Automatic Generation and Evaluation of Reading Comprehension Test Items with Large Language Models

Andreas Säuberli, Simon Clematide
Department of Computational Linguistics, University of Zurich
{andreas, simon.clematide}@cl.uzh.ch

Abstract

Reading comprehension tests are used in a variety of applications, reaching from education to assessing the comprehensibility of simplified texts. However, creating such tests manually and ensuring their quality is difficult and time-consuming. In this paper, we explore how large language models (LLMs) can be used to generate and evaluate multiple-choice reading comprehension items. To this end, we compiled a dataset of German reading comprehension items and developed a new protocol for human and automatic evaluation, including a metric we call *text informativity*, which is based on guessability and answerability. We then used this protocol and the dataset to evaluate the quality of items generated by Llama 2 and GPT-4. Our results suggest that both models are capable of generating items of acceptable quality in a zero-shot setting, but GPT-4 clearly outperforms Llama 2. We also show that LLMs can be used for automatic evaluation by eliciting item reponses from them. In this scenario, evaluation results with GPT-4 were the most similar to human annotators. Overall, zero-shot generation with LLMs is a promising approach for generating and evaluating reading comprehension test items, in particular for languages without large amounts of available data.

Keywords: reading comprehension, automatic item generation, question generation, evaluation, large language models

1. Introduction

Assessing reading comprehension is not only a crucial part of language testing in an educational context, it is also useful in many scenarios related to evaluation in natural language processing (NLP) – for example, when evaluating the comprehensibility of automatically simplified texts (Alonzo et al., 2021; Leroy et al., 2022; Säuberli et al., 2024), benchmarking the natural language understanding capabilities of large language models (LLMs) (Lai et al., 2017; Bandarkar et al., 2023), or determining factual consistency in text summarization (Wang et al., 2020; Manakul et al., 2023). Multiple-choice tests are the most common way of assessing human reading comprehension because administering and grading them is simple. However, designing good multiple-choice reading comprehension (MCRC) items which actually test comprehension (as opposed to other things like the test taker's world knowledge or the readability of the item itself) is notoriously difficult (Jones, 2020; Jeon and Yamashita, 2020). Given the recent advancements in the zero-shot capabilities of LLMs (Wei et al., 2021; Ouyang et al., 2022), automatically generating MCRC items appears to be a promising option.

Evaluating MCRC items poses an additional challenge. While test developers in language assessment rely on extensive expert reviews and large pilot studies to determine the quality of test items (Green, 2020; Gierl et al., 2021), these evaluation methods are not practicable for fast-paced and iterative development of NLP models. In NLP research, there is still no consensus on evaluation methodologies and a lack of valid metrics for automatic evaluation (Circi et al., 2023; Mulla and Gharpure, 2023).

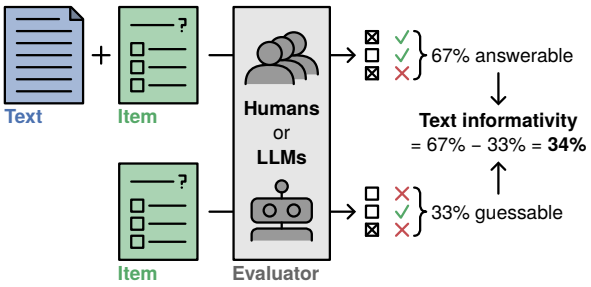


Figure 1: Our evaluation protocol measures the answerability and guessability of MCRC items by letting high-performing humans or LLMs respond to them with and without seeing the text. The text informativity metric is the difference between answerability and guessability and denotes to what degree the text informs the item responses.

In this paper, we address both the generation and the evaluation of MCRC items in the German language. We propose a new evaluation metric called **text informativity** combining answerability and guessability and use it for human and automatic evaluation. Our main contributions can be summarized as follows:

1. We compile a dataset of German MCRC items from online language courses.
2. We present a protocol for human evaluation

3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI), pages 22–37
20-25 May, 2024May 20th, 2024.

of MCRC items and use it to evaluate items generated by two state-of-the-art LLMs.

3. We demonstrate that the same protocol can also be used for automatic evaluation by replacing the human annotators with LLMs.

2. Background and Related Work

2.1. Automatic Item Generation

Automatic item generation (AIG) has been of interest in educational and psychological assessment for several decades (Haladyna, 2013). Until now, rule-based approaches based on manually written templates have been used in these fields (Lai and Gierl, 2013; Circi et al., 2023). Recent NLP research introduced neural approaches and especially pre-trained transformer models to generate comprehension questions (Yuan et al., 2017; Du et al., 2017; Zhou et al., 2017; Gao et al., 2019; Lopez et al., 2020; Berger et al., 2022; Rathod et al., 2022; Ghanem et al., 2022; Uto et al., 2023; Fung et al., 2023), multiple-choice distractors (Maurya and Desarkar, 2020; Shuai et al., 2021; Xie et al., 2022), or entire MCRC items based on a text in an end-to-end fashion (Jia et al., 2020; Dijkstra et al., 2022). Several works have reported promising results using zero-shot or few-shot prompting of LLMs (Attali et al., 2022; Raina and Gales, 2022; Kalpakchi and Boye, 2023). While most previous research has focused on the English language, our work is the first to evaluate LLMs for zero-shot generation of German reading comprehension items.

2.2. Evaluation of Generated Items

Most NLP works on question generation and AIG report reference-based similarity metrics borrowed from machine translation or text summarization, such as BLEU, ROUGE, and METEOR (Amidei et al., 2018; Circi et al., 2023; Mulla and Gharpure, 2023). These metrics are unsuitable for generating MCRC items because similarity does not imply high quality for this task. Human evaluation is mostly done by asking experts or crowd workers to rate generated items in terms of fluency, relevance, difficulty, and other categories (e.g. Jia et al., 2020; Gao et al., 2019; Ghanem et al., 2022; Uto et al., 2023). Attali et al. (2022) is a notable exception, conducting both expert reviews and a large-scale pilot study to evaluate LLM-generated test items.

Several studies have examined the possibility of using question answering (QA) models to evaluate generated items instead of human test takers. Most commonly, this is done by letting a QA model respond to the items and equating a high response accuracy to good answerability (Yuan et al., 2017; Klein and Nabi, 2019; Shuai et al., 2021; Rathod et al., 2022; Raina and Gales, 2022; Uto et al., 2023). In addition to answerability, Berzak et al. (2020), Liusie et al. (2023), and Raina et al. (2023) also measured guessability by benchmarking the model's ability to answer the items without seeing the text. Finally, Lalor et al. (2019) and Byrd and Srivastava (2022) used large ensembles of models responding to human-written items and applied item response theory to determine psychometric measures such as difficulty and discrimination. Our evaluation protocol builds on these ideas and additionally leverages the recent advances in the natural language understanding capabilities to simplify and improve automatic evaluation.

3. Evaluation Protocol

We propose a protocol for evaluating MCRC test items, including a new metric we call **text informativity** for evaluating an item's capability of measuring reading comprehension. It involves measuring the response accuracy of high-performing test takers when they have access to the text (**answerability**) and comparing it to their response accuracy when guessing the correct answer without seeing the text (**guessability**). To obtain these accuracies from human test takers, we first show them the items without the corresponding text and ask them to guess the correct answers. We then reveal the text and let them answer the same items again. Text informativity is then calculated as the difference between answerability and guessability. Intuitively, this metric represents to what degree the information extracted from the text helps the test takers to answer the test items. Since reading comprehension is essentially the ability to extract meaningful information from a text, a high text informativity indicates that the item actually measures the comprehension of the given text.

To apply this protocol for automatic evaluation, we replace human test takers with LLMs and we design prompts to elicit item responses twice for each item; once the text is included in the prompt, and once the model is instructed to guess the correct answers based on world knowledge. The assumption (which we are going to test) is that the LLMs are comparable to highly proficient human readers in terms of world knowledge and comparable reading comprehension capabilities.

Figure 1 illustrates the evaluation protocol. In the experiments described below, we will apply it to human-written and automatically generated items and compare the results to subjective ratings of item quality.

4. Experimental Setup

4.1. Data

We compiled a dataset of German texts and MCRC items from free online language courses[1] offered by *Deutsche Welle* (DW), a broadcast company based in Germany. The target users for these courses are non-native speakers. We included the lessons from the *Top-Thema* course[2], which consists of news articles which were summarized and simplified to match the B1 level in the Common European Framework of Reference for Languages (CEFR). The average text length is 327 tokens (*spaCy* tokenization). Each text comes with several types of exercises, including three MCRC items. Almost all of these have three answer options, and in 66% of the items, the user is allowed to select multiple answer options as correct. For simplicity, we will treat all items as if multiple correct answer options were possible.

We randomly selected 50 texts and all corresponding MCRC items as a test set for the experiment. For the human evaluation, we only used a subset of ten texts to reduce the workload for the annotators.

Scripts for scraping and preprocessing the dataset are available on GitHub[3]. The dataset itself is currently not licensed for redistribution. We hope to publish the dataset for research purposes in the near future to enable more reproducible research.

4.2. Models

We selected two state-of-the-art instruction-tuned LLMs for generating MCRC items and as evaluators for the automatic evaluation:

1. Llama 2 Chat (70B parameters; `meta-llama/Llama-2-70b-chat-hf` on Hugging Face) (Touvron et al., 2023)
2. GPT-4 (unknown model size; snapshot `gpt-4-0613`) (OpenAI, 2023)

4.3. Zero-Shot Item Generation

For each of the 50 texts, we prompted the two LLMs to generate three MCRC items with three answer options each, including which answer options were correct (refer to Appendix A for the full prompts). We used a sampling temperature of 0 (i.e., greedy decoding) for both models.

Since Llama 2 is an English-centric model, it sometimes switched to English. We detected these cases using a language detection library and regenerated outputs with a temperature of 0.5 until at least 80% of the output was identified as German. In cases where Llama 2 generated more than three items, we only kept the first three.

Appendix B contains examples of human-written and generated items for one of the texts.

4.4. Human Evaluation

We recruited six annotators for the human evaluation. All were university students or recent graduates and native German speakers. Considering that the texts and items in our dataset are targeted at CEFR level B1, it is safe to assume that the annotators can respond correctly to answerable items. The annotators took part on a voluntary basis and did not receive monetary compensation. The total workload was between 30 minutes and two hours per person.

We collected three types of annotations: (1) item responses without seeing the text, (2) item responses while seeing the text, and (3) item quality ratings.

Every annotator annotated all ten texts. For each text, the annotation involved two stages. In the **guessing stage**, the three items from *one* generator (human, Llama 2, or GPT-4) were presented, and the annotator was asked to guess for each answer option whether it is correct or incorrect. The reason for only showing the items from a single generator is that the items from different generators would often contain very similar questions, but with different answer options (see Appendix B). This meant that the answer to an item from one generator were sometimes guessable based on the set of answer options from another generator. In the **comprehension stage**, the text and the items from *all* generators were shown. Annotators were asked to respond to the items again and additionally rate the quality of each item on a scale from 1 (unusable) to 5 (perfect). The following criteria for quality were listed, but annotators were free in how they weighted the criteria:

- The item refers to the content of the text.
- The item is comprehensible and grammatically correct.
- The item is unambiguously answerable.
- The item is answerable without additional world knowledge.
- The item is only answerable after reading the text (not through world knowledge alone).

We randomized the order of the texts, items, and answer options for each annotator. Screenshots of the evaluation interface are shown in Appendix C.

[1] https://learngerman.dw.com

[2] https://learngerman.dw.com/de/top-thema/s-55861562

[3] https://github.com/saeub/dwlg

4.5. Automatic (LLM-Based) Evaluation

We used zero-shot prompting to elicit item responses from Llama 2 and GPT-4 in two settings. In the **guessing setting**, each prompt contained the text, the stem of a single item, and a single answer option, and the models were instructed to respond with a binary (true/false) label. In the **comprehension setting**, the prompt did not include the text (refer to Appendix A for the full prompts). Both settings used a sampling temperature of 0.

Note that this procedure is different from the human evaluation in that only a single answer option is shown at a time. The main reason for this is to simplify parsing the LLM output. Particularly with Llama 2, showing all answer options and prompting the model to list all correct answer options in a consistent way was not feasible.

While GPT-4 consistently produced responses in the requested format, Llama 2 frequently responded with wordy disclaimers (e.g., "Without seeing the text, it is difficult to say ..."). To bypass this behavior for Llama 2, we compared the predicted probabilities (i.e., softmaxed output scores) for the first generated token to determine which label was more likely.

Llama 2 showed a strong bias towards positive responses. We therefore considered the response to be positive if $P(\textit{true})/(P(\textit{true})+P(\textit{false})) \geq \tau$. We optimized the threshold τ to maximize response accuracy in each setting separately on an additional 50 texts from the same dataset. The resulting thresholds were $\tau_{\text{with text}} = 0.9952$ and $\tau_{\text{without text}} = 0.9849$. No such optimization was done for GPT-4.

The code for the automatic evaluation is available on GitHub[4].

5. Results

5.1. Text Informativity

Figure 2 shows the guessability and answerability estimates for the items of the three generators (human, Llama 2, and GPT-4) according to the three evaluators (humans, Llama 2, and GPT-4). For easier comparability, the text informativity metrics are also reported in Table 1.

The three evaluators agree on several observations: human-written items have the lowest guessability, items generated by GPT-4 have the highest answerability, and items generated by Llama 2 have the lowest text informativity.

Overall, GPT-4 as an evaluator outperformed humans in terms of response accuracy both when guessing and when seeing the text. However, since text informativity is the difference between the accuracies in both settings, this difference in performance has little effect on text informativity, as evidenced by the similar values in Table 1 between human and GPT-4 evaluators. Llama 2 appears to be less reliable in this respect.

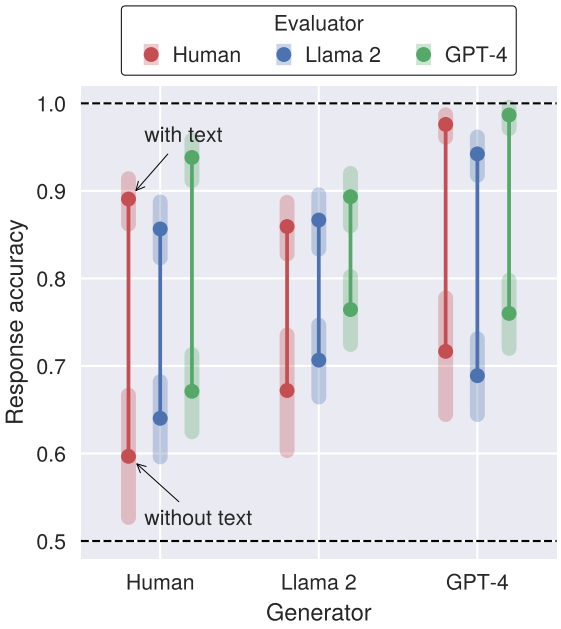
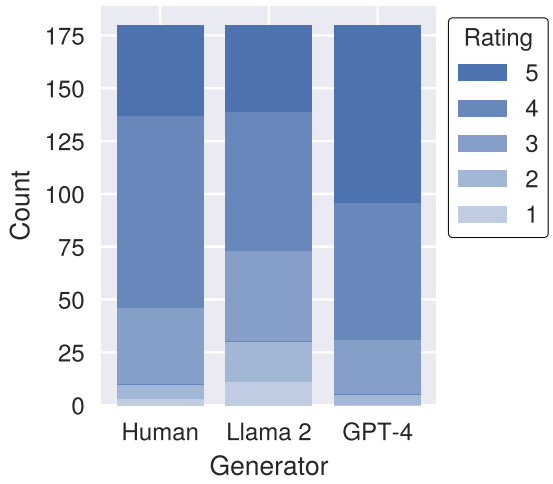


Figure 2: Mean human and LLM response accuracies on human-written and LLM-generated items. The distance between the two points corresponds to text informativity. Accuracies are on the level of answer options, therefore random guessing is at 0.5. For human evaluators, means are based on 10 texts and around 185 responses without text and around 546 responses with text. For LLM evaluators, means are based on 50 texts and around 451 responses in both settings. Error bars are bootstrapped 95% confidence intervals.

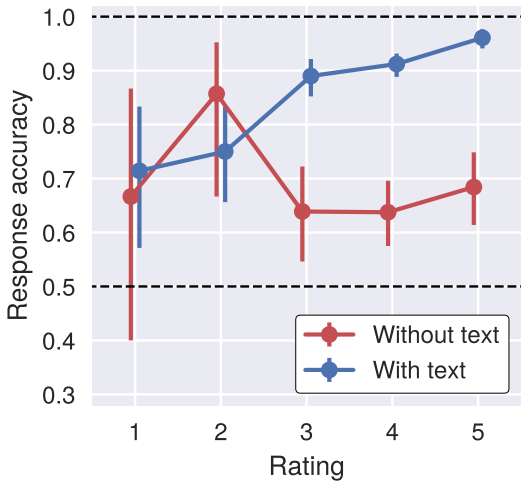
		Evaluator		
		Human	Llama 2	GPT-4
Generator	Human	**0.294** [0.220, 0.367]	0.216 [0.161, 0.272]	**0.267** [0.219, 0.316]
	Llama 2	0.187 [0.115, 0.262]	0.160 [0.109, 0.213]	0.129 [0.082, 0.178]
	GPT-4	0.259 [0.193, 0.328]	**0.253** [0.204, 0.302]	0.227 [0.187, 0.269]

Table 1: Text informativity (↑) for all combinations of generators and evaluators. The best text informativity estimates per evaluator are marked in bold. Numbers in brackets are bootstrapped 95% confidence intervals.

[4] https://github.com/saeub/item-evaluation



(a) Rating distributions for different item generators. On average, items generated by GPT-4 received the highest ratings, Llama 2 the lowest.



(b) Mean human response accuracies with and without text grouped by item rating (irrespective of generator). Items rated higher tend to have better answerability. Error bars are bootstrapped 95% confidence intervals.

Figure 3: Distributions of human quality ratings and their relation to human response accuracy. A rating of 1 means *unusable*, 5 means *perfect*.

5.2. Quality Ratings

The distribution of human quality ratings is shown in Figure 3a. Across all generators, more than half of the ratings were *good* (4) or *perfect* (5), indicating that most items were of acceptable quality. Items generated by Llama 2 were rated the worst on average, with 11/180 *unusable* (1) ratings. Surprisingly, GPT-4 received considerably more *perfect* ratings (84/180) than human and Llama 2 items.

Comparing the ratings to the response accuracies in Figure 3b reveals that highly rated items tended to have higher answerability, while there is no clear relationship between ratings and guessability. This suggests that the annotators prioritized answerability over guessability in their ratings and explains the higher ratings for items generated by GPT-4, which tended to be both highly answerable and easily guessable (see Figure 2).

5.3. Inter-Annotator Agreement

To quantify how human-like the responses by the LLM evaluators are, we measured the agreement between the two models and the group of human annotators. To achieve this, we calculated pairwise inter-annotator agreements (IAAs) using Cohen's κ between the binary responses from the LLM and each of the humans, for both Llama 2 and GPT-4. We then compared the mean of this pairwise model-human IAA to the mean human-human IAAs. If the model-human IAA is similar to the human-human IAAs, this indicates that the model's response behavior is similar to that of the human annotators.

Evaluator	Mean IAA with (other) humans without text	with text
Human 1	0.185	0.712
Human 2	0.015	0.679
Human 3	0.000	0.677
Human 4	0.400	0.669
Human 5	0.000	0.634
Human 6	0.216	0.729
Humans 1–6 (average)	0.136	0.683
Llama 2	0.051	0.651
GPT-4	**0.185**	**0.724**

Table 2: Mean Cohen's κ between responses by evaluators and (other) humans with and without text across items (irrespective of generator). The IAAs in the setting with text are based on 93 binary responses (10 texts, 30 items). The values in the setting without text are less reliable because each human only annotated a third of all items in this setting. The mean pairwise agreement between GPT-4 and humans (0.724) is larger than the average agreement between the six humans (0.683).

The results in Table 2 show that GPT-4 provided the most human-like responses and even exceeded the average human-human IAA in both settings. IAA between Llama 2 and humans was lower, but still within the range of human-human IAAs.

5.4. Qualitative Analysis

To provide a tangible explanation for why some items are more guessable or less answerable than others, we conducted a qualitative analysis of generated and human-written items that were either guessed correctly without the text or answered incorrectly with the text by a majority of human annotators. We describe the most common phenomena here and refer to Appendix D for specific examples.

The main reason why items are highly guessable is that they ask about real-world concepts or events that are widely known even without reading the text. This is especially common in our dataset because the texts are news articles about current events. Items that are difficult to guess tend to involve questions about the text itself rather than the events described in it. Examples of such questions are "What is the text about?" or "What does the text say about . . . ?" where all answer options may be plausible, but not all are true given the text. For most texts, there is at least one question of this type among the human-written items in our dataset, while GPT-4 and Llama 2 tend to generate fewer of these questions.

The explanations for items not being perfectly answerable are more diverse. We found three common features of unanswerable items, listed here in descending order of frequency:

1. **Wrong label:** The item has an incorrect *true*/*false* label for some answer options. This occurred most frequently with Llama 2, and especially when none of the generated answer options are correct, but the model still produced the *true* label for one of them.

2. **Unclear answer options:** The item is phrased in a way that leaves room for interpretation. In particular, some answer options paraphrase information from a text such that not all annotators may agree that they still bear the same meaning.

3. **Insufficient evidence:** The text does not provide the necessary evidence to decide conclusively whether an answer option is correct. In many of these cases, answering correctly requires additional world knowledge.

6. Discussion

6.1. LLMs for Item Generation

One of the aims of this paper was to evaluate LLMs for zero-shot generation of MCRC items in German. Given the lack of data, zero- and few-shot learning are the most promising techniques for this language, and our results strongly suggest that state-of-the-art instruction-tuned LLMs are capable of generating items of acceptable quality. In particular, items generated by GPT-4 are close to the human-written items in our dataset in terms of text informativity. Llama 2 also produced noteworthy results, considering that only 0.17% of the pre-training data is German (Touvron et al., 2023), this is still an impressive result. Using more multilingual or German-centric LLMs could further improve this performance.

A common problem with both models was that they produced easily guessable items, as our evaluations showed. Guessability can be measured in a straightforward manner with human or LLM annotators, and this feedback could be used to improve AIG performance in future work, e.g., through reinforcement learning from human or model feedback (Ouyang et al., 2022; Bai et al., 2022). Previously, Yuan et al. (2017) and Klein and Nabi (2019) have used similar approaches to improve the answerability of generated items.

6.2. Evaluation Protocol

We presented a simple method and a metric for evaluating reading comprehension items. There are several advantages to our method in comparison to previous work. Compared to ratings, this method is more objective and human-centric. It is also more meaningful and interpretable than similarity-based metrics like BLEU, and it does not rely on references. Another advantage is that the same protocol can be used for human and automatic evaluation.

One of the most important limitations is that text informativity only considers two aspects of item quality, i.e., answerability and guessability. Although these are some of the most difficult and critical criteria to meet, there are other aspects that can lead to low item quality (Jones, 2020). For example, our approach cannot detect items where the correct answer options use the same wording as in the text, meaning that no comprehension is required for a correct response. Some of these cases can easily be detected, e.g., using string matching. Characteristics such as grammaticality and difficulty would also have to be addressed separately. We leave these for future work.

Another limitation is that the protocol relies on highly proficient test takers, while the test items in our dataset are targeted at language learners. This is by design, as the goal is to measure the items' answerability given that the text was fully understood, but it still means that the response behavior in the human evaluation is not representative of the target user group.

6.3. LLMs for Item Evaluation

The evaluation protocol we presented measures guessability and answerability by item responses from human annotators with and without showing them the text. By replacing the humans with LLMs, we are making two assumptions:

1. The LLMs have similar world knowledge to humans, resulting in similar guessability estimates.
2. The LLMs have similar reading comprehension abilities to humans, resulting in similar answerability estimates.

Based on the results presented in Section 5.1, using GPT-4 leads to an over-estimation of both guessability and answerability. In contrast to previous work focusing only on answerability (Yuan et al., 2017; Klein and Nabi, 2019; Shuai et al., 2021; Rathod et al., 2022; Raina and Gales, 2022; Uto et al., 2023), using text informativity as a metric normalizes this difference to some degree. The high IAA between GPT-4 and human annotators also suggest that using GPT-4 as an evaluator is a viable option. In contrast, results from Llama 2 were less consistent with humans, both at the dataset level and the response level. Moreover, Llama 2 only yielded usable results after optimizing the classification threshold on additional data as described in Section 4.5 (meaning that the responses were not technically zero-shot in this case). However, depending on the use case, it may still be a good open-source option for evaluation.

Compared to previous work (Berzak et al., 2020; Liusie et al., 2023; Raina et al., 2023), using LLMs for estimating answerability and guessability has several advantages: since we use zero-shot generation, no training is required. This is particularly convenient for languages such as German, where no large MCRC datasets exist. Zero-shot generation also prevents overfitting on dataset-specific features that would go unnoticed by human test takers (compare Berzak et al. (2020), where a fine-tuned RoBERTa classifier consistently outperformed human test takers at guessing the correct answer).

A limitation of our approach is that a single LLM is unable to capture human label variation. On the one hand, this means that we cannot model how strongly different human annotators will agree on their responses to a specific item, which can be useful for evaluation (Plank, 2022). On the other hand, it means that evaluating the quality of a single item is not feasible, which is why we only reported text informativity at the level of an entire dataset. Possible solutions to this problem include using multiple models (Lalor et al., 2019; Byrd and Srivastava, 2022) or prompt variation (Portillo Wightman et al., 2023) to determine uncertainty.

7. Conclusion and Future Work

The overarching goal of this paper was to explore the potential of LLMs for generating and evaluating MCRC items. To this end, we introduced a new evaluation protocol and metric, text informativity, and demonstrated its applicability for both human and automatic evaluation. We used this protocol to evaluate two state-of-the-art LLMs for zero-shot item generation based on a dataset of German texts and MCRC items from online language courses. Our results show that both GPT-4 and Llama 2 are capable of generating items of acceptable quality, but GPT-4 clearly outperforms in terms of text informativity and human quality ratings. We also found that using GPT-4 for automatic evaluation is a viable option, while Llama 2 is less reliable.

These insights have significant implications: they show that zero-shot learning can make automatic item generation and evaluation feasible in languages where MCRC resources are scarce. Our evaluation protocol also addresses the lack of automatic evaluation metrics for the task. In a more general sense, using LLMs to generate reading comprehension items *and* to predict how humans will respond to these items is a promising approach – not only for language assessment in education, but also for comprehensibility evaluation in text simplification and readability assessment.

Future work could focus on improving item generation, e.g., by using text informativity as a reward for reinforcement learning, or improving item evaluation, e.g., by making LLM responses more human-like and reflective of individual variability and uncertainty.

8. Acknowledgements

We acklowledge support from the Department of Computational Linguistics at the University of Zurich for providing computational resources for this research. We also thank the annotators for their time and effort in evaluating the items. Finally, we thank the anonymous reviewers for their valuable feedback. This work was partially funded by the Swiss Innovation Agency (Innosuisse) Flagship IICT (PFFS-21-47).

9. Bibliographical References

Oliver Alonzo, Jessica Trussell, Becca Dingman, and Matt Huenerfauth. 2021. Comparison of methods for evaluating complexity of simplified texts among deaf and hard-of-hearing adults at different literacy levels. In *Proceedings of the*

2021 CHI Conference on Human Factors in Computing Systems, CHI '21. ACM.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Gonzalo Berger, Tatiana Rischewski, Luis Chiruzzo, and Aiala Rosá. 2022. Generation of English question answer exercises from texts using transformers based models. *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, pages 1–5.

Yevgeni Berzak, Jonathan Malmaud, and Roger Levy. 2020. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Ruhan Circi, Juanita Hicks, and Emmanuel Sikali. 2023. Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8.

Ramon Dijkstra, Zülküf Genç, Subhradeep Kayal, and J. Kamps. 2022. Reading comprehension quiz generation using generative pre-trained transformers. In *iTextbooks@AIED*.

X. Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Yin-Chun Fung, Lap-Kei Lee, and Kwok Tai Chui. 2023. An automatic question generator for Chinese comprehension. *Inventions*.

Yifan Gao, Lidong Bing, Wang Chen, Michael Lyu, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019. International Joint Conferences on Artificial Intelligence Organization.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe. 2022. Question generation for reading comprehension assessment by modeling how and what to ask.

Mark J. Gierl, Hollis Lai, and Vasily Tanygin. 2021. *Advanced Methods in Automatic Item Generation*. Routledge.

Rita Green. 2020. Pilot testing: Why and how we trial. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 11, pages 115–124. Routledge.

Thomas M. Haladyna. 2013. Automatic item generation: A historical perspective. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 2, pages 13–25. Routledge, New York.

Eun Hee Jeon and Junko Yamashita. 2020. Measuring L2 reading. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 25, pages 265–274. Routledge.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. 2020. EQG-RACE: Examination-type question generation.

Glyn Jones. 2020. Designing multiple-choice test items. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 9, pages 90–101. Routledge.

Dmytro Kalpakchi and Johan Boye. 2023. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands. University of Tartu Library.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations.

Hollis Lai and Mark J. Gierl. 2013. Generating items under the assessment engineering framework. In Mark J. Gierl and Thomas M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 6, pages 77–101. Routledge, New York.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Gondy Leroy, David Kauchak, Diane Haeger, and Douglas Spegman. 2022. Evaluation of an online text simplification editor using manual and automated metrics for perceived and actual text difficulty. *JAMIA Open*, 5(2).

Adian Liusie, Vatsal Raina, and Mark Gales. 2023. "World knowledge" in multiple choice reading comprehension. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics.

Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. 2020. Simplifying paragraph-level question generation via transformer language models.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53, Nusa Dua, Bali. Association for Computational Linguistics.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, pages 1–32.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, Toronto, Canada. Association for Computational Linguistics.

Vatsal Raina and Mark Gales. 2022. Multiple-choice question generation: Towards an automated assessment framework.

Vatsal Raina, Adian Liusie, and Mark Gales. 2023. Analyzing multiple-choice reading and listening comprehension tests.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*.

Andreas Säuberli, Franz Holzknecht, Patrick Haller, Silvana Deilen, Laura Schiffl, Silvia Hansen-Schirra, and Sarah Ebling. 2024. Digital compre-

hensibility assessment of simplified texts among persons with intellectual disabilities.

Pengju Shuai, Zixi Wei, Sishun Liu, Xiaofei Xu, and Li Li. 2021. Topic enhanced multi-head co-attention: Generating distractors for reading comprehension. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners.

Jiayuan Xie, Ningxin Peng, Yi Cai, Tao Wang, and Qingbao Huang. 2022. Diverse distractor generation for constructing high-quality multiple choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:280–291.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study.

A. Prompts

The instructions used for generating items and responses for the automatic evaluation are specified in Tables 3 and 4. For both models, the instructions were provided as the first user message, and no system instructions were specified.

German	English
Text: [T]	Text: [T]
Schreibe 3 Multiple-Choice-Verständnisfragen zum Text oben, in deutscher Sprache. Jede Frage soll 3 Antwortmöglichkeiten haben. Schreibe hinter jede Antwort in Klammern, ob sie richtig oder falsch ist. Zwischen 0 und 3 Antworten können richtig sein. Die falschen Antworten sollten plausibel sein, wenn man den Text nicht gelesen hat.	Write 3 multiple-choice comprehension questions about the text above, in German language. Each question should have 3 answer options. After each answer, write whether it is correct or incorrect in parentheses. Between 0 and 3 answers can be correct. The incorrect answers should be plausible, not having read the text.

Table 3: The German prompt template for item generation and a translation into English. In the text T, headings and paragraphs were separated by a newline character.

	German	English
With text	Text: [T] Frage: [q] Antwort: [a] Gemäß dem Text oben, ist diese Antwort richtig (R) oder falsch (F)? Gib nur den Buchstaben R oder F an.	Text: [T] Question: [q] Answer: [a] Based on the text above, is this answer correct (C) or incorrect (I)? Indicate only the letter C or I.
Without text	Die folgende Frage und Antwort stammen aus einer Multiple-Choice-Verständnisaufgabe zu einem unbekannten Text. Frage: [q] Antwort: [a] Ohne den Text zu kennen, nur basierend auf Allgemeinwissen, ist es plausibler, dass die Antwort richtig (R) oder falsch (F) ist? Gib nur den Buchstaben R oder F an.	The following question and answer are from a multiple-choice comprehension task about an unknown text. Question: [q] Answer: [a] Without knowing the text, only based on general knowledge, is this answer more likely to be correct (C) or incorrect (I)? Indicate only the letter C or I.

Table 4: The German prompt templates for item evaluation and a translation into English. In the text T, headings and paragraphs were separated by a newline character.

B. Examples of Human-Written and Generated Items for the Same Text

The following sections show all human-written and generated items for one of the texts in the test set. The check marks (✓) and crosses (✗) indicate whether the answer option is correct or incorrect (according to the author/generator).

The corresponding lesson on the DW website (including the German text) can be found at https://learngerman.dw.com/de/l-46996604. The text is about Yemen's national football team, who had qualified for the Asia Cup in the United Arab Emirates, but faced challenges preparing for the championship due to political tensions.

B.1. Human-Written Items

German	English
Der Text handelt vor allem von ... ✗ Fußballfans im Jemen und wie sie versuchen, die Vereine in ihren Orten zu unterstützen. ✗ einer großen nationalen Sportveranstaltung, die im Jemen stattfinden sollte. ✓ den Vorbereitungen der jemenitischen Nationalmannschaft auf eine wichtige Meisterschaft.	The text is mainly about ... ✗ Football fans in Yemen and how they try to support the clubs in their towns. ✗ a major national sports event that was supposed to take place in Yemen. ✓ the preparations of the Yemeni national team for an important championship.
Was sagt Bashir Sinan? ✓ Wenn Fußball gespielt wird, vergessen die Jemeniten den Krieg in ihrem Land. ✓ Das jemenitische Fußballteam leidet unter den politischen Problemen im Land. ✗ Für die Fußballer ist das nächste Ziel, die Asienmeisterschaften zu gewinnen.	What does Bashir Sinan say? ✓ When playing football, Yemenis forget about the war in their country. ✓ The Yemeni football team suffers from the political problems in the country. ✗ The next goal for the football players is to win the Asia Cup.
Was ist richtig? ✓ Viele Jemeniten fliehen vor dem Bürgerkrieg in ihrem Land. ✓ Die jemenitische Mannschaft will, dass es Friedensgespräche gibt. ✗ Katar und Saudi-Arabien kämpfen im Krieg gemeinsam gegen den Jemen.	What is correct? ✓ Many Yemenis are fleeing the civil war in their country. ✓ The Yemeni team wants peace talks to take place. ✗ Qatar and Saudi Arabia are fighting together in the war against Yemen.

B.2. Items Generated by Llama 2

German	English
Warum dürfen keine Fußballspiele mehr in Jemen stattfinden? ✓ Weil das Land zu unsicher ist ✗ Weil es ein Bürgerkrieg gibt ✗ Weil die Bevölkerung nicht interessiert ist	Why are no more football games allowed in Yemen? ✓ Because the country is too unsafe ✗ Because there is a civil war ✗ Because the population is not interested
Wie haben die jemenitischen Fußballspieler sich für die Asienmeisterschaften qualifiziert? ✓ Sie haben in verschiedenen arabischen Ländern trainiert ✗ Sie haben ein wildcard [*sic*] erhalten ✗ Sie haben sich nicht qualifiziert	How did the Yemeni football players qualify for the Asia Cup? ✓ They trained in various Arab countries ✗ They received a wildcard ✗ They did not qualify
Was ist die Botschaft der jemenitischen Fußballmannschaft an die Bevölkerung und die Politik? ✓ Lassen wir die Gewehre schweigen und einen vernünftigen Dialog und Frieden an ihre Stelle treten ✗ Lassen wir die Gewalt weitergehen und uns nichts ausreden ✗ Lassen wir uns auf die politischen Spannungen einigen und die Fußballmeisterschaft boykottieren	What is the message of the Yemeni football team to the population and the politicians? ✓ Let's silence the guns and have a reasonable dialogue and peace take their place ✗ Let's let the violence continue and not talk ourselves out of anything ✗ Let's agree on the political tensions and boycott the football championship

B.3. Items Generated by GPT-4

German	English
Warum dürfen im Jemen keine Fußballspiele mehr stattfinden? ✓ Wegen des Bürgerkriegs und der damit verbundenen Unsicherheit. ✗ Weil die Fußballstadien zerstört wurden. ✗ Weil die jemenitischen Fußballspieler alle das Land verlassen haben.	Why are no more football games allowed in Yemen? ✓ Because of the civil war and the associated insecurity. ✗ Because the football stadiums were destroyed. ✗ Because all Yemeni football players have left the country.
Wo fand das Training für die Asienmeisterschaften statt? ✓ In verschiedenen arabischen Ländern, wie Saudi-Arabien und Katar. ✓ In Malaysia. ✗ Im Jemen.	Where did the training for the Asia Cup take place? ✓ In various Arab countries, such as Saudi Arabia and Qatar. ✓ In Malaysia. ✗ In Yemen.
Was ist die Botschaft der jemenitischen Fußballmannschaft an die Bevölkerung und die Politik? ✗ Sie sollten mehr Geld in den Fußball investieren. ✓ Sie sollten die Gewehre schweigen lassen und einen vernünftigen Dialog und Frieden an ihre Stelle treten. ✗ Sie sollten die Fußballspieler besser unterstützen.	What is the message of the Yemeni football team to the population and the politicians? ✗ They should invest more money in football. ✓ They should silence the guns and have a reasonable dialogue and peace take their place. ✗ They should support the football players better.

C. User Interface for Human Evaluation

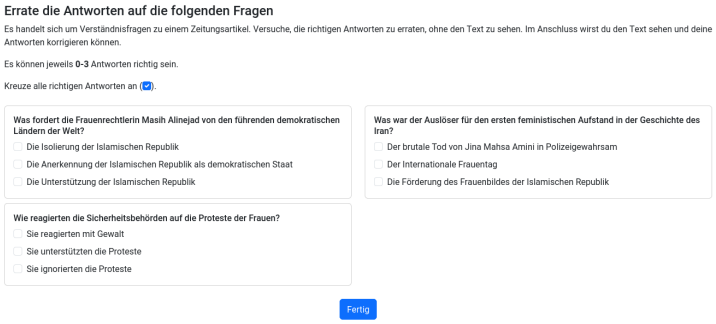


Figure 4: Screenshot of the user interface for the human evaluation, without text.

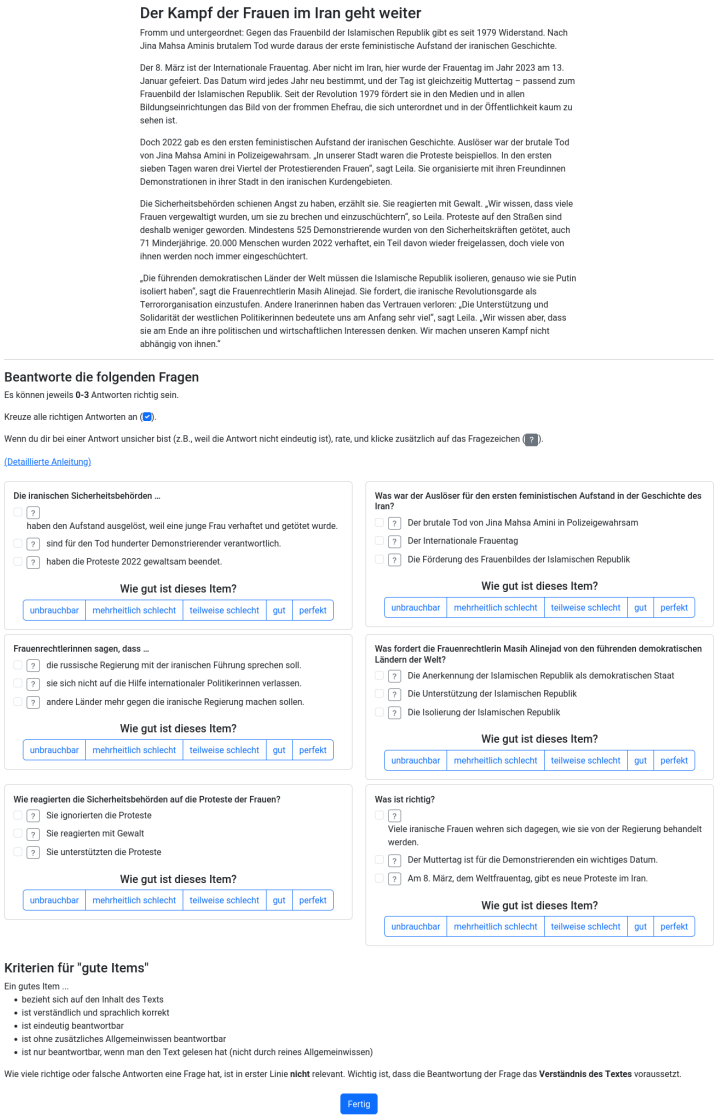


Figure 5: Screenshot of the user interface for the human evaluation, with text and quality ratings.

D. Examples of Guessable and Unanswerable Items

D.1. Guessable Items

	German	English
Item (human-written)	Die Saturnalien ... ✗ wurden an den längsten Tagen im Jahr gefeiert. ✓ waren ein Fest, bei dem unfreie Menschen mit den Herrschern die Rollen wechselten. ✓ fanden unter anderem in der römischen Stadt Köln statt.	The Saturnalia ... ✗ were celebrated on the longest days of the year. ✓ were a festival where unfree people switched roles with the rulers. ✓ took place in the Roman city of Cologne, among other places.
Item (generated by GPT-4)	Was bedeutet der Name „Karneval“ aus dem Lateinischen übersetzt? ✗ „Fest der Freude“ ✓ „Fleisch, leb wohl“ ✗ „Tanz der Narren“	What does the name “Carnival” mean when translated from Latin? ✗ “Festival of joy” ✓ „Flesh, farewell“ ✗ „Dance of the fools“

D.2. Unanswerable Items

D.2.1. Wrong label

	German	English
Text excerpt	[...] *Die Musikwissenschaftlerin Marina Schwarz meint dazu: „Das ist Teil der immer noch patriarchalischen Gesellschaft, in der wir leben.“ Offenbar finden auch viele Frauen, die in dieser Gesellschaft aufgewachsen sind, solche Texte normal.* [...]	[...] *Musicologist Marina Schwarz says: “This is part of the still patriarchal society in which we live.” Apparently, many women who have grown up in this society also find such lyrics normal.* [...]
Item (generated by Llama 2)	Was ist laut Text Marina Schwarz’ Meinung zu sexistischen Texten im Schlager? ✓ Sie findet sie inakzeptabel. [should be ✗] ✗ Sie findet sie normal, weil es Teil der patriarchalischen Gesellschaft ist. ✗ Sie findet sie nicht sexistisch, sondern nur humorvoll.	According to the text, what is Marina Schwarz’ opinion on sexist lyrics in Schlager? ✓ She finds them unacceptable. [should be ✗] ✗ She finds them normal, because it is part of the patriarchal society. ✗ She does not find them sexist, just humorous.

D.2.2. Unclear answer options

	German	English
Text excerpt	[...] *Für viele Deutsche zählt beim Kiosk eher die Atmosphäre – besonders in der warmen Jahreszeit.* [...]	[...] *For many Germans, the atmosphere is more important at the kiosk – especially in the warm season.* [...]
Item (human-written)	Viele Menschen kaufen Alkohol am Kiosk, weil ... ✓ er dort billiger ist als in Bars und Kneipen. ✓ sie die schöne Stimmung vor Ort mögen. [unclear if ✓or ✗] ✓ sie auf dem Weg zu einer Party etwas trinken möchten.	Many people buy alcohol at the kiosk because ... ✓ it is cheaper there than in bars and pubs. ✓ they like the nice atmosphere on site. [unclear if ✓or ✗] ✓ they want to drink something on the way to a party.

D.2.3. Insufficient evidence

	German	English
Text excerpt	[...] *Besonders im Rheinland sind die Straßen voll mit kostümierten Menschen, die tanzen, singen und feiern* [...]	[...] *Especially in the Rhineland, the streets are full of people in costumes who dance, sing and celebrate* [...]
Item (generated by Llama 2)	Wo finden die meisten Karnevalsumzüge und -feiern statt? ✓ In Köln ✗ In Rom ✗ In Berlin	Where do most carnival parades and celebrations take place? ✓ In Cologne ✗ In Rome ✗ In Berlin