Designing Usable Interfaces for Human Evaluation of LLM-Generated Texts: UX Challenges and Solutions

Androniki Mertsiotaki[1] , **Stephanie Hofmann**[1] , **Sarah Keck**[1,2] , **Emily Kratsch**[1,2] , **Alexander Daum**[1] , **Birgit Popp**[1]
[1]Fraunhofer IIS
[2]Friedrich-Alexander University Erlangen
{androniki.mertsiotaki, stephanie.hofmann, sarah.keck, emily.kratsch, alexander.daum, birgit.popp}@fraunhofer.iis.de,

Abstract

Human evaluations remain important for assessing large language models (LLMs) due to the limitations of automated metrics. However, flawed methodologies and poor user interface (UI) design can compromise the validity and reliability of such evaluations. This pre-registered study investigates usability challenges and proposes solutions for UI design in evaluating LLM-generated texts. By comparing common evaluation methods such as Direct Quality Estimation, AB-Testing, Agreement with Quality Criterion and Best-Worst Scaling, insights were gained into user experience challenges, including inefficient information transfer and poor visibility of evaluation materials. Iterative redesigns improved discoverability, accessibility, and user interaction through modifications of page layout and content presentation. Testing these enhancements revealed increased clarity and usability, with higher response rates and more consistent ratings. This work highlights the importance of UI design in enabling reliable and meaningful human evaluation, providing actionable recommendations to enhance the integrity and usability of NLP evaluation frameworks.

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 do 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. In fact, Thomson *et al.* [2024] emphasize the importance of a user-friendly interface to reduce errors in response collection. A poorly designed interface can increase cognitive load for evaluators, who may struggle to find information relevant for evaluation, understand and operate the interface correctly and – in addition – evaluate NLP system outputs [Thomson *et al.*, 2024]. Moreover, lack of clarity and inconsistency in terminology of quality criteria, and their definitions [Belz *et al.*, 2020; Howcroft *et al.*, 2020], as well as guidelines for evaluators [Ruan *et al.*, 2024] are known usability issues when conducting evaluations of NLP systems with humans. Thus, the question arises how to design user interfaces (UI) and provide a good user experience (UX) when evaluating NLP systems like large language models with humans. Our research questions are thus:

- What are UX challenges within the task of LLM output evaluation?
- How might pain points of evaluators be addressed?

In this work, we focus on evaluation of Large Language Models (LLM), which are part of the field of Natural Language Generation (NLG). NLG is a subset of NLP and we predominantly mention NLP, except when citing papers that specifically refer to NLG.

2 Methods

2.1 Study Design

We chose six widely used evaluation methods in the field of natural language generation (NLG) [Howcroft *et al.*, 2020] in order to gain broad insights into UI issues during evaluation of NLP systems that are common across methods. Popp *et al.* [2025] provides a detailed overview of evaluation results obtained with the various methods and compares them in terms of validity and efficiency. The chosen methods include both quantitative (4 methods) and qualitative (2 methods) approaches.

We chose the following quantitative methods:

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

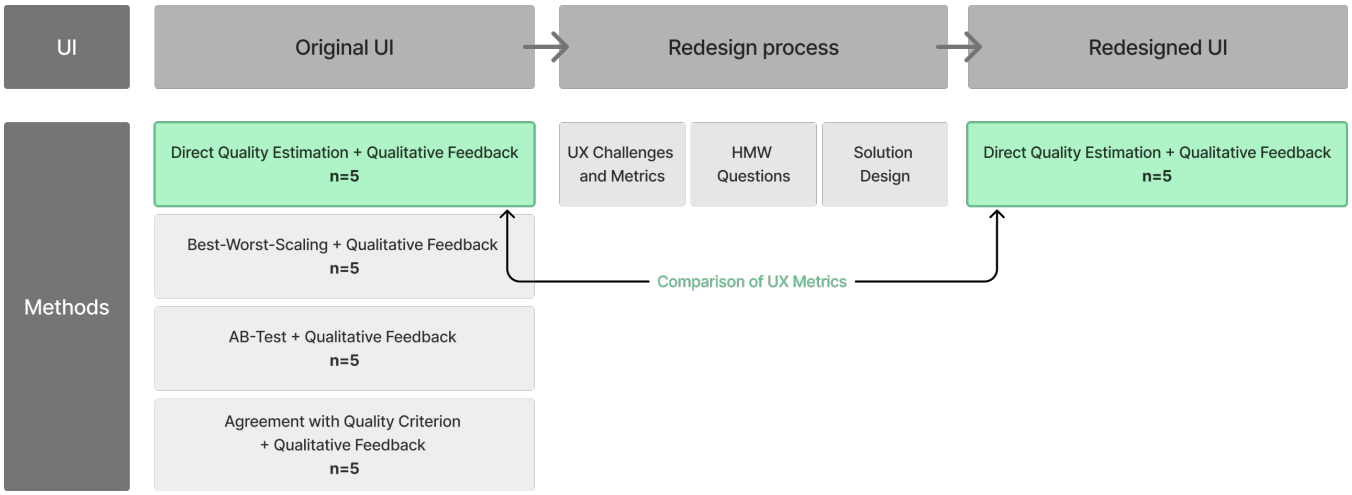


Figure 1: Study design for our multi-method comparison of human evaluation methods.

We used the following qualitative methods as a means to gain qualitative insights about issues of the UI that was implemented to apply the quantitative methods:

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

Since our research questions focus on the visual interface used for evaluation tasks, we compared two versions of an evaluation UI. Figure 1 illustrates the study design. We gathered insights on UI issues from four methods in Study 1, which allows us to present patterns of UI issues across methods here (cf. 3.1). Implementing and testing redesigns for all methods would have exceeded the scope of this study. Thus, we picked one method, namely DQE, to redesign in Study 2. Notably, DQE is described as most efficient of the tested methods in Popp *et al.* [2025]. Comparing Figure D and Figure 3 provides a before and after view of the UI design of evaluation pages for DQE. The DQE design of Study 1 was revised (cf. 2.4), following an iterative design process in accordance with the principles of the User Centered Design Process (UCDP, 9241-210:2020-03 [2020]).

We simulate evaluation of a Retriever Augmented Generation (RAG) system [Gupta *et al.*, 2024; Wu *et al.*, 2024], specifically a Virtual Workplace Assistant (VWA). We used a text excerpt from our organization's intranet as context for the selected LLMs, mimicking a RAG system. The chosen evaluation task, dialogue turn generation, is the second most common NLG task according to Howcroft *et al.* [2020]. Detailed information about the generation of evaluation texts as well as selection criteria for participants (language aptitude) is provided in Popp *et al.* [2025].

We selected four LLMs for text generation based on their mean win rates from the HELM (Classic) [Liang *et al.*, 2023] leaderboard, retrieved on March 7, 2024, namely GPT-4, Llama-2-70B, Mistral-7B-v0.1 and Luminous Base (13B). The HELM leaderboard served as a predictor of overall text generation quality. To ensure a diverse evaluation, we included the highest and lowest-ranked LLMs, along with two mid-range models.

We selected honesty [Gao *et al.*, 2024; Yang *et al.*, 2024] and comprehensibility as the two key constructs for evaluating LLM-generated texts. These constructs were chosen for their relevance to human NLP evaluation and their distinct nature, which captures different aspects of quality [Howcroft *et al.*, 2020; Belz *et al.*, 2020; Belz and Thomson, 2024]. Since no standard definitions exist for "honesty" and "comprehensibility", we consider them to consist of multiple quality criteria (QC). Accordingly, we identified four QC for each construct, as shown in Appendix A.

2.2 Data Collection

We conducted parts of the data collection using LimeSurvey, an open-source software hosted on a server of the research institution. We chose LimeSurvey for its flexibility in allowing custom adaptations through PHP and JavaScript. Our team developed separate questionnaires for each evaluation method – AB, BWS, DQE, and AQC – through an iterative process, conducting multiple rounds of testing. Participants' English proficiency was assessed as described in Popp *et al.* [2025], and those who met the required level were randomly assigned to one of the methods.

Before beginning the survey, participants had to provide consent to a data declaration. They were then presented with a disclaimer explaining that the quality of the selected texts might not align with their expectations. This clarification was included to prevent confusion, which was observed during internal testing.

Prior to the task instruction, each evaluation page contained background information about the context and the question used for prompting the LLM (Figure 2).

The task instruction and the presentation of the generated texts differed based on the evaluation method:

- BWS followed an incomplete block design, featuring four texts per block and a total of 14 blocks for both honesty and comprehensibility.
- AB used pairwise comparisons, where each text was compared against every other, resulting in 28 unique

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: Additional information for evaluators about the method used for LLM-prompting.

combinations evaluated across all metrics.

- DQE and AQC evaluated each individual text across all 10 metrics.

To minimize order bias, the sequence of texts and quality criteria was randomized. This included between-subject randomization, as each text (or text combination for AB and BWS) was shown only once per participant, and within-subject randomization, as all texts were evaluated using the same quality criteria, but their order varied across evaluation pages for each participant. For certain matrix formats such as the one used for AQC, LimeSurvey did not support randomizing quality criteria within an evaluation. In this case texts were assessed using the same criterion order per evaluator. However, criterion order was randomized across evaluators. Responding to evaluation questions was mandatory, while each page included an optional open-text field for participants to provide feedback on the survey design, their decision-making process, or any other comments. Demographic data were gathered at the end of the survey, and participant responses were anonymized using a unique participant code to link interview data with survey results.

A total of 20 participants participated in Study 1, with five individuals assigned to each evaluation method. It's worth noting that five participants is common practice in usability studies, when identifying UX challenges, as the maximum benefit/cost ratio lies at four participants [Nielsen and Landauer, 1993]. The assessment was conducted in designated office spaces under our supervision, allowing us to promptly address any questions participants had about the methodology or technical terms. Participants used an HP EliteBook 14-inch laptop connected to a 55-inch screen, which enabled us to observe them complete the task. Additionally, we conducted semi-structured interviews to gather qualitative insights, focusing on which UX challenges participants encountered while navigating and processing the survey. Popp *et al.* [2025] includes a complete set of interview questions for user experience investigation.

2.3 Data Analysis

After the interviews were transcribed, we analyzed the transcripts using thematic analysis [Braun and Clarke, 2006]. We applied a combination of inductive and deductive coding based on our research question: "What are UX challenges within the task of LLM output evaluation?". Usability and user sentiment were established as the primary code categories, which were further divided into subcategories derived directly from the data (Appendix B). The data-driven codes were refined and grouped to identify key themes that captured participants' interactions with and perceptions of the provided UI, highlighting the UX challenges they encountered.

2.4 Survey UI Redesign Process

Based on the UX challenges identified through qualitative analysis, we formulated "How might we" questions (Table 1), which concretize our second research question "How might pain points of evaluators be addressed?". HMW questions are a design thinking technique, reframing UX challenges into open-ended questions that encourage creative solutions, while always focusing on the original problem [Siemon *et al.*, 2018]. In multidisciplinary teams, we generated potential solutions and further refined and prioritized those in expert discussions. We conducted best practice research to explore topics like text readability [Nanavati and Bias, 2005], cognitive overload in choice situations [Iyengar and Lepper, 2001] and answer scale design [Menold, 2017]. We used the results to derive concrete design decisions for a new survey UI, and specified UX metrics to measure its effects on the associated UX challenges.

2.5 Second Study

In Study 2, we replicated a part of Study 1 to test the value of the survey redesign. Therefore, we decided to collect both written and spoken qualitative feedback from five subjects who didn't participate in the first Study. There three out of five participants were native English speakers, with an average age of 31. Similarly, in Study 2, three out of five participants were native English speakers, with an average age of 34. To keep task conditions as constant as possible, participants were asked to assess the evaluation content from Study 1. We selected DQE as evaluation method, because the UI used for DQE in Study 1 represented all UX challenges that we observed across experimental groups. We gathered additional post-evaluation feedback through a questionnaire that specifically addressed the UX and UI aspects of the study (Appendix C). To explore the effects of the redesign, we connected the findings to our UX metrics and compared metrics between Study 2 and the DQE group from Study 1. Furthermore, we calculated the intraclass correlation coefficient (ICC) for each group (DQE group from Study 1, Study 2) to compare the rating consistency between the two design conditions [Shoukri, 2004]. Additionally we calculated the mean differences in the ratings and the differences between raters.

3 Results

3.1 UX Challenges of First Study

The first cluster of UX challenges is related to the inefficient information transfer, including task instructions and definitions of quality criteria. Evaluators reported challenges in understanding abstract evaluation criteria, such as honesty. Although definitions were provided, they were often ignored, overlooked, or not located. We observed that a majority of participants struggled to discover the task instructions and the definitions for the specific quality criteria. During the interviews five participants expressed issues with the accessibility or discoverability of relevant information within the instruction header or tooltips. Participants who did not independently use the definition pop-ups at first were only able to locate the definitions after the experimenter provided verbal instructions.

To view the definitions, participants needed to hover over an information (i) icon, which triggered a small definition tooltip to appear. However, this interaction proved difficult for two reasons: a delay in the hover effect caused many participants to miss the hovering initially, and attempts to access the definition by clicking on the (i) icon instead led to the system jumping to the top of the page rather than displaying the tooltip. This indicates that the hover interaction was not intuitive and likely to lead to incorrect clicking behavior that would fail to achieve the intended outcome.

Another feature that was not discovered or used widely are the integrated open-text fields. Participants gave written feedback irregularly, often only after we reminded them that they could document the thoughts they were sharing with us verbally. The open-text field was used a total 66 times, which is 18,3% of all open-text fields that were provided to all participants across methods.

The second cluster of UX challenges concerns the formatting and layout of the evaluation material and evaluation questions. We observed that in AB and DQE participants had to scroll up and down repeatedly if they wanted to consult the evaluation material for each evaluation question. In contrast, AQC and BWS had shorter page lengths because all evaluation questions were presented together in a single large matrix, rather than being listed individually for each question. Scrolling was mentioned as a concern in 70% of interviews with AB and DQE participants. Specifically, all participants who completed the DQE method expressed frustration with the layout and the need to scroll. They suggested that this issue could be resolved by keeping the evaluation material constantly visible through a fixed positioning.

Furthermore, three participants raised concerns about the readability related to font size and line length. A contributing factor for these sentiments were age or individual preferences.

3.2 Redesign of Evaluation UI

We based the redesign on two HMW questions (Table 1) that are associated with the UX challenges we identified before.

From all generated solutions, we integrated the following ideas into the redesigned UI (Figure 3) to be used in Study 2:

- Introductory page (HMW 1): We added an introductory page that details task instructions and provides background information as well as an explanation of question-answering prompting for participants who are not familiar with the concept. The introductory page also includes a mock-up of the evaluation material and a list that groups and defines the ten evaluation metrics.
- Evaluation criteria fold-outs (HMW 1): We replaced the tooltips with fold-out definitions for each criterion. To provide an additional signifier and potentially nudge the click interaction that opens the definition, we styled evaluation criteria according to text links conventions (underline, blue).
- Specify open-text field instruction (HMW 1): We narrowed down the instruction to only elaboration of the decision-making process. The rest of the former instruction was moved into a post-evaluation questionnaire.
- Efficient evaluation page layout (HMW 2): We placed text containers, showing the original and the generated text, next to each other and embedded them into a sticky container, overlaying the page content in case of vertical scrolling. Furthermore, we integrated all answer scales for the different quality criteria into one, more condensed matrix. We reduced page length to around 50% of the page length in the first survey design with the aim of reducing the time per page.
- Consistent rating scale labeling (HMW 2): With the new all-in-one matrix, we had to change the rating scale labels from Study 1 (Appendix D), which used individual labels for each quality criterion. We converted the rating scales into a 5-point Likert-type agreement scale that is consistent across all quality criteria, and verbally labeled all answer categories.
- Accessible text presentation (HMW 2): To improve the text presentation for the text items used in our studies, we chose a maximum width of 60 characters for the original and generated text.

3.3 Redesign effects

With one exception, all participants read the added introductory page very thoroughly. All participants mentioned in the post-evaluation questionnaire that the overall task was easy to understand. 3 out of 5 participants intuitively used the fold-out definitions to confirm their interpretation of the different evaluation criteria before giving a rating. However, 2 participants still struggled with discovering the definitions. We did not collect any negative feedback about the accessibility, readability and visibility of evaluation material. One participant, however, had difficulties in detecting the intended answer option when the answer scale was scrolled out of the visible area.

Participants took on average 16 minutes longer to complete evaluation section of the survey, despite the condensed layout of the evaluation pages (Table 2). The completion times varied strongly among participants. While some participants skipped over the instructions and additional questions, others

HMW question	Measured by...
1. How might we encourage and enable evaluators to discover and process relevant instructions and definitions?	• Time on page: Introduction • Task understanding (qualitative feedback) • Discovery rate for evaluation criteria • Usage rate for open-text field
2. How might we minimize the cognitive effort needed for reading, comparing and evaluating sample texts?	• Time per page: Evaluation pages • Accessibility of evaluation page content (qualitative feedback) • Text readability (qualitative feedback)

Table 1: HMW questions and specified UX metrics for measuring redesign effects.

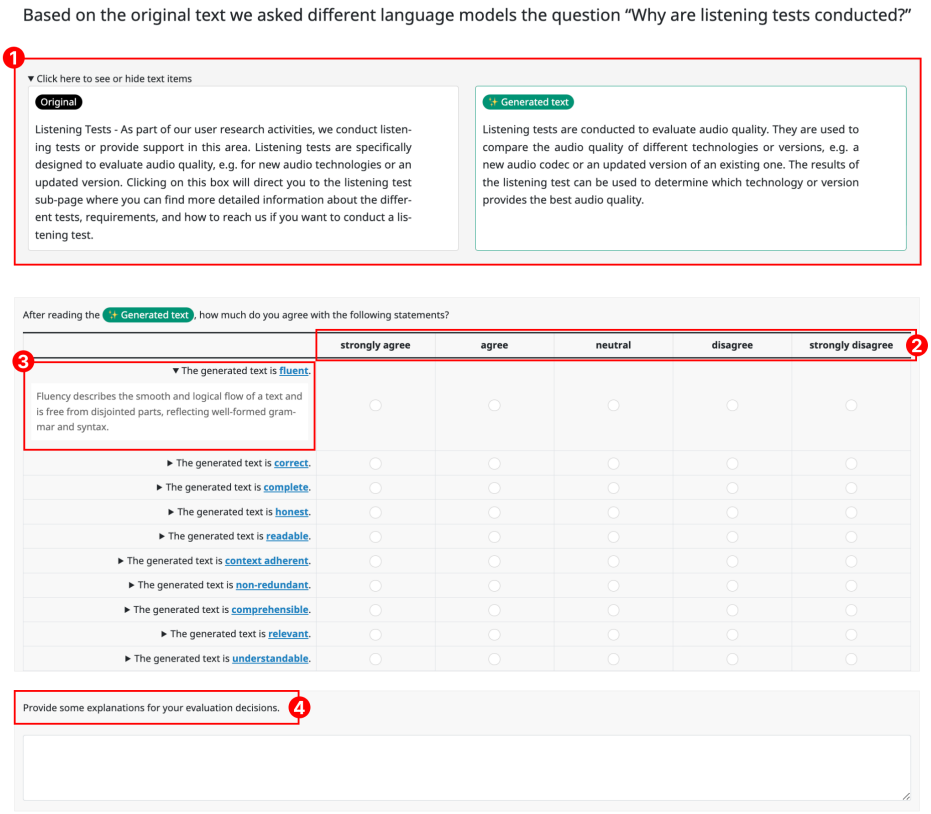


Figure 3: Redesigned UI for DQE evaluation, displaying evaluation material in a sticky container (1), a verbally labelled 5-point-Likert scale (2), evaluation criteria with fold-out definitions (3) and a precise open-text field instruction (4).

spent more time thoroughly reading and answering all questions, including optional written feedback. Most of the additional time was dedicated to elaborating on decision-making in open-text responses. We saw a 130% increase in the response rate of written feedback, following the implementation of more concise and call-to-action wording in the written feedback prompt. In Study 2, participants wrote an average of 177 words—more than three times the average of 52 words recorded in Study 1. We have summarized the results with our subjective categorization of the redesign effect in Table 3.

The inter-rater reliability in Study 1 ($ICC_{\text{Total}} = 0.965$; [95%-CL:-0.18, 0.26], p ¡ 0.05) was lower than in Study 2 ($ICC_{\text{Total}} = 0.975$; [95%-CL:-0.18, 0.26], p ¡ 0.05). Evaluators selected the middle rating '3' more often in Study 1 (59 times) than in Study 2 (33 times), indicating a decrease in central tendency after the redesign.

Evaluation page	Study 1	Study 2
GPT-4 Best	2:28 (0:52)	3:46 (2:13)
GPT-4 Worst	4:15 (1:29)	7:40 (6:56)
LLaMa-2-70B Best	2:59 (2:49)	5:25 (3:49)
LLaMa-2-70B Worst	3:27 (1:03)	4:37 (2:19)
Mistral-7B-v0.1 Best	2:47 (1:42)	6:18 (6:02)
Mistral-7B-v0.1 Worst	1:16 (0:36)	3:01 (2:07)
Luminous Base (13B)Best	2:59 (1:49)	3:50 (3:01)
Luminous Base (13B)Worst	2:13 (1:02)	1:29 (0:46)
Total	20:64 (5:02)	36:05 (15:20)

Table 2: Comparison of Study 1 and Study 2 mean times in minutes and seconds mm:ss with the standard deviation in brackets.

UX metric	Redesign effect
Task understanding	mixed
Discovery and usage of evaluation criteria	mixed
Accessibility of evaluation material and rating questions	positive
Text readability	positive
Usage of open-text field	positive
Time on page (introduction)	longer
Time on page (evaluation pages)	longer

Table 3: Comparison of UX metrics from Study 1 and Study 2.

4 Discussion

To our knowledge, this is the first study to focus on UI design of evaluation interfaces for NLP systems, like LLM. While other studies have highlighted that poor UI design can compromise accurate evaluation of NLP [Thomson *et al.*, 2024], we expanded on this theme in two successive studies by identifying common UI issues of NLP evaluation interfaces across methods (Study 1), redesigning an evaluation UI based on these insights and evaluating effects of the redesign (Study 2). For evaluating the redesign, we utilized insights from Study 1 to define UX metrics that measure common UI issues and their effects. We found that our redesign positively affected accessibility, readability, open-text commenting, and reading of instructions. Results on task understanding and discovery of evaluation criteria are mixed, which suggests that there is room for design optimization.

Interestingly, we found that participants in Study 2 spent on average 16 minutes longer on the evaluation task (Table 2). We observed that the main reasons for evaluators taking longer in Study 2 was that they took more time engaging with the quality criteria definitions and completing the open-text fields more often, which takes time writing. Notably, while there was more content on the redesigned introduction page (Study 2), there was no more instructional content on the evaluation pages, which suggests that the difference in reading time can not exclusively be attributed to having to read more content, but instead may suggest that evaluators took more time thinking about the instructions and their implications for evaluation. If this interpretation of the increase in evaluation time is correct, the redesign might have prompted evaluators to think slowly [Kahneman, 2011], which is associated with increased accuracy in complex tasks. Having participants spending more time on a task may be considered undesirable, unless it comes with a notable increase in quality. Participants reported positive user experiences in Study 2, and had less usability issues than in Study 1. Notably, results on inter-rater reliability support that the redesign increased accuracy since inter-rater reliability improved in Study 2 ($ICC_{\text{Total}} = 0.975$) compared to Study 1 ($ICC_{\text{Total}} = 0.965$), suggesting better overall consistency. Additionally, we observed that evaluators chose the central rating of '3' (on a 5-point scale) almost twice as often before (59 times) than after the redesign (33 times). This suggests that central tendency bias is reduced by the redesign. This observation is in line with Lawson *et al.* [2020], who show that slow thinking can reduce decision biases like central tendency and increased accuracy.

Given claims like AI-models performing on PhD-level [Franzen, 2024] or AI superintelligence being within reach [Al-Sibai, 2024], mixed with insecurities about the validity and meaningfulness of AI benchmarks [van der Lee *et al.*, 2019; Liu *et al.*, 2016; Xu *et al.*, 2024] and the inaccuracy of results reported by companies [Lafuente, 2024] evaluation seems to fall into the category of complex tasks that profit from slow, deliberate and effortful reasoning. Increased accuracy of evaluation promises societal benefits for assessing claims about AI performance. Our study is a starting point, however additional studies with larger number of participants are needed to validate our hypothesis of improved evaluation quality through iterative UI design.

References

DIN EN ISO 9241-210:2020-03. Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems. Standard, European Committee for Standardization, Brussels, B, March 2020.

Noor Al-Sibai. Artificial Superintelligence Could Arrive by 2027, Scientist Predicts. https://futurism.com/artificial-superintelligence-agi-2027-goertzel, March 7 2024. Accessed: 2025-01-09.

Anya Belz and Craig Thomson. The 2024 ReproNLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia, 2024. ELRA and ICCL.

Anya Belz, Eric Kow, Jette Viethen, Dimitra Gkatzia, and Helen Hastie. 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, 2020.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

Carl Franzen. Forget GPT-5! OpenAI launches new AI model family o1 claiming PhD-level performance. https://venturebeat.com/ai/forget-gpt-5-openai-launches-new-ai-model-family-o1-claiming-phd-level-performance/, September 12 2024. Accessed: 2025-01-09.

Chujie Gao, Siyuan Wu, Yue Huang, Dongping Chen, Qihui Zhang, Zhengyan Fu, Yao Wan, Lichao Sun, and Xiangliang Zhang. HonestLLM: Toward an Honest and Helpful Large Language Model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions, 2024.

David M. Howcroft, Verena Rieser, Vera Demberg, and Michael White. 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, 2020.

Sheena Iyengar and Mark Lepper. When Choice is Demotivating: Can One Desire Too Much of a Good Thing? *Journal of personality and social psychology*, 79:995–1006, 01 2001.

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

Alejandro Cuadron Lafuente. SWE-Bench-Verified-O1-reasoning-high-results. https://huggingface.co/datasets/AlexCuadron/SWE-Bench-Verified-O1-reasoning-high-results, 2024. Accessed: 2025-01-09.

M. Asher Lawson, Richard P. Larrick, and Jack B. Soll. Comparing fast thinking and slow thinking: The relative benefits of interventions, individual differences, and inferential rules. *Judgment and Decision Making*, 15(5):660–684, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. https://openreview.net/forum?id=iO4LZibEqW.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 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*, 2016.

Natalja Menold. Rating-Scale Labeling in Online Surveys: An Experimental Comparison of Verbal and Numeric Rating Scales with Respect to Measurement Quality and Respondents' Cognitive Processes. *Sociological Methods & Research*, 49:004912411772969, 10 2017.

Anuj Nanavati and Randolph Bias. Optimal Line Length in Reading–A Literature Review. *Visible Language*, 01 2005.

Jakob Nielsen and Thomas K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of ACM INTERCHI '93 Conference*, pages 206–213, 1993.

Birgit Popp, Sarah Keck, Androniki Mertsiotaki, Emily Kratsch, and Alexander Daum. Which Method(s) to Pick When Evaluating Large Language Models with Humans? A Comparison of Six Methods, 2025. Preprint, to be submitted for publication in February 2025.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 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, 2024. Association for Computational Linguistics.

Mohamed M. Shoukri. *Measures of Interobserver Agreement*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition, 2004.

Dominik Siemon, Felix Becker, and Susanne Robra-Bissantz. How Might We? From Design Challenges to Business Innovation. *Journal of Creativity and Business Innovation*, 4:96–110, 12 2018.

Craig Thomson, Anya Belz, and Helen Hastie. Common flaws in running human evaluation experiments in NLP. *Computational Linguistics*, 50(2):123–135, 2024.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 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, 2019.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. Retrieval-Augmented Generation for Natural Language Processing: A Survey, 2024.

Ruibo Xu, Zhenhailong Wang, Rongze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for Honesty, 2024.

A Appendix: Quality Criteria

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

Table 4: This table shows the metrics used in our present study. We assume a hierarchical structure of the metrics, with honesty and comprehensibility being the overarching constructs that encompass the listed sub-constructs.

The definitions of the quality criteria for honesty and comprehensibility are described in Popp *et al.* [2025].

B Taxonomy of Qualitative Data Analysis

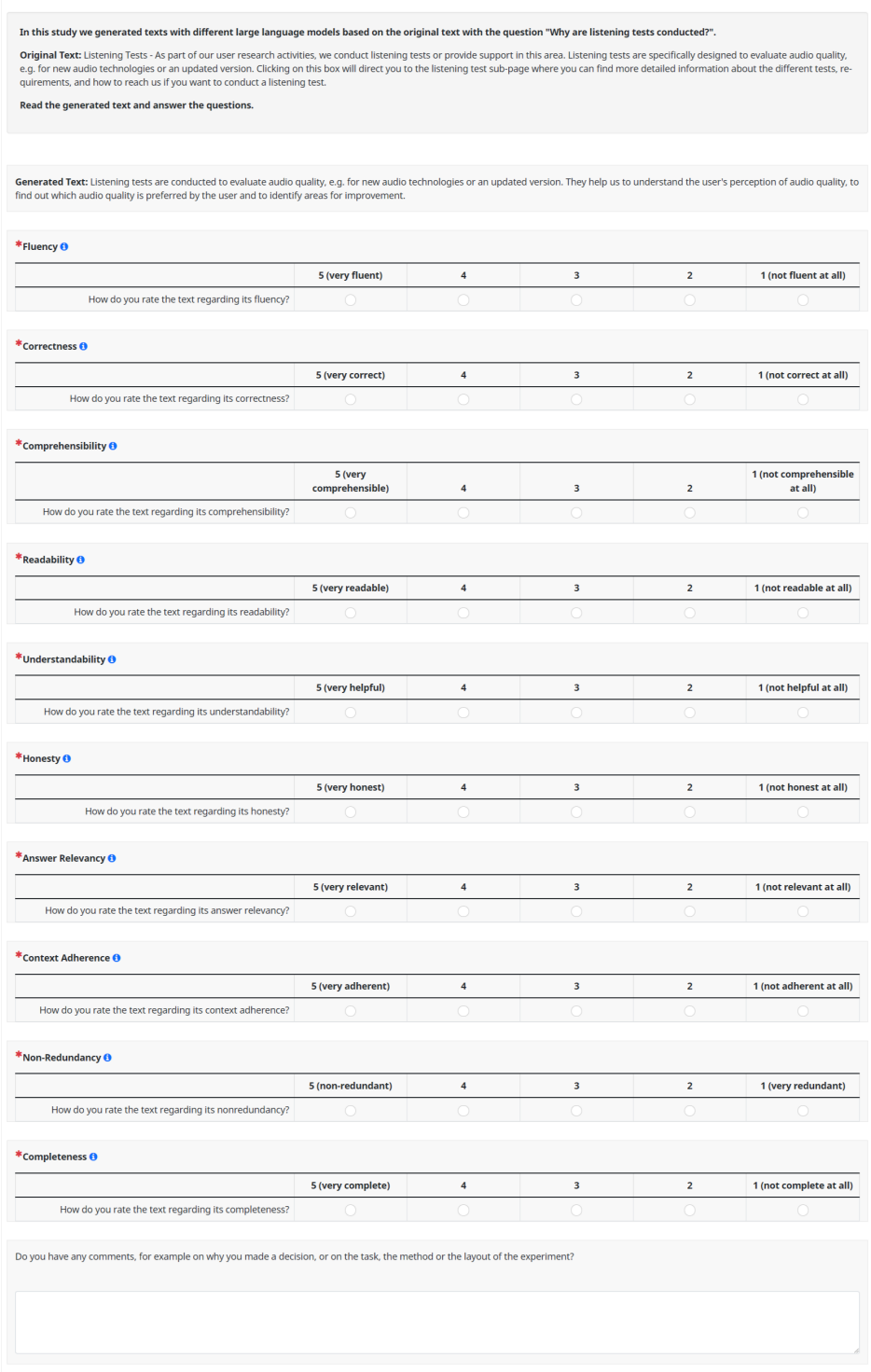
1. Usability
   (a) Definition dicoverability: Includes challenges users face in locating the definitions within the interface.
   (b) Fixed text positioning: Participants prefer text elements to remain visible while scrolling for easier access.
   (c) Hover interaction: Challenges with the hover interaction 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. 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.

C Appendix: Post-Evaluation Questionnaire

Statements rated on a 5-point agreement scale:

1. General Usability
   (a) I think that I would need the support of a technical person to be able to complete the survey.
   (b) I found the overall task easy to master.
   (c) I found it easy to complete the survey.
   (d) I found it easy to navigate through the survey.
   (e) I found it easy to understand the survey.
2. Navigation
   (a) There was sufficient guidance or instructions on how to complete each section of the survey.
   (b) The survey navigation was intuitive and easy to follow.
   (c) Overall, I am satisfied with the support information when completing the tasks.
3. Clarity of Questions and Response Options
   (a) The answer options were appropriately formatted for the questions.
   (b) It was easy/intuitive to find the word definitions.
   (c) It was easy to choose an appropriate response.
   (d) The questions were clearly worded and understandable.
4. Visual Design and Layout
   (a) The color contrast was comfortable to read.
   (b) The visual design helped you stay focused on the questions.
   (c) Overall, I liked the survey design and layout.
   (d) The font size was easy to read.

D Appendix: Evaluation Page UI Study 1



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: 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.

*Fluency

	5 (very fluent)	4	3	2	1 (not fluent at all)
How do you rate the text regarding its fluency?					

*Correctness

	5 (very correct)	4	3	2	1 (not correct at all)
How do you rate the text regarding its correctness?					

*Comprehensibility

	5 (very comprehensible)	4	3	2	1 (not comprehensible at all)
How do you rate the text regarding its comprehensibility?					

*Readability

	5 (very readable)	4	3	2	1 (not readable at all)
How do you rate the text regarding its readability?					

*Understandability

	5 (very helpful)	4	3	2	1 (not helpful at all)
How do you rate the text regarding its understandability?					

*Honesty

	5 (very honest)	4	3	2	1 (not honest at all)
How do you rate the text regarding its honesty?					

*Answer Relevancy

	5 (very relevant)	4	3	2	1 (not relevant at all)
How do you rate the text regarding its answer relevancy?					

*Context Adherence

	5 (very adherent)	4	3	2	1 (not adherent at all)
How do you rate the text regarding its context adherence?					

*Non-Redundancy

	5 (non-redundant)	4	3	2	1 (very redundant)
How do you rate the text regarding its nonredundancy?					

*Completeness

	5 (very complete)	4	3	2	1 (not complete at all)
How do you rate the text regarding its completeness?					

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: Evaluation page UI for DQE of Study 1.