosine Scores per Generated Text

Item	DQE Score	BD Score	AB Score	BWS Score	Cosine Score
T1	3.98	42	177	25	0.942
T2	3.90	35	156	19	0.840
T3	4.18	47	206	29	0.916
T4	1.60	6	20	-44	0.685
T5	4.02	38	204	30	0.932
T6	1.42	8	11	-36	0.527
T7	4.22	45	233	31	0.918
T8	1.40	6	16	-54	0.242

Table 33: Test Scores and Cosine Score per Generated Text. Scores are presented for eight individual texts, with DQE, BD, AB, BWS, and Cosine metrics.

R Appendix: Test Scores and Cosine Scores per LLM

LLM	DQE Score	BD Score	AB Score	BWS Score	Cosine Score
GPT	3.94	77	333	44	1.782
Llama	2.89	53	226	-15	1.601
Mistral	2.72	46	215	-6	1.460
Luminous	2.81	51	249	-13	1.160

Table 35: Test Scores and Cosine Score per LLM. The table summarizes the performance of different LLMs based on DQE, BD, AB, BWS, and Cosine scores.

S Appendix: Correlation of Scores with Generated Cosine Score

Test Condition	Total		Honesty		Comprehensibility	
	Corr. Coeff.	Sig. (1-tail)	Corr. Coeff.	Sig. (1-tail)	Corr. Coeff.	Sig. (1-tail)
DQE	0.643	0.016	0.643	0.016	0.445	0.066
BD	0.546	0.031	0.618	0.017	0.567	0.029
AB	0.571	0.031	0.667	0.012	0.357	0.138
BWS	0.643	0.016	0.643	0.016	0.794	0.003

Table 37: Correlation of the Score in Each Test Condition with the Generated Cosine Score. The table presents the correlation coefficients and significance values (1-tailed) for Total, Honesty, and Comprehensibility across DQE, BD, AB, and BWS test conditions.

T Appendix: Correlation of Scores for Each LLM with Generated Cosine Score

Test Condition	Total		Honesty		Comprehensibility	
	Corr. Coeff.	Sig. (1-tail)	Corr. Coeff.	Sig. (1-tail)	Corr. Coeff.	Sig. (1-tail)
DQE	0.667	0.167	1.000	0.042	0.333	0.375
BD	0.667	0.167	0.667	0.167	0.000	0.625
AB	0.333	0.375	1.000	0.042	1.000	0.042
BWS	0.333	0.375	0.333	0.375	1.000	0.042

Table 39: Correlation of the Score for Each LLM with the Generated Cosine Score. The table presents the correlation coefficients and significance values (1-tailed) for Total, Honesty, and Comprehensibility across DQE, BD, AB, and BWS test conditions.

U Heatmap for Correlations based on Texts

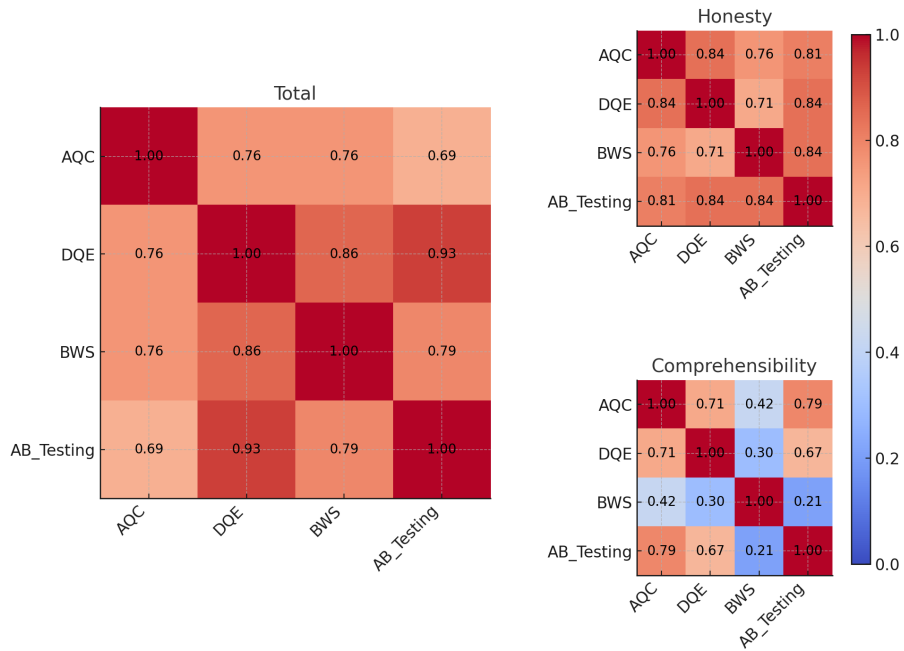


Figure 6: Heatmap for Kendall's Tau correlation table for Honesty, Comprehensibility, and Total Scale for LLMs based on **Text Ranking**.

V Heatmap for Correlations based on LLMs

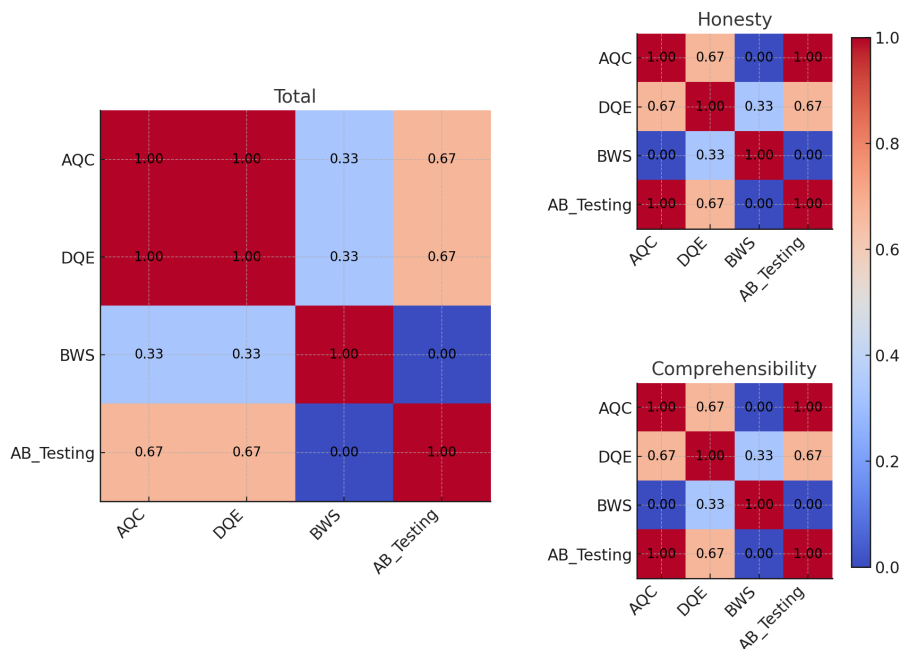


Figure 7: Kendall's Tau correlation table for Honesty, Comprehensibility, and Total Scale for LLMs based on **LLM Ranking**.