Annual Review of Applied Linguistics (2025), **45**, 222–252
doi:10.1017/S0267190525000066

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

How good are LLMs in generating input texts for reading tasks in German as a foreign language?

Anastasia Drackert[1,2], Andrea Horbach[3,4] and Anja Peters[1]

[1]Gesellschaft für akademische Studienvorbereitung und Testentwicklung (g.a.s.t.), Bochum, Germany; [2]Department of German Studies, Ruhr-Universität, Bochum, Germany; [3]Leibniz-Institute for Science and Mathematics Education, Kiel, Germany and [4]Institute for Psychology of Learning and Instruction, Christian Albrecht University, Kiel, Germany
Corresponding author: Anastasia Drackert; Email: drackert@gast.de

Abstract
To study the potential of generative AI for generating high-quality input texts for a reading comprehension task on specific CEFR levels in German, we investigated the comparability of reading texts from a high-stakes German exam used as benchmarks for the purpose of this study and those generated by ChatGPT (3.5 and 4). These three types of texts were analyzed according to a variety of linguistic features and evaluated by three assessment experts. Our findings indicate that AI-generated texts provide a valuable starting point for the production of test materials, but they require adjustments to align with benchmark texts. Computational analysis and expert evaluations identified key discrepancies that necessitate careful control of certain textual features. Specifically, modifications are needed to address the frequency of nominalizations, lexical density, the use of technical vocabulary, and non-idiomatic expressions that are direct translations from English. To enhance comparability with benchmark texts, it is essential to incorporate features such as examples illustrating the discussed phenomena and the use of passive constructions in the AI-generated content. We discuss the consequences of the usage of ChatGPT for input text generation and point out important aspects to consider when using generated texts as input materials in assessment tasks.

Abstract in German
Die Erstellung von Prüfungstexten für rezeptive Prüfungsteile ist ein zeitaufwendiger und ressourcenintensiver Prozess. Um das Potenzial generativer KI für die Erstellung von Input-Texten für eine high-stakes-Prüfung für Deutsch als Fremdsprache zu ermitteln, haben wir Lesepassagen, die von geschulten Autor*innen erstellt wurden, mit ChatGPT-generierten Texten verglichen. Die Texte wurden in Hinblick auf eine Reihe von linguistischen Merkmalen mittels einer computerbasierten Analyse ausgewertet und von drei Testerstellungsexpertinnen beurteilt. Unsere Ergebnisse zeigen, dass KI-generierte Texte einen wertvollen Ausgangspunkt für die Erstellung von Prüfungstexten bieten, aber Anpassungen erforderlich sind, um eine Vergleichbarkeit mit den Texten der

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Autor*innen zu erreichen. Insbesondere sind Modifikationen hinsichtlich der folgenden Aspekte notwendig: Veranschaulichung der dargestellten Inhalte durch Beispiele, lexikalische Dichte, Gebrauch von Fachvokabular, Idiomatik, Nominalisierungen und Passivkonstruktionen. Abschließend diskutieren wir die Konsequenzen der Nutzung von ChatGPT zur Erstellung von Input-Texten.

Keywords: readability assessment; reading comprehension; academic language proficiency; TestDaF exam

Introduction

The assessment of receptive language skills in language tests is intricately linked to the consideration of input texts and their linguistic properties. Next to the item types and the proficiency level of a learner, it is the characteristics of the input texts that have a significant impact on the difficulty level of assessment tasks (Révész & Brunfaut, 2013; Toyama, 2021). Ensuring the comparability of input texts across different versions of examinations, particularly in the context of high-stakes standardized tests, is therefore crucial for test providers (Fitzgerald et al., 2016). Consequently, readability research focusing on the analysis of the linguistic complexity of input materials has garnered significant attention within the field of language testing, particularly concerning English (Chen & Sheehan, 2015; Freedle & Kostin, 1993).

The process of identifying and revising suitable authentic texts for assessment tasks as a pivotal step of item-writing is challenging and time-consuming (Green & Hawkey, 2011; Salisbury, 2005). With the increasing refinement of artificial intelligence (AI) tools based on large language models (LLMs), such as ChatGPT, Llama and so on, generating input texts with the help of generative AI seems to be a promising option for developing test materials (Bolender et al., 2023; O'Sullivan, 2023). So far, it remains under-researched whether LLMs can produce texts for specific assessment contexts that are of the same quality and linguistic complexity as non-AI texts selected and adapted for assessment purposes by professional test writers. Especially for German, research has yet to systematically explore the capabilities and limitations of LLMs in generating texts for assessment purposes.

While larger English-speaking testing institutions have begun developing customized test development engines powered by LLMs (Attali et al., 2022; Bolender et al., 2023), there is still limited research on the extent to which currently available LLMs can produce high-quality input texts that align with specific levels of the CEFR levels, particularly in languages other than English. Given that input texts play a crucial role in determining task difficulty, understanding the linguistic characteristics and limitations of AI-generated texts is essential – not only for test developers but also for broader language learning and teaching contexts. This study contributes to this research gap by systematically analyzing the linguistic properties of AI-generated texts in comparison to benchmark texts used in a high-stakes German language exam, with a focus on their suitability for assessing academic reading proficiency at the B2/C1 level.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Literature review

LLMs

Generative LLMs are a particular type of language model with the capability to both encode and decode human language. Due to massive amounts of training data and the complexity of the model architecture, they can handle various topics in many languages (Min et al., 2023). They learn their knowledge about language from the co-occurrence of words in huge text corpora. Thus, they can produce fluent, coherent, and mostly error-free texts (Adesso, 2023). However, LLMs tend to hallucinate, that is, they make up facts (Alkaissi & McFarlane, 2023) without indicating so, and references, if given in a text, are often also made up and non-existent in reality (Ray, 2023).

LLMs work better for languages for which they have seen more training data and often show a bias towards a US-centric view (Feng et al., 2023). Although it is unknown what exactly the training data for ChatGPT looks like, one can estimate through comparisons with web content that it might have been exposed to about 10 times more content in English than in German (Petrosyan, 2024).

In the context of reading comprehension, LLMs have been used to produce simplified versions of authentic texts for their use in English-as-a-foreign-language settings (Young & Shishido, 2023) or reading comprehension tasks (Shin & Lee, 2023; Xiao et al., 2023), as detailed below.

Criteria of good input texts for assessment tasks

The routine of item writers encompasses not only the adaptation and revision of stimulus texts and the production of items but also the challenging process of text sourcing, which involves finding potential texts (Green & Hawkey, 2011; Salisbury, 2005). In developing reading tasks for assessment purposes, it is essential to follow certain rules in the selection of written texts. A critical requirement for the texts is their relevance to and representativeness of the target use domain and/or alignment with the designated learning objectives (Chapelle & Lee, 2021). In the context of evaluating reading proficiency for university admission purposes, it is, for example, imperative to employ texts reflective of those types of written texts that prospective students are likely to encounter during their studies, thereby ensuring content validity of the test (Green & Hawkey, 2011). While the information contained in the input texts should be understandable to non-experts, it should not be self-evident or common knowledge of the target group. Furthermore, the texts must align closely with the test specifications, which include, among other things, the assessed level of language proficiency, the target group, the genre, style, length, and subject matter of the text (Brunfaut, 2021).

The potential suitability of a text to facilitate the generation of specific item types to measure certain constructs should also be considered. For instance, from the perspective of a test construct, the ability to recognize text structure and overarching causal relationships is evaluated by requiring test-takers to arrange sections of a text in the correct order. Such a text, therefore, must be divisible into clear sections marked by appropriate connectors. Additionally, from the standpoint of item format, if multiple-choice questions are used, the text should encompass a breadth of information points, which are important for the development of plausible distractors (Brunfaut, 2021).

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Another important aspect regarding the suitability of a text as an input text is its complexity. This line of research has been so far conducted primarily for English tests (Chen & Sheehan, 2015; Freedle & Kostin, 1993). For example, using TextEvaluator, Chen and Sheehan (2015) compared the stimulus material of TOEFL *Primary, Junior*, and *iBT* according to eight groups of features that refer to syntactic complexity, vocabulary difficulty, academic orientation, argumentation, concreteness, cohesion, degree of narrativity, and style. The calculated ranges of scores, derived from the distributions of overall complexity as well as the component scores for each passage at various test levels, are used as benchmarks for the development or selection of new passages.

For the German language, there are several developmental projects in the area of German language learning that aim at assigning a CEFR level to German reading texts, for example, the DaFLex project, the CEFRSERV project as part of the European Language Grid (Rehm et al., 2020) or the Level-Adequate Texts in Language Learning Project (Vázquez-Ingelmo et al., 2023). However, to our best knowledge, there is no published research in the context of language assessment for the German language that compares human and AI-generated input texts for different CEFR proficiency levels.

LLMs and generation of reading assessment materials

Despite the expanding body of work on using AI for item generation (Bolender et al., 2023; Pugh et al., 2020), little empirical research in the area of language testing so far has focused on the capability of generative AI to produce high-quality input texts for specific assessment purposes. In an EFL context, Attali et al. (2022) investigated the quality of reading passages and items, which were generated using the GPT-3 model family and few-shot conditioning, through psychometric characteristics of the items as well as content and fairness reviews. While the authors outlined the criteria underpinning the evaluation of the LLM-generated test material, namely, content appropriateness, cohesion, clarity, and logical consistency, and reported that a total of 58% of passages were retained as a result of "all reviews and adjudication," they did not explicitly specify the shortcomings of the AI-generated passages that resulted in the exclusion of certain passages. Furthermore, and crucially, the study did not include benchmark texts as the basis for the analysis nor perform a linguistic analysis of the textual features. Additionally, it is noteworthy that the longest passages produced in their test context were capped at 175 words. This limitation is particularly relevant when compared to the longer texts usually employed in language admission exams within the German context, suggesting an area for further exploration.

Two other studies, each focusing on the English language as well, have contributed to our understanding of the capabilities of LLMs in generating reading comprehension test items by comparing them with benchmark texts. Specifically, this research examined the applications within the English section of the College Scholastic Ability Test in South Korea, as reported by Shin and Lee (2023), and in the context of middle school English learning in China, as detailed by Xiao et al. (2023). Despite utilizing different LLMs – Shin & Lee employed ChatGPT-3.5, while Xiao et al. evaluated a range of content generation models including a fine-tuned GPT-2, ChatGPT in a zero-shot configuration, and ChatGPT in a one-shot scenario – both studies discovered that the

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

LLM-generated texts were of a quality that was at least on par with, if not superior to, existing human-authored materials. For several reasons, the findings from these studies may not be directly applicable to our research context. Not only does our study focus on a different language and target proficiency, but it also takes a different methodological approach.

Shin and Lee's (2023) human evaluation focused on two aspects of the quality of the generated texts only (the natural flow of the passages and the naturalness of the English expressions) and did not incorporate any computational analysis. In contrast, Xiao et al.'s (2023) approach can be praised for its robust design, as it combined human and computational analyses. Their computational analysis focused on five features: negative log-likelihood loss, SMOG and Flesch grade, type-token ratio (TTR), and the proportion of repeated n-grams. Additionally, human evaluation assessed five aspects of the generated texts: readability, correctness, coherence, engagement, and overall quality. However, the study did not include several features relevant to academic reading passages, such as alignment with the target genre or analysis of syntactic complexity. Moreover, both studies relied on previous versions of ChatGPT available at the time of their research.

In her discussion of key validity issues for using generative AI in test development, Xi (2023) formulated an important question that test users should ask and test providers should answer, namely "is there research evidence that shows AI-generated test content edited by trained test developers can emulate the quality of test content created by human developers entirely?" (p. 369). Our exploratory study seeks to address a closely related question, namely how AI-generated material compares with content created by test developers based on non-AI resources, such as online magazines and other websites. This is done in the context of a reading comprehension task of a German exam for university admission purposes through the systematic analysis of the generated texts. In doing so, our study addresses a preliminary step towards answering the question posed by Xi. By examining the differences between these two types of material, we hope to identify those text features which test developers should pay attention to when editing AI-generated content for testing purposes. In this study, we focus on longer input texts. We combine a detailed computational analysis using a wide variety of linguistic features with a human review and original assessment texts as a benchmark.

Purpose of the study and research questions

The study aims at investigating the potential of generative AI to produce high quality input texts for reading comprehension by exploring the nuances of language production by AI in comparison to human test development, particularly focusing on two main research questions (RQs):

RQ1: How do AI-generated texts and texts created by human developers (benchmark texts) differ in their linguistic features?

RQ2: What differences do experienced test developers see between the two text types?

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Methodology

Study context and target task

The study was conducted in the context of TestDaF, a standardized language test for university admission purposes in Germany (Norris & Drackert, 2018). Since its primary purpose is to determine whether a candidate's German language proficiency in four language skills is sufficient for participation in German university studies, the exam includes authentic tasks and texts that students encounter at the starting phase of their studies and are similar to those which students might read in a university course or on campus.

The reading section of the paper-based TestDaF comprises three tasks (for a sample test, see g.a.s.t., TestDaF-Institut, 2020). Task 2, chosen for this study, is based on a reading passage of around 500 words with 10 multiple-choice items and its level corresponds to levels B2.2./C1.1 of the CEFR. The reading passages for this task consist of a report on an academic topic, outlining either a scientific study or the latest academic findings on a scientific phenomenon. They are based on texts that are taken from popular science magazines, or the websites of universities and scientific institutions. Ten multiple-choice items test the understanding of main ideas as well as detailed information from the input text. The last item is a "macro" item relating to the text as a whole.

Data

Benchmark texts

A total of 30 input texts for reading comprehension Task 2 written by experienced human test writers and used in the TestDaF exams were randomly selected for this study and served as a benchmark.

AI-generated texts

Based on the test specifications provided for item writers, we generated 30 texts on the topics of the benchmark texts using ChatGPT-3.5 and 4 each with the following prompt: **Write a report about a scientific study on the state of the art of research about "a name-of-a-specific-phenomenon" with many details. The text should be between 450 and 550 words long and must not include any lists or headings.** This prompt was a result of prior try-outs with different prompts. For example, we initially did not include the specification that the text should not contain any lists or headings. However, upon generating texts across five topics, it became apparent that the output frequently featured excessive lists and multiple subheadings, rendering them less suitable for task creation due to their overly detailed organization. This instruction was provided in German without further specifications: We did not provide the target CEFR-levels since previous research has shown that LLMs do not have accurate knowledge of the CEFR (Benedetto et al., 2025).

A total of 90 texts belonging to three text types were used for the analysis in this study as presented in Figure 1.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

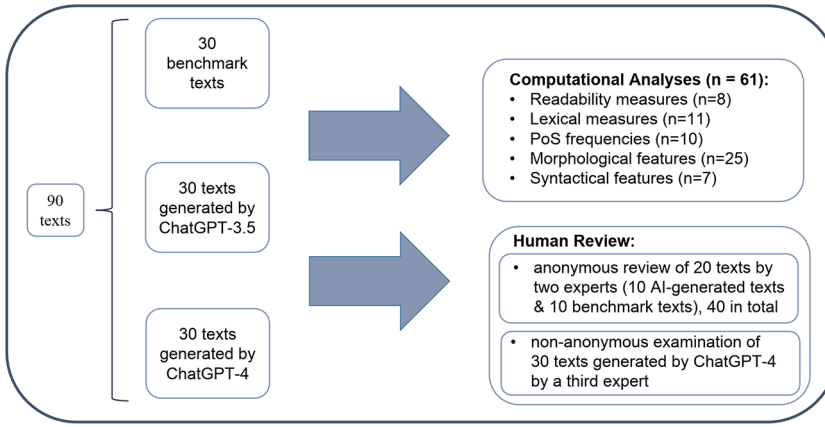


Figure 1. Illustration of the research design.

Computational analysis of the texts

Analyzed features

In our comparison of the three different text types, we used a set of linguistic features ($N = 61$) on different linguistic granularity levels (see Table 1 for an overview). Features from these categories were identified as indicators of linguistic complexity in previous work (see, for example, Hancke et al., 2012; Weiß, 2024; Weiss & Meurers, 2018). We explored a broad variety of textual features to identify possible differences between the three text types. We included features with a clear reference to important characteristics of the task at hand such as syntactic complexity, breadth of vocabulary or style (nominal vs verbal), making sure that the target construct is sufficiently covered.

Traditional readability measures. We used several traditional readability measures originally developed for English, which have been applied to German (Hancke et al., 2012), as well as the Wiener Sachtextformel (Bamberger & Vanecek, 1984), a readability metric specifically designed for German. Most readability formulas combine word complexity, often approximated through the number of syllables or characters per token, with sentence complexity, often measured by sentence length in tokens.

Lexical features. Our lexical features measured vocabulary breadth by using a corrected variant of TTR. TTR in its basic form is known to be dependent on text length (Koizumi & In'nami, 2012), therefore we used Moving Average TTR (MATTR, Covington & McFall, 2010), where equally-sized segments (100 tokens in our case) are repeatedly extracted from a text, and the average over the individual TTR values for these segments is used to model lexical variation. We further computed TTR for individual part-of-speech (POS) classes as well as lexical variation, that is, TTR for content words (nouns, verbs, adjectives) only, and measured lexical density as the relative frequency of content words among all words (see also Lu, 2014, p. 80).

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 1. Overview of the analyzed features per category

Category	#	Analyzed features
(1) Traditional Readability measures	8	ARI, COLEMAN_LIAU, KINCAID, SMOG, WSTF1, Average Number of Syllables per Word, Average Number of Characters per Token, Average Number of Tokens per Sentence
(2) Lexical measures	11	Lexical Variation, Lexical Density, Type-Token Ratio, TTR_ADJ, TTR_ADP, TTR_ADV, TTR_CONJ, TTR_DET, TTR_NOUN, TTR_PRON, TTR_VERB
(3) POS measures	10	ADJ, ADP, ADV, CONJ, DET, NOUN, NUM, PRON, PROPN, VERB
(4) Morphological features	25	Finite Verb Ratio, Frequency of Passive Sentences, Frequency of Passive Sentence with Bekommen, Frequency of Passive Sentence with Impersonal Pronoun, Frequency of Passive Sentence with Sich Lassen, Frequency of Passive Sentence with Zu, Frequency of Passive Sentence with Adjective, Frequency Of Typical Passive Sentence, Frequency of All Suffixes, Nominalization per Infinite Verb, Percentage of Nouns in Accusative, Percentage of Nouns in Dative, Percentage of Nouns in Genitive, Percentage of Nouns in Nominative, Frequency of suffixes *-HEIT, -IE, -KEIT, -MENT, -MUS, -NIS, -SCHAFT, -TION, -TUM, -TÄT, -UNG*
(5) Syntactical features	7	Average Syntax Tree Depth, Maximum Syntax Tree Depth, Average Number of Subordinate Clauses per Sentence, Average Number of Connectives per Sentence, Average Number of Subordinate Clauses with Conjunction per Sentence, Average Number of Infinitive Clauses per Sentence, Average Number of Relative Clauses per Sentence

POS features. Measuring the distribution of individual POS, such as verbs, nouns, adjectives, and so on, allows inferences on the relation between nominal versus verbal style in a text, the frequency of subordinate clauses (introduced by relative pronouns or conjunctions) or filler words, such as particles. It also gives us information about the usage of pronouns in comparison to common nouns or proper nouns. For this group of features, we evaluated the relative frequency of individual POS tags. As tag set, we used the coarse-grained Universal Dependency Tagset (Petrov et al., 2012) with 12 different POS tags, of which we included 10 in our analyses.

Morphological features. Morphological features measure aspects of word formation. We analyzed the texts using the Mate Morphological Tagger (Björkelund et al., 2010) and counted the relative frequency of nouns in different cases (nominative, genitive, dative, and accusative) among all nouns. As nominalizations play an important role in our text genre, we measured the frequency of different derivational suffixes used to form nouns in German (such as "-heit," "-keit," "-ung") in relation to the overall number of nouns (see Hancke et al., 2012). We further computed the ratio of finite verbs among all verbs (as a proxy for the complexity of the verb phrase) and the frequency of passive constructions.

Syntactic features. Our syntactic features model the grammatical complexity of sentences in a text. We measured the maximal and average depth of parse trees, which were obtained from the CoreNLP Parser (Manning et al., 2014), for individual sentences in a text with more deeply nested sentences receiving higher values (Chen & He, 2013).

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

We further computed the average number of relative clauses per sentence, the number of infinitive clauses per sentence, and the number of subordinate clauses introduced by a subordinate conjunction. Table 1 gives an overview of the analyzed features per category.

Text processing

All texts were linguistically processed using the LiFT toolkit (Zesch et al., 2021). This java-based toolkit makes use of various NLP preprocessing components provided through DkPro Core (Eckart de Castilho & Gurevych, 2014), such as tokenization, POS tagging, lemmatization, and parsing; and integrates the individual feature extractors in an UIMA pipeline (Ferrucci & Lally, 2004).

Statistical analysis

We computed the descriptive statistics for the features grouped by categories. To investigate the presence of statistical significance among individual features across the three distinct types of texts, we employed the Kruskal–Wallis test. This non-parametric method serves as an alternative to the Analysis of Variance when the assumptions of normal distribution and homogeneity of variance are not met, making it particularly suited for our dataset, which comprises frequency occurrences. The assumption of unequal variances was confirmed through Levene's test. For analyzing statistically significant differences between the three types of texts, we applied the Mann–Whitney U-test as a post-hoc analysis with two degrees of freedom ($df = 2$).

Human review

To enrich the computational quantitative analysis with an additional qualitative perspective, we asked two experienced TestDaF team members with extensive expertise in assessment and item writing within the relevant context to assess a subset of texts used in the automated analysis. In total, each of them evaluated 20 texts, comprising 10 generated by the latest version of ChatGPT4 and 10 benchmark TestDaF texts, against a set of predefined criteria. The human reviewers were unaware of the study's objectives and the fact that half of the texts were AI-generated. In total, they provided 40 evaluations – 20 for the AI-generated texts and 20 for the benchmark texts – for the specified criteria regarding vocabulary (Qs 1–2), syntactic complexity (Q3), and content/genre (Qs 4–7). They were also asked to identify any peculiarities within the texts and to provide specific examples of those (see Appendix 1 for the questionnaire). We conducted a Mann–Whitney U-test to investigate if the ratings for two text types differ statistically. Responses to the open-ended question where manually categorized with particular attention paid to multiple mentions of specific text features.

In addition, a third expert responsible for testing receptive skills at the TestDaF-Institute undertook a thorough examination of 30 texts generated by ChatGPT-4. In particular, this reviewer analyzed the texts regarding the correspondence to the target CEFR level as well as the known problems of LLM-generated texts such as biases and hallucinations.

Figure 1 gives an overview of the data and its analyses.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

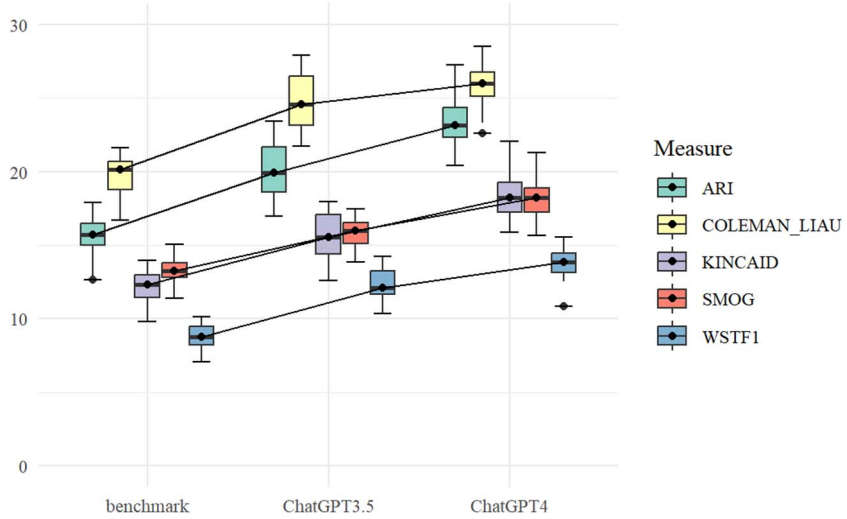


Figure 2. Box-plots for five readability measures for the three text types.

Results

In this section, we present a detailed comparison of linguistic features across the three text types employing computational analysis and the human evaluation of a sample of texts.

Computational analysis of texts

The results of the computational analysis of texts will be presented according to the feature categories analyzed.

Readability measures

The descriptive statistics provided in Appendix 2 (Table 2.1), along with the box-plot illustrations in Figure 2 depicting a selection of readability indices, demonstrate that texts generated by both versions of ChatGPT exhibit a higher score, that is, are more complex, compared to the benchmark texts. With the exception of the Coleman–Liau index, the observed differences in readability scores are statistically significant for all three text types as shown in significance tests for pairwise comparisons in Table 2.3 in Appendix 2. The biggest difference in the post-hoc Mann–Whitney *U*-tests for readability measures between the benchmark texts and ChatGPT4 texts was found for the ARI index ($U = -56.667$, $p < .001$).

Lexical measures

The descriptive statistics provided in Appendix 3, along with the box-plot illustrations in Figure 3, indicate that benchmark texts tend to have a higher MATTR than both types of ChatGPT texts ($M_{benchmark} = 0.76$; $M_{ChatGPT\text{-}3.5} = 0.70$; $M_{ChatGPT\text{-}4} = 0.72$), with the differences being statistically significant as shown in the post-hoc Mann–Whitney *U*-tests in Table 3.3 in Appendix 3. Regarding lexical variation, benchmark texts tend to

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

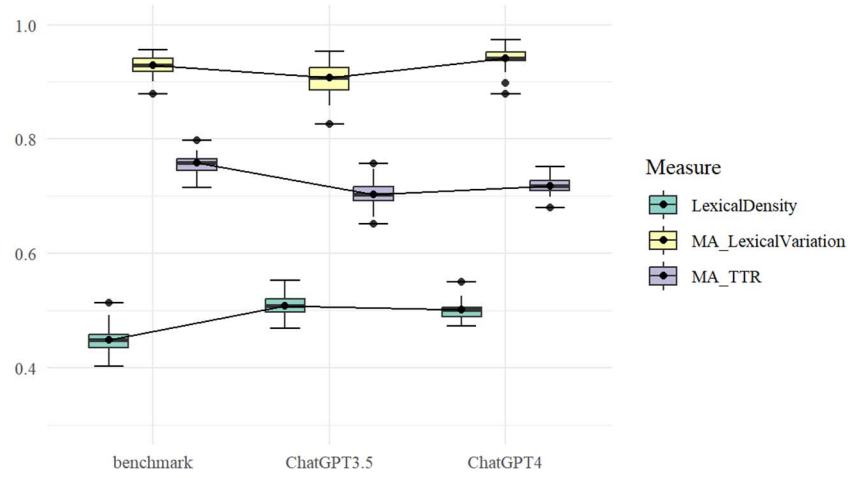


Figure 3. Box-plots for lexical measures for the three text types.

be closer to ChatGPT4 texts, but the texts do differ significantly on this measure. At the same time, ChatGPT texts tend to have more content words in comparison to all words and are thus more lexically dense, as seen in mean comparisons and significance tests in Tables 3.1 and 3.3 in Appendix 3 ($M_{benchmark} = 0.45$; $M_{ChatGPT\text{-}4} = 0.50$). In examining lexical density across text types, a Mann–Whitney U-test revealed significant differences between the benchmark and ChatGPT-4 texts ($U = -37{,}400$, $p < .001$), as well as between the benchmark texts and ChatGPT-3.5 texts ($U = -45{,}800$, $p < .001$).

POS analysis

As visualized in the box-plots in Figures 4 and 5 and reported in the Kruskal–Wallis test results, significant differences in the relative frequency of individual POS tags between three text types were found for all parts-of-speech but for prepositions.

In particular, as subsequent post-hoc Mann–Whitney U-tests (Table 4.3 in Appendix 4) showed, ChatGPT-4 texts tend to contain significantly more adjectives ($U = -41{,}067$, $p < .001$), more conjunctions ($U = -30{,}833$, $p < .001$), more determiners ($U = -38{,}067$, $p < 001$), and more nouns ($U = -45{,}883$, $p < .001$) than benchmark texts. At the same time, benchmark texts tend to have more adverbs, more numerals, more pronouns, more proper nouns, and more verbs (see Table 4.3 in Appendix 4).

Morphological complexity

In the investigation of 25 morphological features encompassing various dimensions of word formation related to verbs and nouns, the analysis revealed significant differences in several measures. Notably, as illustrated in Figure 6 and corroborated by statistical evidence in Appendix 5, the finite verb ratio and the prevalence of passive sentences were significantly higher in the benchmark texts compared to those generated by ChatGPT-4. This difference was substantiated by post-hoc Mann–Whitney U-tests, which yielded values of $U = 30.683$, $p < .001$ for the finite verb ratio and

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

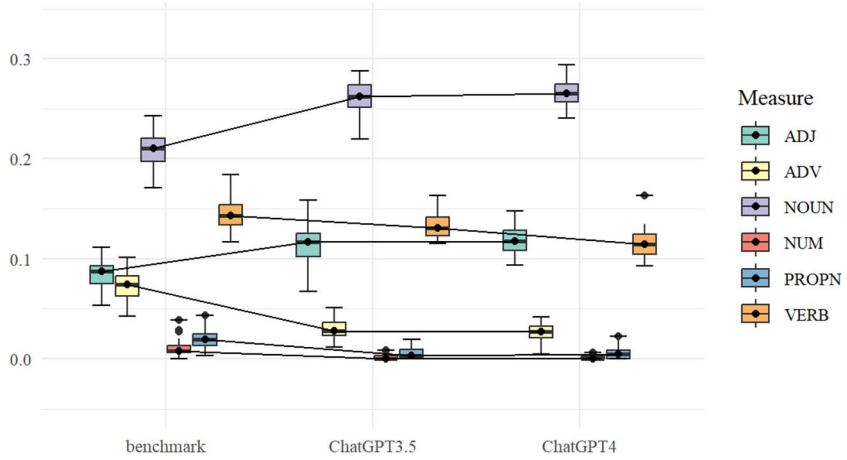


Figure 4. Box-plots for relative frequency of content words for the three text types.

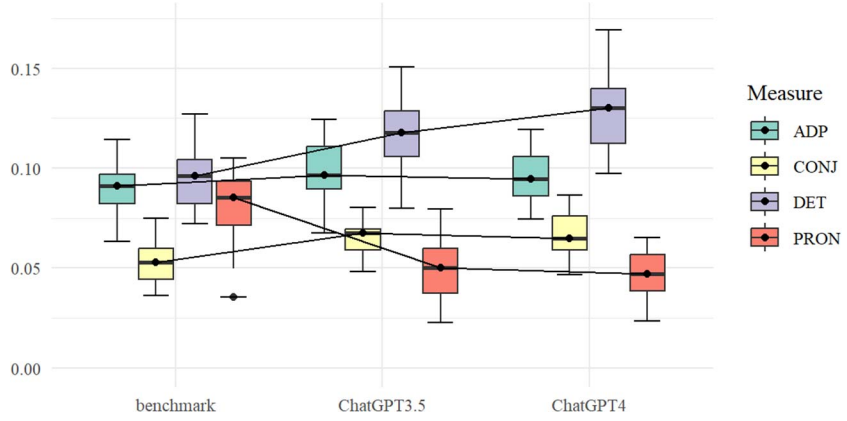


Figure 5. Box-plots for relative frequency of function words for the three text types.

$U = 21.200$, $p = .002$ for the frequency of passive sentences. By contrast, ChatGPT texts of both types tend to contain more nominalizations as measured by the feature frequency of all suffixes ($M_{benchmark} = 0.17$; $M_{ChatGPT\text{-}3.5} = 0.39$; $M_{ChatGPT\text{-}4} = 0.41$) with the differences being statistically significant as seen in Mann–Whitney U-tests in Table 6.3 in Appendix 6.

The analysis of case usage across three text types revealed distinct patterns: ChatGPT-generated texts utilize more nouns in the genitive and accusative cases, while benchmark texts contain more instances of the nominative case. The most pronounced difference was noted in the genitive case, as confirmed by the Kruskal–Wallis test ($H = 20.067$, $p < .001$). No significant differences were observed in the usage of the dative case.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

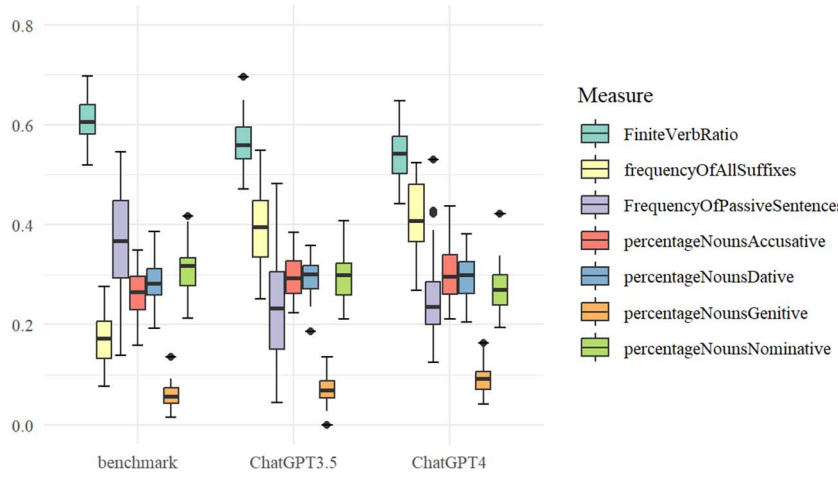


Figure 6. Box-plot for morphology features for the three text types.

Syntactic complexity

Significant differences between the three text types were also found for the measures of syntactic complexity. In particular, ChatGPT-4-generated texts tend to have on average a more deeply-nested sentence structure than benchmark texts and ChatGPT-3.5-generated texts ($M_{benchmark} = 5.08$; $M_{ChatGPT\text{-}4} = 5.89$), $U = -39.350$, $p < .001$. At the same time, the most complex sentence in benchmark texts is on average deeper nested than the most complex sentence in AI-generated texts ($M_{benchmark} = 11.33$; $M_{ChatGPT3.5} = 8.03$; $M_{ChatGPT4} = 9.60$) and confirmed by Mann–Whitney U-tests in Table 6.3 in Appendix 6.

When comparing the average number of different types of clauses per sentence as visualized in Figure 7 and summarized in Tables 6.1–6.3 in Appendix 6, ChatGPT-4-generated texts have the highest proportion of three types of clauses: subordinate clauses, infinitive clauses, and relative clauses. Furthermore, ChatGPT-generated texts have a higher average number of connectives per sentence ($M_{benchmark} = 1.24$; $M_{ChatGPT\text{-}3.5} = 1.55$; $M_{ChatGPT\text{-}4} = 1.92$), with Mann–Whitney U-tests showing statistical significance between three types of texts at the p-level lower than .003.

Human review

The analysis of human evaluations (see Table 2) revealed that ChatGPT-4-generated texts received higher scores across two linguistic categories (vocabulary and syntax), suggesting they are perceived as more complex by experts. However, the differences were not statistically significant. Regarding content, benchmark texts were found to contain more examples that illustrate the described phenomena ($M_{benchmark} = 1.95$, $SD = 0.61$) compared to ChatGPT-4-generated texts ($M_{ChatGPT\text{-}4} = 0.9$, $SD = 0.55$) on a scale from 0 to 4, with a U-value of 127 and a significance level of $p < 0.05$.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

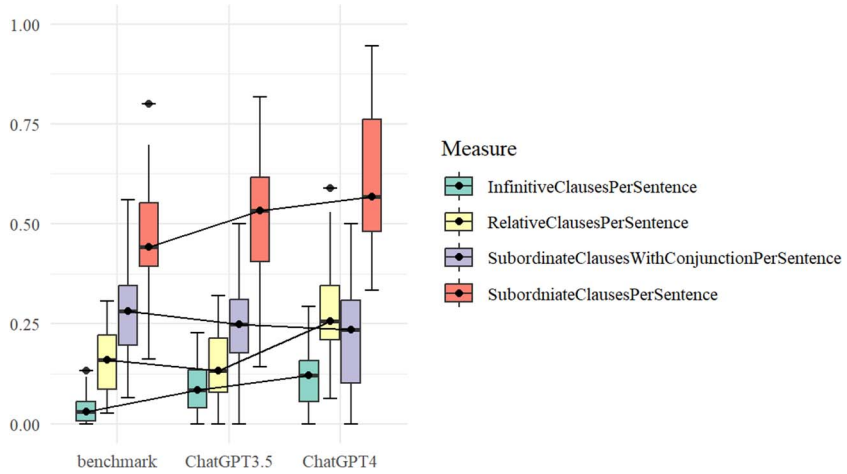


Figure 7. Box-plot for syntactic features for the three text types.

Table 2. Human reviewers' mean ratings of two text types

	Text type	M	SD	Min	Max
Specialized vocabulary	Benchmark	1.05	0.76	0	2
	ChatGPT-4	1.45	0.89	0	3
Breadth of vocabulary	Benchmark	2.15	0.49	1	3
	ChatGPT-4	2.45	0.51	2	3
Syntactic complexity	Benchmark	1.70	0.80	0	3
	ChatGPT-4	2.15	0.67	1	3
Text source	Benchmark	2.40	0.75	1	3
	ChatGPT-4	2.40	0.75	1	3
Predictable content	Benchmark	1.65	0.58	1	3
	ChatGPT-4	1.50	0.51	1	2
Examples*	Benchmark	1.95	0.61	1	3
	ChatGPT-4	0.90	0.55	0	2
Coherence	Benchmark	2.30	0.47	2	3
	ChatGPT-4	2.20	0.77	1	3

Furthermore, evaluations of text coherence, the origin of texts, and content predictability revealed no statistically significant differences between benchmark and ChatGPT4-generated texts.

The analysis of the comments in the open-ended questions revealed further differences between two text types. Human reviewers pointed to some cases where the use of some expressions was not idiomatic and seem to include translations from English, for example, *Rollenmodelle* (role models) and *nicht-CO2-Effekte* (non-CO2

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

effects). For these terms, more idiomatic German expressions exist (*Vorbilder; Effekte, die nicht durch CO2 verursacht werden*). Sometimes English expressions were used in the texts (e.g., "ride-hailing," "cetacean stranding") or technical terms that should ideally not be used or at least be explained, for example, *Sonar-Navigationssysteme* and *geomagnetische Anomalien* ("sonar navigation systems" and "geomagnetic anomalies"). Furthermore, human reviewers emphasized a large number of nominalizations in ChatGPT-generated texts and their rigid structure.

A third expert in test development conducted a qualitative analysis of 30 ChatGPT-4-generated sample texts, focusing on CEFR level alignment, bias, and hallucinations. The expert found a good correspondence with the B2.2/C1.1 levels and no blatantly incorrect information. Fact-checking confirmed the accuracy of content, which provided informative overviews accessible to readers from diverse academic backgrounds. However, the expert identified some weaknesses. First, all ChatGPT-4 texts followed the same general structure. A short introduction lists the main ideas that are to be discussed in the main body of the text. The texts conclude with a brief summary of the main ideas and a rather generic outlook on the future relevance of the topic. In contrast, the benchmark texts showed more structural variety. In addition, the AI-generated texts tended to list information rather than explain causal relationships. In some cases, especially in scientific outlines, the texts lacked the necessary detail, focusing on general information rather than specific study findings. Yet, when reporting recent scientific phenomena, these discrepancies were absent.

Discussion

This article delves into the role of LLMs in language testing and contributes to the understanding of the nuanced interplay between input texts, linguistic features, and the efficacy of LLMs. Specifically, the study explored the comparability between benchmark texts employed in a high-stakes German examination and LLM-generated texts, addressing a context that has been hitherto underrepresented in the domain of language testing research – that of German language exams.

The readability analysis revealed that ChatGPT-4 texts are more complex than those generated by ChatGPT-3.5 and benchmark texts, with longer sentences and words potentially posing increased cognitive demands on readers. Lexical analysis showed that, while benchmark texts have a broader lexical diversity across all parts-of-speech, ChatGPT-4 texts show greater variation in content words and a higher lexical density, indicating a denser concentration of information. Human reviewers noted no significant differences in vocabulary breadth but pointed out the AI-generated texts' use of technical and infrequent vocabulary.

In terms of parts-of-speech, it was found that ChatGPT-4 texts contain significantly more adjectives, conjunctions, determiners, and nouns, whereas benchmark texts have higher frequencies of adverbs, numerals, pronouns, proper nouns, and verbs. Morphologically, benchmark texts show a higher use of finite verbs and passive constructions, while ChatGPT texts have more nominalizations. Notably, ChatGPT texts contain more genitive and accusative cases, adding complexity to the text. Syntactically, ChatGPT-4 texts are characterized by deeper nested structures and a higher use of various clauses and connectives, compared to benchmark and ChatGPT-3.5 texts.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Despite the more rigid structure of AI-generated texts, human reviews evaluated their coherence as similar to that of benchmark texts.

Our findings suggest that AI-generated texts can be used as a starting point for creating new text inputs for Reading Task 2 of the paper-based TestDaF, especially for those instances of Task 2 that describe the state of the art of a scientific phenomenon. However, these texts should always be refined by experts. The differences in language, structure, and content revealed by the analyses provide clear targets for refinement of the AI-generated texts during the revision stage and for the refinement of the prompt used. For example, linguistically, the texts have to be checked for English expressions and technical terms as well as non-idiomatic expressions resulting from inadequate translations from English, a finding that is to be expected, taking into account that ChatGPT was trained primarily on English data (Zhang et al., 2023) and tends to use English as their internal pivot language (Wendler et al., 2024). Also, passages with clusters of nominalizations would have to be rephrased by using verbs and subordinate clauses, and as higher readability indices indicated, AI-generated texts might benefit from some simplification. The rigid structure that was observed in our data would benefit from a greater variety. Most importantly, however, the texts require more detailed information and examples in order to provide sufficient material for writing plausible and unambiguous multiple-choice items (Brunfaut, 2021).

The analysis revealed that the texts did not provide false information as suggested in the literature (Alkaissi & McFarlane, 2023) or bias (Feng et al., 2023) but included references that were made up primarily with regard to the scientific studies (Ray, 2023). Both aspects would need to be checked when using AI-generated texts.

The need for such revisions could be reduced by adapting the prompt. For instance, the prompt could ask for more details and relevant examples, and the text length could be increased in order to generate more material from which non-suitable sections could be removed. Furthermore, few-shot prompting, which involves providing an LLM with a few example texts to guide its generation, might improve their quality. By providing an LLM with samples of benchmark texts, and asking it to produce similar texts on a different topic, the generated texts will likely have more of the desired features than texts generated by zero-shot prompting. However, it is still to explore whether these revised prompts would have the desired effects since it has been shown that LLMs "cannot leverage (the information about the CEFR) to accurately perform educational tasks" (Benedetto et al., 2025, p. 13). Furthermore, in order to ensure test security, this method would only be suitable for use with customized LLMs or the workspace version of ChatGPT. In any case, a systematic follow-up analysis and evaluation of the texts by first language speakers are necessary.

When considering the implications for the differences found, we have to differentiate between those that can directly inform or instruct item developers how to change a text and those which point at small differences that might have little relevance in real life even if they are statistically significant. A finding that an AI-generated text has 50% more adverbs than a human-written text could lead to specific revisions of the generated text. In contrast, a slightly higher TTR value of a specific text type, though statistically significant, might not require immediate action.

The research reported in this article builds upon prior studies on AI-generated content suitability for test purposes (Attali et al., 2022; Shin & Lee, 2023; Xiao et al., 2023)

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

by expanding the evaluation criteria and parameters through a more comprehensive linguistic analysis and a direct comparison with benchmark texts. These criteria may vary in importance depending on the proficiency level, target language, or genre. For instance, passive constructions may be less relevant at lower proficiency levels.

Our findings emphasize the need for a hybrid evaluation approach that combines human review with computational analysis. For instance, the computational analysis has demonstrated that AI-generated texts often contain fewer proper nouns or numerals compared to benchmark texts. However, it is only through human analysis that these characteristics can be explained by the lack of specific examples and detailed information in the descriptions of research studies within the generated texts.

Limitations

This study represents a cross-sectional analysis, that is, it captures the development at a certain point in time. The findings might not be applicable to future versions of ChatGPT. The intransparency of commercial LLM solutions and the to-date inferior performance of open-source solutions (Gudibande et al., 2023) makes it infeasible to track changes in an AI system with certainty. Therefore, more research is needed to confirm the findings of this study. Also, a direct comparison between both ChatGPT versions was out of scope for our study, although such an analysis could be done based on the statistical data provided in Appendices 2–6.

In our study, we opted for a fixed prompt and did not systematically explore the vast option space for prompt engineering and refining the obtained results through multiple dialogue turns. This strategy has been successfully applied in, for example, the domains of question answering (Liu et al., 2024) or information extraction (Wei et al., 2023), and we are aware that this might have a large influence on text quality. Furthermore, we did not fine-tune the LLM using human-generated texts as training material – also for the reason that such texts should not be made public or exposed to a language model.

Concluding remarks

This study offers empirical evidence that can assist test developers and developers of learning materials in refining AI-generated text content to match the quality of content created by human developers based on non-AI sources. By transparently highlighting the differences between AI-generated texts and benchmark texts (Burstein, 2023) within the framework of a German reading comprehension exam used for admission purposes, this research enhances our understanding of the potential applicability of LLMs in new contexts. Notably, we utilized a widely accessible LLM for text generation, making our findings particularly relevant for test developers and language teachers who may not have extensive resources available. The importance of conducting further replication studies cannot be overstated, as these would not only validate our results but also aid LLM users in comprehending ongoing changes in the technology. Additionally, extending this research to languages other than English could provide broader insights into the adaptability and effectiveness of LLMs across different linguistic landscapes.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Acknowledgments. We would like to thank colleagues from g.a.s.t. e.V. and anonymous reviewers for constructive feedback on the previous versions of the manuscript.

References

Adesso, G. (2023). Towards the ultimate brain: Exploring scientific discovery with ChatGPT AI. *AI Magazine*, *44*(3), 328–342. https://doi.org/10.1002/aaai.12113

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, *15*(2), e35179. https://doi.org/10.7759/cureus.35179

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, *5*, 1–13. https://doi.org/10.3389/frai.2022.903077

Bamberger, R., & Vanecek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben: Die Schwierigkeitsstufen von Texten in deutscher Sprache* [Reading–comprehension–learning–writing: The levels of difficulty of texts in the German language Jugend und Volk. https://doi.org/10.2307/3530491

Benedetto, L., Gaudeau, G., Caines, A., & Buttery, P. (2025). Assessing how accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, *8*, 100353. https://doi.org/10.1016/j.caeai.2024.100353

Björkelund, A., Bohnet, B., Hafdell, L., & Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. *Proceedings of COLING 2010: Demonstrations* (pp. 33–36). COLING 2010 Organizing Committee. http://www.aclweb.org/anthology/C/C10/C10-3009.pdf

Bolender, B., Foster, C., & Vispoel, S. (2023). The criticality of implementing principled design when using AI technologies in test development. *Language Assessment Quarterly*, *20*(4–5), 512–519. https://doi.org/10.1080/15434303.2023.2288266

Brunfaut, T. (2021). Assessing reading. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 254–267). Routledge.

Burstein, J. (2023). The Duolingo English test responsible AI standards. Retrieved March 29, 2024. Available at https://go.duolingo.com/ResponsibleAI

Chapelle, C. A., & Lee, H. (2021). Understanding argument-based validity in language testing. In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 19–44). Cambridge University Press. https://doi.org/10.1017/9781108669849.004

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1741–1752). Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1180.pdf

Chen, J., & Sheehan, K. M. (2015). Analyzing and comparing reading stimulus materials across the TOEFL® Family of Assessments. *ETS Research Report Series*, *2015*(1), 1–12. https://doi.org/10.1002/ets2.12055

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100. https://doi.org/10.1080/09296171003643098

Eckart de Castilho, R., & Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT* (pp. 1–11). Association for Computational Linguistics and Dublin City University. https://doi.org/10.3115/v1/W14-5201

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11737–11762). https://doi.org/10.18653/v1/2023.acl-long.656

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3–4), 327–348. https://doi.org/10.1017/S1351324904003523

Fitzgerald, J., Elmore, J., Relyea, J. E., Hiebert, E. H., & Stenner, A. J. (2016). Has first-grade core reading program text complexity changed across six decades? *Reading Research Quarterly*, *51*(1), 7–28. https://doi.org/10.1002/rrq.115

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items. *ETS Research Report Series*, *1993*(1), i–48. https://doi.org/10.1002/j.2333-8504.1993.tb01524.x

g.a.s.t., TestDaF-Institut (2020). *Modelltest 01 - Leseverstehen*. https://www.testdaf.de/fileadmin/testdaf/downloads/Modelltests_papierbasierter_TestDaF/Modelltest_1/Lesen/Modelltest_01_LV_Heft.pdf Retrieved May 8, 2025.

Green, A., & Hawkey, R. (2011). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, *29*(1), 109–129. https://doi.org/10.1177/0265532211413445

Gudibande, A., Wallace, E., Snell, C., Geng, X., Liu, H., Abbeel, P., … Song, D. (2023). The false promise of imitating proprietary LLMs. *arXiv preprint arXiv:2305.15717*. https://doi.org/10.48550/arXiv.2305.15717

Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. *Proceedings of COLING 2012* (pp. 1063–1080). The COLING 2012 Organizing Committee. https://aclanthology.org/C12-1065/

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

Liu, Z., Ping, W., Roy, R., Xu, P., Lee, C., Shoeybi, M., & Catanzaro, B. (2024). ChatQA: Surpassing GPT-4 on conversational QA and RAG. *Advances in Neural Information Processing Systems*, *37*, 15416–15459. https://proceedings.neurips.cc/paper_files/paper/2024/file/1c0d54ebd0a6e58c4eca7d591e374b9d-Paper-Conference.pdf

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Springer.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60). https://doi.org/10.3115/v1/P14-5010

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., … Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, *56*(2), 1–40. https://doi.org/10.1145/3605943

Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing*, *35*(1), 149–157. https://doi.org/10.1177/0265532217715848

O'Sullivan, B. (2023). Reflections on the application and validation of technology in language testing. *Language Assessment Quarterly*, *20*(4–5), 501–511. https://doi.org/10.1080/15434303.2023.2291486

Petrosyan, A. (2024). Most used languages online by share of websites 2023. *Statista*. Retrieved April 29, 2024, Available at https://www.statista.com/statistics/262946/most-common-languages-on-the-internet

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). European Language Resources (ELRA). http://lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning*, *15*(1), 1–13. https://doi.org/10.1186/s41039-020-00134-8

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rehm, G., Berger, M., Elsholz, E., Hegele, S., Kintzel, F., Marheinecke, K., Klejch, O. (2020). European language grid: An overview. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3366–3380). European Language Resources Association (ELRA). https://aclanthology.org/2020.lrec-1.413.pdf

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, *35*(1), 31–65. https://doi.org/10.1017/S0272263112000678

Salisbury, K. (2005). *The edge of expertise? Towards an understanding of listening test item writing as professional practice* [Doctoral dissertation, King's College London]. Kings Research Portal. https://kclpure.kcl.ac.uk/portal/files/2936144/419477.pdf

Shin, D., & Lee, J. H. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology*, *27*(3), 27–40. https://hdl.handle.net/10125/73530

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Toyama, Y. (2021). What makes reading difficult? An investigation of the contributions of passage, task, and reader characteristics on comprehension performance. *Reading Research Quarterly*, *56*(4), 633–642. https://doi.org/10.1002/rrq.440

Vázquez-Ingelmo, A., García-Holgado, A., Therón, R., Shoeibi, N., & García-Peñalvo, F. J. (2023). Design and development of the LATILL platform for retrieving adequate texts to foster reading skills in German. In T. G. I. Saltiveri, M. S. Veloso, J. E. G. Navarro, R. G. González, M. T. Cairol, M. O. Solé, J. V. Gomà (Eds.), *Proceedings of the XXIII International Conference on Human-Computer Interaction* (pp. 1–9). ACM. https://dl.acm.org/doi/pdf/10.1145/3612783.3612796

Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., … Han, W. (2023). Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*. https://doi.org/10.48550/arXiv.2302.10205

Weiß, Z. L. (2024). *An integrative approach to linguistic complexity analysis for German* [Doctoral dissertation, Universität Tübingen]. https://bibliographie.uni-tuebingen.de/xmlui/bitstream/handle/10900/152467/Weiss-2024.pdf?sequence=1

Weiss, Z., & Meurers, D. (2018). Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. *Proceedings of the 27th International Conference on Computational Linguistics*, 303–317. Association for Computational Linguistics. https://aclanthology.org/C18-1026/

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do llamas work in English? On the latent language of multilingual transformers. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15366–15394). https://doi.org/10.18653/v1/2024.acl-long.820

Xi, X. (2023). Advancing language assessment with AI and ML – Leaning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, *20*(4–5), 357–376. https://doi.org/10.1080/15434303.2023.2291488

Xiao, C., Xu, X. S., Zhang, K., Wang, Y., & Xia, L. (2023). Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 610–625). https://doi.org/10.18653/v1/2023.bea-1.52

Young, J. C., & Shishido, M. (2023). Evaluation of the potential usage of ChatGPT for providing easier reading materials for ESL students. In T. Bastiaens (Ed.), *Proceedings of EdMedia + Innovate Learning* (pp. 155–162). Association for the Advancement of Computing in Education (AACE).

Zesch, T., Horbach, A., Weiss, Z., Aggarwal, P., Bewersdorff, J., Bexte, M., … Westphal, M. (2021). *Linguistic Features in Text (LiFT)*. GitHub. https://github.com/zesch/linguistic-features-in-text

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023). Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7915–7927). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.491

Appendix 1

Questionnaire for Evaluating Reading Texts for Authors

1. The text contains terms that require specialized knowledge.
2. The vocabulary used is broad.
3. The text includes complex linguistic structures (e.g., passive voice, subjunctive II, nominal style, subordinate clause structures, participial structures, infinitive constructions).
4. This text could appear in a popular science magazine.
5. The content can be understood based on general knowledge.
6. There are examples that illustrate the topic and the content.
7. The text is coherent.

Response options:

1 (strongly disagree) 2 3 4 (strongly agree)

Open-ended question: Are there any peculiarities in TEXT 1-20? Provide specific examples from the text.

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Appendix 2 Statistics for readability measures

Table 2.1. Descriptive statistics for readability measures for the three types of texts

Feature	Text Type	N	M	SD	Min.	Max.
ARI	benchmark	30	15.57	1.38	12.65	17.91
	ChatGPT3.5	30	20.05	1.88	16.99	23.41
	ChatGPT4	30	23.28	1.54	20.45	27.26
COLEMAN_LIAU	benchmark	30	19.73	1.20	16.71	21.65
	ChatGPT3.5	30	24.71	1.82	21.73	27.92
	ChatGPT4	30	25.98	1.36	22.60	28.52
KINCAID	benchmark	30	12.11	1.17	9.82	13.98
	ChatGPT3.5	30	15.59	1.56	12.59	17.98
	ChatGPT4	30	18.34	1.46	15.88	22.04
SMOG	benchmark	30	13.32	0.88	11.42	15.05
	ChatGPT3.5	30	15.90	1.00	13.85	17.49
	ChatGPT4	30	18.05	1.15	15.67	21.28
WSTF1	benchmark	30	8.72	0.91	7.07	10.12
	ChatGPT3.5	30	12.30	1.09	10.34	14.24
	ChatGPT4	30	13.82	1.00	10.83	15.56
Average Number of Syllables per Word	benchmark	30	1.77	0.07	1.61	1.92
	ChatGPT3.5	30	2.04	0.10	1.88	2.22
	ChatGPT4	30	2.12	0.09	1.93	2.30
Average Number of Characters per Token	benchmark	30	5.67	0.21	5.17	5.97
	ChatGPT3.5	30	6.53	0.31	6.05	7.07
	ChatGPT4	30	6.65	0.22	6.06	7.02
Average Number of Tokens per Sentence	benchmark	30	19.87	2.38	16.18	25.12
	ChatGPT3.5	30	20.53	2.24	16.00	25.00
	ChatGPT4	30	25.58	2.75	22.10	35.36

Table 2.2. Kruskal–Wallis test of statistical significance for readability measures ($df = 2$)

	ARI	COLEMANN LIAU	KINCAID	SMOG	WSTF1	Average Number of Syllables per Word	Average Number of Characters per Token	Average Number of Tokens per Sentence
H	70.893	62.541	69.243	71.305	69.281	62.461	60.724	49.045
p	.000	.000	.000	.000	.000	.000	.000	.000

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 2.3. Mann–Whitney *U* Tests of the significant results for readability measures

	ARI	COLEMANN LIAO	KINCAID	SMOG	WSTF1	Average Number of Syllables per Word	Average Number of Characters per Token	Average Number of Tokens per Sentence
ChatGPT4-ChatGPT3.5	0.000	0.074	0.000	0.000	0.002	0.048	0.240	0.000
ChatGPT4-benchmark	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ChatGPT3.5-benchmark	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.390

Appendix 3 Statistics for lexical measures

Table 3.1. Descriptive statistics for lexical measures for the three types of texts

Feature	Text Type	N	M	SD	Min.	Max.
Lexical Variation	benchmark	30	0.93	0.02	0.88	0.96
	ChatGPT3.5	30	0.91	0.03	0.83	0.95
	ChatGPT4	30	0.94	0.02	0.88	0.97
Lexical Density	benchmark	30	0.45	0.02	0.40	0.51
	ChatGPT3.5	30	0.51	0.02	0.47	0.55
	ChatGPT4	30	0.50	0.02	0.47	0.55
Type-Token Ratio	benchmark	30	0.76	0.02	0.71	0.80
	ChatGPT3.5	30	0.70	0.02	0.65	0.76
	ChatGPT4	30	0.72	0.02	0.68	0.75

Table 3.2. Kruskal–Wallis test of statistical significance for lexical measures ($df = 2$)

	MA Lexical Variation	Lexical Density	MA Type-Token Ratio
H	27.62	52.264	53.546
p	.000	.000	.000

Table 3.3. Mann–Whitney *U* tests of the significant results for lexical measures

	MA Lexical Variation	Lexical Density	MA Type-Token Ratio
ChatGPT4-ChatGPT3.5	0.000	0.213	0.041
ChatGPT4-benchmark	0.013	0.000	0.000
ChatGPT3.5-benchmark	0.006	0.000	0.000

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 3.4. Descriptive statistics for POS TTR for the three types of texts

Feature	Text Type	N	M	SD	Min.	Max.
TTR_ADJ	benchmark	30	0.92	0.03	0.85	0.98
	ChatGPT3.5	30	0.83	0.07	0.67	0.93
	ChatGPT4	30	0.88	0.06	0.73	0.98
TTR_ADP	benchmark	30	0.42	0.05	0.33	0.51
	ChatGPT3.5	30	0.33	0.06	0.24	0.47
	ChatGPT4	30	0.37	0.05	0.29	0.50
TTR_ADV	benchmark	30	0.75	0.07	0.63	0.89
	ChatGPT3.5	30	0.71	0.14	0.47	1.00
	ChatGPT4	30	0.78	0.11	0.55	1.00
TTR_CONJ	benchmark	30	0.38	0.09	0.21	0.57
	ChatGPT3.5	30	0.21	0.06	0.12	0.32
	ChatGPT4	30	0.23	0.05	0.11	0.32
TTR_DET	benchmark	30	0.21	0.03	0.13	0.29
	ChatGPT3.5	30	0.22	0.05	0.14	0.40
	ChatGPT4	30	0.20	0.03	0.15	0.28
TTR_NOUN	benchmark	30	0.75	0.04	0.63	0.84
	ChatGPT3.5	30	0.69	0.05	0.55	0.79
	ChatGPT4	30	0.76	0.05	0.67	0.84
TTR_PRON	benchmark	30	0.59	0.08	0.45	0.75
	ChatGPT3.5	30	0.62	0.10	0.48	0.92
	ChatGPT4	30	0.60	0.10	0.43	0.85
TTR_VERB	benchmark	30	0.78	0.05	0.66	0.86
	ChatGPT3.5	30	0.72	0.07	0.55	0.87
	ChatGPT4	30	0.80	0.08	0.64	0.96

Table 3.5. Kruskal–Wallis test of POS TTR for lexical measures ($df = 2$)

	TTR_ADJ	TTR_ADP	TTR_ADV	TTR_CONJ	TTR_DET	TTR_NOUN	TTR_PRON	TTR_VERB
H	27.664	26.674	5.799	45.913	5.186	23.937	1.312	16.488
p	.000	.000	.055	.000	.075	.000	.519	.000

Table 3.6. Mann–Whitney *U* tests of the significant results for POS TTR

	TTR_ADJ	TTR_ADP	TTR_CONJ	TTR_NOUN	TTR_VERB
ChatGPT4-ChatGPT3.5	0.020	0.031	0.448	0.000	0.000
ChatGPT4-benchmark	0.003	0.003	0.000	0.927	0.655
ChatGPT3.5-benchmark	0.000	0.000	0.000	0.000	0.001

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Appendix 4 Statistics for POS measures

Table 4.1. Descriptive statistics for POS measures for the three types of texts

Feature	Text Type	N	M	SD	Min.	Max.
ADJ	benchmark	30	0.08	0.01	0.05	0.11
	ChatGPT3.5	30	0.11	0.02	0.07	0.16
	ChatGPT4	30	0.12	0.01	0.09	0.15
PREP	benchmark	30	0.09	0.01	0.06	0.11
	ChatGPT3.5	30	0.10	0.02	0.07	0.12
	ChatGPT4	30	0.10	0.01	0.07	0.12
ADV	benchmark	30	0.07	0.02	0.04	0.10
	ChatGPT3.5	30	0.03	0.01	0.01	0.05
	ChatGPT4	30	0.03	0.01	0.00	0.04
CONJ	benchmark	30	0.05	0.01	0.04	0.07
	ChatGPT3.5	30	0.07	0.01	0.05	0.08
	ChatGPT4	30	0.07	0.01	0.05	0.09
DET	benchmark	30	0.10	0.01	0.07	0.13
	ChatGPT3.5	30	0.12	0.02	0.08	0.15
	ChatGPT4	30	0.13	0.02	0.10	0.17
NOUN	benchmark	30	0.21	0.02	0.17	0.24
	ChatGPT3.5	30	0.26	0.02	0.22	0.29
	ChatGPT4	30	0.27	0.01	0.24	0.29
NUM	benchmark	30	0.01	0.01	0.00	0.04
	ChatGPT3.5	30	0.00	0.00	0.00	0.01
	ChatGPT4	30	0.00	0.00	0.00	0.01
PRON	benchmark	30	0.08	0.02	0.04	0.10
	ChatGPT3.5	30	0.05	0.01	0.02	0.08
	ChatGPT4	30	0.05	0.01	0.02	0.07
PROPN	benchmark	30	0.02	0.01	0.00	0.04
	ChatGPT3.5	30	0.01	0.01	0.00	0.02
	ChatGPT4	30	0.01	0.01	0.00	0.02
VERB	benchmark	30	0.14	0.02	0.12	0.18
	ChatGPT3.5	30	0.13	0.01	0.11	0.16
	ChatGPT4	30	0.12	0.01	0.09	0.16

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 4.2. Kruskal-wallis test of statistical significance for POS measures ($df = 2$)

	ADJ	PREP	ADV	CONJ	DET	NOUN	NUM	PRON	PROPN	VERB
H	42.809	5.910	56.662	26.079	33.367	56.821	40.310	42.101	39.636	38.523
p	.000	.086	.000	.000	.000	.000	.000	.000	.000	.000

Table 4.3. Mann–Whitney *U* tests of the significant results for POS measures

	ADJ	ADV	CONJ	DET	NOUN	NUM	PRON	PROPN	VERB
ChatGPT4-ChatGPT3.5	0.333	0.502	0.754	0.079	0.556	0.982	0.632	0.732	0.000
ChatGPT4-benchmark	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ChatGPT3.5-benchmark	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017

Appendix 5 Statistics for morphological measures

Table 5.1. Descriptive statistics for morphological measures for the three types of texts

Feature	Text type	N	M	SD	Min.	Max.
Finite Verb Ratio	benchmark	30	0.61	0.04	0.52	0.70
	ChatGPT3.5	30	0.56	0.05	0.47	0.69
	ChatGPT4	30	0.54	0.05	0.44	0.65
Frequency of Passive Sentences	benchmark	30	0.36	0.12	0.14	0.55
	ChatGPT3.5	30	0.24	0.12	0.04	0.48
	ChatGPT4	30	0.26	0.10	0.13	0.53
Frequency of Passive Sentence with *Bekommen*	benchmark	30	0.00	0.01	0.00	0.06
	ChatGPT3.5	30	0.00	0.00	0.00	0.00
	ChatGPT4	30	0.00	0.00	0.00	0.00
Frequency of Passive Sentence with Impersonal Pronoun	benchmark	30	0.06	0.06	0.00	0.18
	ChatGPT3.5	30	0.00	0.01	0.00	0.05
	ChatGPT4	30	0.00	0.00	0.00	0.00
Frequency of Passive Sentence with *Sich Lassen*	benchmark	30	0.01	0.02	0.00	0.09
	ChatGPT3.5	30	0.01	0.02	0.00	0.06
	ChatGPT4	30	0.01	0.02	0.00	0.06
Frequency of Passive Sentence with *Zu*	benchmark	30	0.02	0.02	0.00	0.08
	ChatGPT3.5	30	0.01	0.03	0.00	0.11
	ChatGPT4	30	0.01	0.03	0.00	0.12

(*Continued*)

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 5.1. (*Continued.*)

Feature	Text type	N	M	SD	Min.	Max.
Frequency of Passive Sentence with Adjective	benchmark	30	0.03	0.03	0.00	0.11
	ChatGPT3.5	30	0.01	0.02	0.00	0.08
	ChatGPT4	30	0.02	0.03	0.00	0.10
Frequency Of Typical Passive Sentence	benchmark	30	0.24	0.09	0.03	0.42
	ChatGPT3.5	30	0.20	0.10	0.04	0.41
	ChatGPT4	30	0.22	0.09	0.12	0.43
Frequency of All Suffixes	benchmark	30	0.17	0.05	0.08	0.28
	ChatGPT3.5	30	0.39	0.07	0.25	0.55
	ChatGPT4	30	0.41	0.07	0.27	0.52
Nominalization per Infinite Verb	benchmark	30	0.41	0.14	0.11	0.69
	ChatGPT3.5	30	1.41	0.39	0.71	2.38
	ChatGPT4	30	1.81	0.49	1.05	2.71
Percentage of Nouns in Accusative	benchmark	30	0.26	0.05	0.16	0.35
	ChatGPT3.5	30	0.30	0.04	0.22	0.39
	ChatGPT4	30	0.30	0.05	0.21	0.44
Percentage of Nouns in Dative	benchmark	30	0.29	0.04	0.19	0.39
	ChatGPT3.5	30	0.30	0.04	0.19	0.36
	ChatGPT4	30	0.29	0.05	0.20	0.38
Percentage of Nouns in Genitive	benchmark	30	0.06	0.03	0.01	0.14
	ChatGPT3.5	30	0.07	0.03	0.00	0.14
	ChatGPT4	30	0.10	0.03	0.04	0.16
Percentage of Nouns in Nominative	benchmark	30	0.31	0.05	0.21	0.42
	ChatGPT3.5	30	0.29	0.05	0.21	0.41
	ChatGPT4	30	0.27	0.05	0.19	0.42
Frequency of *HEIT*	benchmark	30	0.01	0.01	0.00	0.04
	ChatGPT3.5	30	0.01	0.01	0.00	0.05
	ChatGPT4	30	0.02	0.01	0.00	0.06
Frequency of *IE*	benchmark	30	0.02	0.03	0.00	0.12
	ChatGPT3.5	30	0.05	0.04	0.00	0.14
	ChatGPT4	30	0.05	0.04	0.00	0.13
Frequency of *KEIT*	benchmark	30	0.01	0.01	0.00	0.04
	ChatGPT3.5	30	0.03	0.03	0.00	0.15
	ChatGPT4	30	0.03	0.02	0.01	0.08

(*Continued*)

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 5.1. (*Continued.*)

Feature	Text type	N	M	SD	Min.	Max.
Frequency of *MENT*	benchmark	30	0.00	0.01	0.00	0.03
	ChatGPT3.5	30	0.01	0.01	0.00	0.04
	ChatGPT4	30	0.01	0.01	0.00	0.06
Frequency of *MUS*	benchmark	30	0.00	0.01	0.00	0.03
	ChatGPT3.5	30	0.00	0.00	0.00	0.01
	ChatGPT4	30	0.00	0.01	0.00	0.03
Frequency of *NIS*	benchmark	30	0.01	0.01	0.00	0.02
	ChatGPT3.5	30	0.02	0.01	0.00	0.04
	ChatGPT4	30	0.02	0.01	0.00	0.05
Frequency of *SCHAFT*	benchmark	30	0.01	0.01	0.00	0.03
	ChatGPT3.5	30	0.02	0.02	0.00	0.10
	ChatGPT4	30	0.02	0.02	0.00	0.08
Frequency of *TION*	benchmark	30	0.02	0.02	0.00	0.06
	ChatGPT3.5	30	0.04	0.03	0.00	0.19
	ChatGPT4	30	0.04	0.02	0.01	0.11
Frequency of *TUM*	benchmark	30	0.00	0.01	0.00	0.03
	ChatGPT3.5	30	0.00	0.01	0.00	0.04
	ChatGPT4	30	0.00	0.00	0.00	0.02
Frequency of *TÄNT*	benchmark	30	0.01	0.01	0.00	0.02
	ChatGPT3.5	30	0.02	0.02	0.00	0.05
	ChatGPT4	30	0.02	0.02	0.00	0.06
Frequency of *UNG*	benchmark	30	0.08	0.03	0.00	0.15
	ChatGPT3.5	30	0.20	0.05	0.10	0.31
	ChatGPT4	30	0.21	0.05	0.11	0.31

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 5.2. Kruskal–Wallis test of statistical significance for morphological measures ($df = 2$)

	Finite Verb Ratio	Frequency of Passive Sentences	Frequency of Passive Sentence with *Bekommen*	Frequency of Passive Sentence with Impersonal Pronoun	Frequency of Passive Sentence with *Sich Lassen*	Frequency of Passive Sentence with *Zu*	Frequency of Passive Setence with Adjective	Frequency of Typical Passive Sentence
H	22.331	16.656	4.045	52.755	.933	3.575	8.185	2.944
p	.000	.000	.132	.000	.627	.167	.017	.229
	Frequency of All Suffixes	**Nominalization per Finite Verb**	**Percentage of Nouns in Accusative**	**Percentage of Nouns in Dative**	**Percentage of Nouns in Genitive**	**Percentage of Nouns in Nominative**		
H	59.4	63.249	8.929	1.192	20.067	10.691		
p	.000	.000	.012	.551	.000	.005		
	Frequency of *HEIT*	**Frequency of *IE***	**Frequency of *KEIT***	**Frequency of *MENT***	**Frequency of *MUS***	**Frequency of *NIS***		
H	9.49	13.887	18.827	2.964	1.03	22.116		
p	.009	.001	.000	.227	.597	.000		
	Frequency of *SCHAFT*	**Frequency of *TION***	**Frequency of *TUM***	**Frequency of *TÄNT***	**Frequency of *UNG***			
H	12.773	18.628	.889	11.559	55.712			
p	.002	.000	.641	.003	.000			

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 5.3. Mann–Whitney *U* tests of the significant results for morphological measures

	Finite verbs	Frequency of Passive Sentence with Impersonal Pronoun	Frequency of Passive Sentences	Frequency of Passive Sentence with Adjective	Frequency of All Suffixes	Nominal-ization per Finite Verb	Percentage of Nouns in Accusative	Percentage of Nouns in Genitive	Percentage of Nouns in Nominative
ChatGPT4-ChatGPT3.5	0.243	0.796	0.495	0.725	0.398	0.048	0.812	0.007	0.048
ChatGPT4-benchmark	0.000	0.000	0.002	0.022	0.000	0.000	0.007	0.000	0.001
ChatGPT3.5-benchmark	0.001	0.000	0.000	0.008	0.000	0.000	0.014	0.085	0.204
	Frequency of *HEIT*	**Frequency of *IE***	**Frequency of *KEIT***	**Frequency of *NIS***	**Frequency of *SCHAFT***	**Frequency of *TION***	**Frequency of *TÄNT***	**Frequency of *UNG***	
ChatGPT4-ChatGPT3.5	0.057	0.837	0.214	0.263	0.116	0.345	0.772	0.658	
ChatGPT4-benchmark	0.002	0.001	0.000	0.001	0.000	0.000	0.002	0.000	
ChatGPT3.5-benchmark	0.0252	0.002	0.003	0.000	0.046	0.001	0.005	0.000	

Appendix 6 Statistics for syntactic measures

Table 6.1. Descriptive statistics for syntactic measures for the three types of texts

Feature	Text Type	N	M	SD	Min.	Max.
Average Syntax Tree Depth	benchmark	30	5.08	0.34	4.55	5.77
	ChatGPT3.5	30	5.17	0.44	4.23	6.20
	ChatGPT4	30	5.89	0.45	5.30	7.14
Maximum Syntax Tree Depth	benchmark	30	11.33	2.80	8.00	20.00
	ChatGPT3.5	30	8.07	1.26	6.00	10.00
	ChatGPT4	30	9.60	1.65	7.00	15.00
Average Number of Subordinate Clauses per Sentence	benchmark	30	0.48	0.14	0.16	0.80
	ChatGPT3.5	30	0.49	0.17	0.14	0.82
	ChatGPT4	30	0.62	0.17	0.33	0.94
Average Number of Connectives per Sentence	benchmark	30	1.24	0.32	0.78	2.12
	ChatGPT3.5	30	1.55	0.34	1.00	2.20
	ChatGPT4	30	1.92	0.38	1.18	2.93
Average Number of Subordinate Clauses with Conjunction per Sentence	benchmark	30	0.28	0.12	0.06	0.56
	ChatGPT3.5	30	0.25	0.13	0.00	0.50
	ChatGPT4	30	0.21	0.13	0.00	0.50
Average Number of Infinitve Clauses per Sentence	benchmark	30	0.04	0.04	0.00	0.13
	ChatGPT3.5	30	0.09	0.07	0.00	0.23
	ChatGPT4	30	0.12	0.08	0.00	0.29
Average Number of Relative Clauses per Sentence	benchmark	30	0.15	0.08	0.03	0.31
	ChatGPT3.5	30	0.15	0.09	0.00	0.32
	ChatGPT4	30	0.29	0.13	0.06	0.59

Table 6.2. Kruskal–Wallis test of statistical significance for syntactic measures ($df = 2$)

	Average Syntax Tree Depth	Maximum Syntax Tree Depth	Average Number of Subordniate Clauses per Sentence	Average Number of Connectives per Sentence	Average Number of Subordinate Clauses with Conjunction per Sentence	Average Number of Infinitive Clauses per Sentence	Average Number of Relative Clauses per Sentence
H	41.101	31.615	9.88	35.558	3.225	17.644	21.308
p	.000	.000	.007	.000	.196	.000	.000

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press

Table 6.3. Mann–Whitney *U* tests of the significant results for syntactic measures

	Average Syntax Tree Depth	Maximum Syntax Tree Depth	Average Number of Subordinate Clauses per Sentence	Average Number of Connectives per Sentence	Average Number of Infinitive Clauses per Sentence	Average Number of Relative Clauses per Sentence
ChatGPT4-ChatGPT3.5	0.000	0.001	0.029	0.003	0.185	.818
ChatGPT4-benchmark	0.000	0.016	0.002	0.000	0.000	.000
ChatGPT3.5-benchmark	0.538	0.000	0.387	0.002	0.005	.000

Cite this article: Drackert, A., Horbach, A., & Peters, A. (2025). How good are LLMs in generating input texts for reading tasks in German as a foreign language?. *Annual Review of Applied Linguistics*, *45*, 222–252. https://doi.org/10.1017/S0267190525000066

https://doi.org/10.1017/S0267190525000066 Published online by Cambridge University Press