Using k-Shot Prompting with Large k for the Automated Scoring of a German Written Elicited Imitation Test

Malte Sternik[1], Ronja Laarmann-Quante[1] and Anastasia Drackert[2,3]
Ruhr University Bochum, Germany
firstname.lastname@rub.de
[1]Faculty of Philology, Department of Linguistics
[2]Faculty of Philology, Institute for German Language and Literature
[3]g.a.s.t. (Society for Academic Study Preparation and Test Development)

Abstract

This paper explores the application of a Large Language Model (LLM) using k-shot prompting with large k for automatically scoring a German Written Elicited Imitation Test (WEIT), a test for assessing literacy-dependent procedural knowledge in German as a foreign language. In this test, test-takers are briefly presented with written sentences which they then have to reproduce in writing as accurately as possible. The responses are scored on an ordinal scale which differentiates between different types of errors (e.g. lexical vs. grammatical). We find that with increasing k (in a range from 1 to 700) accuracy increases significantly but it also depends on the drawn sample and varies across different runs of the same prompt. Overall, the k-shot setting which relies on in-context learning without being provided with the scoring rubric outperforms a baseline where only the scoring rubric is provided to the model. However, the LLM does not outperform previous results based on rule-based or BERT-based models.

1 Introduction

The Written Elicited Imitation Test (WEIT) is a test used for measuring procedural linguistic knowledge in writing (Timukova et al., 2026). In this computer-based test, participants view sentence after sentence for a short time duration and after a brief pause, they are asked to reproduce the sentence by typing it in a text field. The responses are rated on an ordinal scale ranging from 0 to 4 based on a scoring rubric. Responses receive higher scores when they are closer to the original while lower scores are given to grammatically, syntactically or semantically altered sentences (see Section 3 for a more detailed description of the scoring rubric).

Chifligarov et al. (2025) were the first to explore the automatic scoring of a German WEIT. They compared two approaches: (i) a rule-based scoring model which directly implements categories from the scoring rubric and (ii) a deep-learning-based approach where a pre-trained transformer model from the BERT family (DistilBERT; Sanh et al., 2019), henceforth referred to as BERT-based model, was finetuned on pairs of annotated stimulus and response sentences in order to learn (implicitly) which scores to apply. The results showed that both approaches had different strengths and weaknesses. While the rule-based model performed better on unseen stimulus sentences and scores at the edges of the rating scale, the BERT-based model was better at mid-range scores.

Overall, the results appear not yet robust enough to be applied in practice on the level of the individual response. At the same time, language learners and instructors would greatly benefit from an automatic scoring of WEITs, which would reduce the manual evaluation effort and thus make this instrument more widely usable. Developing a more robust scoring model is thus still a desideratum.

The present paper is a follow-up to the study of Chifligarov et al. (2025), exploring the use of a Large Language Model (LLM) for automatically scoring a German WEIT. It focuses on one specific prompting strategy and systematically investigates different sources of variance connected to this strategy. The strategy we use is k-shot prompting, an in-context learning approach where the LLM is presented with k scored samples and in our case no explicit scoring rubric. The idea is that based on the provided examples (shots), the LLM derives the scoring principles on its own. Besides the choice of k, we study the impact of different samples drawn. Another source of variance applying to all LLM-based studies is the non-determinism of an LLM, i.e. different runs of the same prompt can produce different results even if the temperature parameter is set to 0 (Atıl et al., 2025). This source of variance has rarely been studied systematically (Song et al., 2025, but see e.g. Ouyang et al., 2025).

The **aim of the paper** is twofold: (1) compar-

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 272–280

ing the performance of an LLM-based model with the rule-based and BERT-based models used by Chifligarov et al. (2025), thereby contributing to the development of automatic scoring models for WEITs and (2) an in-depth investigation of the k-shot prompting strategy used for the LLM, contributing to the understanding of variability in LLM performance for this particular task. We pursue the following **research questions** related to (2):

RQ1: What is the optimal number of examples k for scoring a WEIT?

RQ2: What variance in performance arises through the use of different samples for the in-context learning examples?

RQ3: Given the non-determinism of LLMs, what variance is to be expected only by running the same prompt multiple times?

The LLM used in this paper is GPT-4o by OpenAI, which was chosen due to its popularity. Section 2 discusses related work and Section 3 presents the data and scoring rubric. In Experiment 1 (Sec. 4), we explore a large range of k ($10 \leq k \leq 700$) while in Experiment 2 (Sec. 5), we focus on small $k \leq 25$. In Section 6 we compare the results to those obtained by Chifligarov et al. (2025). The code and data from this study are available on `https://gitlab.ruhr-uni-bochum.de/vamos-cl/german-weit-k-shot-prompting`.

2 Related Work

Chifligarov et al. (2025) were the first to automatically score a written EIT. Previously, only the automated scoring of oral EITs (OEIT) had been explored. However, the OEITs were typically rated on a binary scale (Millard, 2011) or interval scale (Graham et al., 2008; Lonsdale and Christensen, 2011), not an ordinal scale with a rubric that distinguishes different error categories (see Sec. 3).

In other areas of the educational domain involving automated scoring, LLMs have already become popular. Different studies investigated the use of k-shot prompting, especially few-shot prompting, coming to mixed results. In the domain of automatic essay scoring, for example, Mansour et al. (2024) found that adding one example (1–shot prompting) to the scoring rubric improves the results but not across all models and all writing tasks. Yoshida (2024) investigated 1–3 shots, finding that the choice and order of examples impacts the results. Huang and Wilson (2025) found that few-shot prompting improves the results. They added five examples to a prompt providing context and chain-of-thought instructions. In the area of short-answer scoring, Chamieh et al. (2024) explored k-shot prompting wih k varying between 1 and 10. They found that increasing the number of shots was not beneficial for all models and datasets.

For other domains, there has already been evidence that in-context learning with a large k can be powerful. Bertsch et al. (2025) used long context models on different classification tasks, e.g. question classification or intent classification. They found accuracy gains of up to 50.8 points when scaling from 10 to 1,000 examples. Furthermore, with increasing k, it became less important which examples were chosen. Note that their prompts *only* contained the shots. In our experiments, we follow this approach of using large values of k and, while we provide a little bit of context in the prompt, we do not include the full scoring rubric either, which means that the model predominantly relies on the provided examples. In this respect, our approach differs from the k-shot prompting studies in the areas of short-answer scoring or essay scoring discussed above, where a rather small number of shots (≤ 10) was used to support rather than replace the scoring guidelines in the prompt. Furthermore, in the context of a WEIT, a single shot contains only one sentence pair, whereas for essay scoring, a single shot is equivalent to a whole text.

3 Data

We use the WEIT dataset published by Chifligarov et al. (2025). Its collection, the creation of a scoring rubric and the manual scoring of the data was part of a larger research project (Timukova et al., 2026). The dataset consists of 3,900 stimulus-response pairs (195 participants responded to 20 different stimulus sentences each). Each response was scored based on an ordinal scoring rubric. In the following, we quote a scored example (Table 1) and the brief summary of the scoring rubric from Chifligarov et al. (2025). The full rubric can be found in their supplementary material.

Score 4 The response matches the stimulus sentence exactly or 1–2 typos are present.[1]

Score 3 Changes in grammar or lexical changes that preserve the original structure and result

[1]Typos include: transposed letters (all present), one letter replaced by a QWERTZ-adjacent key, one letter added/omitted next to an adjacent key, or a missing space between words.

Score	Example	Explanation
0	Bein praktikum	less than half of the words repeated correctly
1	Bei einem Praktikum * * *	half of the words repeated correctly but most of the meaning lost
2	Bei ein Praktikum lernt man viel.	case wrongly marked, ungrammatical sentence
3	Beim Praktikum lernt man viel.	change in grammar (contraction of preposition + article) but still grammatical
4	Bei einem Praktikum lenrt man viel.	one typo

Table 1: Example of the scoring rubric for the stimulus sentence *Bei einem Praktikum lernt man viel.* 'At an internship, one learns a lot'. Table taken from Chifligarov et al. (2025).

in grammatically correct and meaningful sentences, e.g. confusing definite and indefinite articles (where interchangeable), or (near-)synonymic substitutions of words.

Score 2 Changes in grammar that result in ungrammatical sentences or grammatical sentences which are not meaningful, e.g. violated agreement between subject and verb, structural omissions or wrong plural formation.

Score 1 More than half of the words are repeated but a considerable part of the original meaning or structure is lost or changed.

Score 0 Less than half of the words are repeated.

We use the same split into training, development and test data as used by Chifligarov et al. (2025).

4 Experiment 1: Large Samples

In the following experiments, the LLM GPT-4o (henceforth referred to as GPT model) is prompted to score the data described in Section 3 only based on examples from the training data, without explicitly providing the scoring rubric. In Experiment 1, three sources of variance are explored: (1) the number of examples k, (2) different samples drawn from the training set (consisting of 3,260 stimulus-response pairs in total) (3) different runs of the same prompt using the same examples.

4.1 Method

We vary the number of examples k between 10 and 700. Between $k = 10$ and $k = 50$, k is increased in steps of 10, above $k = 50$ in steps of 50, yielding 18 variations of k in total. For each k, two different samples (based on a fixed random seed) are drawn from the training set. For each k and each sample, four runs of the same prompt are carried out, yielding a total of 144 runs (18 different $k \times 2$ different samples $\times$ 4 runs). To keep the variance as low as possible, we set the temperature parameter to 0.

```
Identity:
You score answers from a German Elicited Imitation test. You score between 0 and 4. The test is in german language.
Instructions:
- The training data contains answers and the original items, that are already scored between 0 and 4
- The training data has an ID at the beginning, the original second, the answer as third and the score as fourth
- Try to understand why and how the scores of the training data relate to the answer original pairs
- The training data is here: EXAMPLES
- You will get the same answer original pairs but with no scores, score each pair based on the training data.
- Output only the answer original pairs with the respective scores and ID of the answer original pair
- Format the output in a way that the ID, the answer, the original and the score are divided by '\t'
```

Figure 1: Prompt given to GPT-4o in the developer/system role.

Prompt The model is given a set of stimulus-response pairs to score from the user role. A simple instruction prompt from the developer/system role is added, asking the model to analyze the given examples (termed training data), see Figure 1. The examples are provided as a simple string with a new line for each stimulus-response pair and divided into four columns named *id*, *original* (=stimulus), *answer* (=response) and *score* divided by a tab. The string is included within the prompt (marked by **EXAMPLES** in Figure 1). A little bit of manual adjusting of the prompt was done in order to get the model to just output the scored items and not add an unwanted information (reasons, questions, comments etc.). We are aware that subtle prompt formatting, e.g. regarding casing or spacing can impact the result (Sclar et al., 2024). Investigating this systematically was beyond the scope of this study though.

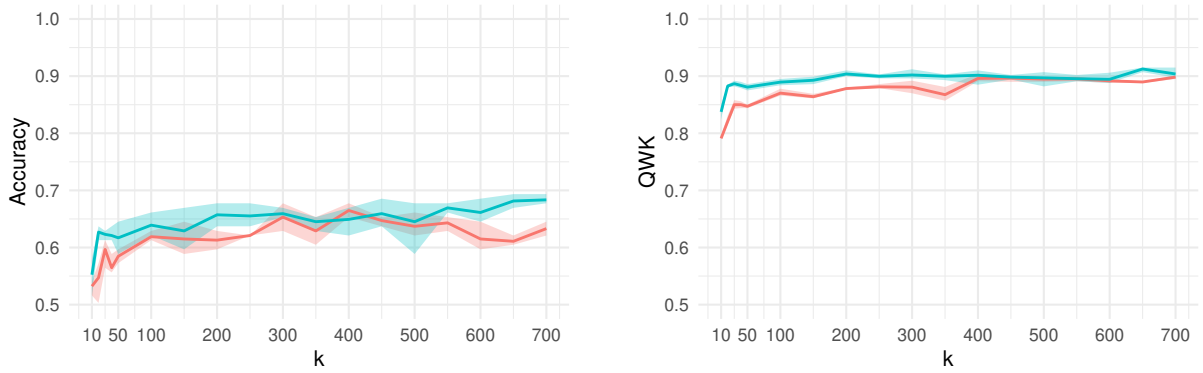

Figure 2: Results for **Experiment 1**: Accuracy (left) and QWK (right) for each number of shots k. The red and blue line represent the two different samples. The solid line represents the mean of the four runs, the shaded area indicates the minimum and maximum value per run.

Evaluation The results are evaluated on the development set consisting of 125 stimulus-response pairs that are balanced across scores. We use accuracy and quadratically weighted kappa (QWK) as evaluation metrics. While accuracy is a strict metric, only counting exact score matches, QWK quantifies the magnitude of deviations, punishing greater deviations from the gold score more severely than smaller deviations.

4.2 Result

Figure 2 illustrates how accuracy and QWK vary with the number of shots k, the chosen sample and the different runs. Overall, QWK is rather stable across k, especially with $k > 50$, and there is almost no influence of the different runs. When $k < 400$, there is a notable difference between the two samples. Accuracy shows much more variability involving all factors. There is a clear drop in performance with very low values of k, which will be investigated more closely in Experiment 2 in Section 5.

In order to assess the influence of the different factors, we conducted a binomial logistic regression analysis with accuracy as the dependent variable and k (numeric), *sample* (categorical) and *run* (categorical) and the interaction of k and *sample* as predictors.[2] Rather than modeling *sample* and *run* as random effects in a generalized linear mixed-effects model, which would conceptually correspond better with their randomness, we decided to include them as fixed effects since we wanted to study whether they have a systematic influence on scoring accuracy. Furthermore, their small number of factor levels (two for *sample* and four for *runs*) does generally not make them suitable random effects (Gomes, 2022). In our analysis, the variable *run* was effects-coded so that the coefficients reflect deviations from the overall mean of the dependent variable (in contrast to the default treatment coding, where the comparison is made against one reference level of the predictor). We found two significant effects, namely the main effects of k ($\beta < .01, SE < .01, z = 4.46, p < .001$), and *sample* ($\beta < .13, SE < .05, z = 2.54, p = .01$). This means that a higher k yields a higher accuracy and also one of the samples yields a significantly higher accuracy than the other.

One aspect that certainly influences the performance of the models is whether they have seen all 18 stimulus sentences in the examples (see also Section 6 for the performance on known vs. unknown stimulus sentences). In fact, only with $k \geq 100$, all stimulus sentences are present in the examples. To investigate this further, we conducted two separate regression analyses on the data with $k < 100$ and $k \geq 100$. In these separate regressions, none of the factors are significant anymore (all $\beta < .18, SE < .14, -.64 < z < 1.8, p > .08$). This indicates that the significant effects of k and *sample* arise only when considering the whole range of k between 10 and 700.

5 Experiment 2: Small Samples

In Experiment 1, we found that overall, a higher number of shots k improves the accuracy and QWK and that there is a notable drop in performance when only few examples are given. To investigate this further, Experiment 2 uses even smaller values of k.

[2]The regression was carried out using the `glm` function of R (version 4.5.3)

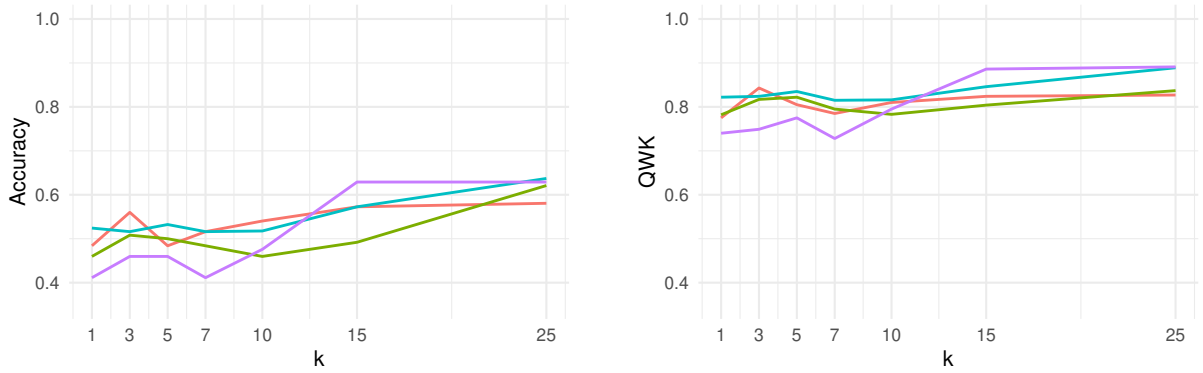

Figure 3: Results for **Experiment 2**: Accuracy (left) and QWK (right) for each number of examples k. Each line represents a different sample.

5.1 Method

In this experiment, $k \in \{1, 3, 5, 7, 10, 15, 25\}$ is used. We assume that the smaller the number of examples, the more it matters which ones are chosen. We therefore compare 4 different samples drawn from the training data with a fixed random seed per value of k. Since the different runs were not a significant factor in Experiment 1, we run each of the 28 configurations in this exeriment (7 variations of $k \times 4$ different samples) only once. The prompt and the evaluation setting are the same as in Experiment 1.

5.2 Result

Figure 3 illustrates the accuracy and QWK for the different values of k and the four samples. As expected, for small k values, the chosen sample can make a big difference of over ten percentage points in accuracy. It is notable, though, that even with very few examples, the GPT model performs way above chance level (which would be 20% percent in the balanced development set on which the evaluation is carried out).

We again conducted a binomial logistic regression analysis with accuracy as the dependent variable and k (numeric) and *sample* (categorical) and their interaction as predictors. The variable *sample* was effects-coded. Only k turns out as a significant predictor ($\beta = .02, SE = .004, z = 5.53, p < .001$), with higher k yielding higher accuracy. Higher values of k are strongly correlated with the number of stimulus sentences seen (Pearson's $r = 0.95$), which certainly influences model performance.

Model	Known It.		Unknown It.		Combined	
	acc	qwk	acc	qwk	acc	qwk
Rule-based	.71	.89	**.60**	.77	**.64**	.82
BERT	**.77**	**.93**	.58	**.81**	**.64**	**.86**
GPT-rubric	.51	.72	.46	.74	.47	.73
GPT-kshot	.66	.89	.55	.80	.59	.84

Table 2: Accuracy and QWK for the rule-based and BERT-based models from Chifligarov et al. (2025) compared to the GPT-rubric and GPT-k-shot model on the test sets with known items and unknown items, respectively, and the combined test set. Numbers in bold indicate the best model per set and metric.

6 Comparison with Rule-Based and BERT-Based Models

To assess the overall performance of the GPT k-shot model, we compare it to the rule-based and BERT-based models from Chifligarov et al. (2025).[3] In this evaluation, we use the k-shot model that performed best on the development set in Experiment 1 ($k = 700$). Furthermore, we add another GPT-based baseline, where the prompt for GPT-4o contains the scoring rubric used by the human raters (henceforth referred to as GPT-rubric). In this setting, no scored examples (shots) are given but the rubric contains some concrete examples illustrating each score. The exact prompt with the rubric can be found in the online repository. Systematic tuning of this rubric-based prompt was beyond the scope and goal of this paper, so this setting is only meant as a basic point of reference for comparing the k-shot setting to.

For the evaluation, we use the two test sets from Chifligarov et al. (2025). The first test set contains a

[3] In the following, the BERT-based model refers to the weighted model used by Chifligarov et al. (2025).

Score			Precision				Recall				F1-Score			
	# answ.	%	Rule	BERT	GPT-r	GPT-k	Rule	BERT	GPT-r	GPT-k	Rule	BERT	GPT-r	GPT-k
0	92	24%	.85	.84	.63	**.86**	**.91**	.82	.29	.55	**.88**	.83	.40	.68
1	99	26%	.51	**.66**	.41	.54	**.86**	.61	.44	.64	**.64**	.63	.43	.59
2	77	20%	.29	**.47**	.28	.40	.17	**.74**	.26	.30	.21	**.57**	.27	.34
3	79	21%	**.80**	.58	.50	.57	.35	.42	**.71**	.68	.49	.49	.58	**.62**
4	34	9%	**.94**	.82	.72	.62	.94	.53	**.97**	**.97**	**.94**	.64	.82	.76
macro avg			**.68**	.67	.51	.60	**.65**	.62	.54	.63	**.63**	**.63**	.50	.60
micro avg			.64	**.66**	.48	.60	**.64**	**.64**	.47	.59	.61	**.64**	.46	.58

Table 3: Distribution of (gold) scores in the combined test set and precision, recall and F1-score per (gold) score as well as macro average and micro (=weighted) average for each model (GPT-r = GPT-rubric, GPT-k = GPT-k-shot). The numbers in bold indicate the highest value in precision, recall, and F1-score, respectively, per score.

random held-out sample of answers comprising all 18 different stimulus sentences that are also part of the training and development set (henceforth *known test set*). This test set reflects the performance on new answers to known stimulus sentences. It is balanced across all scores 0-4. The second test set (*unknown test set*) contains all answers to two completely new stimulus sentences, reflecting the performance of the models when a new test with new stimuli is used.

In contrast to Chifligarov et al. (2025), we remove duplicates from the test set. Duplicate answers mainly concern answers that are practically identical to the stimulus sentence, therefore mostly answers with the highest score 4 are removed. Furthermore, due to a token limit for the GPT model, not all answers from the unknown test set could be processed but only a random sample. Consequently, our test set comprises the intersection of answers processed by all models and excluding the duplicates. This yields 116 answers (out of 125) in the known test set and 265 (out of 390) in the unknown test set.

Table 2 shows accuracy and QWK for each model for the known test set, the unknown test set and the combination of both test sets.[4] We can see that GPT-k-shot outperforms GPT-rubric on all test sets and the difference in correctly scored items is significant on both sets (both $\chi^2(1) > 4.71$, $p < .03$, two-sided). However, GPT-k-shot does not perform better than the models presented in Chifligarov et al. (2025), neither on the known nor on the unknown set. In fact, the differences in correctly scored items between GPT-k-shot on the one hand and the rule-based and BERT-based model on the other hand are not significant (all $\chi^2(1) < 2.57$, $p > .10$, two-sided).

All models tend to perform worse on the unknown than on the known test set. The performance drop is greater for the BERT-based model than for the GPT-k-shot model although both models rely on training samples. However, the number of correctly scored items only differs significantly for the BERT-based model ($\chi^2(1) = 11.31$, $p < .001$, two-sided) not for the other models (all $\chi^2(1) < 3.53$, $p > .06$, two-sided). The GPT-rubric model has the smallest performance drop. This result is intuitive because the scoring rubric describes more general scoring rules applying to all stimulus sentences. Still, GPT-rubric shows the worst performance on the unknown test set compared to the other models.

Table 3 breaks down the results, showing precision, recall and F1-score per score for each model on the combined test set. We see that due to duplicate removal, score 4 only applies to 9% of the answers, according to the gold standard, whereas the other scores are more evenly distributed. As already noted by Chifligarov et al. (2025), the rule-based model outperforms the other models on the edges of the rating scale, i.e. scores 0, 1 and 4, whereas the BERT-based model clearly performs better than the others on score 2. The rule-based model does not perform so well on scores 2 and 3, partly because not all categories from the scoring rubric pertaining to these scores have been implemented as they are not straightforward to be captured in rules. On score 3, GPT-k-shot has the best overall performance based on F1-score, although the rule-based model has a higher precision and GPT-rubric a slightly higher recall. As mentioned in Section 3, score 2 mainly captures grammatical

[4]Note that in Table 2, the performance of the rule-based model is sightly worse and that of the BERT-based model slightly better than reported in Chifligarov et al. (2025). This is mainly due to the removal of duplicates, which drastically reduces the number of answers with score 4, a category where the rule-based model outperforms the BERT-based model.

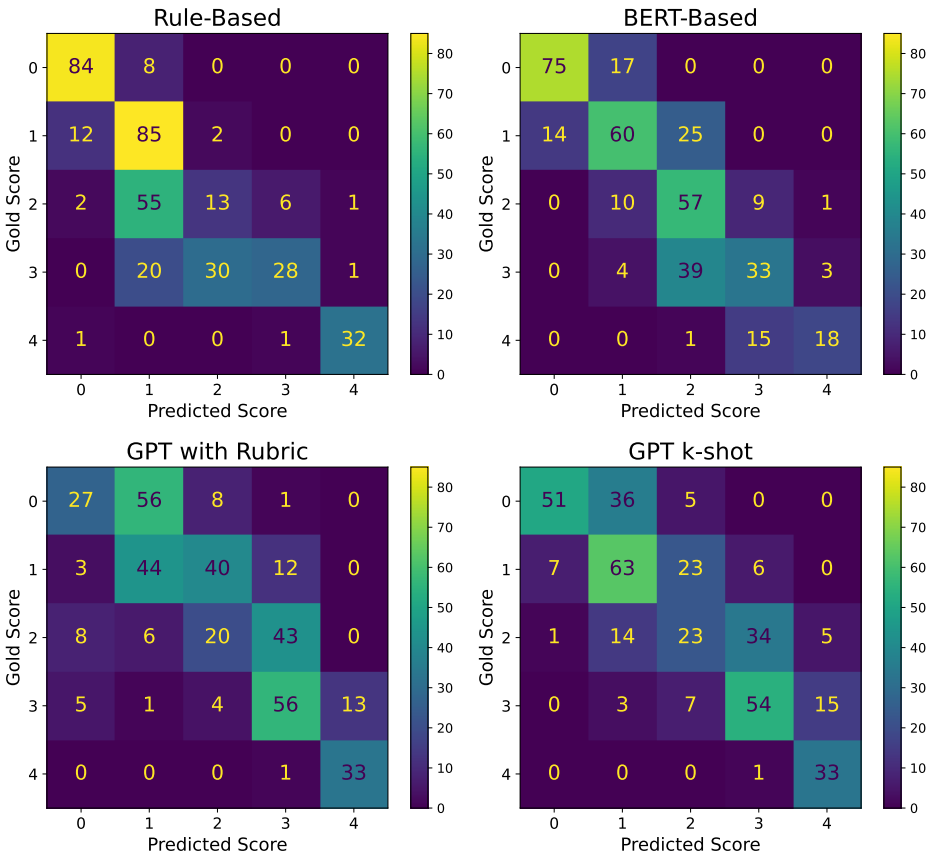

Figure 4: Confusion matrix of predicted vs. gold score for each model on the combined test set.

deviations whereas score 3 applies to semantic deviations. The results therefore suggest that the GPT-based models rather pick up semantic differences between stimulus and answer sentence compared to the BERT-based model. Looking at the score confusions in Figure 4 shows that the BERT- and rule-based models tend to undervalue answers in this score range, often giving answers with gold score 3 a score of 2 whereas the GPT-based models show the opposite pattern in that answers with gold score 2 often receive a score of 3.

7 Conclusion and Future Work

This paper explored the use of an LLM for scoring a German WEIT. The experiments focused on in-context learning, using k-shot prompting with large k values up to $k = 700$. We considered three sources of variance: The number of examples k, the chosen sample and the non-determinism of LLMs when running the same prompt multiple times. We showed that accuracy varies significantly with k (higher values yielding higher accuracy) and the chosen sample but with sufficiently large k, the effects become non-significant. Furthermore, while having no significant systematic effect in our regression model, one could see that accuracy is not stable across different runs of the same model. QWK turned out much more robust in this respect, reflecting that the deviations in the scores arising through multiple runs are only small.

Using its best configuration, our GPT k-shot model outperformed a zero-shot GPT model that uses the scoring rubric and no examples. Furthermore, we compared the GPT models to previously obtained results from Chifligarov et al. (2025), who used a rule-based and a BERT-based model. While the GPT k-shot model could not generally outperform the other models, an interesting pattern emerged when looking at scores 2 and 3, which were previously shown to be the ones that are the most difficult to predict. The GPT k-shot model tends to overestimate responses with score 2 (responses with grammatical errors) while the BERT-based model tends to underestimate responses with score 3 (responses with lexical deviations which are still grammatical). A possible explanation could be that the GPT model relies mostly on semantic similarity between the stimulus and the response sen-

tence, ignoring other kinds of deviations, whereas the BERT-based model pays too little attention to the semantics of a sentence. In future work, we want to investigate this further.

Our study focused on one particular LLM, thus the results may not generalize to other models but may also depend on the architecture, context window etc. Rather, our results show that an LLM does not *necessarily* outperform a rule-based or BERT-based model and that its observed variability, especially when running the same prompt multiple times, may limit its practical applicability. Further investigations are necessary here to find out whether this variability occurs randomly or affects e.g. certain scores in particular.

For future work, we see two promising avenues to improve the automated scoring of a WEIT, firstly experimenting with further prompting strategies for LLMs, especially combining the k-shot approach with the scoring rubric and secondly combining different kinds of models (rule-based, BERT-based and LLMs), exploiting their different strengths and weaknesses across the different scores.

Limitations

The focus of this paper was to study in-depth one prompting strategy (in-context learning with k-shot prompting) and the associated variability when using a particular LLM (GPT-4o) to score a German WEIT. The focus on one LLM limits the generalizability of the results as other models with other architectures and characteristics, e.g. in terms of parameters and context windows, may behave differently. Although our study already captures many factors that influence the performance, there can still be even more sources of variance such as periodic differences in performance of GPT-4o at different points in time (Tschisgale and Wulff, 2026). Furthermore, we did not aim to find the overall best LLM setup for scoring a German WEIT. From our results, one cannot draw the conclusion that k-shot prompting generally works better than rubric-based prompting as we did not systematically tune the rubric-based prompt. Finding the best LLM setup would involve trying different models and more prompting strategies, especially combining the scoring rubric with k-shot prompting, as is state of the art for many applications.

Acknowledgments

The WEIT for German was conceived, developed, and the data collected as part of a research project funded by the Deutsche Forschungsgemeinschaft (DFG), grant number 462766474. We would also like to thank the anonymous reviewers for their very helpful comments.

References

Berk Atıl, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. Non-determinism of "deterministic" LLM system settings in hosted environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pages 135–148, Mumbai, India. Association for Computational Linguistics.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149, Albuquerque, New Mexico. Association for Computational Linguistics.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Mihail Chifligarov, Jammila Laâguidi, Max Schellenberg, Alexander Dill, Anna Timukova, Anastasia Drackert, and Ronja Laarmann-Quante. 2025. Automated scoring of a German written elicited imitation test. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 237–247, Vienna, Austria. Association for Computational Linguistics.

Dylan G E Gomes. 2022. Should I use fixed effects or random effects when I have fewer than five levels of a grouping factor in a mixed-effects model? *PeerJ*, 10(e12794):e12794.

C. Ray Graham, Deryle Lonsdale, Casey Kennington, Aaron Johnson, and Jeremiah McGhee. 2008. Elicited imitation as an oral proficiency measure with ASR scoring. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Yue Huang and Joshua Wilson. 2025. Evaluating LLM-based automated essay scoring: Accuracy, fairness, and validity. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, pages 71–83, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).

Deryle Lonsdale and Carl Christensen. 2011. Automating the scoring of elicited imitation tests. In *Proc. Machine Learning in Speech and Language Processing (MLSLP 2011)*, pages 16–20.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can Large Language Models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italia. ELRA and ICCL.

Benjamin J Millard. 2011. Oral proficiency assessment of French using an Elicited Imitation Test and Automatic Speech Recognition. Master's thesis, Brigham Young University.

Shuyin Ouyang, {Jie M.} Zhang, Mark Harman, and Meng Wang. 2025. An empirical study of the non-determinism of ChatGPT in code generation. *ACM Transactions on Software Engineering and Methodology*, 34(2). Publisher Copyright: © 2025 Copyright held by the owner/author(s).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4195–4206, Albuquerque, New Mexico. Association for Computational Linguistics.

Anna Timukova, Oleksandra Yazdanfar, and Anastasia Drackert. 2026. So viele Lücken, so wenig Zeit: Die Rolle der Zeit im Konstrukt des deutschen C-Tests anhand der Analyse der Verarbeitungsprozesse. *Zeitschrift für Interkulturellen Fremdsprachenunterricht (ZIF)*, 31:81–108.

Paul Tschisgale and Peter Wulff. 2026. Daily and weekly periodicity in large language model performance and its implications for research. *Preprint*, arXiv:2602.15889.

Lui Yoshida. 2024. The impact of example selection in few-shot prompting on automated essay scoring using GPT models. In Andrew M. Olney, Irene-Angelica Chounta, Zitao Liu, Olga C. Santos, and Ig Ibert Bittencourt, editors, *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, volume 2150, pages 61–73. Springer Nature Switzerland, Cham. Series Title: Communications in Computer and Information Science.