



**University of
Zurich**^{UZH}

Master's thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of
Master of Arts

Automatic Generation and Evaluation of Multiple-Choice Reading Comprehension Items with Large Language Models

Author: Andreas Säuberli

Student ID: 17-705-161

Supervisor: Dr. Simon Clematide

Department of Computational Linguistics

Date of submission: 31 December 2023

Abstract

Designing effective reading comprehension tests is a labor-intensive process involving experts manually writing and reviewing test items, as well as running trials with large numbers of test-takers to ensure their quality. This thesis investigates how large language models (LLMs) can be used to automatically generate and evaluate multiple-choice reading comprehension items. I present DWLG, a German dataset consisting of 454 simplified news articles and 1361 human-written reading comprehension items. I also introduce a new evaluation metric called text informativity, which measures how answerable and how guessable reading comprehension items are based on a small number of item responses. I used this metric in a human evaluation study to compare the quality of items generated by Llama 2 (70B) and GPT-4. Results showed that GPT-4 clearly outperforms Llama 2 in zero-shot item generation, but does not reach human-level text informativity. I then demonstrate that this metric can also be used for automatic evaluation by substituting human responses with responses from LLMs. Both GPT-4 and Llama 2 showed good agreement with human responses. Finally, I experimented with fine-tuning Llama 2 to improve its item generation capabilities. Although item quality did not improve, the results shed a light on the challenges associated with adapting LLMs to the requirements of test items. Overall, although modern LLMs are able to generate usable reading comprehension items, there is still room for improvement, and data scarcity for German remains a problem.

Zusammenfassung

Die Entwicklung effektiver Leseverständnistests ist ein arbeitsintensiver Prozess, der das manuelle Schreiben und Überprüfen von Test-Items durch Expert:innen sowie die Durchführung von Testläufen mit einer grossen Anzahl von Teilnehmenden erfordert, um die Qualität der Aufgaben sicherzustellen. Diese Arbeit untersucht, wie grosse Sprachmodelle (*large language models*, LLMs) zur automatischen Generierung und Evaluation von Multiple-Choice-Leseverständnis-Items eingesetzt werden können. Ich präsentiere DWLG, einen deutschen Datensatz aus 454 vereinfachten Nachrichtenartikeln und 1361 von Menschen geschriebenen Leseverständnis-Items. Ausserdem stelle ich Textinformativität als neue Evaluationsmetrik vor, die basierend auf einer geringen Anzahl von Antworten misst, wie beantwortbar und wie erratbar Leseverständnis-Items sind. Diese Metrik habe ich in einer Humanevaluationsstudie verwendet, um die von Llama 2 (70B) und GPT-4 generierten Items in Bezug auf ihre Qualität zu vergleichen. Die Ergebnisse zeigen, dass GPT-4 bei der Zero-Shot-Item-Generierung Llama 2 deutlich übertrifft, aber das menschliche Niveau für Textinformativität nicht erreicht. Dann demonstriere ich, dass diese Metrik auch für die automatische Evaluation verwendet werden kann, indem menschliche Antworten durch Antworten von LLMs ersetzt werden. Sowohl GPT-4 als auch Llama 2 zeigten eine gute Übereinstimmung mit den menschlichen Antworten. Schliesslich habe Fine-Tuning-Experimente mit Llama 2 durchgeführt, um dessen Item-Generierung zu verbessern. Obwohl sich die Qualität der generierten Items nicht verbessert hat, werfen die Ergebnisse ein Licht auf die Herausforderungen, die mit der Anpassung von LLMs an die Anforderungen von Test-Items verbunden sind. Insgesamt sind moderne LLMs zwar in der Lage, brauchbare Leseverständnis-Items zu generieren, aber es gibt noch Verbesserungspotential, und die Datenknappheit für die deutsche Sprache bleibt ein Problem.

Acknowledgements

First and foremost, I would like to thank Simon Clematide for supervising my thesis and giving me the freedom to explore my own interests. He was the LLM enthusiast I needed in order to stay motivated, and his expertise and guidance throughout the project were invaluable.

At every stage of this project, there were many people who contributed ideas, feedback, and encouragement, and helped shape this thesis into what it is. In particular, I would like to thank Cui Ding, Jan Brasser, Yevgeni Berzak, and Lena Jäger for many inspiring discussions about item evaluation.

I am also indebted to Niclas Bodenmann and Deborah Jakobi for their feedback on ideas and visualizations, and for picking me up whenever frustration got the better of me.

In addition, I would like to acknowledge Franz Holzknecht for first introducing me to item response theory and the world of language testing, and Olena Rossi for sharing her valuable insights on item writing.

I am grateful to Sarah Ebling for continuously supporting and encouraging me throughout my studies. She is the person who sparked my interest in the human factors in computational linguistics, which is also at the heart of this thesis.

Finally, I would like to thank the six annotators who took time out of their busy lives in order to take part in the human evaluation.

Contents

Abstract	i
Acknowledgements	iii
Contents	iv
List of figures	viii
List of tables	x
List of acronyms	xi
List of datasets	xii
1 Introduction	1
1.1 Relevance of reading comprehension	1
1.2 Anatomy of a multiple-choice item	2
1.3 Challenges in test development	3
1.4 Research gaps in automatic item generation	4
1.5 Structure of this thesis	4
2 Background: theory of testing	6
2.1 Measuring test-taker proficiency	6
2.2 Measuring item characteristics	7
2.2.1 Difficulty	7
2.2.2 Discrimination	7
2.2.3 Guessability	8
2.2.4 Answerability	8
2.3 Common IRT models	9
2.4 Information functions	11
2.5 Text informativity for reading comprehension items	13
2.6 Applications in NLP	14
2.7 Summary	15

3	Datasets of reading comprehension items	16
3.1	A taxonomy of MRC datasets	17
3.1.1	Text properties	17
3.1.1.1	Text length and linguistic unit	17
3.1.1.2	Text source, genre and domain	18
3.1.1.3	Text difficulty	19
3.1.2	Item properties	19
3.1.2.1	Item type	19
3.1.2.2	Item source	21
3.1.2.3	Item difficulty and skill	21
3.1.3	Language	22
3.1.4	Purpose	22
3.2	Existing German MCRC datasets	23
3.3	DWLG: a new German MCRC dataset	23
3.3.1	Data source	24
3.3.2	Scraping and preprocessing	24
3.3.3	Characteristics and statistics	26
3.3.3.1	Readability	27
3.3.3.2	Question types	27
3.4	Limitations	28
3.5	Summary	29
4	Zero-shot item generation	31
4.1	Related work	31
4.1.1	Zero-shot capabilities of LLMs	31
4.1.2	Automatic item generation	32
4.1.3	Human evaluation of generated items	33
4.2	Task description	34
4.3	Experimental setup	34
4.3.1	Data	34
4.3.2	Models	35
4.3.3	Prompting	35
4.3.4	Postprocessing	36
4.3.5	Human evaluation	37
4.3.5.1	Estimating guessability, answerability, and text informativity	37
4.3.5.2	Annotators	38
4.3.5.3	Design and procedure	38
4.4	Results	41
4.4.1	Surface-level features of generated items	41

4.4.2	Item responses	42
4.4.3	Quality ratings	43
4.4.4	Qualitative analysis	44
4.5	Discussion	46
4.5.1	Item quality	46
4.5.2	Evaluation methodology	46
4.5.3	Limitations	48
4.6	Summary	49
5	Automatic item evaluation	50
5.1	Related work	50
5.1.1	Automatic evaluation of generated items	50
5.1.2	Simulating test-takers	51
5.2	Evaluation protocol	52
5.3	Experimental setup	53
5.3.1	Data	53
5.3.2	Models	53
5.3.3	Prompting	54
5.3.4	Threshold optimization	56
5.4	Results	57
5.4.1	System-level guessability, answerability, and text informativity . .	57
5.4.2	Response-level inter-annotator agreement	58
5.5	Discussion	59
5.5.1	Validity of the item evaluation protocol	59
5.5.2	The case for LLM-based simulation of test-takers	61
5.5.3	Limitations	61
5.6	Summary	62
6	Improving item generation through fine-tuning	63
6.1	Background and related work	64
6.1.1	Efficient LLM fine-tuning	64
6.1.2	Training on LLM-generated data	65
6.2	Experimental setup	65
6.2.1	Training data	65
6.2.2	Fine-tuning	66
6.2.3	Inference and postprocessing	66
6.2.4	Evaluation	66
6.3	Results	67
6.3.1	Surface-level features of generated items	67

6.3.2	Guessability, answerability, and text informativity	67
6.3.3	Qualitative analysis	67
6.4	Discussion	71
6.4.1	Learning item quality from data	71
6.4.2	Is Llama 2 a lost cause?	72
6.4.3	Limitations	72
6.5	Summary	73
7	Conclusion and outlook	74
7.1	Answers to research questions	74
7.2	Open questions and future work	75
	References	77
A	User interface for human evaluation	95
B	Fine-tuning details	97
B.1	Training sample format	97
B.2	QLoRA configuration	98

List of figures

1.1	Example of a MCRC item from the Belebele dataset	2
2.1	IRCs comparing the effects of changing the difficulty, discrimination, guessability, and answerability parameters	10
2.2	IRCs and IIFs for three fictional items in a 4PL IRT model	12
2.3	Relation between the text informativity metric and the maximum of the IIF	14
3.1	Screenshot of a multiple-choice comprehension item on <i>DW Learn German</i>	25
3.2	Readability of texts in DWLG compared to Belebele and QA4MRE	28
3.3	Distribution of question types in DWLG and of question words in DWLG, QA4MRE, and Belebele	29
4.1	Sources of evidence for responding to comprehension items in the human evaluation	40
4.2	Statistics on item lengths and the number of correct answers in human-written and generated items	41
4.3	Mean response accuracy with and without text for human-written and generated items	42
4.4	Quality ratings and unclear answer-options in human-written and generated items	43
4.5	Relation of quality ratings with response accuracy and unclear answer options	44
4.6	Examples of generated items and human response accuracies	47
5.1	Sources of evidence for responding to comprehension items in the automatic evaluation	55
5.2	Mean human and LLM response accuracies on human-written and generated items	57
6.1	Changes in surface-level characteristics of generated items while fine-tuning Llama 2 on human-written and GPT-4-generated items	68
6.2	Changes in guessability and answerability while fine-tuning Llama 2 on human-written and GPT-4-generated items	69
A.1	Screenshot of the user interface for the human evaluation, without text .	95

A.2 Screenshot of the user interface for the human evaluation, with text and
quality ratings 96

List of tables

2.1	Summary of the parameters in IRT models up to 4PLM	11
3.1	Dataset characteristics of DWLG compared to Belebele and QA4MRE . .	26
3.2	Median numbers and lengths for texts, items, and answer options in DWLG compared to the German parts of Belebele and QA4MRE	27
4.1	The German prompt template for item generation and a translation into English	36
4.2	Mean response accuracy with and without text for human-written and generated items, and their difference.	43
5.1	The German prompt templates for item evaluation and a translation into English	54
5.2	Text informativity estimates for all combinations of M_{gen} and M_{eval} for DWLG	58
5.3	Mean IAA (Cohen's κ) between evaluators and (other) humans with and without text	59

List of acronyms

- 1PLM** one-parameter logistic model. 9, 11
- 2PLM** two-parameter logistic model. 9, 11
- 3PLM** three-parameter logistic model. 10, 11, 14
- 4PLM** four-parameter logistic model. x, 10, 11, 13, 14, 38, 48
- AIG** automatic item generation. 1, 3–5, 23, 32–34, 50, 51, 74–76
- API** application programming interface. 24, 35, 56, 61, 69
- CEFR** Common European Framework of Reference for Languages. 1, 24, 27, 48
- CTT** classical test theory. 6–8, 15, 33, 34
- DW** *Deutsche Welle*. 5, 24, 29, 30
- IAA** inter-annotator agreement. x, 58, 59, 62
- IIF** item information function. viii, 11–15
- IRC** item response curve. viii, 9–12, 38, 48
- IRT** item response theory. x, 4, 6–9, 11, 14, 15, 33, 37, 38, 51, 74
- LLM** large language model. i–iii, viii, 1, 5, 31, 33, 35, 41, 49, 50, 52, 53, 55–58, 60–65, 71, 72, 74, 75
- LoRA** low-rank adaptation. 64, 66, 73
- MCRC** multiple-choice reading comprehension. viii, 2, 4, 5, 16, 22, 23, 30–32, 34, 46, 51, 52, 63, 71, 72, 74
- MRC** machine reading comprehension. 16–23, 26, 27, 29, 32, 51, 52, 58, 60, 61
- NLP** natural language processing. 3, 4, 6, 14–16, 22, 23, 29, 32–34, 37
- QA** question answering. 16, 17, 20, 22, 32

List of datasets

Belebele Bandarkar et al. (2023). viii, x, 2, 21–23, 26–30, 41, 53, 57, 60, 61, 73, 76

BookTest Bajgar et al. (2017). 18–21

C³ Sun et al. (2020). 22

CNN/Daily Mail Hermann et al. (2015). 18

DaNetQA Glushkova et al. (2021). 22

DREAM Sun et al. (2019). 18

DuReader He et al. (2018). 21, 22

Entrance Exams Peñas et al. (2014). 23

FairytaleQA Xu et al. (2022). 19, 21

FLoRes-200 NLLB Team et al. (2022). 23

GermanQuAD Möller et al. (2021). 22, 23, 33

HotpotQA Yang et al. (2018). 21, 52

LAMBADA Paperno et al. (2016). 18–20

LiveQA Liu et al. (2020). 22

MC-AFP Soricut and Ding (2016). 20, 21

MultiRC Khashabi et al. (2018). 16, 17

NarrativeQA Kočiský et al. (2018). 18, 20

Natural Questions Kwiatkowski et al. (2019). 17, 18, 21, 22

NewsQA Trischler et al. (2017). 18–21

OneStopEnglish Vajjala and Lucic (2018). 19

OneStopQA Berzak et al. (2020). 19, 21

- Pirà** Paschoal et al. (2021); Pirozelli et al. (2023). 22
- QA4MRE** Peñas et al. (2011, 2012). viii, x, 21–23, 26–30, 41, 76
- RACE** Lai et al. (2017). 16–22, 27, 32, 33, 52
- RACE++** Liang et al. (2019). 19
- RecipeQA** Yagcioglu et al. (2018). 19, 20
- ReClor** Yu et al. (2020). 18, 21
- ReCO** Wang et al. (2020). 22
- ReviewQA** Grail and Perez (2018). 20
- SQuAD** Rajpurkar et al. (2016, 2018). 17, 18, 20–22, 32, 33
- SuperGLUE** Wang et al. (2019). 16
- ViMMRC** Nguyen et al. (2020); Luu et al. (2023). 22
- WikiQA** Yang et al. (2015). 18
- WikiReading** Hewlett et al. (2016). 20

1 Introduction

Many important decisions in society are made based on tests. From receiving the right to drive a car to admission to a university or obtaining citizenship – tests have a tremendous impact on an individual’s life and opportunities. As a result, test developers have a great responsibility to ensure that the tests they create are valid and actually measure the intended traits and skills. This responsibility is also reflected in the cost of test development: Writing, reviewing, and piloting a high-stakes test can cost up to several thousand dollars per item, and developing an item bank for large-scale assessment costs millions (Gierl and Haladyna, 2013). In an effort to speed up development and reduce costs, researchers have been looking for ways to automatically generate test items since the 1970s (Haladyna, 2013). Recently, the advent of generative large language models (LLMs) has presented new opportunities and challenges for automatic item generation (AIG) (Circi et al., 2023). The present thesis is an attempt to tap into this newly gained potential and develop methods to support the test development process. Specifically, I will investigate how LLMs can be effectively used to automatically generate and evaluate multiple-choice items for German reading comprehension tests.

1.1 Relevance of reading comprehension

Reading comprehension is one of the primary communicative language activities defined by the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2020) and an established part of standardized language testing. It is commonly assessed by letting test-takers read a text and answer questions about its content (Jeon and Yamashita, 2020). This type of test is a good candidate for AIG, because reading comprehension items can be written only based on the text that is shown to the test-taker, unlike tests assessing external factual knowledge (e.g., a history test). However, writing such items is also a challenging task, because it requires in-depth semantic understanding, awareness of what is and what isn’t implied by the text, the ability to ask relevant questions, and in the case of multiple-choice items the additional ability to find distractors that are plausible but unambiguously false.

Beyond testing human language proficiency, there are also other applications for reading comprehension items. For example, they can be used to benchmark the natural language understanding capabilities of language models (Bandarkar et al., 2023), to evaluate factual consistency in summarized texts (Manakul et al., 2023), or to assess the comprehensibility of texts in simplified language (Säuberli et al., 2023). In all of these applications, the goal is to measure how well information can be extracted and inferred from a written text, but the specific requirements for the test items may differ substantially depending on the use case (Dunietz et al., 2020). In the present thesis, I will focus on the classical language assessment scenario, where the goal is to determine the level of reading comprehension skills of human test-takers.

1.2 Anatomy of a multiple-choice item

Multiple-choice tests have been the most common choice of format in standardized reading assessments (Jeon and Yamashita, 2020). The main advantage of multiple-choice items is that they are exceptionally simple to administer and mark compared to, for example, open-ended questions, making them an ideal format for fully computerized testing (Jones, 2020).

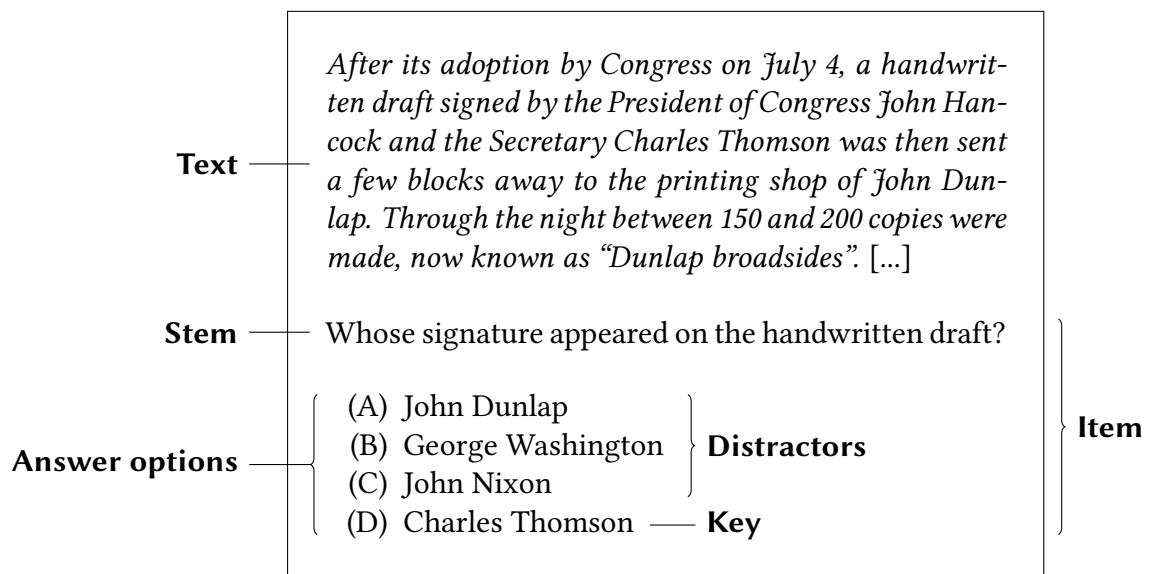


Figure 1.1: Example of a multiple-choice reading comprehension (MCRC) item from the Belebele dataset (Bandarkar et al., 2023).

A multiple-choice item consists of a **stem** (usually a question) and a list of **answer options** from which the correct one is to be selected (see Figure 1.1 for an example). In the conventional case, exactly one of the answer options is correct (the **key**) and the others are incorrect (the **distractors**). However, variants with multiple correct answer

options also exist. Three has been suggested as the ideal number of answer options (Jones, 2020), but four is another common choice. While a higher number of answer options reduces the probability of randomly guessing the correct answer, it also makes it more difficult to find plausible distractors. The plausibility of distractors (meaning that the distractors should not be identifiable as such without having read the text) is one of the major factors in the quality of a test item, and measuring plausibility will be a key aspect in the evaluation of generated items in this thesis.

1.3 Challenges in test development

Developing high-quality items is challenging for many reasons. I will name just three challenges I have faced while writing and reviewing test items in the context of language education and research. First, a number of factors have to be balanced, such as difficulty, clarity, or the time and cognitive effort required to respond. These factors are often in conflict with each other, and test developers have to decide how to prioritize them. Second, it is difficult to predict how actual test-takers will behave when they respond to an item. Especially for multiple-choice items, which heavily rely on natural language, ambiguity or vagueness can cause differences between test-takers' interpretations. Third, tests are usually designed to measure a single and rather specific trait, but in reality, it is rarely possible to isolate the effect of this trait in the response to a test item. For example, developers of reading comprehension tests have to make sure that a test-taker's result is only affected by their ability to read and understand texts, and not by confounding factors like the amount of world knowledge they have.

In the development of high-stakes tests, these challenges are tackled through extensive quality control, including expert reviews and pilot studies with large numbers of test-takers, in order to make sure that the test is valid and reliable (Green, 2020). For example, the *TOEFL Essentials* test (an English language test required by many universities) underwent three stages of trials: first with 570, then with 700, and finally with 5000 test-takers (Papageorgiou et al., 2021). At each stage, items that did not meet the necessary quality criteria were dropped from the test. It goes without saying that running pilot studies at this scale is extremely time-consuming and expensive.

I will investigate ways in which natural language processing (NLP) can potentially contribute to improving and accelerating the test development from two perspectives: first, by supporting the item writing process with AIG, and second, by developing a method for automatically evaluating item quality.

1.4 Research gaps in automatic item generation

Even though AIG is by no means a new idea, NLP research so far has mostly been unsuccessful in developing appropriate solutions that could actually be applied in large-scale assessment (Gierl et al., 2021; Circi et al., 2023). I argue that this is due to three major issues that have received too little attention in the NLP community:

1. **Lack of non-English data:** There are almost no datasets of reading comprehension items available in languages other than English. For training and evaluating AIG systems for languages like German, high-quality datasets are indispensable.
2. **Lack of valid automatic evaluation metrics:** The development and improvement of NLP systems relies on automatic evaluation metrics. For AIG, NLP research has mainly resorted to similarity-based metrics designed for machine translation or text summarization, which is far from the objective of generating test items. Meanwhile, running human evaluations at the scale that is common in language assessment is not viable. As long as there are no empirically validated methods for automatic evaluation that actually reflect item quality, the field is unlikely to make significant progress.
3. **Lack of interdisciplinarity:** It appears that the communities of NLP and language testing researchers are to a large degree mutually exclusive and unaware of the state of research in each other's fields. Although some more interdisciplinary work has been done in recent years (e.g. Attali et al., 2022), most of the work in NLP is done without referencing the large body of literature about AIG and the methodologies developed in other fields. In order to make measurable progress in this challenging task, a better exchange of requirements, knowledge, methods, and data between the involved disciplines is vital.

The present thesis contributes towards addressing these issues by (1) compiling and analyzing a new dataset of German multiple-choice reading comprehension (MCRC) items, (2) developing a protocol for automatic item evaluation, and (3) bridging the gap between test theory and NLP by applying the former to AIG evaluation.

1.5 Structure of this thesis

This main part of this thesis is thematically divided into five chapters:

- **Chapter 2** introduces some relevant theoretical background on test theory, including a primer on item response theory (IRT). It also presents a new metric for

item quality called *text informativity*, which will then be applied in the experimental chapters.

- **Chapter 3** provides an overview of existing datasets of reading comprehension items, highlighting their differences. Moreover, it introduces DWLG, a new German dataset of news articles and MCRC items authored by the German broadcasting company *Deutsche Welle* (DW). The experiments in Chapters 4 to 6 are centered around this dataset.
- **Chapter 4** presents an experiment evaluating zero-shot generation of MCRC items with two LLMs, Llama 2 and GPT-4.
- **Chapter 5** introduces and empirically validates a protocol for automatically evaluating the quality of MCRC items.
- **Chapter 6** attempts to improve on the AIG results from Chapter 4 by experimenting with LLM fine-tuning.

The thesis will close with an overall conclusion in Chapter 7, summarizing and synthesizing the insights from the experiments.

2 Background: theory of testing

There are two statistical frameworks that are commonly applied in test development and assessment: classical test theory (CTT) and item response theory (IRT). In test development, these frameworks can be used to evaluate different properties of test items such as difficulty or reliability. When applying the test items in assessments, they can be used to estimate traits of test-takers, for example, language proficiency. In many ways, IRT extends CTT and is the more powerful and versatile of the two theories. Therefore this chapter will focus on IRT and only draw comparisons to CTT to highlight some important differences.

IRT is very well established with a large body of literature in the field of psychometrics, where it has also been used to evaluate automatically generated items (Sinharay and Johnson, 2013). However, the method has barely received any attention in natural language processing (NLP). More interdisciplinarity in this area could contribute to much-needed advancements in human and automatic evaluation methodology in NLP.

This chapter will present central concepts of IRT, in an attempt to bridge the gap between established theories in language assessment and the challenge of evaluating the quality of items generated by NLP models. These concepts will be useful in discussing the design and results of the evaluations in Chapters 4 to 6.

2.1 Measuring test-taker proficiency

The ultimate purpose of language testing is to measure a person’s language proficiency. However, proficiency is not a trait that is directly observable. In a test, the only observations we can make is how the person responds to items in that test. Thus, in IRT, proficiency is modeled as a **latent trait** that is inferred from item responses.

In order to do this, the items need to be able to distinguish test-takers with a higher proficiency from those with a lower proficiency. In other words, the responses to an item need to be sensitive to the respondent’s proficiency. For example, in a reading comprehension test, a specific item should elicit correct responses from test-takers above a certain

proficiency threshold, and incorrect responses from test-takers below that threshold. If we want to measure a person's proficiency as precisely as possible, we (as test designers) need to make sure that there are enough items with a threshold around that person's proficiency, and that these items are sufficiently good at discriminating between persons above and below that threshold.

Importantly, we do not judge the person's proficiency merely based on the number of correct responses, but we take into account each item's inherent characteristics such as difficulty and discrimination to estimate it. The scale on which we measure proficiency is not fixed or predefined, but determined by how the items behave in relation to the test-takers. Therefore, the question remaining is how to measure these item characteristics.

2.2 Measuring item characteristics

While the test-takers are only characterized in a single dimension (i.e., proficiency), the test items can potentially differ from each other in many ways. This section will list the ones which are most commonly considered when evaluating items using IRT, focusing on intuitive explanations. More formal definitions will be given in Section 2.3.

2.2.1 Difficulty

One of the most basic properties of a test item is its **difficulty**. We can think of this as the threshold that separates test takers who are proficient enough to respond correctly to this item from test-takers who are not (see Section 2.1). Essentially, when item difficulty is high, a higher proficiency is required to respond correctly, and vice-versa. In CTT, an item's **facility** (the opposite of difficulty) is defined as the percentage of test-takers who responded correctly to that item. This is simple to calculate, but it requires that every test-taker responds to every item, and it does not consider that the proficiencies of the test-takers may not be evenly distributed and therefore not every correct response should contribute equally to the item's facility. IRT takes this into account by jointly modeling proficiency and item difficulty.

2.2.2 Discrimination

Discrimination is a measure of how effective an item is at distinguishing between more and less proficient test-takers. Every item separates the test-takers into two disjoint sets: the ones who responded correctly to that item and the ones who did not. If the discrimi-

nating power of the item is maximal, this means that the proficiencies in the first set are all higher than the proficiencies in the second set. In other words, the item is perfectly capable of telling whether some test-taker is above or below a certain threshold. The lower the discrimination, the blurrier the boundaries of the two sets become, and the less reliably the item can tell apart highly proficient from lowly proficient test-takers. If the discrimination is minimal, the item is completely insensitive to proficiency, rendering it useless. In CTT, discrimination is calculated as the correlation between the test-takers' scores in the item and the test-takers' total scores across all items.

2.2.3 Guessability

Although a person with a low proficiency is not likely to respond correctly to a difficult item, they may still do so by chance. For example, in multiple-choice tests, even the worst performing test-takers will get some items correctly, either just by randomly picking answer options or by excluding obviously implausible distractors without actually knowing the correct answer. Thus, **guessability** can be considered an item characteristic, and it is obvious that an item of high quality should have a low guessability.

In practice, it is very difficult to empirically quantify how guessable an item is, because we cannot easily determine whether a test-taker responded correctly because they knew the answer or because they were simply guessing. There is also no equivalent measure in CTT. However, IRT still allows us to infer an item's guessability through statistical modeling.

2.2.4 Answerability

By analogy with guessability, it is also possible that highly proficient test-takers get some very easy items wrong. This phenomenon can also be modeled as an item characteristic in IRT, and is sometimes referred to as **carelessness** (Barton and Lord, 1981). However, I argue that carelessness would be a property of the person and not the item, since it is plausible that certain test-takers tend to respond more carelessly than others but implausible that some items evoke more carelessness than others across all test-takers. Instead, I suggest to label this phenomenon as **answerability**: some items mislead test-takers into giving an incorrect response, even if they know the correct answer, and this affects all test-takers equally. In multiple-choice tests, a reason for this might be that the item is written in a confusing way, such that it is unclear what the correct answer option really is. From this perspective, it is of course desirable to maximize item answerability.

2.3 Common IRT models

To summarize Sections 2.1 to 2.2, IRT lets us infer each test-taker's proficiency and each item's characteristics by looking at the test-takers' item responses. Specifically, IRT models the test-taker's probability of responding correctly to an item based on the test-taker's proficiency and the item characteristics.¹ This section describes the most commonly used IRT models, which differ only in what item characteristics they include or exclude.

This simplest model is the one-parameter logistic model (1PLM), which only uses a single parameter to characterize items, namely difficulty. It is commonly called the Rasch model (Rasch, 1960). Formally, it is equivalent to logistic regression:

$$P(\text{correct response to item } i \text{ by person } j) = \text{logit}^{-1}(\theta_j - b_i),$$

where $\theta_j \in \mathbb{R}$ is a person parameter representing the proficiency level of person j , $b_i \in \mathbb{R}$ is an item parameter representing the difficulty of item i . Note that b_i is simply subtracted from θ_j , which means that person proficiency and item difficulty use the same unit of measurement (logits), and we can compare them on a common scale (Wright and Stone, 1979). On this scale, when some person parameter is equal to some item's difficulty, that person has a predicted probability of 50% of responding correctly to the item. If the person parameter is higher than the item difficulty, the person is expected to have a higher probability of responding correctly.

The probabilities can be visualized as a function of the person parameter θ . The resulting sigmoid curve is called an item response curve (IRC). As shown in the top left panel of Figure 2.1, a higher difficulty parameter causes the IRC to shift to the right.

The two-parameter logistic model (2PLM) adds a second item parameter a_i representing discriminating power, which is fixed to 1.0 in the 1PLM:

$$P(\text{correct response to item } i \text{ by person } j) = \text{logit}^{-1}(a_i (\theta_j - b_i))$$

Items with a higher discrimination parameter have a higher slope in the logistic sigmoid curve (see Figure 2.1, top right panel). In other words, the item is better at distinguishing test-takers with small differences in proficiency, as long as the proficiencies are close to the difficulty of the item.

¹For the sake of simplicity, I only consider dichotomous IRT models here, which are applicable for predicting binary outcomes (e.g., correct/incorrect response). There are also polytomous models, which can predict the probabilities for responses on an ordinal scales.

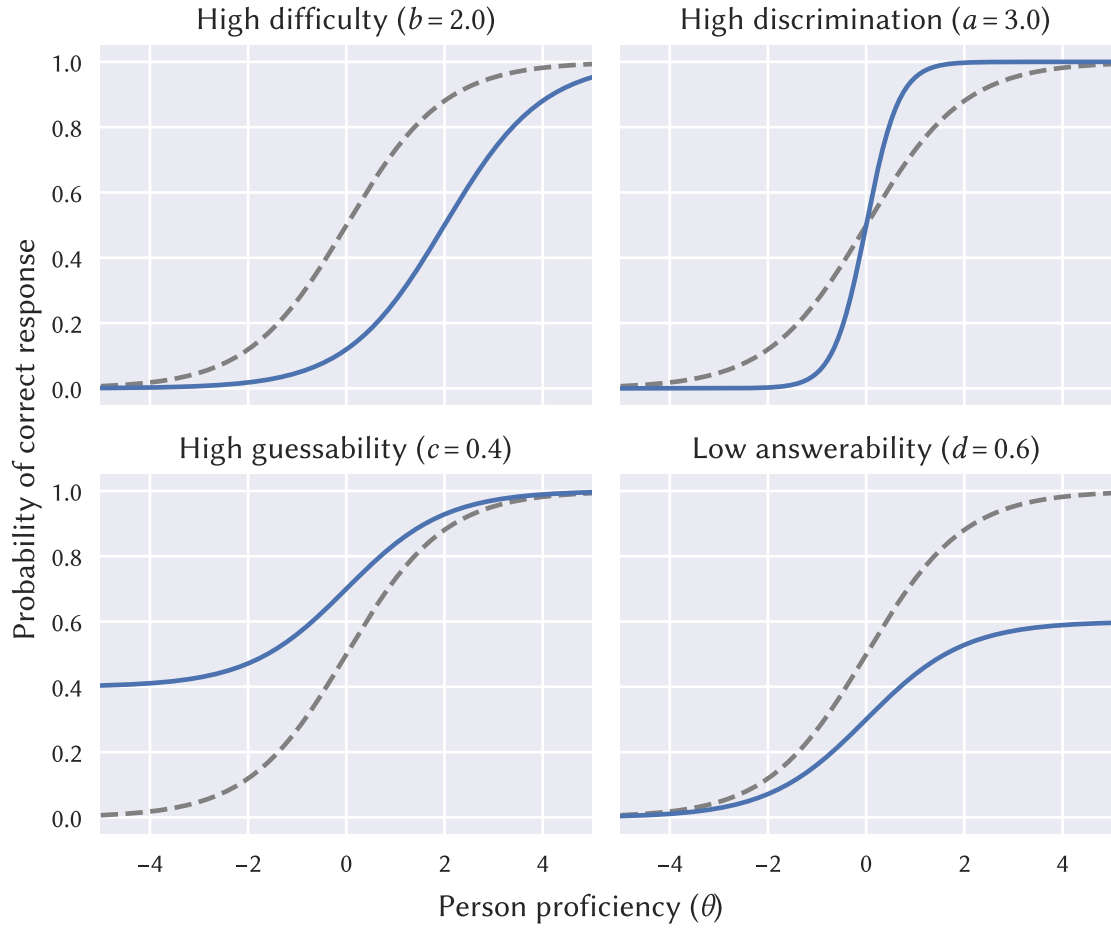


Figure 2.1: IRCs comparing the effects of changing the difficulty, discrimination, guessability, and answerability parameters. The dashed gray IRC always corresponds to an item with $b = 0.0$, $a = 1.0$, $c = 0.0$, and $d = 1.0$, from which each solid blue IRC differs in a single parameter.

In the three-parameter logistic model (3PLM), another item parameter c_i is introduced, which changes the lower asymptote of the sigmoid curve (Lord, 1980, p. 12–14):

$$P(\text{correct response to item } i \text{ by person } j) = c_i + (1 - c_i) \text{logit}^{-1}(a_i(\theta_j - b_i))$$

As a result, the probability of correct response to this item can never be lower than c_i . Similarly, the four-parameter logistic model (4PLM) introduces d_i as the upper asymptote (Barton and Lord, 1981):

$$P(\text{correct response to item } i \text{ by person } j) = c_i + (d_i - c_i) \text{logit}^{-1}(a_i(\theta_j - b_i))$$

These modifications are visualized in the bottom panels of Figure 2.1. According to the

Model	Person parameter	Item parameters			
	Proficiency θ	Difficulty b	Discrimination a	Guessability c	Answerability d
1PLM	Est.	Est.	1	0	1
2PLM	Est.	Est.	Est.	0	1
3PLM	Est.	Est.	Est.	Est.	1
4PLM	Est.	Est.	Est.	Est.	Est.

Table 2.1: Summary of the parameters in IRT models up to 4PLM. *Est.* indicates that this parameter is estimated based on observed data. Note that the conventional variable names for item parameters are not alphabetically ordered.

intuition described in Sections 2.2.3 and 2.2.4 the lower and upper asymptote can be interpreted as the item’s guessability and answerability.

Table 2.1 summarizes the four models and all of the parameters involved. The 4PLM with the upper asymptote parameter is only rarely applied in practice. However, we will see that it is particularly useful for conceptualizing the evaluation of automatically generated questions, where poor answerability and guessability are a common issue.

2.4 Information functions

The ultimate goal of a test is to measure proficiency as accurately and precisely as possible. As discussed in Section 2.2, the items in a test can differ from each other in several ways, and as a result, each item’s contribution to the precision of measurement is also different. Some items may be good at measuring test-takers at a specific proficiency, and some may be more reliable than others. An item’s contribution to measurement precision is quantified by its item information function (IIF) (Lord, 1980, p. 72–73):

$$I_i(\theta) = \frac{P'_i(\theta)^2}{P_i(\theta)(1 - P_i(\theta))},$$

where P_i is the IRC of item i , and P'_i is its derivative. A more intuitive description of the IIF is that it tells us how sensitive the item is to small changes in proficiency. As can be seen from Figure 2.2, different item characteristics have specific effects on the IIFs:

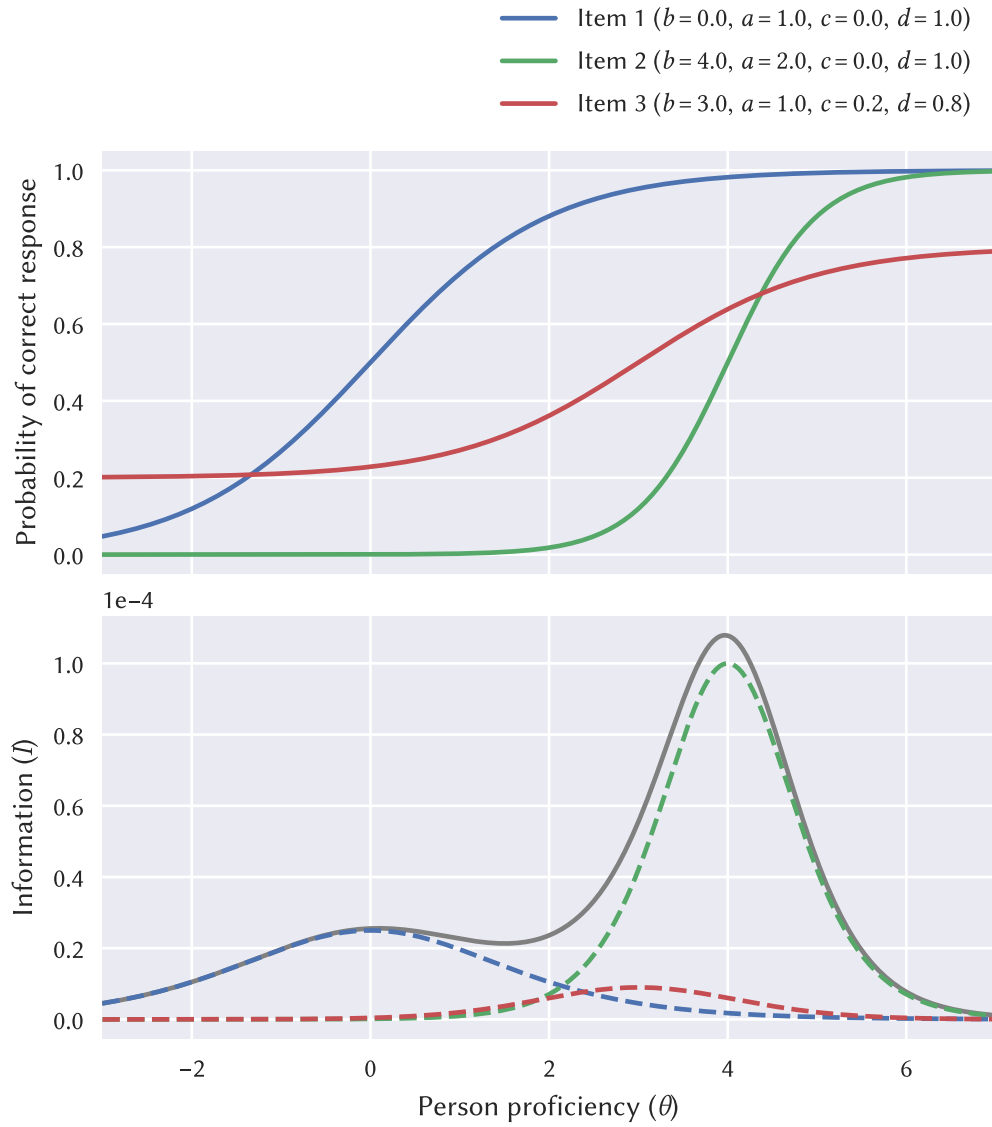


Figure 2.2: IRCs and IIFs for three fictional items in a 4PL IRT model. Dashed curves are IIFs. The solid gray curve is the test information function, which is the sum of all IIFs.

- An item provides the largest amount of information when its difficulty is close to the test-taker's proficiency.
- An item provides more information when its discriminating power is high.
- An item provides more information when the difference between its lower and upper asymptote is high.

The information function of the entire test is defined as the sum of all IIFs, visualized by the solid gray curve in Figure 2.2 (Lord, 1980, p. 70–71):

$$I_{\text{test}}(\theta) = \sum_i I_i(\theta)$$

The amount of information in a test is directly related to the standard error of estimation in the latent trait. Specifically, the more information a test provides, the more precisely we are able to estimate the proficiency level of a test-taker (Lord, 1980, p. 71). In this sense, information functions are an appropriate way of evaluating the quality of single items or an entire test.

2.5 Text informativity for reading comprehension items

For reading comprehension items, I propose to use the difference between guessability and answerability ($d_i - c_i$) as a metric for evaluating item quality. I call this metric **text informativity**.

Intuitively, the guessability parameter for reading comprehension items tells us how likely test-takers are at guessing the correct answer without looking at or understanding the text, while answerability represents the probability for a correct response given that the text was understood perfectly. Therefore, text informativity is a measure of how strongly the evidence available in the text can inform the test-taker's response accuracy. In other words, it measures by how much at best a test-taker can improve their response accuracy by reading the text. Since the aim of the item is to test reading comprehension, it is obvious that maximizing text informativity is desirable.

Theoretically, the text informativity metric can be justified using the definition of the IIF of the 4PLM. Magis (2013) derived the maximum information for a given 4PLM item as a function of its item characteristics. Based on their formula, Figure 2.3 shows that the maximum information increases monotonically and (at least for high answerability

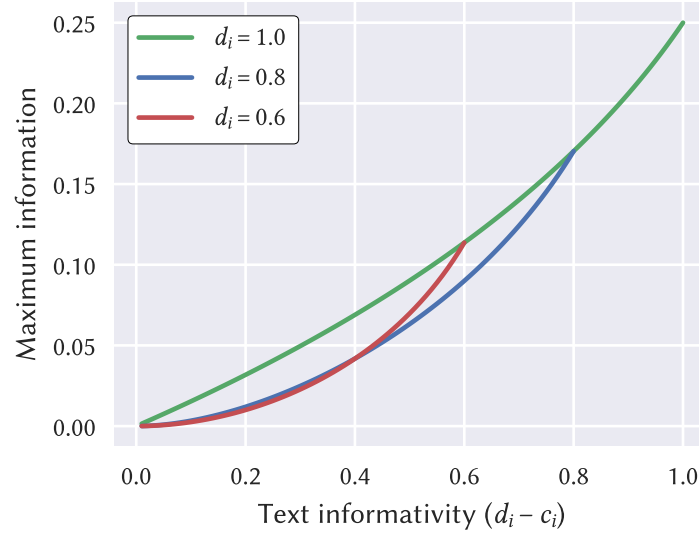


Figure 2.3: Relation between the text informativity metric and the maximum of the IIF of an item i for different answerability values (d_i). Difficulty (b_i) is set to 0 and discrimination (a_i) is set to 1 in all cases.

values) close to linearly with increasing text informativity and constant difficulty and discrimination parameters.

2.6 Applications in NLP

As mentioned in the introduction to this chapter, IRT is well established in various fields of research and widely used for estimating proficiency and evaluating test items. There are also many potential use cases in NLP research, for instance, to analyze responses in human evaluations, to evaluate the quality of automatically generated items, or to test the performance of question answering models. Recently, some researchers have made use of IRT in various NLP contexts (Lalor et al., 2019; Vania et al., 2021; Rodriguez et al., 2021, 2022; Byrd and Srivastava, 2022; Uto et al., 2023), but these remain exceptions.

An important disadvantage of IRT is that especially the higher-parameter models (e.g., 3PLM and 4PLM) require large sample sizes, both in terms of the number of test-takers and the number of items, in order to get precise parameter estimations. According to recommendations in the literature, estimating item parameters in a 4PLM requires several thousand test-takers (Cuhadar, 2022). This may contribute to the reluctance to adopt this method in human evaluations in NLP, where it is common to use rather small sample sizes.

In addition, the assumption that the response data fit the logistic models defined by IRT – although thoroughly tested for human responses – might not necessarily hold for artificial intelligence applications (see Rodriguez et al., 2022). All the more, further research is needed to tap the full potential of IRT and make effective use of it in NLP.

2.7 Summary

Classical test theory (CTT) and item response theory (IRT) are two statistical frameworks for analyzing test items. IRT models the probability of a given test-taker responding correctly to a given item, based on the test-taker’s proficiency and the item’s characteristics, which we can infer by fitting the model to item response data. Different IRT models include different types of item characteristics like difficulty, discrimination, guessability, and answerability. The item information function (IIF) tells us how precisely an item is able to measure a test-taker’s proficiency and essentially reflects the quality of that item.

I introduced the text informativity metric, which is defined as the difference between guessability and answerability. For reading comprehension items, this metric measures how much reading the text can inform a proficient test-takers’ response to that item. We will use text informativity as an indicator of item quality in the experiments in the following chapters.

3 Datasets of reading comprehension items

Most standardized language tests contain items for assessing reading comprehension. Consequently, an abundance of item banks with high quality items in many languages have been professionally developed. However, these item banks are usually confidential, partly due to their commercial value, but also because the items lose their validity if test-takers can access them beforehand. This makes them unsuitable for use in natural language processing (NLP), where publicly available test sets are essential for developing and evaluating models. Therefore, research into machine reading comprehension has made use of datasets which were collected through various means and have never been properly validated (with few exceptions, e.g., Xu et al., 2022).

The idea of testing a machine’s understanding of texts by letting it answer questions is not recent. Initial experiments date back at least to the 1970s (Lehnert, 1977), and some resources were developed by Hirschman et al. (1999). However, the field only started picking up pace by the mid-2010s, and since then a large number of question answering (QA) and machine reading comprehension (MRC) benchmark datasets have been published (Zeng et al., 2020; Dziedzic et al., 2021). More recently, also multiple-choice MRC has become a standard task in natural language understanding, in part due to the wide-spread use of the RACE dataset (Lai et al., 2017) and the addition of MultiRC (Khashabi et al., 2018) to the SuperGLUE benchmark (Wang et al., 2019).

This chapter will give an overview of the different types of currently available MRC datasets (without a restriction on language or item types), followed by a more detailed comparison of German multiple-choice reading comprehension (MCRC) resources. Then, I will present a new German MCRC dataset, which will serve as the main dataset for the experiments in Chapters 4 to 6.

3.1 A taxonomy of MRC datasets

In the following, I will describe several dimensions in which MRC datasets differ, give examples, discuss advantages and disadvantages, with a particular focus on the suitability of datasets for human and machine reading comprehension. The taxonomy is largely based on the surveys by Zeng et al. (2020), Dziedzic et al. (2021), and Rogers et al. (2023).

3.1.1 Text properties

The texts which a system (or human) is required to understand in order to answer items correctly are a central part of MRC. They are the main difference to QA, which may also rely on other sources of evidence such as knowledge bases or images.¹

3.1.1.1 Text length and linguistic unit

The text may be an entire document, a short passage from a document, a single paragraph, or just a few sentences.

RACE and Natural Questions (Kwiatkowski et al., 2019) are examples of datasets on the document level. Text lengths can vary widely for document-level datasets. On average, texts in RACE are 330 tokens long, while Natural Questions is based on Wikipedia pages with an average text length of 7312 (Dziedzic et al., 2021). SQuAD (Rajpurkar et al., 2016, 2018) and MultiRC are paragraph-level datasets with an average text length of 137 and 263 tokens, respectively.

This variability in the length of the text has several consequences:

- **Number of items:** The longer the text, the more items can be written about it.
- **Skill/difficulty level of items:** Longer texts can provide more opportunities for writing items that require combining facts from different parts of the text (see, e.g., MultiRC), while items for short texts may be more oriented towards retrieval and therefore less difficult.
- **Processing effort:** In general, longer texts are more challenging and/or time-consuming to process and extract information from for both machines and humans.

¹Although there are multimodal QA datasets which may require understanding of written text (e.g., text and image: Kembhavi et al., 2017), I will only consider datasets which fully rely on the text as a source of evidence. The item itself, however, may still be multimodal.

However, not all datasets with long documents make use of these advantages. For instance, the questions in Natural Questions originate from real online search queries, and thus tend to be more focused on retrieval from large documents. In assessment of human reading comprehension, reading time and therefore cost increases with the length of the text. As a result, having a large number of items compared to the text length is desirable.

3.1.1.2 Text source, genre and domain

Datasets created from language tests often contain texts written by experts specifically for the test, in order to precisely control for difficulty level and content. RACE and ReClor (Yu et al., 2020) are examples of such resources. Datasets designed a priori for MRC more often use publicly available texts, e.g., Wikipedia articles (SQuAD, Natural Questions, WikiQA (Yang et al., 2015)), books from Project Gutenberg or BookCorpus (BookTest (Bajgar et al., 2017), LAMBADA (Paperno et al., 2016)), or news articles (CNN/Daily Mail (Hermann et al., 2015), NewsQA (Trischler et al., 2017)). Several datasets contain written dialogs or movie scripts, such as DREAM (Sun et al., 2019) and NarrativeQA (Kočíský et al., 2018).

Genre and domain can affect the difficulty of the comprehension task, both for humans and machines, and the suitability of using a specific genre/domain is highly dependent on the research question at hand. However, a generally important factor in reading comprehension assessment is the reader's prior knowledge and familiarity with the topic or domain of the text: A test-taker is more likely to answer an item correctly if they have prior knowledge about the content of the text (Smith et al., 2021; Spyridakis and Wenger, 1991). Recently, Liusie et al. (2023) and Raina et al. (2023a) have found that MRC models are also subject to this effect. This means that some items may measure the reader's/system's world knowledge instead of comprehension, and that some readers/systems may have an advantage due to individual differences in world knowledge. These problems can be reduced (but likely not completely avoided) by carefully choosing the topic according to the reader. For instance, news of current events are a bad choice for human test-takers, but may be good for language models with a knowledge cutoff before the events have happened.

The effect of prior knowledge will be of particular importance in Chapters 4 and 5, and I will discuss them in more detail there.

3.1.1.3 Text difficulty

Several MRC datasets (especially those created as educational resources) contain texts written to match a specific difficulty level. For example, RACE++ (Liang et al., 2019) contains texts from English examinations on three educational levels: middle school, high school, and college. However, the texts and items are not related across the three difficulty levels. OneStopQA (Berzak et al., 2020) is based on the OneStopEnglish corpus (Vajjala and Lucic, 2018) containing news articles in their original version, in addition to two simplified versions (*Elementary* and *Intermediate*), with the same items across all versions. FairytaleQA (Xu et al., 2022) excludes texts above a certain difficulty threshold, according to surface-level readability metrics.

A higher level of difficulty is expected to lead to a lower performance in comprehension by humans. Berzak et al. (2020) showed that this is also the case for MRC models.

A disadvantage of choosing texts with a lower linguistic complexity is that the comprehension questions may then be more difficult to understand than the text. This is problematic because we would end up measuring the comprehension of the items instead of the text. Consequently, writing items that are linguistically simple enough but still accurately test comprehension is challenging. In second language acquisition, this could be avoided by asking comprehension questions in the native language of the test-taker, while in first language acquisition, different modalities (e.g., images) could be used (see, e.g., RecipeQA (Yagcioglu et al., 2018)).

3.1.2 Item properties

3.1.2.1 Item type

Zeng et al. (2020) propose a taxonomy of item types in MRC datasets, which classifies items with respect to three dimensions:

- **Question type**

- **Natural:** a question in natural language without manipulations
Example: *What does the president think about the new law?*
Example datasets: NewsQA, RACE
- **Cloze:** usually a sentence with a placeholder to be filled with an appropriate word or a phrase
Example: *The president thinks that the new law should be ____ immediately.*
Example datasets: LAMBADA, BookTest

- **Synthetic:** a sequence of words which do not necessarily follow natural language grammar, such as keyword queries

Example: *president, opinion about the new law, ?*

Example datasets: WikiReading (Hewlett et al., 2016), MC-AFP (Soricut and Ding, 2016)

- **Answer type**

- **Natural:** a single natural language word, phrase, sentence, or image

Example: *abolished* (in response to the cloze question above)

Example datasets: SQuAD, LAMBADA

- **Multiple-choice:** several answer options from which a subset is to be selected

Example: *abolished, passed, withdrawn* (in response to the cloze question above)

Example datasets: RACE, BookTest

- **Answer source**

- **Spans:** the answer is a substring of the text

Example datasets: SQuAD, ReviewQA (Grail and Perez, 2018)

- **Free-form:** the answer is not (necessarily) a substring of the text

Example datasets: NarrativeQA, RecipeQA

Many large crowdsourced QA datasets such as SQuAD and NewsQA are span-based, because span annotations are easier to compare between annotators and checking for grammar or spelling errors in the answer is not required. In addition, responses to span-based items are easier to evaluate (e.g., through span overlap or exact matching) compared to free-form natural answers. Evaluating multiple-choice items is even more straight-forward, but writing good multiple-choice items is notoriously difficult (Jones, 2020), and even professionally written items (such as in RACE) sometimes suffer from low quality, for example, due to implausible distractors (Berzak et al., 2020).

Cloze-style questions are often extracted directly from the text and can therefore be generated automatically, although human filtering may be necessary to ensure quality (e.g., LAMBADA). As gap-filling is essentially the pre-training objective for masked and causal language models, this task is particularly well-suited for MRC, although it is unclear whether it actually tests reading comprehension, as opposed to world knowledge or language production.

3.1.2.2 Item source

In most cases, datasets are generated based on texts from existing corpora, and items are created by crowdworkers (e.g., SQuAD, NewsQA) or experts (e.g., OneStopQA, ReClor), automatically generated (e.g., MC-AFP, BookTest), or a combination of automatic generation and manual annotation (e.g., Natural Questions, DuReader (He et al., 2018)). When the items are also taken from existing resources like exams, the item writing process is often unknown (e.g., RACE), whereas newly created datasets can involve well-documented guidelines and extensive quality control (e.g., Belebele (Bandarkar et al., 2023), QA4MRE (Peñas et al., 2011, 2012)).

Although employing crowdworkers or non-experts allows creating datasets at a much larger scale, the quality of the resulting items may suffer (Dunietz et al., 2020; Rogers et al., 2020). For item types that are more difficult to write, even professionally designed items can contain quality issues (Berzak et al., 2020).

Dunietz et al. (2020) made a compelling argument as to why the method through which items are obtained is particularly important in the context of MRC. They advocate for a better definition of what researchers mean by *comprehension* and argue that the item generation process should be aligned with this definition. For the most part, the guidelines provided to crowdworkers (e.g., for SQuAD) as well as the requirements for items designed in an educational settings (e.g., for RACE) do not reflect this, limiting the use of these datasets for measuring MRC.

3.1.2.3 Item difficulty and skill

Not many datasets specify the level of difficulty or the type of skill to be tested by the items. One notable exception is FairytaleQA, which contains texts and items written for students between kindergarten and eighth grade. The questions target different skills involved in narrative comprehension by asking about various aspects of the stories, such as characters, feelings, causal relationships, or outcomes of events. QA4MRE also includes items that take the skill level into consideration. For example, they specify which items require combining information from different parts of the text and categorize them according to the type of information they target (e.g., cause, purpose, or degree of truth). HotpotQA (Yang et al., 2018) is a dataset of so-called multi-hop question-answer pairs, which require reasoning in multiple steps and including several parts of the text, increasing the level of skill being tested.

3.1.3 Language

MRC and QA tasks are currently highly English-centric (Rogers et al., 2023). Some resources exist for Chinese (DuReader, ReCO (Wang et al., 2020), LiveQA (Liu et al., 2020); C³ (Sun et al., 2020)), Vietnamese (ViMMRC (Nguyen et al., 2020; Luu et al., 2023)), Russian (DaNetQA (Glushkova et al., 2021)), and other languages. For German, the only original monolingual dataset appears to be GermanQuAD (Möller et al., 2021), which is a span-based QA dataset similar to SQuAD.

There are also several datasets containing texts and comprehension questions in multiple languages in parallel. QA4MRE is a dataset developed for a shared task, containing parallel texts and items in English, German, Italian, Romanian, and Spanish. Pirà (Paschoal et al., 2021; Pirozelli et al., 2023) is a domain-specific Portuguese-English dataset. The most recent and the most substantial addition is Belebele, a parallel MCRC dataset for 122 language variants, including multiple writing systems for some languages.

One issue to consider when creating parallel multilingual datasets is that some item types may not be easily translatable between languages. For example, span-based or cloze items heavily rely on the specific syntactic structure of the question or the text (e.g., the continuity of verb phrases), which may not always be replicable in another language.

3.1.4 Purpose

Reading comprehension items can be designed for different purposes. When items are intended for human users, their purpose can be to assess reading skills in examinations (RACE), to support the language learning process in textbooks, or merely to motivate readers to pay attention to the text in order to study their reading behavior (Dillon et al., 2013). In contrast, the vast majority of the datasets presented in this chapter are specifically created as NLP tasks for machines. Many of these datasets are created from resources that are completely unrelated to reading comprehension, such as search engine queries (Natural Questions) or live commentary during basketball games (LiveQA). Evidently, these are fundamentally different in quality and in what they are capable of testing.

In test design, an important distinction is made between high-stakes and low-stakes tests. A high-stakes test is one where there is a clear line between passing and failing, and passing or failing the test has real consequences for test-takers, for instance, of financial, career-related, or legal nature. Typical examples for high-stakes tests are exit examinations and language certifications required for admission to a university or ob-

taining citizenship. Items from these tests are usually not openly available, or only in small quantities and with little information on the creation and validation process (e.g., Entrance Exams (Peñas et al., 2014)).

MRC is generally concerned with assessing the abilities of NLP systems, and are therefore low-stakes by definition. In contrast, automatic item generation (AIG) can directly support test developers as end-users in creating items for high-stakes testing. As a result, more high-quality items are required for developing and evaluating AIG systems.

3.2 Existing German MCRC datasets

To my knowledge, no originally German MCRC datasets have been published in NLP literature (excluding automatically or manually translated datasets). As mentioned in Section 3.1.3, a span-based MRC dataset exists with GermanQuAD, and theoretically, distractors could be added by automatic or manual means. However, since the key is always an exact substring of the text, it would be very difficult to write distractors that are not easily detected by looking for matching substrings in the text.

The two public multilingual datasets including texts and manually written multiple-choice items in German are Belebele and QA4MRE. Belebele was created based on a pre-existing multilingual dataset for machine translation FLoRes-200 (NLLB Team et al., 2022). The texts in QA4MRE are TED talks and manually translated English web pages. Both datasets are explicitly designed for MRC and not piloted with human test-takers. Bandarkar et al. (2023) only reported expert performance on a subset of the items in the English part of Belebele with a response accuracy 97.6%. A more detailed comparison of the two datasets will be given in Section 3.3.3.

3.3 DWLG: a new German MCRC dataset

As a contribution to the issue of data scarcity in German, I present a new dataset called DWLG (*Deutsche Welle – Learn German*) containing human-written German multiple-choice items designed for human reading comprehension. It is based on texts and items from online German language courses which, to my knowledge, have not been exploited for item generation or evaluation to this date.

3.3.1 Data source

The data originate from the website *DW Learn German* (Deutsche Welle, 2023), which offers free online courses in the German language. *Deutsche Welle* (DW) is a German state-owned broadcasting company funded by tax revenue. Its main purpose is to offer radio and television broadcasting in German and other languages for audiences in foreign countries. The law about its statutory mission states that it should “promote understanding of Germany as an independent nation with its roots in European culture and as a liberal, democratic, constitutional state based on the rule of law” (*Deutsche-Welle-Gesetz*²; translation cited from Deutsche Welle, 2008), and mentions that the German language should be promoted in particular. In line with this mission statement, the courses on *DW Learn German* are targeted at foreign or second language learners at all levels and frequently feature cultural topics, as well as German and European news.

The texts and items included in DWLG are taken from the *Top-Thema* course. In every lesson, this course features a short news article (in spoken and written form) about current events and topics at level B1 according to the Common European Framework of Reference for Languages (CEFR), and interactive exercises testing comprehension, vocabulary, and grammar. Each text is based on and linked with an original news article published by DW, summarized and simplified to match the B1 level. The material has been developed since 2018, with new lessons being added twice per week. Figure 3.1 shows an example of an exercise item on the website.

3.3.2 Scraping and preprocessing


The *DW Learn German* website uses an undocumented GraphQL application programming interface (API), which can be reverse-engineered to fetch content in a JSON format. However, the site uses persisted queries, meaning that the full GraphQL queries are stored server-side and referenced from the client with hashes, which appear to be invalidated from time to time. Therefore, I used *Selenium*³ to automatically render a lesson page in a headless browser and scraped the query hashes from the rendered HTML elements.

I extracted the article text and the exercises for each lesson from the API responses, and filtered the exercises to only include multiple-choice items. These items frequently include questions that do not refer to the content of the text, but test grammar or vo-

²Gesetz über die Rundfunkanstalt des Bundesrechts “Deutsche Welle”, 14 September 2021: <https://www.gesetze-im-internet.de/dwg/BJNR309410997.html>

³<https://www.selenium.dev/>

Neues Gesetz für ausländische Fachkräfte



02:40

1 / 4

Wovon handelt der Text?

Hör dir das Audio an. Was ist richtig? Wähl aus.

Manuskript ▾

In dem Text geht es um ein neues Gesetz, das ...

es Nicht-EU-Bürgern erlaubt, in Deutschland in ihrem Beruf zu arbeiten.

es nur noch EU-Bürgern erlaubt, sich in Deutschland auf Jobs zu bewerben.

vorschreibt, dass ausländische Fachkräfte nicht länger als zwei Jahre in Deutschland arbeiten dürfen.

0 von 1 Aufgaben gelöst. 0 erhaltene Punkte.

Weiter

1 / 4

Figure 3.1: Screenshot of a multiple-choice comprehension item on *DW Learn German*. An audio recording of the text can be played by clicking on the play button, and the transcript is shown when clicking on *Manuskript*. Link to page: <https://learngerman.dw.com/de/wovon-handelt-der-text/l-52650096/e-61598476>

Feature	DWLG	Belebele	QA4MRE
Text length/linguistic unit	Short documents	Paragraphs	Long documents
Text source/genre/domain	News	Wiki	TED talks, web
Text difficulty	B1	—	—
Item type			
Question type	Natural, cloze	Natural	Natural
Answer type	Multiple-choice	Multiple-choice	Multiple-choice
Answer source	Free-form	Free-form	Free-form
Item source	Experts (?)	Experts	Experts
Item difficulty/skill	—	—	—
Language	German	Multilingual	Multilingual
Purposes	L2 learning	MRC	MRC

Table 3.1: Dataset characteristics of DWLG compared to Belebele and QA4MRE. A dash (—) indicates that the feature is unspecified or not applicable for that dataset.

cabulary (e.g., *Which of the following words means ‘to repair a piece of art’?*) and are not suitable for testing reading comprehension. Manual inspection showed that the first three multiple-choice items consistently contain questions about the text, therefore I only included those.

Preprocessing involved removing HTML tags from the texts, inserting double newlines at paragraph boundaries (keeping the title as a paragraph), and replacing gaps in cloze-style stems with three underscores.

The scraping and preprocessing scripts are available on GitHub⁴ and the complete containerized pipeline on DockerHub⁵.

3.3.3 Characteristics and statistics

Table 3.1 characterizes the DWLG dataset according to the taxonomy defined in Section 3.1, and Table 3.2 compares basic statistics with the German parts of Belebele and QA4MRE. In terms of size, DWLG is comparable to Belebele, but it contains more items per text. It also differs in that it was designed for language learning and not for MRC. This could mean that, although all three were designed for low-stakes testing, DWLG is less focused on test validity and more on the educational value of the items.

⁴<https://github.com/saeub/dwlg>

⁵<https://hub.docker.com/r/saeub/dwlg>

	# Texts	# Items	# Answers per item	Text length	Stem length	Answer length
DWLG	454	1361	3	327	8	10
Belebele	488	900	4	88	12	3
QA4MRE	28	280	5	1916	11	4

Table 3.2: Median numbers and lengths for texts, items, and answer options in DWLG compared to Belebele and QA4MRE. Lengths are the number of tokens. Numbers are calculated based on the dataset downloaded on May 10, 2023.

In DWLG, most items have three answer options (only five have more than three), while Belebele consistently has four answer options per item, and QA4MRE consistently has five. Another important difference is that in about 66% of the DWLG items, the test-taker is allowed to select multiple answer options, and in 47% of all items, multiple answer options are correct. In Belebele, QA4MRE, and most other multiple-choice datasets, only one answer option is correct. On the one hand, this limits the comparability of MRC performance metrics between the datasets, on the other hand, this potentially increases the information per item in DWLG (see Section 2.4), as it essentially turns an item with three answer options into three items with two answer options (*true* and *false*) each.

3.3.3.1 Readability

As mentioned in Section 3.3.1, the *DW Learn German* website claims that the texts in DWLG are at CEFR level B1. Figure 3.2 shows how this manifests itself in surface-level readability metrics. Flesch reading ease is calculated based on the numbers of sentences, words, and syllables in a text, and a low reading ease corresponds to high difficulty (Flesch, 1948; Amstad, 1978). The texts in DWLG have an average reading ease of 55.3 (“fairly difficult”), compared to 44.6 for Belebele and 46.1 for QA4MRE (“difficult”).

3.3.3.2 Question types

Question types in DWLG are less consistent than in most existing datasets. As can be seen from Figure 3.3, the majority of items are cloze-style, where the stem is an incomplete sentence with a gap at the end and the answer options are continuations. In this respect, DWLG is similar to the RACE dataset. Very few examples in this style also exist in QA4MRE, and none exist in Belebele. A possible reason for this is that it is difficult to create cloze-style items that are translatable across many languages in multilingually parallel datasets (see Section 3.1.3).

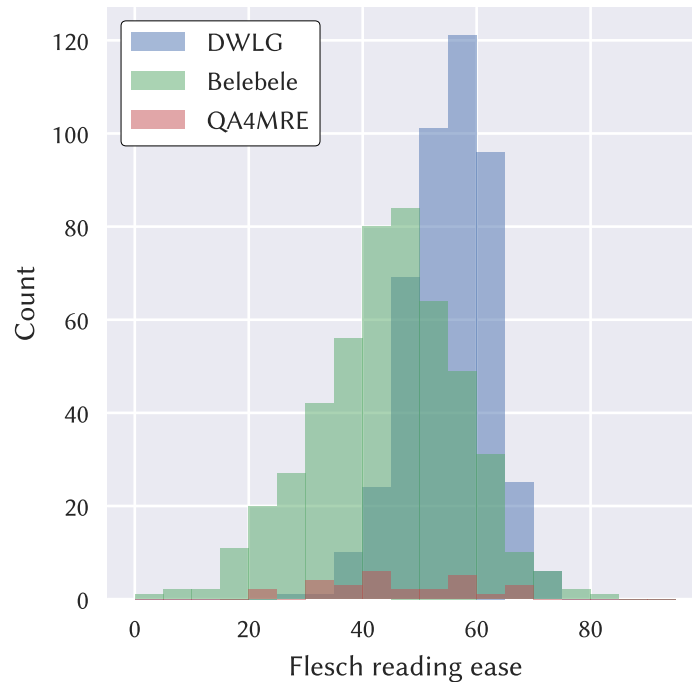


Figure 3.2: Readability of texts in DWLG compared to Belebele and QA4MRE. Flesch reading ease was calculated based on the parameters for German determined by Amstad (1978) using the Python package *textstat*.

Excluding items whose stems are not clearly marked as questions with a question mark, the distribution of question words are also remarkably different across the three datasets (Figure 3.3). While the question word *what* (including German pronominal adverbs such as *wofür* ‘for what’, *worüber* ‘about what’, etc.) is by far the most common in DWLG, question words in QA4MRE are much more evenly distributed. This can be attributed to the item creation process for QA4MRE, where the authors defined several semantic question targets such as location (*where?*), number (*how many?*), or person (*who?*), and aimed for an even distribution of them for each text (Peñas et al., 2012). The annotation guidelines for Belebele did not include such criteria (Bandarkar et al., 2023).

3.4 Limitations

Although the items in DWLG were specifically designed for a relatively well-defined population of human test-takers (unlike Belebele and QA4MRE), the test items were likely not piloted or systematically reviewed, since they are low-stakes and mainly for educational support. This may negatively impact quality, for example in terms of guessability and answerability. I will investigate this in more detail in Chapter 4.

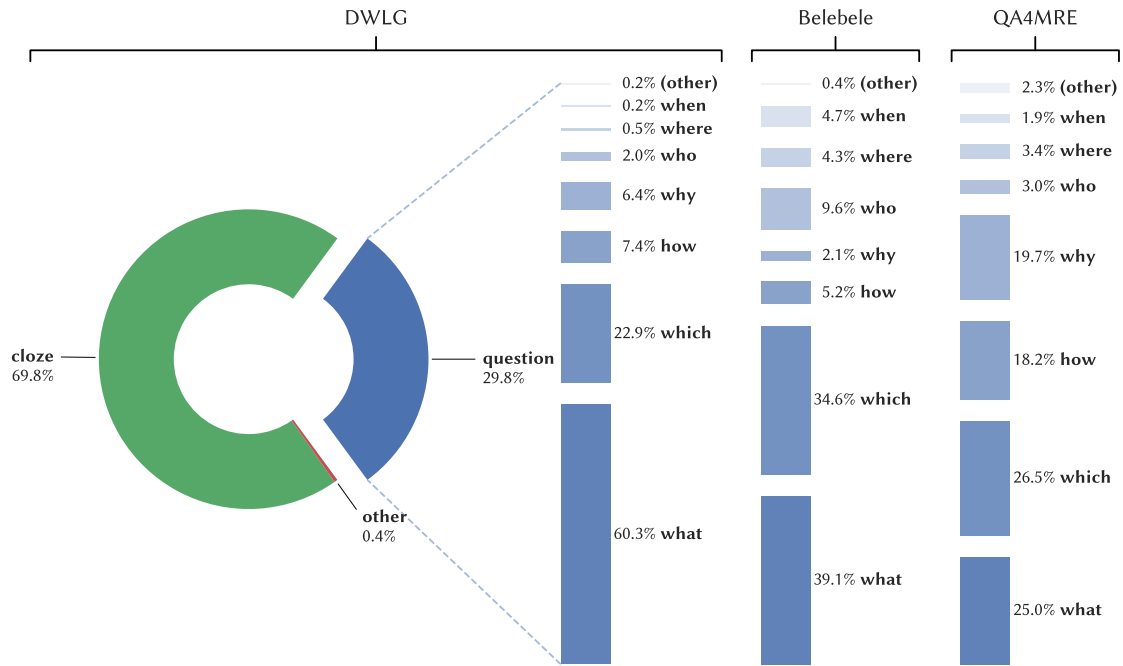


Figure 3.3: Distribution of question types in DWLG and of question words (translated from German) in DWLG, Belebele, and QA4MRE.

In addition, the audio recording is much more prominently presented on the website, and the transcript is hidden by default (see Figure 3.1). This could mean that the items are designed to test listening rather than reading comprehension. Previous work has suggested that if the same item is administered in a listening test instead of a reading test, it differs mainly in its difficulty (Larsen and Feder, 1940).

Finally, DW as the copyright holder of the material has currently not licensed the dataset for redistribution. The ability to publish this dataset for research purposes would be helpful for advancing German NLP, since high-quality human-generated data is still rare in this language.

3.5 Summary

The first part of this chapter provided an overview of how currently available machine reading comprehension (MRC) datasets differ from each other, categorizing them by the types of texts and items in those datasets, as well as in terms of languages and the purposes for which they were created. These properties can have an effect on the quality of the test items, and not all MRC datasets would also be suited for testing human reading comprehension. Compared to English, only little data is available in other languages.

For multiple-choice reading comprehension (MCRC), I identified two relatively small multilingual datasets that also contain German items (Belebele and QA4MRE).

The second part introduced DWLG as a newly compiled dataset of German MCRC items. The texts in DWLG are simplified news articles from *Deutsche Welle* (DW), and the items were written for second language learners. Compared to Belebele and QA4MRE, the texts are more readable, and the multiple-choice items have a slightly different format. Most importantly, most items only have three answer options, and several answer options can be correct.

4 Zero-shot item generation

In this chapter, I will present an experiment to test the ability of state-of-the-art large language models (LLMs) to generate German multiple-choice reading comprehension (MCRC) items for texts in the DWLG dataset. The purpose of this experiment was to get an idea of the baseline performance of both proprietary and open-source LLMs without additional task-specific training data. This setting is especially relevant for German, where data scarcity is a problem. To measure performance, I conducted a small-scale human evaluation study, collecting both subjective quality ratings and item responses for estimating guessability, answerability, and text informativity (see Sections 2.2.3, 2.2.4 and 2.5).

The research question guiding this experiment is: *How good are large language models at generating German multiple-choice reading comprehension items for given texts in a zero-shot setting?*

4.1 Related work

4.1.1 Zero-shot capabilities of LLMs

In recent years, the ability of pre-trained language models to solve tasks solely based on a natural language instruction and without having been shown any training examples was recognized as an emergent capability in LLMs (Wei et al., 2022). Wei et al. (2021) showed that this capability can be improved by fine-tuning the model on instructions for a variety of tasks, and Ouyang et al. (2022) further enhanced this approach using reinforcement learning from human feedback, aligning LLMs with user intents for chat bot applications. Kojima et al. (2022) found that chain-of-thought prompting can be used to further improve zero-shot generation for tasks involving complex reasoning.

Recent instruction-tuned LLMs such as GPT-4 (OpenAI, 2023), Llama 2 (Touvron et al., 2023), and Claude 2 (Anthropic, 2023) have been shown to be highly performant in zero-shot settings across various language understanding and generation tasks, in some cases even surpassing task-specifically trained models (Shaham et al., 2023; Zhang et al., 2023;

Karpinska and Iyyer, 2023).

4.1.2 Automatic item generation

The idea of automatically generating test items has a long history in educational and psychological assessment (Haladyna, 2013). In these fields, there is a strong interest in maintaining the validity of generated test items without having to run trials with every new set of items. As a result, the most commonly applied approaches are rule-based and make use of manually written templates, which purposefully restrict the diversity of the generated items (Lai and Gierl, 2013; Circi et al., 2023). Because of these restrictions, reading comprehension items generated with these approaches tend to only target recall and lexical knowledge, rather than higher-level inferential comprehension skills (Haladyna, 2013; Lai and Gierl, 2013).

In natural language processing (NLP), automatic item generation (AIG) is a less well-known task, often termed *question generation* (Das et al., 2021; CH and Saha, 2020) and considered the counterpart to question answering (QA) or machine reading comprehension (MRC). The earliest MCRC item generation systems were pipelines of mostly rule-based components, typically starting by selecting a sentence from the text to generate a question for, followed by the generation of the question and correct answer, and finally a distractor generation step (Mitkov et al., 2006; Aldabe and Maritxalar, 2010; Papasalouros et al., 2008; Singh Bhatia et al., 2013; Majumder and Saha, 2014, 2015).

The introduction of the SQuAD dataset (Rajpurkar et al., 2016, 2018) has led to a series of works on span-based answer-aware question generation, where the text and an answer span is given as input and the goal is to generate a corresponding question. Several of these approaches rely on recurrent neural networks (Yuan et al., 2017; Du et al., 2017; Zhou et al., 2017; Gao et al., 2019), and more recently on pre-trained transformer models (most commonly, T5 (Raffel et al., 2020); 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). Among these, some have also studied how to enforce certain properties in the generated questions: Gao et al. (2019), Ghanem et al. (2022), Uto et al. (2023), and Wang et al. (2023a) focused on controlling the difficulty or required skill of the generated items, Rathod et al. (2022) generated multiple diverse questions from the same text, and Fei et al. (2022) investigated controlled generation of multi-hop questions.

For MCRC, most research is based on the RACE dataset (Lai et al., 2017). While some approaches separate the task into two steps of generating question-answer pairs and generating distractors (Rodriguez-Torrealba et al., 2022; Maurya and Desarkar, 2020; Shuai et al., 2021, 2022; Xie et al., 2022), end-to-end approaches also exist (Jia et al., 2020; Raina

and Gales, 2022; Dijkstra et al., 2022; Kalpakchi and Boye, 2023a,b). Most recent works apply fine-tuning to language models like T5 or GPT-3 on the RACE dataset (Rodriguez-Torrealba et al., 2022; Xie et al., 2022; Raina and Gales, 2022; Dijkstra et al., 2022). Some have also investigated zero-shot and few-shot generation with GPT-3 or ChatGPT with promising results, also for languages other than English (Attali et al., 2022; Raina and Gales, 2022; Kalpakchi and Boye, 2023a,b).

For generating reading comprehension questions in German, only very little preliminary work has been done. Kolditz (2015) used a rule-based approach for generating open-ended questions, De Kuthy et al. (2020) trained a recurrent encoder-decoder architecture on the resulting synthetic data, and Michel (2022) experimented with fine-tuning mT5 (Xue et al., 2021) on GermanQuAD (Möller et al., 2021) and machine translated SQuAD. Gütl et al. (2011) also described a system for generating various types of reading comprehension items, but did not specify their methods in detail. To my knowledge, this chapter presents the first evaluation of LLMs generating German reading comprehension items in a zero-shot setting.

4.1.3 Human evaluation of generated items

In classical (i.e., template-based) AIG, the evaluation of generated items is focused on two methods: manual review by test developers (also called *subject matter experts*) and piloting with test-takers (Circi et al., 2023). Expert reviews can include checking the logic and content of the templates and the plausibility of distractors, and rating the quality of generated items (Gierl et al., 2021, p. 120–143). Piloting involves collecting item responses from a sufficiently large sample of representative test-takers (usually between 30 and several hundreds), followed by statistical analyses to determine item characteristics such as difficulty and discrimination through item response theory (IRT) or classical test theory (CTT) (Green, 2020).

In NLP research, human evaluation mainly focuses on quality ratings by experts or crowdworkers. The categories evaluated and the type of scale used varies widely between works. The most commonly included categories are fluency and relevance (Gao et al., 2019; Uto et al., 2023; Fei et al., 2022; Ghanem et al., 2022; Shuai et al., 2021, 2022; Xie et al., 2022; Jia et al., 2020), followed by item characteristics like difficulty or answerability (Gao et al., 2019; Uto et al., 2023; Du et al., 2017; Rodriguez-Torrealba et al., 2022; Ghanem et al., 2022; Jia et al., 2020). Depending on the task, ratings for more specific requirements like multi-hop complexity (Fei et al., 2022), distractor plausibility (Maurya and Desarkar, 2020; Shuai et al., 2022) or diversity (Xie et al., 2022) are included. Some works use a rating scale for general item quality (Zhou et al., 2017; Rodriguez-Torrealba

et al., 2022).

Attali et al. (2022) is a notable exception where a more elaborate approach was used to evaluate NLP-based AIG. They conducted expert reviews and a large-scale pilot study, collecting millions of responses to over 5000 items and using CTT to determine difficulty and discrimination, as well as analyzing response time. However, this was only possible because the pilot was administered through *Duolingo*, a free language learning app with millions of active users that also offers paid language tests.

4.2 Task description

The item generation task can be framed as follows: Given a text T and a language model M_{gen} , generate n MCRC items with m answer options each, including a corresponding label for each answer option indicating whether it is correct or incorrect. Any number of answer options may be correct.

There are two reasons to generate n items at once instead of sampling for a single item multiple times. First, the items need to be sufficiently different from each other. When sampling multiple items independently, the model will likely generate items that refer to the same information in T . Second, without sampling, the output is deterministically conditioned on the input, which simplifies reproducing the results.

In the present experiment, n and m are both fixed to 3, in order to be comparable to the human-written items in DWLG.

4.3 Experimental setup

4.3.1 Data

This experiment uses texts and items from DWLG. In order to keep the results comparable between the experiments in Chapters 4 to 6, I randomly sampled 50 lessons (i.e., 50 texts with a total of 150 corresponding items) from the dataset to use as a test set. The remaining 404 lessons will be used for fine-tuning in Chapter 6.

4.3.2 Models

I chose two state-of-the-art LLMs for generating items. Both are transformer-based autoregressive language models predicting the next token based on unidirectional context (i.e., generating text from left to right). To facilitate zero-shot prompting, the models are already instruction-tuned and optimized for a chat environment.

Llama 2 is a family of open-source LLMs developed by *Meta* (Touvron et al., 2023). I chose to use the largest model size (70 billion parameters) in the experiments to optimize performance. The models were pre-trained on 2 trillion tokens of publicly available data. The training data is highly English-centric, with almost 90% English and only 0.17% German (around 3.4 billion tokens). This limits its German language understanding and generation capabilities, but my own preliminary tests showed that it is still more promising than smaller models with larger amounts of German training data such as *Falcon 40B* (Almazrouei et al., 2023). I used the checkpoint `Llama-2-70b-chat-hf` on *Hugging Face* (published in July 2023; pre-training knowledge cutoff September 2022) for all experiments in this thesis.

GPT-4 is a proprietary LLM developed by *OpenAI* (OpenAI, 2023). No details on the architecture, model size, training data, pre-training and fine-tuning are published. OpenAI reported human-level performance on academic exams for a wide range of subjects, showing that it has state-of-the-art natural language understanding capabilities and extensive world knowledge. The model supports multimodal input, but this experiment only uses text as input. GPT-4 can only be accessed via a web application programming interface (API), and the information returned is very limited. For example, at the time of writing, it is not possible to obtain probability distributions for tokens output by GPT-4. An additional limitation is reproducibility: The models offered through the OpenAI API are continuously updated, and older checkpoints become invalidated after a period of time. I used the snapshot `gpt-4-0613` (published in June 2023, pre-training knowledge cutoff September 2021).

4.3.3 Prompting

To elicit items through zero-shot generation, I manually created a prompt template that could be used for both Llama 2 and GPT-4 (shown in Table 4.1). The goal was to produce output in a consistent format and containing all the required information. The prompt also hints at the guessability criterion, by asking for plausible distractors.

Chat-optimized models are fine-tuned on conversations, where user messages and system messages alternate and are separated by special tokens. Both Llama 2 and GPT-4

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

Table 4.1: 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. n and m were both fixed to 3.

also allow providing a system instruction before the conversation, which can contain information about the persona the system should take on in its responses. In the present experiment, the prompt was given to the model as the first user message, without providing a system instruction. The system response was generated using greedy decoding (i.e., no sampling and no beam search).

For generating with Llama 2, I loaded the model parameters in 16-bit floating point precision (`torch.float16`). Inference was performed on five NVIDIA V100 GPUs with 32 GB of memory each.

4.3.4 Postprocessing

Inspecting the output from both models, several issues became apparent. In general, GPT-4 generated output in a much more consistent way and followed formatting instructions more precisely than Llama 2. As a result, parsing the items took more effort for the latter.

Llama 2 did not always produce German output and frequently switched to English, sometimes mid-sentence. Making the desired language explicit in the prompt did not

fully prevent this behavior. Therefore, I used the Python package *lingua*¹ for language identification and rejected outputs where fewer than 80% of the lines were identified as German. In those cases, I regenerated the output with a sampling temperature of 0.5 until a German output was produced.

In cases where Llama 2 generated more than three items (i.e., the end-of-sequence token was produced too late or not at all), I only kept the first three items. For the number of answer options, this was not an issue.

4.3.5 Human evaluation

To compare the quality of the generated items, I randomly selected ten texts and corresponding human-written and generated items (a total of 90 items) from the test set for human evaluation. I collected three different types of annotations from six human annotators:

- **Quality ratings:** Rating the general quality of the items on a 5-point scale.
- **Item responses:** Responding to the items in two different settings: (1) while seeing the text (as in a typical multiple-choice test), and (2) without seeing the text (i.e. guessing the correct answers).
- **Answer clarity:** Marking answer options for which it is unclear whether it is correct or incorrect.

This allowed me to compare items generated by Llama 2 and GPT-4 to human-written items in terms of subjective (non-expert) perception of quality and in terms of item characteristics estimated from response behavior. For the latter, I focused on guessability and answerability.

4.3.5.1 Estimating guessability, answerability, and text informativity

As mentioned in Section 2.6, applying IRT for evaluating item characteristics would require trials with large numbers of representative test-takers. This is not feasible for small-scale NLP experiments. For this reason, I propose a shortcut method for estimating system-level guessability and answerability with fewer annotators, by calculating response accuracies achieved by highly proficient test-takers with and without looking at the text.

¹<https://pypi.org/project/lingua-language-detector/>

As demonstrated in Section 2.2, the item response curve (IRC) flattens out when the proficiency of the test-taker is much higher than the difficulty of the item. In other words, if the item is much too easy for the test-taker, their probability of responding correctly approximates the answerability parameter of that item under the assumptions of the four-parameter logistic model (4PLM). Using the responses from only six test-takers leads to an imprecise estimate of item-level answerability, but it should be good enough for estimating system-level answerability by averaging response accuracies across many items generated by the same model.

In order to apply the same principle to estimate guessability, we can exploit a special property of reading comprehension items: All of the required evidence for successfully answering questions should come from the text, and the proficiency of the test-takers corresponds to their ability to extract this evidence from the text. Therefore, by hiding the text from test-takers and letting them guess solely based on prior world knowledge, we can essentially reduce their proficiency to a minimum (i.e., negative infinity)² and measure their response accuracy to get an estimate of guessability.

After obtaining guessability and answerability estimates, text informativity can simply be calculated as the difference between the two (see Section 2.5).

4.3.5.2 Annotators

The most important requirement for the approach described in the previous section is that the test-takers must be highly proficient in comparison to the difficulty of the items. To achieve this, I recruited six native German speakers from Switzerland who are university students or recent graduates. Considering that the lessons in DWLG are targeted at B1-level German speakers, it is safe to assume that the annotators' proficiency is higher than the difficulty of the items. All annotators took part on a voluntary basis and without monetary compensation.

4.3.5.3 Design and procedure

Each annotator read all ten texts, and for each text they responded to human-written and generated comprehension items in two settings: First without seeing the text, and second while seeing the text. For each text, the two settings were presented immediately after each other.

²This assumes that guessability is an inherent property of an item, and that the test-taker's language proficiency does not help them to respond correctly. For reading comprehension items, this is most certainly not true, since understanding the stem and answer options themselves already require a certain level of proficiency. However, I stick to this assumption here due to its prevalence in IRT.

In the first setting (without text), annotators saw items from only one source (human, GPT-4, or Llama 2), with the instruction to select the correct answer options without seeing the text. Annotators were allowed to select multiple answer options for all items. After submitting their responses, in the second setting (with text), the text was shown to them, and at the same time the items from *all* sources (including the items they had already responded to before). The reason for including items from only a single source in the first setting is that items from different sources would frequently interfere with each other. For example, if two systems generated a similar stem (e.g., *How many people were at the event?*) but with different sets of answer options (e.g., A: 100, B: 200, C: 500 vs. A: 200, B: 300, C: 400), the correct answer option could be easily guessable by looking at the intersection of the two sets. In the second setting, I assumed that the evidence from the text is strong enough that this effect is negligible. Figure 4.1 compares the two settings in terms of how the annotators were able to gather evidence for their responses.

In addition to responses to the items themselves, annotators marked answer options in the second setting to which they could not respond with confidence as unclear. Annotators were also asked to rate each item's quality on a scale from 1 (*unusable*) to 5 (*perfect*). The following criteria were listed for what constitutes a high-quality item:

- 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).

After submitting their responses in the second setting, annotators could see how many items they answered correctly, and a leaderboard showed how they were performing in comparison to other annotators. The purpose of this gamification was only to motivate the annotators, and I did not use these results in the analysis. The annotation then continued with the next text. The order of the texts, the sources of the items in the first setting, and the order of the items and answer options were randomized across annotators. The annotators were not informed which items were human-written or generated.

The user interface for the annotation was a web application implemented using *Django*³ and *Bootstrap*⁴ (refer to Appendix A for screenshots). Annotators were free to use a desktop or mobile device.

³<https://www.djangoproject.com/>

⁴<https://getbootstrap.com/>

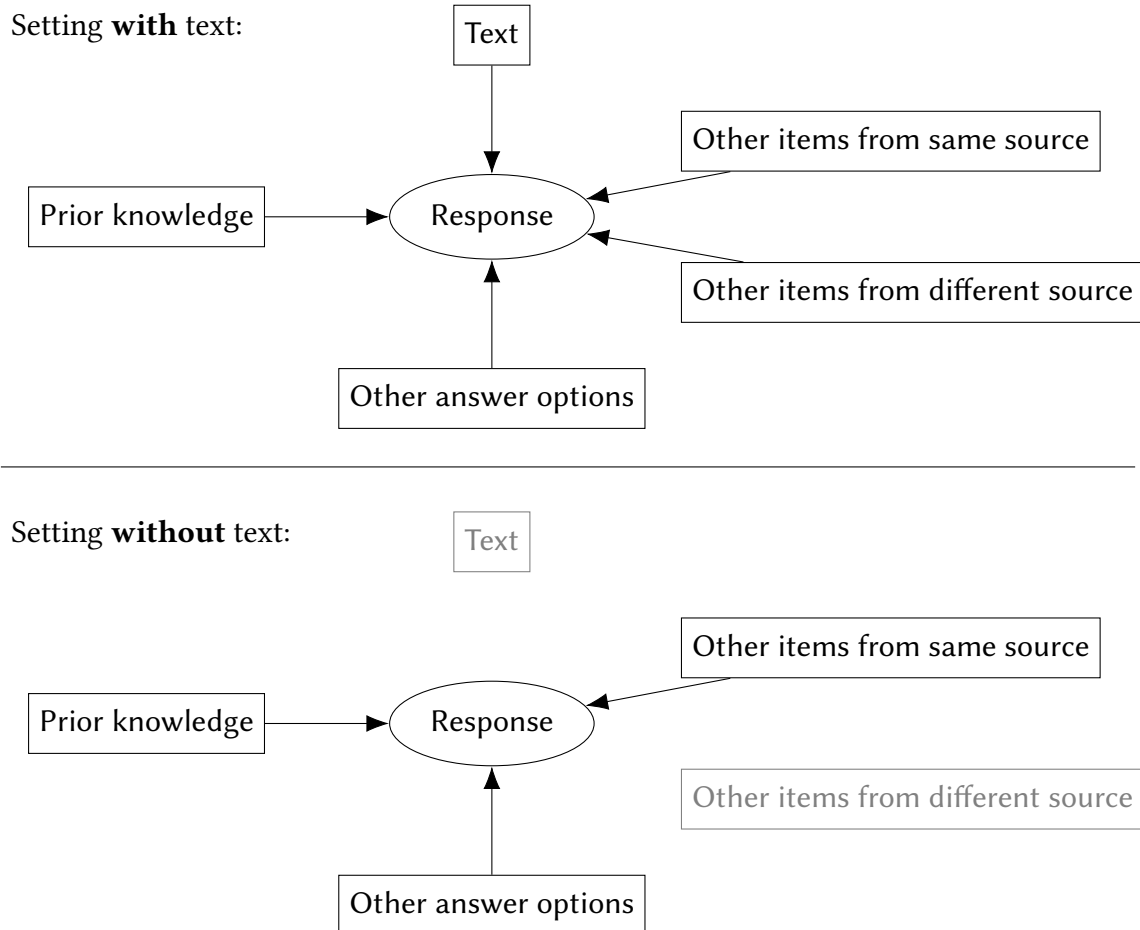
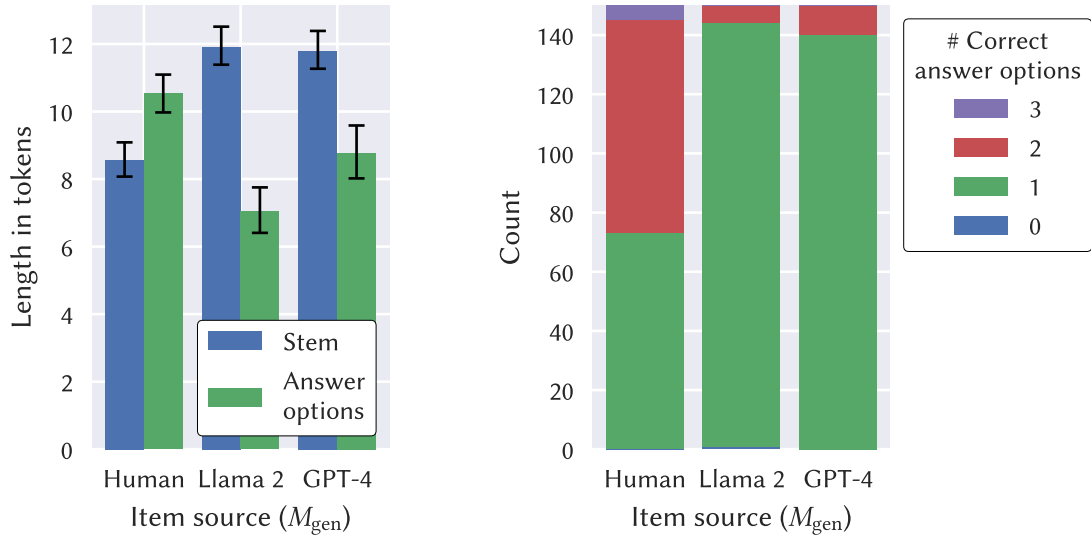


Figure 4.1: Sources of evidence for responding to comprehension items in the human evaluation. In the setting with text, when deciding whether a specific answer option is correct or incorrect, annotators were able to consult the text, their own prior knowledge, other answer options in the same item, and other items from the same or different sources. In the setting without the text, only items from a single source were shown, and the text was hidden.



(a) Mean length of stems and answer options in tokens. All error bars in this thesis are bootstrapped 95% confidence intervals calculated using the BC_a method implemented by *SciPy* (Efron and Tibshirani, 1994; Virtanen et al., 2020).

(b) Distribution of the number of correct answer options per item. The total number of answer options was three in all cases.

Figure 4.2: Statistics on item lengths and the number of correct answers in human-written and generated items. Numbers are based on the test set of 50 texts with three items each.

4.4 Results

4.4.1 Surface-level features of generated items

While about 70% of human-written items in DWLG have cloze-style stems (see Figure 3.3), Llama 2 and GPT-4 exclusively generated items with regular questions as stems. Statistics on item lengths and the number of correct answer options are shown in Figure 4.2. The LLMs tended to generate items with longer stems and shorter answer options, and almost exclusively produced items with a single correct answer option. These characteristics of the generated items are similar to those of datasets such as Belebele (Bandarkar et al., 2023) or QA4MRE (Peñas et al., 2011, 2012) (see Section 3.3.3).

4.4.2 Item responses

The human evaluation resulted in a total of 555 binary responses (one per answer option and annotator) in the setting without text and 1638 in the setting with text. Average response accuracies with and without seeing the text are visualized in Figure 4.3 and reported in Table 4.2. Since each item could have any number of correct answer options, I calculated response accuracy on the level of answer options. Therefore, an ideal item would have a response accuracy of 50% without the text and 100% with the text, resulting in a text informativity of 50%. In Figure 4.3, the two connected points denote guessability (without text) and answerability (with text), while text informativity is represented by the length of the line connecting the two points.

Based on the results, the human items in the evaluation were the least guessable, while items generated by GPT-4 were the most guessable. Llama 2 generated the least answerable and GPT-4 the most answerable items. In terms of text informativity, human-written items performed best, followed by GPT-4.

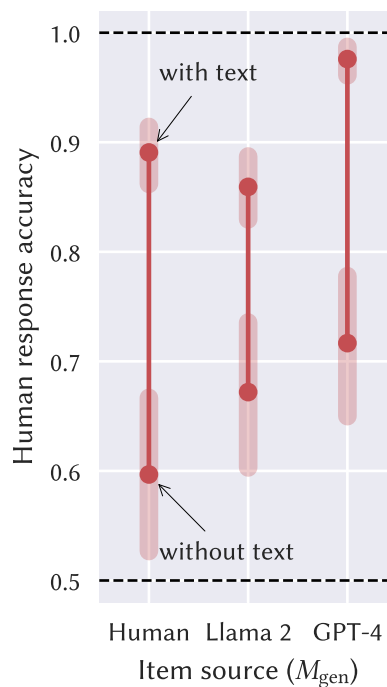
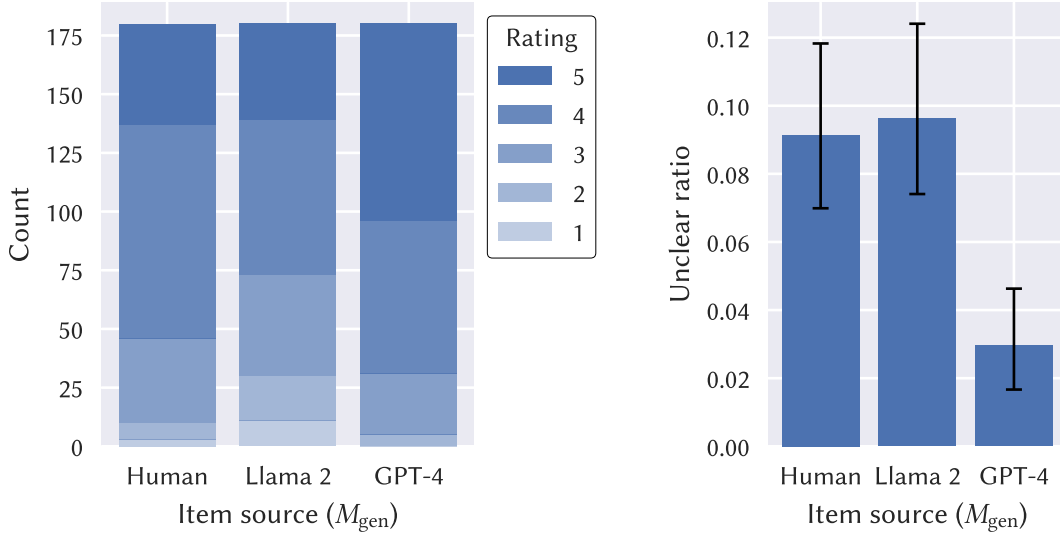


Figure 4.3: Mean response accuracy with and without text for human-written and generated items. Accuracies are on the level of answer options, therefore random guessing is at 50%. Means are based on 10 texts and around 185 responses without text and around 546 responses with text. Error bars are bootstrapped 95% confidence intervals.

Item source (M_{gen})	Guessability \downarrow	Answerability \uparrow	Text informativity \uparrow
Human	0.597	0.891	0.294
Llama 2	0.672	0.859	0.187
GPT-4	0.717	0.976	0.259

Table 4.2: Mean response accuracy with and without text for human-written and generated items, and their difference.



(a) Rating counts for human-written and generated items. The best rating level is 5 (*perfect*), the worst is 1 (*unusable*).

(b) Ratio of answer options marked as unclear in human-written and generated items. Error bars are bootstrapped 95% confidence intervals.

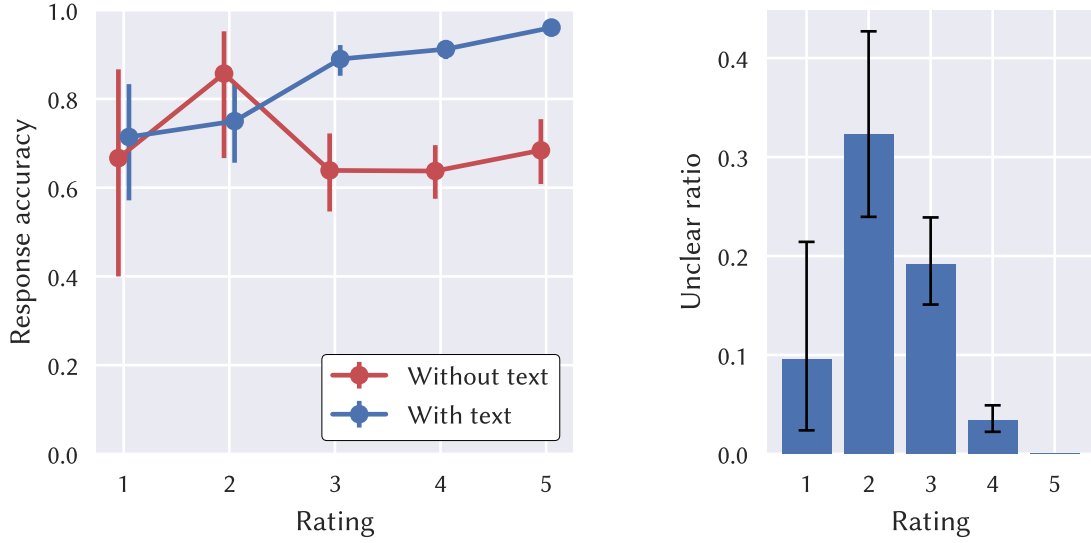
Figure 4.4: Quality ratings and unclear answer-options in human-written and generated items.

4.4.3 Quality ratings

All 90 items were rated by all six annotators, resulting in a total of 540 rating responses. Additionally, in 119 out of 1638 cases, an answer option was marked as unclear. Results from both are visualized in Figure 4.4 for each item source.

GPT-4 received the highest number of *perfect* ratings ($84/180 = 47\%$), while Llama 2 received the highest number of *unusable* ratings ($11/180 = 6\%$). Interestingly, it appears that items generated by GPT-4 were perceived as being of higher quality than human-written items. GPT-4 is also by far the best-performing model in terms of the number of unclear answer options.

To support the interpretation of these results, Figure 4.5 shows how they relate to re-



(a) Mean response accuracy without and with text per rating level. Error bars are bootstrapped 95% confidence intervals.

(b) Ratio of answer options marked as unclear per rating level. Error bars are bootstrapped 95% confidence intervals.

Figure 4.5: Relation of quality ratings with response accuracy and unclear answer options.

sponse accuracies and the ratio of answer options that were marked as unclear. We can see from that items with higher ratings also tend to exhibit higher answerability and fewer unclear answer options. In Figure 4.5a, at rating level 3 and upwards, both answerability and guessability increase, meaning that text informativity remains roughly constant. This suggests that the annotators prioritized answerability over guessability when rating the quality of the items.

4.4.4 Qualitative analysis

Figure 4.6 shows examples of human-written and generated items from the test set and the corresponding response accuracies. We can see that items that ask about definitions or facts about the real world (H1, G1, L1) tend to be highly guessable, while items that explicitly refer to the text (H2, G2) have guessing accuracies close to 50%.

Based on a manual error analysis of all items that were answered incorrectly by the majority of annotators, I identified three common reasons for why items are unanswerable, listed here in descending order of frequency:

- **Wrong label:** The model generated the wrong *correct/incorrect* label for one or

more answer options. This occurred especially when none of the answer options are correct, but the model still produced the label *correct* for one of them.

Example (generated by Llama 2):

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.* [...]

Item:

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.

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

Example (human-written):

Text excerpt:

[...] *Für viele Deutsche zählt beim Kiosk eher die Atmosphäre – besonders in der warmen Jahreszeit.* [...]

Item:

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.

- **Insufficient evidence:** The text does not provide the necessary evidence to decide whether an answer option is correct or incorrect. Sometimes, these items are still answerable based on world knowledge. See L2 in Figure 4.6 for an example.

It appears that Llama 2 is more prone to generate unanswerable questions. However, examples of all of the issues above can also be found in human-written items. These observations support the quantitative results.

4.5 Discussion

4.5.1 Item quality

Overall, the quality ratings from the human evaluation suggest that off-the-shelf instruction-tuned language models are capable of producing good MCRC items in a zero-shot setting: For both models, more than half of the generated items were rated 4 (*good*) or 5 (*perfect*), and less than 10% of the answer options were marked as unclear (see Figure 4.4). Notably, GPT-4 received better ratings than the human-written items, and the response accuracies as well as the qualitative error analysis confirm that items generated by GPT-4 contain fewer issues regarding answerability than those in the original DWLG dataset. However, we can see from Figure 4.3 that GPT-4 trades answerability for guessability, generating more items that can be answered solely based on world knowledge. The text informativity metric reflects this trade-off and shows that GPT-4 does not outperform humans in this respect.

Llama 2 performs significantly worse based on ratings, answerability and guessability. Reasons for this may be the (supposedly) smaller number of model parameters, the low proportion of German in the training data, and less effective instruction tuning, which also manifested itself in the lack of consistency in the output format.

Another important finding is that the data quality in DWLG appears to be suboptimal, with a relatively large number of unanswerable questions. This can certainly be attributed to the fact that the items were not developed for high- or even low-stakes testing but as educational material for language learners. Nevertheless, it is still a competitive human baseline dataset in terms of text informativity.

4.5.2 Evaluation methodology

From a methodological point of view, measuring guessability in the here proposed way is perhaps the less controversial part of the evaluation protocol. While previous work has used responses from humans or machine learning models to evaluate guessability (Berzak et al., 2020; Raina et al., 2023a), directly measuring answerability in a similar way is not commonly done, and to my knowledge, this is the first study proposing the

Karneval – ein Fest mit langer Tradition

Jedes Jahr feiern Millionen Menschen auf der Welt in bunten Kostümen Karneval. An den „tollen Tagen“ ist fast alles erlaubt, was Spaß macht. Die geschichtlichen Ursprünge des Fests haben dagegen viel mit Religion zu tun.

Im Februar fangen sie wieder an, die „tollen Tage“ im Karneval. Besonders im Rheinland sind die Straßen voll mit kostümierten Menschen, die tanzen, singen und feiern – auch und gerade dann, wenn es in der Welt Krieg, Krankheiten und Krisen gibt. Für Christoph Kuckelkorn, den Präsidenten des Festkomitees Kölner Karneval, ist der Karneval „eine Stütze in schwierigen Zeiten, eine Auszeit von den Problemen des Alltags“. Aber seit wann wird der Karneval eigentlich gefeiert?

Zwar sind die genauen Ursprünge nicht klar. Man weiß aber, dass es zum Beispiel in Köln schon vor 2000 Jahren ein Fest gab, das dem heutigen Karneval ähnelte: In der damals römischen Stadt feierte man im Winter die sogenannten „Saturnalien“, um den Gott Saturn zu ehren. Wie heute wurde viel getrunken und getanzt, und die Reichen tauschten mit den Sklavinnen und Sklaven die Kleider und bedienten sie sogar. Auch die Herrscher wurden während der Saturnalien humorvoll kritisiert.

Dass man sich über die Mächtigen lustig macht, ist im Karneval bis heute so. Die Karnevalshochburgen am Rhein sind für ihre „Büttenreden“ bekannt, bei denen ein Redner auf die Bühne kommt und Witze zum Beispiel über Politiker macht. Auch die Wagen auf den Karnevalsumzügen sind oft so geschmückt, dass sie politische Botschaften – oder eben Kritik – enthalten. In Köln übernimmt sogar das sogenannte Dreigestirn für die Karnevalszeit die Regierung über die Jecken.

Karneval in seiner heutigen Form ist aber wahrscheinlich eine Erfindung des frühen Christentums. Für die Kirche war das närrische Treiben ein Sinnbild der Sünde. Ein paar Tage lang durfte man ausgelassen feiern und (fast) alles war erlaubt. Doch nach dem Ende des Karnevals begann eine 40-tägige Fastenzeit, in der die Menschen weniger essen und mehr beten mussten. Daher hat der Karneval wohl seinen Namen – wenn man das Wort aus dem Lateinischen übersetzt, heißt es: „Fleisch, leb wohl.“

(H1)	Human-written items:	(H2)
<p>Die Saturnalien ...</p> <p>✗ wurden an den längsten Tagen im Jahr gefeiert. (100/100%)</p> <p>✓ waren ein Fest, bei dem unfreie Menschen mit den Herrschern die Rollen wechselten. (50/100%)</p> <p>✓ fanden unter anderem in der römischen Stadt Köln statt. (100/83%)</p>		<p>In dem Text geht es um ...</p> <p>✗ einen Ort in Deutschland, in dem der Karneval anders gefeiert wird als anderswo. (50/100%)</p> <p>✗ einen Kölner Karnevalsverein, den es schon sehr lange gibt. (0/100%)</p> <p>✓ die Frage, wie und wann der Karneval entstanden ist. (50/100%)</p>
(G1)	Items generated by GPT-4:	(G2)
<p>Was bedeutet der Name „Karneval“ aus dem Lateinischen übersetzt?</p> <p>✗ „Fest der Freude“ (100/100%)</p> <p>✓ „Fleisch, leb wohl“ (100/100%)</p> <p>✗ „Tanz der Narren“ (100/100%)</p>		<p>Was passiert nach dem Ende des Karnevals laut dem Text?</p> <p>✓ Es beginnt eine 40-tägige Fastenzeit. (0/100%)</p> <p>✗ Es beginnt eine Zeit der Feierlichkeiten und des Essens. (0/100%)</p> <p>✗ Es gibt keine besonderen Ereignisse oder Traditionen. (100/100%)</p>
(L1)	Items generated by Llama 2:	(L2)
<p>Wieviel Tage dauert die Fastenzeit, die direkt nach dem Karneval beginnt?</p> <p>✓ 40 Tage (100/100%)</p> <p>✗ 20 Tage (100/100%)</p> <p>✗ 10 Tage (100/100%)</p>		<p>Wo finden die meisten Karnevalsumzüge und -feiern statt?</p> <p>✓ In Köln (50/33%)</p> <p>✗ In Rom (100/100%)</p> <p>✗ In Berlin (100/100%)</p>

Figure 4.6: Examples of generated items and human response accuracies in parentheses without/with text for each answer option ($n_{\text{without text}} = 2$, $n_{\text{with text}} = 6$). Both models and humans produce items which are highly guessable based on world knowledge (H1, G1, L1). In addition, L1 contains a grammatical error (*Wieviel[e] Tage*), and L2 is not answerable based on evidence from the text.

text informativity metric. Therefore, additional rigor should be imposed on validating this part of the protocol beyond the theoretical justification given in Section 4.3.5.1.

The item response results for GPT-4 confirm that the low answerability estimates for the other item sources are not due to careless response behavior. Theoretically, considering the IRC in the 4PLM as presented in Section 2.3, a low response accuracy with text could also be due to low discrimination or high difficulty instead of low answerability. However, low discrimination would still be a sign of poor item quality, and high difficulty is unlikely, since the texts are at level B1 according to the Common European Framework of Reference for Languages (CEFR), and the items should not be any less comprehensible than the text (see Section 3.1.1.3). Evaluating difficulty would require collecting responses from a larger and more diverse sample of test-takers.

Overall, the results in this chapter provide initial evidence for the validity of the text informativity metric through triangulation with human judgments and qualitative analysis. Although more data should be collected in future work, I consider it to be reliable enough in order to apply and explore it further in the experiments in Chapters 5 and 6.

4.5.3 Limitations

The evaluation methodology used in this experiment has several limitations. As already mentioned in Section 4.3.5.1, due to the small number of annotators, reliably detecting problems on the level of items or answer options is not plausible. In addition, it was not possible to prevent items generated by different systems from influencing each other (see Figure 4.1), as this would further reduce the number of responses. Here, I assumed that this effect is negligible in the setting with text. In an ideal scenario, a large number of annotators would respond to single answer options, without seeing the other answer options or other items, in order to avoid any confounding effects. However, this would explode the number of texts each annotator would need to read, making this unfeasible for human evaluation. In Chapter 5, I will show that this limitation can be circumvented in the automatic evaluation approach.

Another important restriction of the evaluation protocol is that the annotators must be over-proficient compared to the item difficulty. This is only possible here because of the justified assumption that the items are designed for language learners and therefore relatively easy. If this is not the case, a possible approach to overcome this would be to use majority voting among a group of annotators in order to artificially increase their proficiency. This approach was also used to estimate human performance in the SuperGLUE benchmark (Wang et al., 2019, Appendix C). On the other hand, this also reduces the sample size, leading to less precise answerability estimates.

There are also limitations due to the nature of the DWLG dataset. First, since the news texts are publicly available, some of them (or at least texts about the same topics) are likely to be part of the training data of the LLMs. As a result, the models may be able to generate higher-quality items for these texts compared to unseen topics. This test-train leakage could be avoided by including only recent articles published after the knowledge cutoff of the LLMs, but then the human evaluation would suffer from biases in turn, because the annotators are more likely to be familiar with recent events, leading to an over-estimate of guessability (see Section 3.1.1.2).

Finally, because I used zero-shot generation with a simple instruction prompt, the generated items deviate quite a bit from the format of the human-written items, limiting comparability. In particular, Llama 2 and GPT-4 never generated cloze-style items, as shown in Figure 4.6, and rarely generated items with more than one correct answer option (about 5%). The latter might mean that it is often obvious from the question that there can only be one correct answer, which essentially increases random guessing from 50% to 67%. This could explain the high guessability of items generated by GPT-4 in the human evaluation. This problem can also be alleviated by only showing a single answer option at a time.

4.6 Summary

The aim of the experiment in this chapter was to determine the quality of German multiple-choice reading comprehension items generated by Llama 2 and GPT-4 using zero-shot prompting. I conducted a human evaluation with six annotators who rated the items on a general quality scale, marked unclear answer options, and provided item responses without seeing the text (i.e., guessing) and while seeing the text. System-level text informativity was estimated by calculating the difference between the response accuracies with and without text.

Overall, the majority of items generated by both models are of acceptable quality. GPT-4 received better ratings than both Llama 2 and human-written items, most likely because it generated more answerable items. GPT-4 achieved a text informativity of 0.259, clearly outperforming Llama 2 (0.187), but did not reach human performance (0.294). A qualitative error analysis confirmed these results.

5 Automatic item evaluation

Automatic item generation (AIG) is in dire need of standardized automatic evaluation metrics that actually reflect item quality. In contrast to the majority of previous work, Chapter 4 presented an evaluation protocol that does not only take into account the annotators’ subjective perception of item quality but also estimates objective guessability and answerability. In this chapter, I will attempt to automate this procedure by letting large language models (LLMs) assume the role of annotators, resulting in a protocol for automatic and reference-free evaluation of (one aspect of) item quality. I will compare the response behavior between humans and LLMs and show that this approach can be used to evaluate the quality of human and generated items without the need of reference items.

The primary research question here is: *Can responses by large language models to multiple-choice reading comprehension items be used to automatically evaluate item quality?*

5.1 Related work

5.1.1 Automatic evaluation of generated items

Previous works on AIG and question generation have adopted a variety of metrics for automatically evaluating output quality. Most commonly, reference-based metrics originally developed for machine translation or text summarization such as BLEU (Papineni et al., 2001), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), or BLEURT (Sellam et al., 2020) are used (Amidei et al., 2018; Circi et al., 2023; Mulla and Gharpure, 2023). These compare the generated items to one or several human-written references, which is problematic if the AIG system is free to decide which part of the text it should generate items for. Therefore, these metrics are mostly applied in answer-aware span-based question generation (Yuan et al., 2017; Du et al., 2017; Zhou et al., 2017; Gao et al., 2019; Lopez et al., 2020; Berger et al., 2022; Ghanem et al., 2022). Nema and Khapra (2018) extended BLEU with an term calculated based on the recall and precision of content words, function words, question words, and named entities, which they argue correlate with an-

swerability. Wang et al. (2023a) applied this metric for evaluating open-ended question generation. In addition to measuring the similarity to reference items, Gierl et al. (2021, p. 138–141) and Rathod et al. (2022) also described evaluation metrics to measure lexical diversity between items, which is desirable when generating multiple items based on the same context.

Generating multiple-choice items entails the additional challenge of evaluating distractor quality. For answer-unaware generation of multiple-choice reading comprehension (MCRC) items, generic reference-based text generation metrics are not suitable, as Maurya and Desarkar (2020) and Shuai et al. (2022) mention. Still, most authors report BLEU, ROUGE, and METEOR as automatic evaluation metrics (Maurya and Desarkar, 2020; Shuai et al., 2021, 2022; Xie et al., 2022; Dijkstra et al., 2022). Rodriguez-Torrealba et al. (2022) use cosine similarity to measure distractor diversity. Raina and Gales (2022) propose a rather elaborate suite of quality assessments, including rule-based and trained metrics for grammaticality, answerability, diversity and difficulty. Although they did not rely on similarity to human-written references, they require several models to be trained on relatively large amounts of data, and evaluation reliability is likely to be highly dependent on the performance of those models.

5.1.2 Simulating test-takers

Analyzing how human test-takers respond to items plays an essential role in evaluating item quality (Green, 2020). Therefore, a natural step towards automatic evaluation is the automated generation of human-like responses from artificial models.

Several authors have suggested measuring the performance of machine reading comprehension (MRC) models when responding to the generated items, equating high MRC performance with high 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). Yuan et al. (2017) and Klein and Nabi (2019) have even integrated such MRC-based metrics directly into the training process of AIG models, as a reward for reinforcement learning or in the loss function for supervised learning. The underlying assumption in these approaches is that an item is answerable if a MRC model can answer it. However, none of the works above compare the MRC models to human performance or discuss the limitations that come with this assumption, for example, that it may favor very easy or highly guessable items.

In a more differentiated approach, Lalor et al. (2019) used a large ensemble of neural networks to generate responses to natural language inference and sentiment classification items and fitted item response theory (IRT) models to estimate the difficulty parame-

ters of each item. Their results show that the difficulty estimates correlate moderately between humans and machines. Byrd and Srivastava (2022) used a similar procedure to estimate difficulty and discrimination parameters for items in the HotpotQA dataset (Yang et al., 2018). Raina and Gales (2022) proposed the entropy of MRC model response probabilities across the answer options of multiple-choice items as a measure of the uncertainty or unanswerability. Raina et al. (2023b) additionally evaluated distractor plausibility based on the MRC confidence scores on the distractors.

While the approaches in the previous paragraph always provide the full item context as input to the MRC models, Berzak et al. (2020), Liusie et al. (2023), and Raina et al. (2023a) also considered the response accuracy without showing the text, letting the model answer based on world knowledge alone. Results showed that many items in popular benchmark datasets such as RACE (Lai et al., 2017) are highly guessable. All three works used pre-trained language models specifically fine-tuned on multiple-choice items, and Liusie et al. (2023) demonstrated that the results of the guessability evaluation strongly depends on which dataset the model was trained on. In Berzak et al. (2020), the models even consistently outperformed humans in guessing without seeing the text. This suggests that there are dataset-specific clues (e.g., different lengths or word choices in the distractors) that allow the models to learn to guess the correct answer options with relative ease if they are trained on the same dataset, but these clues may not necessarily be detectable by human test-takers who were not exposed to the dataset before.

The evaluation protocol I will present in this chapter builds on and extends the ideas above in two main ways:

1. I use generative LLMs with zero-shot prompting instead of fine-tuning task-specific MRC models, removing the need for training data and reducing dataset-specific biases.
2. I combine guessability and answerability estimates to evaluate text informativity.

5.2 Evaluation protocol

I implement the following protocol for automatically evaluating MCRC item quality: An evaluator model M_{eval} is asked to respond to an item in two settings, with and without the text. In the setting with text, it is given a text T , a stem q and a single answer option a and is tasked with generating a label *true* or *false* indicating whether a is correct or incorrect. In the setting without text, T is not included in the input (see Table 5.1).

Based on the responses by M_{eval} to all answer options in a (human-written or generated)

item dataset D , we can calculate the average response accuracy in the setting with text, serving as an estimate of the answerability of D , and in the setting without text, serving as an estimate of the guessability of D . Computing the difference between the two estimates yields the text informativity, which I use as the main metric for item quality.

In principle, M_{eval} can be any kind of model, as long as it can make binary true/false decisions. However, in order for the answerability and guessability estimates to be representative of human response behavior, M_{eval} must have strong enough comprehension skills and a similar level of world knowledge to human test-takers. The experiment in this chapter will test two LLMs for this assumption.

5.3 Experimental setup

5.3.1 Data

I applied the protocol described above to evaluate the same manually-written and generated DWLG items used in Chapter 4, with the same test split of 50 lessons as described in Section 4.3.1. I will compare these to the results from the human evaluation in Chapter 4. In addition, I will also report results on the complete German part of Belebele (Bandarkar et al., 2023) for comparison, even if it was not part of the human evaluation in Chapter 4.

5.3.2 Models

In the role of M_{eval} , I used the same instruction-tuned LLMs as in Chapter 4, namely Llama 2 and GPT-4. By adding the results from the previous experiment, I effectively added the human as a third M_{eval} for comparison. This means that items generated by all three M_{gen} models are evaluated by all three M_{eval} models.

This setup also allows us to check for interactions between M_{gen} and M_{eval} . For example, it is plausible that the evaluation could lead to an over-estimation of answerability and guessability in cases where $M_{\text{gen}} = M_{\text{eval}}$ (i.e., when a model evaluates items that it had generated itself).

	German	English
With text	Text: [T]	Text: [T]
	Frage: [q] Antwort: [a]	Question: [q] Answer: [a]
	Gemäß dem Text oben, ist diese Antwort richtig (R) oder falsch (F)? Gib nur den Buchstaben R oder F an.	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]	The following question and answer are from a multiple-choice comprehension task about an unknown text. Question: [q] Answer: [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.	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 5.1: 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.

5.3.3 Prompting

Similarly to Chapter 4, I manually engineered a zero-shot prompt for each setting, shown in Table 5.1, and applied greedy decoding for generation. The prompt instructs the models to generate the true/false responses as single character labels to make sure that they are tokenized as a single subword.

Responding to every answer option separately avoids the influence of other items and other answer options, as discussed in Section 4.5.3. From Figure 5.1, we can see that this setup is cleaner than in the human evaluation in that it only measures the effect of the models' prior knowledge and their reading comprehension ability.

While GPT-4 consistently produced output in the specified format, Llama 2 tended to formulate its response as a complete sentence or, in the setting without text, preface

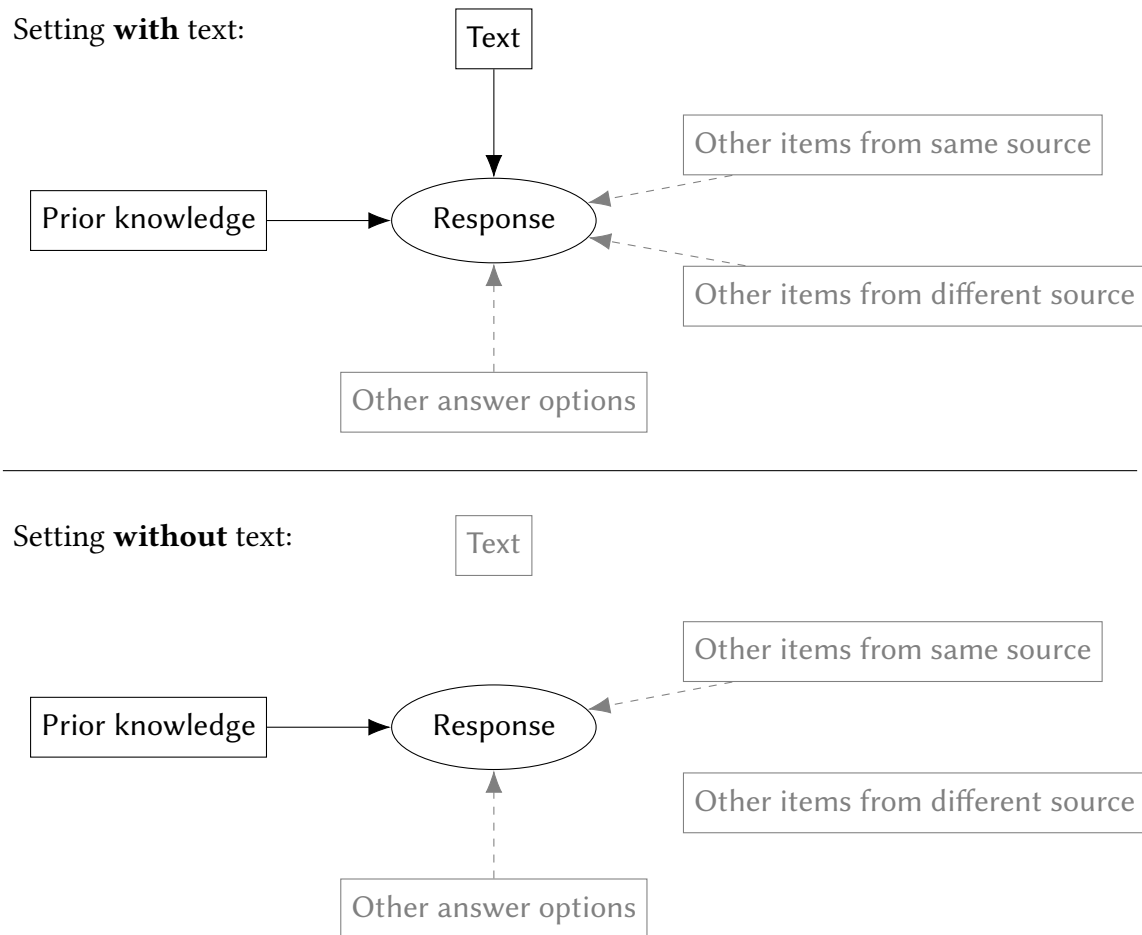


Figure 5.1: Sources of evidence for responding to comprehension items in the automatic evaluation (solid black lines) compared to the human evaluation (dashed gray lines). In the setting with text, when deciding whether a specific answer option is correct or incorrect, the LLM is able to consult the text and its own prior knowledge. In the setting without the text, only prior knowledge is accessible. Only a single answer option from a single item was included in each prompt.

it with a disclaimer like “Without looking at the text, it is difficult to say whether this answer is correct or not” without producing a label at all. I was unable to prevent this behavior through prompt engineering and instead used the predicted log-probabilities to decide which of the two labels was more likely to be generated as the first token in the generated response.

5.3.4 Threshold optimization

After some initial tests, it became clear that Llama 2 had a strong bias towards generating the *true* label, resulting in response accuracies close to random guessing. To counteract this bias, I adapted the classification threshold for positive responses. Specifically, I considered a response to be positive if

$$\frac{P(\text{true})}{P(\text{true}) + P(\text{false})} \geq \tau,$$

where $P(x)$ is the probability Llama 2 assigns to x as the first token of its response. This is equivalent to sampling from the predicted probability distribution of the first generated token, discarding tokens that are not one of the *true* or *false* labels, and then checking if the *true* label was produced with a relative frequency of at least τ .

To get the optimal threshold τ^* , I let Llama 2 respond to human-written items in a separate development split of 50 lessons in DWLG and selected the value for τ that maximizes response accuracy. I did this separately for the settings with and without text, resulting in two optimized thresholds $\tau_{\text{with text}}^* = 0.9952$ and $\tau_{\text{without text}}^* = 0.9849$. These optimized thresholds resulted in response accuracies of 0.880 with text and 0.657 without text on the development split. I used the same two thresholds for evaluating all items, no matter if they were human-written or generated or which dataset they originated from.

For GPT-4, I did not apply threshold optimization, because its application programming interface (API) did not return any token probabilities at the time of writing this thesis, but also because it did not appear to have a strong bias towards any of the labels.

The code I used for generating and parsing LLM responses and optimizing the threshold is available on GitHub¹.

¹<https://github.com/saeub/item-evaluation>

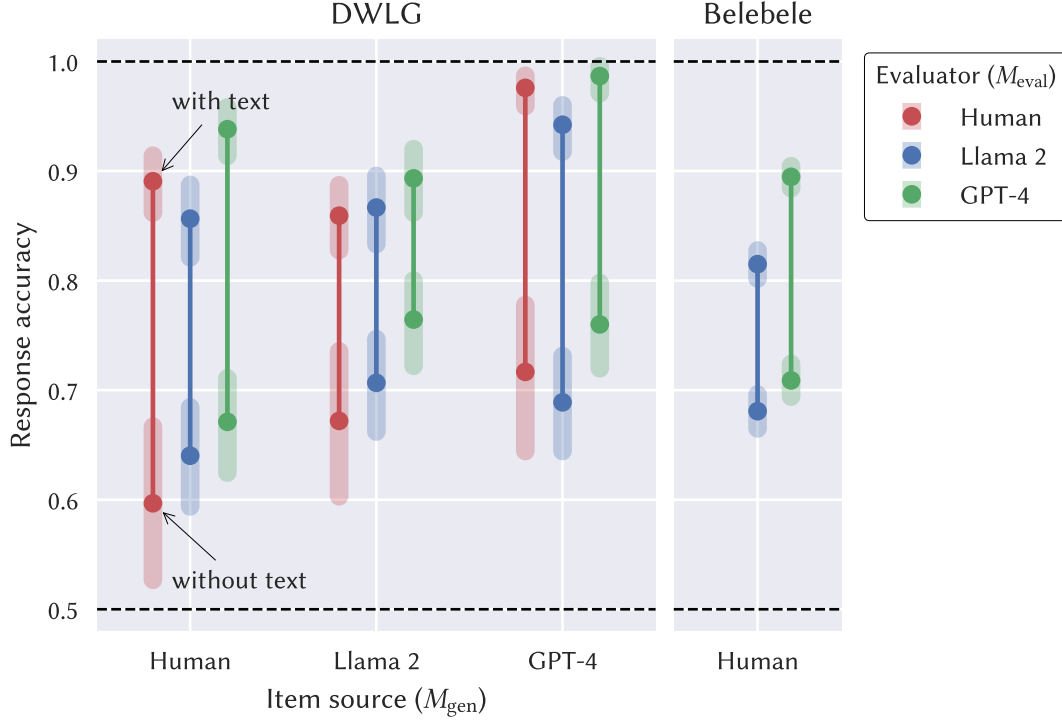


Figure 5.2: Mean human and LLM response accuracies on human-written and generated items. Accuracies are on the level of answer options, therefore random guessing is at 50%. For human evaluators, means are based on 10 texts and around 185 responses without text and around 546 responses with text. For LLM evaluators, means are based on 50 texts and around 451 responses in each setting for DWLG and 488 texts and 3600 responses for Belebele. Error bars are bootstrapped 95% confidence intervals.

5.4 Results

5.4.1 System-level guessability, answerability, and text informativity

Figure 5.2 shows system-level guessability and answerability estimates for all available combinations of M_{gen} and M_{eval} . For comparability, all response accuracies are calculated on the level of answer options, even if Belebele does not allow multiple correct answer options in the same item². The data for the human evaluators in this figure is the same as in Figure 4.3.

We can see that GPT-4 consistently achieved higher response accuracies than both hu-

²This also means that the response accuracies achieved by Llama 2 are not comparable to the 69.4% accuracy on German Belebele as reported by Bandarkar et al. (2023).

		Evaluator (M_{eval})		
		Human	Llama 2	GPT-4
Item source (M_{gen})	Human	0.294	0.216	0.267
	Llama 2	0.187	0.160	0.129
	GPT-4	0.259	0.253	0.227

Table 5.2: Text informativity estimates (\uparrow) for all combinations of M_{gen} and M_{eval} for DWLG. The best text informativity estimates according to each M_{eval} are marked in bold.

mans and Llama 2, with or without seeing the text. This shows that GPT-4 has very strong MRC capabilities as well as rich prior world knowledge that may exceed that of the human annotators. Llama 2 appeared to perform slightly worse or equal to humans in the setting with text, and slightly better or equal in the setting without text. However, most of these differences are not significant given the confidence intervals.

In general, the three evaluators mostly agreed on the observations already reported in Section 4.4.2. They all showed that the items generated by Llama 2 are the worst in terms of text informativity, and items generated by GPT-4 exhibit better answerability and worse guessability than human-written items. Humans and GPT-4 agreed that the human-written items are best in terms of text informativity, while evaluating with Llama 2 lead to a higher text informativity for items generated by GPT-4. Refer to Table 5.2 for a complete comparison of text informativity estimates for all combinations of M_{gen} and M_{eval} .

5.4.2 Response-level inter-annotator agreement

In order to compare the response behavior of Llama 2 and GPT-4 to that of human annotators, I report Cohen’s coefficient κ (Cohen, 1960) as a measure of inter-annotator agreement (IAA). Specifically, for each of the two LLMs, I computed response-level κ between the model and each human annotator separately and then calculated the mean of those values. Since the interpretation of Cohen’s κ is highly dependent on the use case (McHugh, 2012), I also provide the same mean κ values for each human annotator, to measure the average agreement between humans. All results are shown in Table 5.3.

Unsurprisingly, IAA is much higher in the setting with text than without. In both settings, agreement between GPT-4 and humans is comparable to or exceeds agreement between humans. Llama 2 has lower agreement overall, but still well within the range of inter-human agreements.

Evaluator (M_{eval})	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
Human mean	0.136	0.683
Llama 2	0.051	0.651
GPT-4	0.185	0.724

Table 5.3: Mean IAA (Cohen’s κ) between evaluators and (other) humans with and without text. Only human-written and generated DWLG items are included. *Human mean* is the mean of all six human annotators. The values in the setting without text are less reliable because each human only annotated a third of all items in this setting.

5.5 Discussion

5.5.1 Validity of the item evaluation protocol

As mentioned in Section 5.2, two main assumptions about M_{eval} are required for the evaluation protocol to be considered valid:

1. M_{eval} has similar reading comprehension skills to overly proficient humans (i.e., proficiency should exceed item difficulty by a large margin).
2. M_{eval} has similar world knowledge to humans.

We can test these assumptions for Llama 2 and GPT-4 by comparing their response behaviors to human annotators.

Although GPT-4 consistently outperformed humans in terms of response accuracy, its system-level guessability, answerability, and text informativity estimates were comparable to humans. The ranking of the different M_{gen} systems is mostly the same, no matter whether they were evaluated by humans or GPT-4 (see Table 5.2). Llama 2 was less reliable in this respect.

In terms of response-level behavior, both GPT-4 and Llama 2 performed similarly to humans, with IAA values well within the range of human variability (see Table 5.3).

This means that both Llama 2 and GPT-4 could be used to represent or substitute a single human annotator. However, it is important to note that the binary responses produced by the LLMs do not give us any information about label variation or confidence, which are essential aspects in human evaluation (Plank, 2022). Therefore, neither of the two models can be used to represent an entire ensemble of human annotators.

An additional concern mentioned in Section 5.3.2 was that using the same LLMs for both generation and evaluation may lead to biases when $M_{\text{gen}} = M_{\text{eval}}$. Based on Figure 5.2, it is possible that such a bias is present in Llama 2, since it only manages to outperform humans in terms of response accuracy (with text) when responding to items generated by Llama 2. However, compared to the results reported by Liusie et al. (2023) on language models fine-tuned on different MRC datasets, the effect here is rather small, if present at all.

Interestingly, applying the automatic evaluation to Belebele reveals that its text informativity is very low, which is unexpected given that the dataset was created with a rigorous review process to ensure quality. Compared to human-written items in DWLG, Belebele has both higher guessability and lower answerability. The higher guessability may be explained by the fact that items in Belebele always have one correct and three incorrect answer options. Because of this, there is less pressure on the item writers to make every single distractor unguessable, since there are always two other distractors. On the other hand, there should be *more* pressure to make the correct answer option answerable (i.e., unambiguously correct), since not doing so would immediately render the item useless. It is likely, however, that the answerability estimate of 0.895 for Belebele in Figure 5.2 do not represent human performance well. Belebele was created specifically as a challenge for MRC, and as a result, the assumption that item difficulty is much smaller than the test-taker’s proficiency (see Section 4.3.5.1) does not hold even for the highest-performing LLMs. Bandarkar et al. (2023) reported a human response accuracy of 97.6% on Belebele, but this value is not comparable because it was obtained in a conventional multiple-choice setting (i.e., seeing all answer options while selecting the correct one, with random guessing at 25%).

Overall, it appears that text informativity can be a reliable metric for automatic evaluation, if the evaluating LLM has a sufficiently high MRC performance. A consequence of this is that the metric is more suitable for use cases such as second language education, rather than challenge sets for MRC benchmarks. It also means that the evaluating LLM needs to be chosen in accordance with the difficulty of the items. GPT-4 generally seems to be a good choice for estimating answerability and for ranking text informativity, but its world knowledge may be too strong, leading to an over-estimation of guessability.

5.5.2 The case for LLM-based simulation of test-takers

Instruction-tuned LLMs are not designed to behave like humans – they are designed to follow instructions and give useful answers. This raises the question about the justification for using these models to simulate test-takers and evaluate item quality. There are several advantages to using zero-shot LLMs over fine-tuning task-specific models:

- Zero-shotting LLMs does not require additional training data, making them more easily usable for low-resource languages and small datasets.
- The model is dataset-agnostic, meaning that the same model can be used for comparing different datasets with a smaller risk of being biased towards one of them.
- Avoiding training simplifies implementation and reproducibility, especially compared to approaches involving large model ensembles like the one proposed by Byrd and Srivastava (2022).

The main disadvantages of using LLMs are their computational inefficiency and, currently, the dependency on commercial APIs for accessing state-of-the-art models such as GPT-4. While the high MRC performance required by the evaluation protocol is still a limiting factor, proprietary LLMs are inching closer towards both fine-tuned and human performance (Qin et al., 2023), and we can eventually also expect open-source models to become more feasible for item evaluation.

5.5.3 Limitations

While the automatic evaluation protocol has overcome some limitations of the human evaluation conducted in Chapter 4, such as the influence of other items and answer options on the responses (see Section 5.3.3), the experiment in this chapter also comes with its own limitations. First, the threshold optimization required for Llama 2 (see Section 5.3.4) means that it cannot be considered truly zero-shot anymore. Second, the number of items I used to compare human and LLM responses is still relatively low, leading to large confidence intervals in Figure 5.2 and limiting statistical power. A solution to this could be to extract multiple responses per answer option from each model by using several prompts paraphrased in different ways (see Portillo Wightman et al., 2023). Third, the generalizability to datasets other than DWLG needs further investigation. Especially for more challenging datasets such as Belebele, it is possible that the assumption of overly proficient evaluators does not hold anymore (even for GPT-4), and comparable human responses would need to be collected to test this.

5.6 Summary

The goal in this chapter was to develop a protocol for reference-free automatic evaluation of multiple-choice reading comprehension items. The approach I proposed estimates average guessability, answerability, and text informativity of a set of items by letting a highly performant large language model (LLM) respond (zero-shot) to each answer option and measuring response accuracy without the text (guessability) and with the text (answerability). I tested this approach with GPT-4 and Llama 2 and compared the results to the human evaluation from Chapter 4.

Evaluating with GPT-4 gave the most similar results to human annotators, both at the system level and the response level. In fact, inter-annotator agreement (IAA) between GPT-4 and humans exceeded IAA between humans (mean Cohen's $\kappa = 0.724$ vs. 0.683 for responses with text). Results from Llama 2 were also promising, but only after optimizing the classification threshold due to a strong bias towards positive responses. Using LLMs for evaluating generated items could be particularly useful for low-resource scenarios.

6 Improving item generation through fine-tuning

The results presented in Chapter 4 have shown that Llama 2 is capable of generating useful multiple-choice reading comprehension (MCRC) items in a zero-shot setting, but it clearly underperforms in comparison to the larger, closed-source GPT-4. In this chapter, I will investigate whether this performance gap can be reduced by fine-tuning Llama 2 on task-specific data. I will consider both items designed by humans and items automatically generated by GPT-4 as training data.

In accordance with the previous experiments, the evaluation will focus on guessability, answerability, and text informativity as indicators of item quality. The essential question is whether large language models (LLMs) are capable of identifying and learning these requirements implicitly from training examples. Chapter 5 has shown that very large models are necessary to even evaluate these metrics, giving reason to believe that learning to optimize them in generated items is difficult. Given the high level of abstraction at play and insights from benchmarks in previous work (Maynez et al., 2023), supervised fine-tuning is more promising than few-shot learning for this application.

Based on both the theoretical and practical motivations, two research questions guide the experiment in this chapter:

1. *Can large language models learn and implement the meaning of item guessability and answerability implicitly from training examples?*
2. *Can supervised fine-tuning on human- or LLM-generated data improve item generation performance in Llama 2?*

6.1 Background and related work

6.1.1 Efficient LLM fine-tuning

While fine-tuning pre-trained language models on task-specific data has been a very successful approach in many applications, the increasingly large number of parameters presents several efficiency-related challenges when applying the same approach to modern LLMs (Liu et al., 2022; Pfeiffer et al., 2023). In particular, updating the complete set of model parameters during training is computationally expensive. As a result, the research area of parameter-efficient fine-tuning has gained more attention in recent years.

One of the the most commonly used methods for parameter-efficient fine-tuning of LLMs is low-rank adaptation (LoRA) (Hu et al., 2021). In this approach, the pre-trained model weights are frozen and a subset of the matrix multiplications in the model are reparameterized with a small number of additional weights. Specifically, given a pre-trained weight matrix $W \in \mathbb{R}^{d \times k}$ and an input vector $x \in \mathbb{R}^d$, the forward pass $y = Wx$ is reparameterized as

$$y = Wx + BAx,$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are added weight matrices learned during fine-tuning. By choosing a low number for the rank r (much smaller than d and k), the total number of trainable parameters is kept to a minimum. Hu et al. (2021) applied this reparameterization to the attention weights in the transformer model and demonstrated that the approach is effective on a variety of downstream tasks including natural language generation while reducing the number of trainable parameters to 0.1% of the full model.

LoRA is effective at reducing the required computational resources compared to fine-tuning all parameters in the model, but in comparison to in-context (zero-shot or few-shot) learning, memory usage remains high, since the entire frozen base model still has to be stored in memory. Dettmers et al. (2023) presented QLoRA as a solution to this problem, where they applied a lower-precision quantization (e.g., 8-bit or 4-bit floating point precision) to the pre-trained parameters, while keeping the LoRA parameters at a higher precision (16-bit). As a result, fine-tuning a large model requires comparable amounts of memory to running inference with the same model, making experiments with LLM more feasible.

6.1.2 Training on LLM-generated data

Instruction-tuned LLMs have the remarkable ability to perform competitively in a wide range of natural language generation tasks without any training data (see Section 4.1.1), but they are computationally inefficient and potentially over-parameterized for simpler tasks. In response to this, recent research has attempted to prompt LLMs to generate training data for specific tasks and fine-tune smaller language models on that synthetic data. This is essentially a form of knowledge distillation, with the goal of creating a task-specific, more efficient, and possibly better-performing model (Meng et al., 2022).

Meng et al. (2022), Claveau et al. (2022), and Ye et al. (2022a) used GPT-style generative language models to generate training data for various classification tasks and fine-tuned smaller BERT-style masked language models on that data to perform the classification. Ye et al. (2022b) presented a framework where the prompt for generating training samples is iteratively updated in order to optimize data quality. Wang et al. (2023b) showed that instruction-tuning can be done by fine-tuning on filtered generated data, even when the same model is used for generating and fine-tuning. Whitehouse et al. (2023) used LLMs to augment multilingual datasets with low-resource languages.

Overall, these works show that training on LLM-generated data can be beneficial in scenarios where human-generated data is scarce, as long as it is of sufficiently high quality.

6.2 Experimental setup

6.2.1 Training data

I used two target datasets for fine-tuning:

1. **Human target:** The DWLG training set consisting of 354 human-written texts and items (see Section 3.3).
2. **GPT-4 target:** The texts from the DWLG training set with three items per text generated by GPT-4 using the zero-shot method described in Section 4.3.3.

For evaluation, I used the same test set of 50 texts as in Chapters 4 and 5. An additional split of 50 human-written texts and items served as a development set for testing hyperparameters and detecting overfitting during training.

Each training sample contained the same prompt as in Section 4.3.3. The specific format of the items roughly corresponds to the format produced somewhat consistently by

Llama 2 in the zero-shot setting. The full template is shown in Appendix B.1.

6.2.2 Fine-tuning

I used the instruction-tuned checkpoint `Llama-2-70b-chat-hf` as a starting point for fine-tuning. QLoRA fine-tuning was implemented using the `transformers` (Wolf et al., 2020), `peft` (Mangrulkar et al., 2022), and `bitsandbytes`¹ libraries. I loaded the original model parameters in 4-bit (NF4) precision and applied double quantization, while LoRA parameters are kept in 16-bit precision (Dettmers et al., 2023). I used rank $r = 64$ for LoRA.

Training was done on six NVIDIA V100 GPUs with 32 GB of memory each. Both models were trained for two epochs (after which the validation loss did not decrease any further) with a batch size of one, accumulating gradients over four steps. Refer to Appendix B.2 for the complete set of hyperparameters.

6.2.3 Inference and postprocessing

To generate items for texts in the test set, I used the same setup as in Section 4.3.3, with the exception of applying random sampling with a temperature of 0.5 instead of greedy decoding. The reason for this is that the fine-tuned models had a strong tendency to generate very repetitive items, sometimes with duplicate or near-duplicate answer options. Increasing the temperature prevented this behavior to some degree.

The previously observed issue that Llama 2 sometimes abruptly switches to English while generating German text (see Section 4.3.4) did not occur anymore after the first 40 fine-tuning steps, therefore language detection was not necessary.

6.2.4 Evaluation

I applied the automatic evaluation protocol presented in Section 5.2, using both GPT-4 and Llama 2 with zero-shot prompting in the role of M_{eval} to estimate system-level guessability, answerability, and text informativity. If the fine-tuned M_{gen} does in fact learn to imitate the desired characteristics of the items in the target dataset, we should expect those three metrics to shift towards the metrics of the target (human or GPT-4) dataset. To get more insights into the learning process, I evaluated the models at regular intervals during training.

¹<https://github.com/TimDettmers/bitsandbytes>

6.3 Results

6.3.1 Surface-level features of generated items

As mentioned in Section 4.4.1, the items generated by GPT-4 and Llama 2 in the zero-shot setting differ from the human-written items in DWLG at the surface level. Most importantly, the generated items did not contain any cloze-style stems and mostly only contained a single correct answer option. This means that fine-tuning Llama 2 on the human target involves a larger data shift compared to the GPT-4 target. From Figure 6.1, we can see that features like the average number of correct answer options or the percentage of cloze-style items approach the values of the target dataset after about 80 fine-tuning steps, meaning that these surface-level features were effectively learned after about one epoch.

6.3.2 Guessability, answerability, and text informativity

Figure 6.2 shows how guessability, answerability, and text informativity estimates of generated items changed as fine-tuning progressed. When fine-tuning on human-written items, we can see a clear decrease in both guessability and answerability within the first 120 update steps (about 1.5 epochs). Text informativity appears to have decreased slightly. When fine-tuning on GPT-4, all metrics remained relatively stable. The expected trend that the metrics of the generated items should approach the metrics of the target dataset is not visible in these results.

Overall, the two M_{eval} models (GPT-4 and Llama 2) largely agreed on these results, as shown by the large overlap between the green and blue areas in Figure 6.2. This can be considered additional evidence of the reliability of the automatic evaluation protocol.

6.3.3 Qualitative analysis

A closer look at the generated items confirms that the models mainly learned to imitate surface-level features. The issues that had already negatively impacted answerability in the zero-shot setting and require a deeper semantic understanding (see Section 4.4.4) remained present in the fine-tuned models. Both models also still exhibited occasional grammatical errors after fine-tuning.

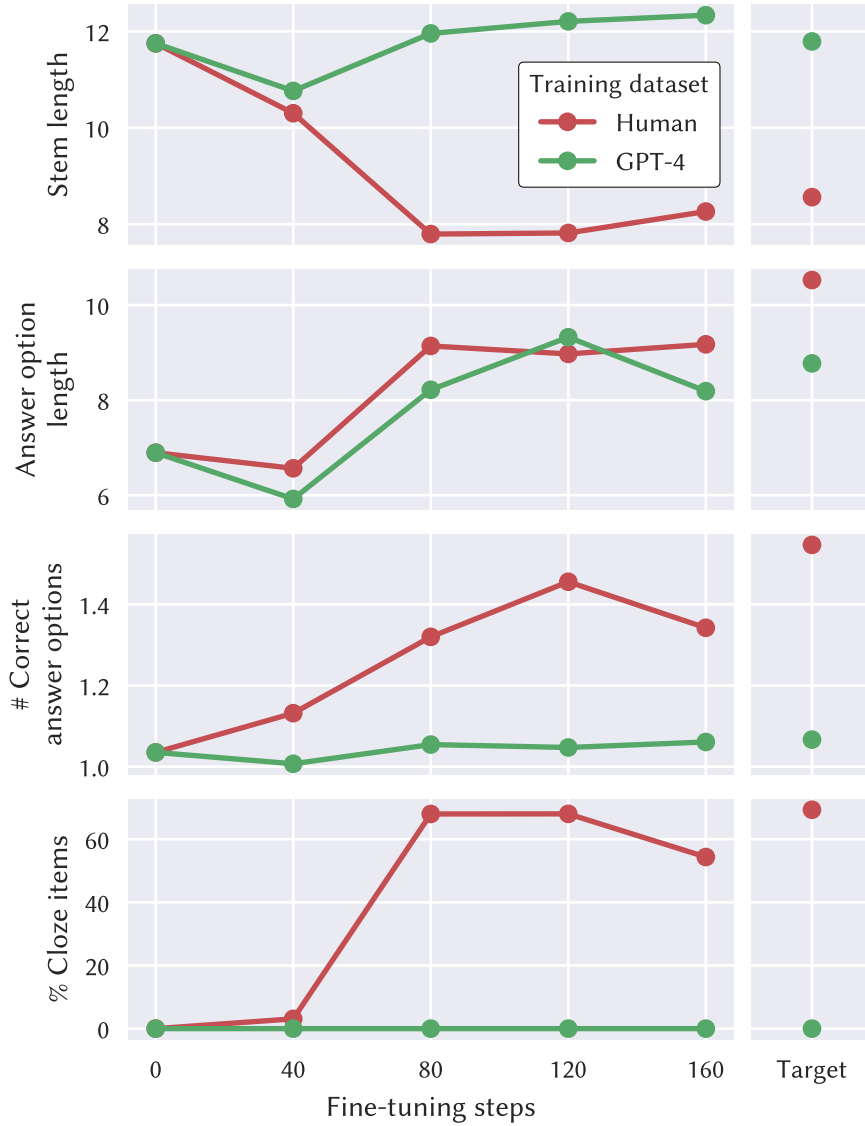


Figure 6.1: Changes in surface-level characteristics of generated items while fine-tuning Llama 2 on human-written and GPT-4-generated items. The y-axes represent the average stem length, the average answer option length, the average number of correct answer options per item, and the percentage of cloze-style items in the output generated by Llama 2. *Target* refers to the dataset the model is trained on. All values are calculated based on the test split of 50 texts.

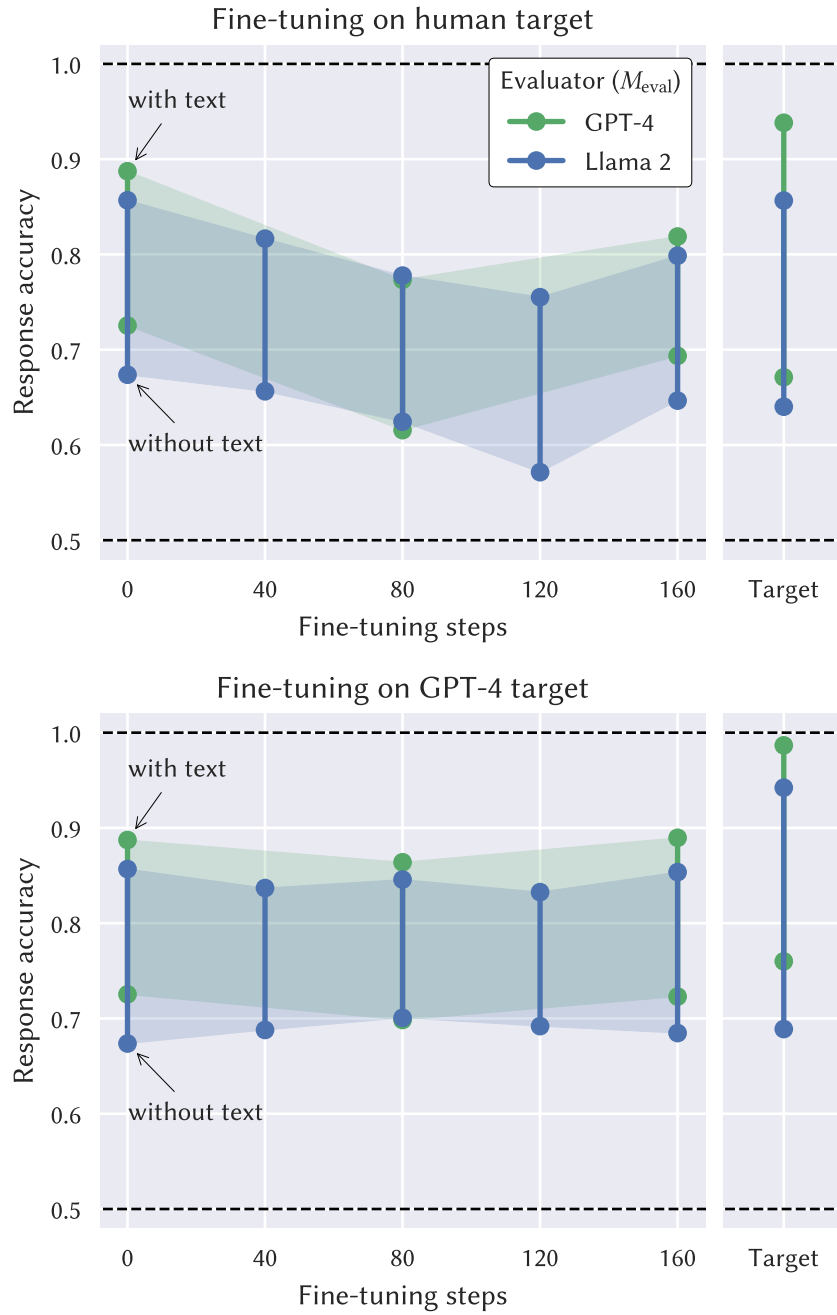


Figure 6.2: Changes in guessability (*without text*) and answerability (*with text*) while fine-tuning Llama 2 on human-written (upper plot) and GPT-4-generated (lower plot) items. *Target* refers to the dataset the model is trained on. All values are calculated based on the test split of 50 texts. Values at fine-tuning step 0 differ from Figure 5.2 because decoding was done using sampling with a temperature of 0.5 here. GPT-4 was only used as M_{eval} every 80 steps in order to reduce API costs.

Fine-tuning on the human target dataset additionally introduced new types of errors:

- **Cloze gap in inappropriate location:** The model generated a cloze-style stem, but the answer options do not syntactically fit in that location. In some cases, the answer options are also semantically incoherent.

Examples:

Item: (generated after 80 fine-tuning steps on human target)

Der Karneval hat ____ Ursprünge.

- ✗ im Christentum
- ✓ im alten Rom
- ✓ in der Zeit der Saturnalien

Item: (generated after 80 fine-tuning steps on human target)

Warum tragen auch Hollywood-Stars jetzt Sandalen und Socken ____

- ✗ weil sie sonst nicht so bequem sind.
- ✗ weil sie sonst nicht so schön sind.
- ✗ weil sie sonst nicht so praktisch sind.

- **Contradictory labels:** The model generated answer options whose correct/incorrect labels contradict each other. This sometimes occurred when the model generated highly repetitive answer options with similar meanings.

Examples:

Item: (generated after 40 fine-tuning steps on human target)

Wie lang hat der Karneval in Köln Tradition?

- ✗ 2000 Jahre [same as the third option, should be ✓]
- ✗ 40 Tage
- ✓ 2000 Jahre

Item: (generated after 80 fine-tuning steps on human target)

Andrea Liekweg ____

- ✗ ist eine Apothekerin [entailed by the third option, should be ✓]
- ✗ ist eine Krankenschwester
- ✓ ist eine Apothekerin, die eine eigene Medikamentenliste führt

In the model fine-tuned on the GPT-4 target dataset, I was unable to locate similar semantic issues, and there were generally fewer noticeable changes as training progressed.

6.4 Discussion

6.4.1 Learning item quality from data

The first research question in this chapter concerned whether it is possible for an LLM to recognize and replicate what makes items answerable or unguessable implicitly by fine-tuning on items that exhibit these features. To find an answer to this question, I fine-tuned Llama 2 on two different sets of items, both of which had a higher average text informativity metric than the ones generated by Llama 2 before fine-tuning. I found that neither model converged towards the guessability, answerability, or text informativity metrics of the target dataset, as shown in Figure 6.2.

While the model fine-tuned on the GPT-4 target did not appear to experience significant changes in item quality, the human target resulted in a noticeable degradation in answerability and introduced several semantic issues in the generated items. A possible explanation for this is the smaller data shift when fine-tuning on items generated by GPT-4, since those are much more similar to the items generated by Llama 2 before fine-tuning. It seems that the model prioritizes surface-level features, which are learned very quickly (as shown in Figure 6.1), sacrificing semantic integrity in return.

Assuming that these interpretations are accurate, there are two practical implications:

1. Fine-tuning should be done on items that have similar surface-level characteristics to the items generated by the model before fine-tuning. A solution could be to first engineer the prompt (e.g., by providing examples) to elicit items that look similar to the target dataset first, and then fine-tuning using that prompt.
2. Improving abstract features such as answerability and guessability through supervised fine-tuning may require larger-scale datasets than DWLG.

Given that there are no large MCRC datasets for most languages except English, alternative training approaches such as reinforcement learning from AI feedback (Bai et al., 2022) or reward ranked fine-tuning (Dong et al., 2023) may be more promising. In these approaches, items are sampled from M_{gen} , a reward quantifying item quality (e.g., text informativity) is calculated using M_{eval} , and M_{gen} is updated to maximize that reward. Such reward-based methods have the potential to solve both of the issues listed above, since sampling items from the model automatically eliminates the surface-level data shift and

removes the need for high-quality training data. However, a prerequisite for applying this would be a reward model that is able to evaluate guessability and answerability at the item level. In this thesis, I relied on binary responses from a small number of humans or zero-shot LLMs in the role of M_{eval} , which means that guessability and answerability can only be estimated reliably on the system-level (see Sections 4.3.5.1 and 5.2). Therefore, a next step would be to develop a method for obtaining human-like response probabilities at the level of single items, for example, by estimating confidence in LLM responses (see, e.g., Portillo Wightman et al., 2023).

6.4.2 Is Llama 2 a lost cause?

The second research question asked specifically whether supervised fine-tuning can improve Llama 2 as an open-source model beyond its zero-shot MCRC item generation performance. Based on the results in this chapter, this question cannot be answered conclusively. In addition to the size of the training dataset, there are two model-specific factors that could have limited the success of my fine-tuning experiments:

1. **Model size:** Although the number of parameters in GPT-4 is not public, it is safe to assume that it is significantly larger than Llama 2. It is possible that the level of abstraction involved in complex semantic features like guessability and answerability requires larger models.
2. **Model pre-training:** Llama 2 is English-centric with only a small fraction of pre-training data in German (see Section 4.3.2). This undoubtedly manifested itself in my experiments, for instance, whenever the model produced English or ungrammatical German output (see Sections 4.3.4 and 6.3.3). It may also mean that the language modeling capabilities for German are insufficient for automatic item generation and continued pre-training on German data may alleviate this problem. Preliminary results on LeoLM (Plüster, 2023) have shown moderate improvements, but German performance is still low compared to English.

Overall, Llama 2 with 70B parameters is still likely to be among the most competitive open-source LLMs currently available for this task, and it is plausible that improvements in item quality can be achieved by pre-training or fine-tuning on more data.

6.4.3 Limitations

There are several aspects limiting the interpretation of the results presented in this chapter. First, the experiment fully relied on automatic evaluation metrics. Although the

informal qualitative analysis gave a similar impression, a quantitative evaluation with human annotators should ideally be conducted to confirm the results. Second, the fine-tuning experiments were restricted to the DWLG dataset. In order to generalize the discussion to other datasets, further experiments would be required. Experiments with Belebele (Bandarkar et al., 2023) could be a good starting point for future work. Third, I did not systematically investigate the effect of hyperparameters on training, and there is a possibility that some improvement can still be achieved through hyperparameter optimization.

6.5 Summary

As a follow-up to Chapter 4, this chapter presented a fine-tuning experiment with Llama 2 in an attempt to improve its ability to generate multiple-choice reading comprehension items. I used low-rank adaptation (LoRA) to fine-tune a model on human-written items and another on GPT-4-generated items and evaluated the quality of the output using the automatic evaluation protocol developed in Chapter 5.

Although the models did learn to reproduce surface-level features such as item type or the number of correct answer options, they failed to produce items with a higher text informativity. When fine-tuning on human-written items, item quality deteriorated, while fine-tuning on GPT-4-generated items did not substantially affect item quality.

7 Conclusion and outlook

The aim of this thesis was to explore different ways of integrating large language models (LLMs) into the process of developing reading comprehension test items, focusing on automatic item generation (AIG) and evaluation. To this end, I compiled DWLG, a dataset of German multiple-choice reading comprehension (MCRC) items, I introduced *text informativity* as an evaluation metric motivated by item response theory (IRT), and I conducted experiments to test the performance of LLMs in generating and evaluating items. These contributions also represent first steps towards closing the three major research gaps presented in Section 1.4: the lack of non-English data, the lack of valid automatic evaluation metrics, and the lack of interdisciplinarity.

7.1 Answers to research questions

In this section, I will revisit the research questions asked in the experiment chapters and summarize the main findings to answer them. For more detailed summaries, refer to the sections at the end of each respective chapter.

How good are large language models at generating German multiple-choice reading comprehension items for given texts in a zero-shot setting? (Chapter 4)

To answer this question, I conducted a human evaluation of Llama 2 and GPT-4, collecting quality ratings and item responses from six annotators. GPT-4 outperformed Llama 2 in terms of text informativity, but did not reach the level of human-written items. On average, GPT-4 received higher quality ratings than human-written items, which is owed to better answerability. However, GPT-4 produced more guessable items.

Can responses by large language models to multiple-choice reading comprehension items be used to automatically evaluate item quality? (Chapter 5)

I compared item responses by Llama 2 and GPT-4 to human responses to answer this question. The experiment showed that guessability, answerability, and text informativity estimates by both models are similar to those by humans when evaluating generated or

human-written items – GPT-4 more so than Llama 2. On the response-level, GPT-4 showed stronger agreement with human annotators than Llama 2, but both were within the range of human variability. In summary, Llama 2 or GPT-4 can be representative of a single human annotator, but cannot replace a larger sample of annotators.

Can large language models learn and implement the meaning of item guessability and answerability implicitly from training examples? (Chapter 6)

To answer this question, I experimented with fine-tuning Llama 2 on human-written and GPT-4-generated items. While the model quickly learned surface-level features such as the item type or length, the guessability and answerability of the generated items did not improve. This suggests that learning guessability and answerability implicitly from examples requires more data.

Can supervised fine-tuning on human- or LLM-generated data improve item generation performance in Llama 2? (Chapter 6)

The quality of generated items did not improve in either case. When fine-tuning on human-written items, quality even deteriorated, possibly because these items have very different surface-level features. It is also possible that more German pre-training data is necessary.

Overall, these results show that LLMs have some potential to make item development more efficient through AIG and item evaluation, but open-source models like Llama 2 still have a long way to go in order to catch up with commercial products like GPT-4.

7.2 Open questions and future work

For AIG, the question remains whether and how the quality of items generated by LLMs can be improved, for example, through more systematic prompt engineering. As the fine-tuning experiments in this thesis were rather limited, future work could study larger or more German-centric open source models or experiment with different learning objectives or paradigms. Another line of research could be controlling the difficulty of generated items (Gao et al., 2019; Uto et al., 2023), since I deliberately excluded this aspect from my experiments. Kalpakchi and Boye (2023a) have done some preliminary work on difficulty-controlled generation with GPT-3 in Swedish, with limited success.

The experiments involving the automatic item evaluation protocol also left some questions open. For instance, it is unclear how well the evaluation works for datasets other than DWLG, and specifically for more difficult items. Moreover, there may be better

choices for open-source evaluator models than Llama 2, which gave poorly calibrated responses and required threshold optimization. Future work could also focus on enabling guessability and answerability evaluation of single items (as opposed to system-level evaluation) by modeling uncertainty and label variation at the response-level. This would also unlock learning paradigms such as reinforcement learning, where text informativity estimates could be used as a reward, similarly to what Yuan et al. (2017) proposed.

The amount of data available for German and other non-English languages remains a problem, and future work should focus on creating more high-quality datasets of test items. Building multilingually parallel datasets in the spirit of Belebele (Bandarkar et al., 2023) or QA4MRE (Peñas et al., 2011, 2012) may be the most efficient way of reducing data scarcity in low-resource languages. For non-English AIG, data scarcity may also be circumventable to some degree by automatically translating texts to English and translating the generated items back into the target language.

References

- I. Aldabe and M. Maritxalar. *Automatic Distractor Generation for Domain Specific Texts*, pages 27–38. Springer Berlin Heidelberg, 2010. ISBN 9783642147708. doi:10.1007/978-3-642-14770-8_5.
- E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo. The Falcon series of open language models. Nov. 2023. doi:10.48550/arXiv.2311.16867.
- J. Amidei, P. Piwek, and A. Willis. Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics, 2018. doi:10.18653/v1/w18-6537.
- T. Amstad. *Wie verständlich sind unsere Zeitungen?* PhD thesis, University of Zurich, Zürich, 1978.
- Anthropic. Model card and evaluations for Claude models, 2023. URL <https://efficient-manatee.files.svdcn.com/production/images/Model-Card-Claude-2.pdf?dm=1689034733>. Accessed: 25 Dec 2023.
- Y. Attali, A. Runge, G. T. LaFlair, K. Yancey, S. Goodwin, Y. Park, and A. A. von Davier. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5, July 2022. doi:10.3389/frai.2022.903077.
- Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan. Constitutional ai: Harmlessness from ai feedback. Dec. 2022. doi:10.48550/arXiv.2212.08073.

- O. Bajgar, R. Kadlec, and J. Kleindienst. Embracing Data Abundance. Feb. 2017. URL <https://openreview.net/forum?id=H1U4mhVFe>.
- L. Bandarkar, D. Liang, B. Muller, M. Artetxe, S. N. Shukla, D. Husa, N. Goyal, A. Krishnan, L. Zettlemoyer, and M. Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. Aug. 2023. doi:10.48550/arXiv.2308.16884.
- S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909>.
- M. A. Barton and F. M. Lord. An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, 1981(1):i–8, 1981. ISSN 2330-8516. doi:10.1002/j.2333-8504.1981.tb01255.x.
- G. Berger, T. Rischewski, L. Chiruzzo, and A. Rosá. 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, 2022.
- Y. Berzak, J. Malmaud, and R. Levy. STARC: Structured annotations for reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.507.
- M. Byrd and S. Srivastava. 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, 2022. doi:10.18653/v1/2022.acl-short.15.
- D. R. CH and S. K. Saha. Automatic multiple choice question generation from text: A survey. *IEEE Transactions on Learning Technologies*, 13(1):14–25, Jan. 2020. ISSN 2372-0050. doi:10.1109/tlt.2018.2889100.
- R. Circi, J. Hicks, and E. Sikali. Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8, May 2023. doi:10.3389/educ.2023.858273.
- V. Claveau, A. Chaffin, and E. Kijak. Generating artificial texts as substitution or complement of training data. In N. Calzolari, F. Béchet, P. Blache, K. Choukri,

- C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4260–4269, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.453>.
- J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960. ISSN 1552-3888. doi:10.1177/001316446002000104.
- Council of Europe. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg, 2020. ISBN 978-92-871-8621-8. URL <https://www.coe.int/lang-cefr>.
- I. Cuhadar. Sample size requirements for parameter recovery in the 4-parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives*, 20(2):57–72, Apr. 2022. doi:10.1080/15366367.2021.1934805.
- B. Das, M. Majumder, S. Phadikar, and A. A. Sekh. Automatic question generation and answer assessment: a survey. *Research and Practice in Technology Enhanced Learning*, 16:1–15, 2021.
- K. De Kuthy, M. Kannan, H. Santhi Ponnusamy, and D. Meurers. Towards automatically generating questions under discussion to link information and discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 2020. doi:10.18653/v1/2020.coling-main.509.
- T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. Technical report, May 2023. arXiv:2305.14314 [cs] type: article.
- Deutsche Welle. From the heart of Europe, 2008. URL <https://p.dw.com/p/E6K3>. Accessed: 25 Dec 2023.
- Deutsche Welle. DW Learn German, 2023. URL <https://learngerman.dw.com>. Accessed: 11 Oct 2023.
- R. Dijkstra, Z. Genç, S. Kayal, and J. Kamps. Reading comprehension quiz generation using generative pre-trained transformers. In *iTextbooks@AIED*, 2022.
- B. Dillon, A. Mishler, S. Sloggett, and C. Phillips. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103, Aug. 2013. doi:10.1016/j.jml.2013.04.003.

- H. Dong, W. Xiong, D. Goyal, R. Pan, S. Diao, J. Zhang, K. Shum, and T. Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. Technical report, May 2023.
- X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics*, 2017. doi:10.18653/v1/P17-1123.
- J. Dunietz, G. Burnham, A. Bharadwaj, O. Rambow, J. Chu-Carroll, and D. Ferrucci. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.701.
- D. Dzendzik, J. Foster, and C. Vogel. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.emnlp-main.693.
- B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, May 1994. ISBN 9780429246593. doi:10.1201/9780429246593.
- Z. Fei, Q. Zhang, T. Gui, D. Liang, S. Wang, W. Wu, and X. Huang. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.acl-long.475.
- R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi:10.1037/h0057532.
- Y.-C. Fung, L.-K. Lee, and K. T. Chui. An automatic question generator for Chinese comprehension. *Inventions*, 2023.
- Y. Gao, L. Bing, W. Chen, M. Lyu, and I. King. 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, Aug. 2019. doi:10.24963/ijcai.2019/690.
- B. Ghanem, L. L. Coleman, J. R. Dexter, S. M. von der Ohe, and A. Fyshe. Question generation for reading comprehension assessment by modeling how and what to ask. Apr. 2022. doi:10.48550/arXiv.2204.02908.

- M. J. Gierl and T. M. Haladyna. Automatic item generation: An introduction. In M. J. Gierl and T. M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 1, pages 3–12. Routledge, New York, 2013. ISBN 9780415897501.
- M. J. Gierl, H. Lai, and V. Tanygin. *Advanced Methods in Automatic Item Generation*. Routledge, Apr. 2021. doi:10.4324/9781003025634.
- T. Glushkova, A. Machnev, A. Fenogenova, T. Shavrina, E. Artemova, and D. I. Ignatov. DaNetQA: A yes/no question answering dataset for the Russian language. In *Lecture Notes in Computer Science*, pages 57–68. Springer International Publishing, 2021. doi:10.1007/978-3-030-72610-2_4.
- Q. Grail and J. Perez. ReviewQA: A relational aspect-based opinion reading dataset. Oct. 2018. doi:10.48550/arXiv.1810.12196.
- R. Green. Pilot testing: Why and how we trial. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 11, pages 115–124. Routledge, 2020. ISBN 9781351034784.
- C. Gütl, K. Lankmayr, J. Weinhofer, and M. Höfler. Enhanced automatic question creator–EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, 9(1): 23–38, 2011. ISSN 1479-4403.
- T. M. Haladyna. Automatic item generation: A historical perspective. In M. J. Gierl and T. M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 2, pages 13–25. Routledge, New York, 2013. ISBN 9780415897501.
- W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang. DuReader: A Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 2018. doi:10.18653/v1/w18-2605.
- K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, pages 1693–1701, Cambridge, MA, USA, Dec. 2015. MIT Press.
- D. Hewlett, A. Lacoste, L. Jones, I. Polosukhin, A. Fandrianto, J. Han, M. Kelcey, and D. Berthelot. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016. doi:10.18653/v1/p16-1145.
- L. Hirschman, M. Light, E. Breck, and J. D. Burger. Deep read. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -. Association for Computational Linguistics, 1999. doi:10.3115/1034678.1034731.
- E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. June 2021. doi:10.48550/arXiv.2106.09685.
- E. H. Jeon and J. Yamashita. Measuring L2 reading. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 25, pages 265–274. Routledge, 2020. ISBN 9781351034784.
- X. Jia, W. Zhou, X. Sun, and Y. Wu. EQG-RACE: Examination-type question generation. Dec. 2020. doi:10.48550/arXiv.2012.06106.
- G. Jones. Designing multiple-choice test items. In *The Routledge Handbook of Second Language Acquisition and Language Testing*, chapter 9, pages 90–101. Routledge, 2020. ISBN 9781351034784.
- D. Kalpakchi and J. Boye. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In T. Alumäe and M. Fishel, editors, *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands, May 2023a. University of Tartu Library. URL <https://aclanthology.org/2023.nodalida-1.48>.
- D. Kalpakchi and J. Boye. Generation and evaluation of multiple-choice reading comprehension questions for Swedish. 2023b. URL <https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-329400>.
- M. Karpinska and M. Iyyer. Large language models effectively leverage document-level context for literary translation, but critical errors persist. Apr. 2023. doi:10.48550/arXiv.2304.03245.
- A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. doi:10.1109/cvpr.2017.571.

- D. Khashabi, S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018. doi:10.18653/v1/n18-1023.
- T. Klein and M. Nabi. Learning to answer by learning to ask: Getting the best of GPT-2 and BERT worlds. Nov. 2019. doi:10.48550/arXiv.1911.02365.
- T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, Dec. 2018. doi:10.1162/tacl_a_00023.
- T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. May 2022. doi:10.48550/arXiv.2205.11916.
- T. Kolditz. Generating questions for German text. Master’s thesis, University of Tübingen, 2015.
- T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, Nov. 2019. doi:10.1162/tacl_a_00276.
- G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy. RACE: Large-scale reading comprehension dataset from examinations. Apr. 2017. doi:10.48550/arXiv.1704.04683.
- H. Lai and M. J. Gierl. Generating items under the assessment engineering framework. In M. J. Gierl and T. M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 6, pages 77–101. Routledge, New York, 2013. ISBN 9780415897501.
- J. P. Lalor, H. Wu, and H. Yu. 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, 2019. doi:10.18653/v1/d19-1434.
- R. P. Larsen and D. D. Feder. Common and differential factors in reading and hearing comprehension. *Journal of Educational Psychology*, 31(4):241–252, Apr. 1940. doi:10.1037/h0060424.

- W. G. Lehnert. *The Process of Question Answering*. PhD thesis, Yale University, New Haven, CT, 1977. URL <https://eric.ed.gov/?id=ED150955>.
- Y. Liang, J. Li, and J. Yin. A new multi-choice reading comprehension dataset for curriculum learning. In W. S. Lee and T. Suzuki, editors, *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 742–757. PMLR, 17–19 Nov 2019. URL <https://proceedings.mlr.press/v101/liang19a.html>.
- C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. May 2022. doi:10.48550/arXiv.2205.05638.
- Q. Liu, S. Jiang, Y. Wang, and S. Li. LiveQA: A question answering dataset over sports live. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1057–1067, Haikou, China, Oct. 2020. Chinese Information Processing Society of China. URL <https://aclanthology.org/2020.ccl-1.98>.
- A. Liusie, V. Raina, and M. Gales. “World knowledge” in multiple choice reading comprehension. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.fever-1.5.
- L. E. Lopez, D. K. Cruz, J. C. B. Cruz, and C. Cheng. Simplifying paragraph-level question generation via transformer language models. May 2020. doi:10.48550/arXiv.2005.01107.
- F. M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980. doi:<https://doi.org/10.4324/9780203056615>.
- S. T. Luu, K. T. Hoang, T. Q. Pham, K. Van Nguyen, and N. L.-T. Nguyen. A multiple choices reading comprehension corpus for Vietnamese language education. Mar. 2023. doi:10.48550/arXiv.2303.18162.
- D. Magis. A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, 37(4):304–315, Feb. 2013. doi:10.1177/0146621613475471.

- M. Majumder and S. K. Saha. Automatic selection of informative sentences: The sentences that can generate multiple choice questions. *Knowledge Management & E-Learning: An International Journal*, pages 377–391, Dec. 2014. ISSN 2073-7904. doi:10.34105/j.kmel.2014.06.025.
- M. Majumder and S. K. Saha. A system for generating multiple choice questions: With a novel approach for sentence selection. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*. Association for Computational Linguistics, 2015. doi:10.18653/v1/w15-4410.
- P. Manakul, A. Liusie, and M. Gales. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, and A. A. Krisnadhi, editors, *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, Nov. 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.ijcnlp-main.4>.
- S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- K. K. Maurya and M. S. Desarkar. 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*, 2020. doi:10.1145/3340531.3411997.
- J. Maynez, P. Agrawal, and S. Gehrmann. Benchmarking large language model capabilities for conditional generation. June 2023. doi:10.48550/arXiv.2306.16793.
- M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3): 276–282, Oct. 2012. ISSN 1330-0962. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>.
- Y. Meng, J. Huang, Y. Zhang, and J. Han. Generating training data with language models: Towards zero-shot language understanding. Feb. 2022. doi:10.48550/arXiv.2202.04538.
- T. W. Michel. Wissensgenerierung für deutschsprachige Chatbots. Master’s thesis, Hochschule Darmstadt, 2022.

- R. Mitkov, L. An Ha, and N. Karamanis. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12(2):177–194, May 2006. ISSN 1469-8110. doi:10.1017/s1351324906004177.
- T. Möller, J. Risch, and M. Pietsch. Germanquad and germandpr: Improving non-english question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.mrqa-1.4.
- N. Mulla and P. Gharpure. Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, pages 1–32, 2023. doi:10.1007/s13748-023-00295-9.
- P. Nema and M. M. Khapra. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi:10.18653/v1/d18-1429.
- K. V. Nguyen, K. V. Tran, S. T. Luu, A. G.-T. Nguyen, and N. L.-T. Nguyen. Enhancing lexical-based approach with external knowledge for Vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417, 2020. doi:10.1109/access.2020.3035701.
- NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. July 2022. doi:10.48550/arXiv.2207.04672.
- OpenAI. GPT-4 technical report. Mar. 2023. doi:10.48550/arXiv.2303.08774.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. Mar. 2022. doi:10.48550/arXiv.2203.02155.
- S. Papageorgiou, L. Davis, J. M. Norris, P. Garcia Gomez, V. F. Manna, and L. Monfils. *Design Framework for the TOEFL® Essentials™ Test 2021*. Educational Testing Service, 2021. URL <https://www.ets.org/Media/Research/pdf/RM-21-03.pdf>.

- A. Papasalouros, K. Kanaris, and K. Kotis. Automatic generation of multiple choice questions from domain ontologies. In *Proceedings of e-Learning 2008*, pages 427–434, Amsterdam, Netherlands, 2008. IADIS. ISBN 9789728924584. URL <https://www.iadisportal.org/digital-library/automatic-generation-of-multiple-choice-questions-from-domain-ontologies>.
- D. Paperno, G. Kruszewski, A. Lazaridou, N. Q. Pham, R. Bernardi, S. Pezzelle, M. Baroni, G. Boleda, and R. Fernandez. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2016. doi:10.18653/v1/p16-1144.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, ACL '02*. Association for Computational Linguistics, 2001. doi:10.3115/1073083.1073135.
- A. F. A. Paschoal, P. Pirozelli, V. Freire, K. V. Delgado, S. M. Peres, M. M. José, F. Nakasato, A. S. Oliveira, A. A. F. Brandão, A. H. R. Costa, and F. G. Cozman. Pirá: A bilingual Portuguese-English dataset for question-answering about the ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. ACM, Oct. 2021. doi:10.1145/3459637.3482012.
- A. Peñas, E. Hovy, P. Forner, Á. Rodrigo, R. Sutcliffe, C. Forascu, and C. Sporleder. Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. volume 1177 of *CEUR Workshop Proceedings*, Amsterdam, The Netherlands, Sept. 2011. CEUR. URL <https://ceur-ws.org/Vol-1177/#CLEF2011wn-QA4MRE-PenasEt2011>.
- A. Peñas, E. Hovy, P. Forner, Á. Rodrigo, R. Sutcliffe, C. Sporleder, C. Forăscu, Y. Benajiba, and P. Osenova. Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. volume 1178 of *CEUR Workshop Proceedings*, Amsterdam, The Netherlands, Sept. 2012. CEUR. URL <https://ceur-ws.org/Vol-1178/#CLEF2012wn-QA4MRE-PenasEt2012>.
- A. Peñas, Y. Miyao, A. Rodrigo, E. Hovy, and N. Kando. Overview of CLEF QA Entrance Exams Task 2014. volume 1180 of *CEUR Workshop Proceedings*, pages 1194–1200, Sheffield, UK, Sept. 2014. CEUR. URL <https://ceur-ws.org/Vol-1180/#CLEF2014wn-QA-PenasEt2014>.
- J. Pfeiffer, S. Ruder, I. Vulić, and E. M. Ponti. Modular deep learning. Feb. 2023. doi:10.48550/arXiv.2302.11529.

- P. Pirozelli, M. M. José, I. Silveira, F. Nakasato, S. M. Peres, A. A. F. Brandão, A. H. R. Costa, and F. G. Cozman. Benchmarks for Pirá 2.0, a reading comprehension dataset about the ocean, the Brazilian coast, and climate change. Sept. 2023. doi:10.48550/arXiv.2309.10945.
- B. Plank. 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, 2022. doi:10.18653/v1/2022.emnlp-main.731.
- B. Plüster. LeoLM: Igniting German-language LLM research, 2023. URL <https://laion.ai/blog/leo-lm/>. Accessed: 25 Dec 2023.
- G. Portillo Wightman, A. Delucia, and M. Dredze. 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, July 2023. Association for Computational Linguistics. doi:10.18653/v1/2023.trustnlp-1.28.
- C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. Is ChatGPT a general-purpose natural language processing task solver? Feb. 2023. doi:10.48550/arXiv.2302.06476.
- C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- V. Raina and M. Gales. Multiple-choice question generation: Towards an automated assessment framework. Sept. 2022. doi:10.48550/arXiv.2209.11830.
- V. Raina, A. Liusie, and M. Gales. Analyzing multiple-choice reading and listening comprehension tests. July 2023a. doi:10.48550/arXiv.2307.01076.
- V. Raina, A. Liusie, and M. Gales. Assessing distractors in multiple-choice tests. Nov. 2023b. doi:10.48550/arXiv.2311.04554.
- P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016. doi:10.18653/v1/d16-1264.

- P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018. doi:10.18653/v1/p18-2124.
- G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Number 1 in Studies in mathematical psychology. Danmarks Paedagogiske Institut, Oxford, England, 1960.
- M. Rathod, T. Tu, and K. Stasaski. Educational multi-question generation for reading comprehension. *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 2022.
- P. Rodriguez, J. Barrow, A. M. Hoyle, J. P. Lalor, R. Jia, and J. Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.acl-long.346.
- P. Rodriguez, P. M. Htut, J. Lalor, and J. Sedoc. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.insights-1.14.
- R. Rodriguez-Torrealba, E. Garcia-Lopez, and A. Garcia-Cabot. End-to-end generation of multiple-choice questions using text-to-text transfer transformer models. *Expert Systems with Applications*, 208:118258, Dec. 2022. doi:10.1016/j.eswa.2022.118258.
- A. Rogers, O. Kovaleva, M. Downey, and A. Rumshisky. Getting closer to AI complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731, Apr. 2020. doi:10.1609/aaai.v34i05.6398.
- A. Rogers, M. Gardner, and I. Augenstein. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45, Feb. 2023. doi:10.1145/3560260.
- A. Säuberli, S. Hansen-Schirra, F. Holzknacht, S. Gutermuth, S. Deilen, L. Schiffel, and S. Ebling. Enabling text comprehensibility assessment for people with intellectual disabilities using a mobile application. *Frontiers in Communication*, 8, Aug. 2023. ISSN 2297-900X. doi:10.3389/fcomm.2023.1175625.

- T. Sellam, D. Das, and A. Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.704.
- U. Shaham, M. Ivgi, A. Efrat, J. Berant, and O. Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. May 2023. doi:10.48550/arXiv.2305.14196.
- P. Shuai, Z. Wei, S. Liu, X. Xu, and L. Li. Topic enhanced multi-head co-attention: Generating distractors for reading comprehension. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.
- P. Shuai, L. Li, S. Liu, and J. Shen. QDG: A unified model for automatic question-distractor pairs generation. *Applied Intelligence*, 53:8275–8285, 2022.
- A. Singh Bhatia, M. Kirti, and S. K. Saha. *Automatic Generation of Multiple Choice Questions Using Wikipedia*, pages 733–738. Springer Berlin Heidelberg, 2013. ISBN 9783642450624. doi:10.1007/978-3-642-45062-4_104.
- S. Sinharay and M. S. Johnson. Statistical modeling of automatically generated items. In M. J. Gierl and T. M. Haladyna, editors, *Automatic Item Generation: Theory and Practice*, chapter 11, pages 183–195. Routledge, New York, 2013. ISBN 9780415897501.
- R. Smith, P. Snow, T. Serry, and L. Hammond. The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, 42(3):214–240, Feb. 2021. doi:10.1080/02702711.2021.1888348.
- R. Soricut and N. Ding. Building large machine reading-comprehension datasets using paragraph vectors. Dec. 2016. doi:10.48550/arXiv.1612.04342.
- J. H. Spyridakis and M. J. Wenger. An empirical method of assessing topic familiarity in reading comprehension research. *British Educational Research Journal*, 17(4): 353–360, 1991. ISSN 0141-1926. URL <https://www.jstor.org/stable/1500645>.
- K. Sun, D. Yu, J. Chen, D. Yu, Y. Choi, and C. Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, Nov. 2019. doi:10.1162/tac1_a_00264.
- K. Sun, D. Yu, D. Yu, and C. Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, Dec. 2020. doi:10.1162/tac1_a_00305.
- H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen,

- G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. July 2023. doi:10.48550/arXiv.2307.09288.
- A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 2017. doi:10.18653/v1/w17-2623.
- M. Uto, Y. Tomikawa, and A. Suzuki. 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, 2023. doi:10.18653/v1/2023.bea-1.10.
- S. Vajjala and I. Lucic. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, 2018. doi:10.18653/v1/w18-0535.
- C. Vania, P. M. Htut, W. Huang, D. Mungra, R. Y. Pang, J. Phang, H. Liu, K. Cho, and S. R. Bowman. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.acl-long.92.
- P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young,

- G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, and Y. Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, Feb. 2020. ISSN 1548-7105. doi:10.1038/s41592-019-0686-2.
- A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.
- B. Wang, T. Yao, Q. Zhang, J. Xu, and X. Wang. ReCO: A large scale chinese reading comprehension dataset on opinion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9146–9153, Apr. 2020. doi:10.1609/aaai.v34i05.6450.
- X. Wang, B. Liu, S. Tang, and L. Wu. SkillQG: Learning to generate question for reading comprehension assessment. May 2023a. doi:10.48550/arXiv.2305.04737.
- Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023b. doi:10.18653/v1/2023.acl-long.754.
- J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. Sept. 2021. doi:10.48550/arXiv.2109.01652.
- J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. Emergent abilities of large language models. June 2022. doi:10.48550/arXiv.2206.07682.

- C. Whitehouse, M. Choudhury, and A. Aji. LLM-powered data augmentation for enhanced cross-lingual performance. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore, Dec. 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.emnlp-main.44>.
- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- B. Wright and M. Stone. Best test design. *Measurement and statistics*, Jan. 1979. URL <https://research.acer.edu.au/measurement/1>.
- J. Xie, N. Peng, Y. Cai, T. Wang, and Q. Huang. Diverse distractor generation for constructing high-quality multiple choice questions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:280–291, 2022. doi:10.1109/taslp.2021.3138706.
- Y. Xu, D. Wang, M. Yu, D. Ritchie, B. Yao, T. Wu, Z. Zhang, T. Li, N. Bradford, B. Sun, T. Hoang, Y. Sang, Y. Hou, X. Ma, D. Yang, N. Peng, Z. Yu, and M. Warschauer. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.acl-long.34.
- L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.naacl-main.41.
- S. Yagcioglu, A. Erdem, E. Erdem, and N. Ikizler-Cinbis. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi:10.18653/v1/d18-1166.
- Y. Yang, W. tau Yih, and C. Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in*

- Natural Language Processing*. Association for Computational Linguistics, 2015. doi:10.18653/v1/d15-1237.
- Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018. doi:10.18653/v1/d18-1259.
- J. Ye, J. Gao, Q. Li, H. Xu, J. Feng, Z. Wu, T. Yu, and L. Kong. Zerogen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022a. doi:10.18653/v1/2022.emnlp-main.801.
- J. Ye, J. Gao, Z. Wu, J. Feng, T. Yu, and L. Kong. Progen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, 2022b. doi:10.18653/v1/2022.findings-emnlp.269.
- W. Yu, Z. Jiang, Y. Dong, and J. Feng. ReClor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HJgJtT4tvB>.
- X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Zhang, S. Subramanian, and A. Trischler. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Association for Computational Linguistics, 2017. doi:10.18653/v1/w17-2603.
- C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640, Oct. 2020. doi:10.3390/app10217640.
- T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto. Benchmarking large language models for news summarization. Jan. 2023. doi:10.48550/arXiv.2301.13848.
- Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou. Neural question generation from text: A preliminary study. Apr. 2017. doi:10.48550/arXiv.1704.01792.

A User interface for human evaluation

Errate die Antworten auf die folgenden Fragen

Es handelt sich um Verständnisfragen zu einem Zeitungsartikel. Versuche, die richtigen Antworten zu erraten, ohne den Text zu sehen. Im Anschluss wirst du den Text sehen und deine Antworten korrigieren können.

Es können jeweils **0-3** Antworten richtig sein.

Kreuze alle richtigen Antworten an (☑).

Was fordert die Frauenrechtlerin Masih Alinejad von den führenden demokratischen Ländern der Welt?

- ☐ Die Isolierung der Islamischen Republik
- ☐ Die Anerkennung der Islamischen Republik als demokratischen Staat
- ☐ Die Unterstützung der Islamischen Republik

Was war der Auslöser für den ersten feministischen Aufstand in der Geschichte des Iran?

- ☐ Der brutale Tod von Jina Mahsa Amini in Polizeigewahrsam
- ☐ Der Internationale Frauentag
- ☐ Die Förderung des Frauenbildes der Islamischen Republik

Wie reagierten die Sicherheitsbehörden auf die Proteste der Frauen?

- ☐ Sie reagierten mit Gewalt
- ☐ Sie unterstützten die Proteste
- ☐ Sie ignorierten die Proteste

Fertig

Figure A.1: Screenshot of the user interface for the human evaluation, without text.

Fromm und untergeordnet: Gegen das Frauenbild der Islamischen Republik gibt es seit 1979 Widerstand. Nach Jina Mahsa Aminis brutalem Tod wurde daraus der erste feministische Aufstand der iranischen Geschichte.

Doch 2022 gab es den ersten feministischen Aufstand der iranischen Geschichte. Auslöser war der brutale Tod von Jina Mahsa Amini in Polizeigewahrsam. „In unserer Stadt waren die Proteste beispiellos. In den ersten sieben Tagen waren drei Viertel der Protestierenden Frauen“, sagt Leila. Sie organisierte mit ihren Freundinnen Demonstrationen in ihrer Stadt in den iranischen Kurdengebieten.

„Die führenden demokratischen Länder der Welt müssen die Islamische Republik isolieren, genauso wie sie Putin isoliert haben“, sagt die Frauenrechtlerin Masih Alinejad. Sie fordert, die iranische Revolutionsgarde als Terrororganisation einzustufen. Andere Iranerinnen haben das Vertrauen verloren: „Die Unterstützung und Solidarität der westlichen Politikerinnen bedeutete uns am Anfang sehr viel“, sagt Leila. „Wir wissen aber, dass sie am Ende an ihre politischen und wirtschaftlichen Interessen denken. Wir machen unseren Kampf nicht abhängig von ihnen.“

Es können jeweils **0-3** Antworten richtig sein.

Wenn du dir bei einer Antwort unsicher bist (z.B., weil die Antwort nicht eindeutig ist), rate, und klicke zusätzlich auf das Fragezeichen (?).

Die iranischen Sicherheitsbehörden ...

☐ ☐ haben den Aufstand ausgelöst, weil eine junge Frau verhaftet und getötet wurde.

☐ ☐ sind für den Tod hunderter Demonstrierender verantwortlich.

☐ ☐ haben die Proteste 2022 gewaltsam beendet.

Wie gut ist dieses Item?

unbrauchbar mehrheitlich schlecht teilweise schlecht gut perfekt

Frauenrechtlerinnen sagen, dass ...

☐ ☐ die russische Regierung mit der iranischen Führung sprechen soll.

☐ ☐ sie sich nicht auf die Hilfe internationaler Politikerinnen verlassen.

☐ ☐ andere Länder mehr gegen die iranische Regierung machen sollen.

Wie gut ist dieses Item?

unbrauchbar mehrheitlich schlecht teilweise schlecht gut perfekt

Wie reagierten die Sicherheitsbehörden auf die Proteste der Frauen?

☐ ☐ Sie ignorierten die Proteste

☐ ☐ Sie reagierten mit Gewalt

☐ ☐ Sie unterstützten die Proteste

Wie gut ist dieses Item?

unbrauchbar mehrheitlich schlecht teilweise schlecht gut perfekt

Was war der Auslöser für den ersten feministischen Aufstand in der Geschichte des Iran?

☐ ☐ Der brutale Tod von Jina Mahsa Amini in Polizeigewahrsam

☐ ☐ Der Internationale Frauentag

☐ ☐ Die Förderung des Frauenbildes der Islamischen Republik

Wie gut ist dieses Item?

unbrauchbar mehrheitlich schlecht teilweise schlecht gut perfekt

Was fordert die Frauenrechtlerin Masih Alinejad von den führenden demokratischen Ländern der Welt?

☐ ☐ Die Anerkennung der Islamischen Republik als demokratischen Staat

☐ ☐ Die Unterstützung der Islamischen Republik

☐ ☐ Die Isolierung der Islamischen Republik

Wie gut ist dieses Item?

unbrauchbar mehrheitlich schlecht teilweise schlecht gut perfekt

Was ist richtig?

☐ ☐ Viele iranische Frauen wehren sich dagegen, wie sie von der Regierung behandelt werden.

☐ ☐ Der Muttertag ist für die Demonstrierenden ein wichtiges Datum.

☐ ☐ Am 8. März, dem Weltfrauentag, gibt es neue Proteste im Iran.

Wie gut ist dieses Item?

unbrauchbar mehrheitlich schlecht teilweise schlecht gut perfekt

- bezieht sich auf den Inhalt des Texts
- ist verständlich und sprachlich korrekt
- ist eindeutig beantwortbar
- ist ohne zusätzliches Allgemeinwissen beantwortbar
- ist nur beantwortbar, wenn man den Text gelesen hat (nicht durch reines Allgemeinwissen)

Fertig

96

B Fine-tuning details

B.1 Training sample format

<s> [INST] 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. [/INST] Frage 1:

[q₁]

A) [a_{1,1}] ([richtig/falsch])

B) [a_{1,2}] ([richtig/falsch])

C) [a_{1,2}] ([richtig/falsch])

Frage 2:

[q₂]

A) [a_{2,1}] ([richtig/falsch])

B) [a_{2,2}] ([richtig/falsch])

C) [a_{2,2}] ([richtig/falsch])

Frage 3:

[q₃]

A) [a_{3,1}] ([richtig/falsch])

B) [a_{3,2}] ([richtig/falsch])

C) [a_{3,2}] ([richtig/falsch]) </s>

T is a placeholder for the text, q_i for item stems, and $a_{i,j}$ for answer options. `<s>`, `</s>`, `[INST]`, `[/INST]` are tokens used as message separators in the instruction-tuned Llama 2 model. Note that in the human target dataset, there is a small number of items with more than three answer options (see Section 3.3.3).

B.2 QLoRA configuration

```
bnb_config = BitsAndBytesConfig(  
    load_in_4bit=True,  
    bnb_4bit_use_double_quant=True,  
    bnb_4bit_quant_type="nf4",  
    bnb_4bit_compute_dtype=torch.float16,  
)
```

```
peft_config = LoraConfig(  
    lora_alpha=16,  
    lora_dropout=0.1,  
    r=64,  
    bias="none",  
    task_type="CAUSAL_LM",  
)
```

```
training_args = TrainingArguments(  
    per_device_train_batch_size=1,  
    gradient_accumulation_steps=4,  
    learning_rate=1e-4,  
    logging_steps=10,  
    num_train_epochs=2,  
    evaluation_strategy="steps",  
    eval_steps=20,  
    save_strategy="steps",  
    save_steps=20,  
)
```