

How Do I Say It Best? What Prompt Optimisation Can Reveal About Optimal Prompt Formulation In LLM-based Automated Short Answer Scoring

Sebastian Gombert¹, Sonja Hahn¹, Fabian Zehner^{1,2}, Longwei Cong¹,
Zhifan Sun¹, Leon Camus¹, Fabíola Gonçalves C. Ribeiro^{1,3},
Hendrik Drachsler^{1,4}

¹DIPF — Leibniz-Institute for Research and Information in Education,
Frankfurt, Germany.

²Centre for International Student Assessment (ZIB), Munich, Germany.

³Federal Institute Goiano (IF Goiano), Goiás, Brasil.

⁴Studiumdigitale & Computer Science Department,
Goethe-University, Frankfurt, Germany.

{s.gombert, s.hahn, f.zehner, l.cong,

z.sun, l.camus, f.goncalves, h.drachsler}@dipf.de

Abstract. *Large language models (LLMs) have shown promise for automated short-answer scoring (ASAS), particularly in cold-start settings where task-specific training data are unavailable. However, prior studies typically relied on a small number of manually designed prompts, making it unclear whether the observed performance reflects the capabilities of the model or just the quality of the prompt. In this paper, we investigate whether prompt optimisation improves ASAS performance and what it can reveal about effective prompt design. Using three open-weight LLMs, three task formulations, and English language benchmark datasets from SemEval-2013, we compare human-written prompts with prompts optimised using the ProTeGi framework. Results show that prompt optimisation yields consistent performance gains across 135 paired configurations. Linguistic analyses further reveal that optimised prompts contain more discourse markers and shorter noun phrases, and these properties are predictive of downstream scoring performance. The findings suggest that prompt optimisation can serve not only as a performance-enhancement technique, but also as a tool for identifying linguistic characteristics of effective prompts for educational NLP tasks.*

1. Introduction

Open-ended responses are widely considered a strong fit for assessing learners' active knowledge [Livingston 2009], and they also lend themselves well to providing formative feedback [Gombert et al. 2024]. However, state-of-the-art methods for scoring short answers typically require models to be trained on large task-specific datasets [Bai and Stede 2023, Bexte et al. 2022, Dada et al. 2025], preventing scalability in contexts where such datasets are not available, which includes the majority of all practical educational contexts, also known as the cold-start problem [Bexte and Zesch 2025]. The

transfer-, multi-task-, and in-context-learning capabilities of generative LLMs promise to diminish this crucial cost factor that limits the widespread scalability of intelligent educational systems, in theory. Generative LLMs often require none or only a small number of examples in combination with a strong prompt to achieve acceptable performance on tasks for which they were not explicitly trained for [Brown et al. 2020]. It is this promise that makes generative AI particularly attractive for the automated scoring of free-text responses in settings where training data are unavailable [Bexte and Zesch 2025, Kasneci et al. 2023].

Early studies have already explored the potential of generative LLMs for automated short-answer scoring (ASAS) [Chang and Ginter 2024, Kortemeyer 2023, Ferreira Mello et al. 2025]. These studies show that, at least for GPT-4, performance is not yet competitive with fully supervised systems, yet it can still yield surprisingly strong baseline scores via pure prompting. A key limitation of these studies is that they evaluate LLM scoring performance using arbitrary hand-crafted prompts. However, because linguistic productivity allows language users to express the same proposition in countless ways [Kreidler 1998], it is unclear whether these prompts are the optimal way to phrase the task. There is concrete evidence that not every possible prompt formulation is optimal, and even replacing single words in a prompt with synonyms can significantly impact downstream performance [Leidinger et al. 2023]. As a consequence, techniques for automatically searching for improved prompts via various heuristic strategies have been shown to outperform hand-crafted prompts in many cases [Pryzant et al. 2023, Fernando et al. 2023]. This raises the question of whether past studies on LLM-based ASAS used optimal prompts. Moreover, given linguistic productivity, which exact features make a given prompt optimal for ASAS is an open question. In this study, our objective was therefore to address these gaps with the following research questions:

- **RQ1:** Does prompt optimisation improve LLM performance in automated short-answer scoring?
- **RQ2:** What linguistic changes distinguish optimised prompts for automated short answer scoring from their human-written counterparts?
- **RQ3:** Which linguistic properties of prompts are predictive of scoring performance in automated short answer scoring?

2. Background

Automated short-answer scoring (ASAS) is one of the earliest established NLP applications in education [Burrows et al. 2015]. It refers to the automated evaluation of short student responses using computational methods, also known as automated short-answer assessment (ASAA) or automated short-answer grading (ASAG). Early approaches relied on concept matching, keyword lexica, rule-based pattern matching between student answers and sample solutions, or clustering techniques based on Latent Semantic Analysis (LSA) or embedding representations [Klein et al. 2011, Saha et al. 2018]. More recent work predominantly uses machine learning-based scoring models [Bai and Stede 2023]. The effectiveness of ASAS fundamentally depends on handling linguistic and content-related variability in student responses [Zesch et al. 2023].

Historically, machine-learning approaches to ASAS have mirrored broader NLP trends. Earlier systems combined hand-crafted features with statistical learners

[Leacock and Chodorow 2003, Jimenez et al. 2013]. With the introduction of BERT [Devlin et al. 2019], most subsequent ASAS research adopted transformer-based architectures [Sung et al. 2019, Li et al. 2021, Dada et al. 2025, Kaya and Cicekli 2024, Poulton and Eliens 2021]. Commonly, these systems frame scoring as a textual entailment task by cross-encoding the answer and reference solutions. In contrast, others rely on SBERT-style embeddings for clustering [Bexte et al. 2022] or for downstream learners [Padó et al. 2023].

The rise of generative LLMs raises the question of whether ASAS can be solved via pure prompting, without task-specific training. Initial evidence from studies such as [Ferreira Mello et al. 2025, Chang and Ginter 2024, Kortemeyer 2023] suggests that this is not yet the case, although evaluations often rely on a single manually crafted prompt. Given the rapid emergence of prompt engineering techniques [Sahoo et al. 2024, Ramnath et al. 2025] and their substantial impact on few-shot performance, it remains unclear whether an appropriately designed prompt could enable generative LLMs to approach or exceed the state-of-the-art performance of supervised systems. Recent findings [Lee et al. 2024] indicate that few-shot prompting and chain-of-thought reasoning [Wei et al. 2022] can improve automated scoring, but effects vary substantially across items, and evaluations remain limited in scope. Thus, the degree to which prompting techniques can reliably enhance LLM-based ASAS remains an open question.

3. Method

3.1. Experimental Setup

To answer **RQ1** and collect data basis for answering **RQ2** and **RQ3**, we evaluated how *task formulation*, *number of in-context examples*, and *prompt optimisation* affect ASAS performance. Each configuration consists of:

1. one of three task formulations: naïve ASAS, paraphrase detection, or textual entailment;
2. a zero-, one-, or five-shot setup;
3. either a hand-crafted or an optimised task prompt.

All models were prompted via standard in-context learning [Brown et al. 2020]. Examples took the form of quadruples containing the question, the sample solution, the student answer, and the gold label; only the task prompt was replaced during optimisation. Few-shot examples are sampled randomly from the training set at each inference step to avoid overfitting to a fixed exemplar set.

3.2. Human-written Prompts

We compared three task formulations: (1) a naïve scoring prompt that directly asks the model to score a student answer based on the sample solution, (2) a paraphrase formulation inspired by [Kortemeyer 2023], which frames correctness as semantic equivalence between student answer and reference answer, and (3) a textual entailment formulation following [Levy et al. 2013], where correctness is interpreted as the degree to which the student answer entails the sample solution.

3.3. Datasets

To ensure comparability with prior work, we relied on established benchmark datasets. The SemEval-2013 shared task [Dzikovska et al. 2013] introduced two widely used short-answer scoring datasets: **SciEntsBank** and **Beetle**. Both datasets contain training sets and multiple test partitions, most notably *unseen answers* (new responses to seen questions) and *unseen questions* (responses to new questions). SciEntsBank also provides an *unseen domains* split. In this paper, we focused on the 2-way variant, which distinguishes correct and incorrect as possible scores. Table 1 shows an example datapoint from the *SciEntsBank* dataset.

Question
You used several methods to separate and identify the substances in mock rocks. How did you separate the salt from the water?
Reference Answer
The water was evaporated, leaving the salt.
Student Answer
We put the water and salt in the Petri dish. The water evaporated and the salt became crystals.
Score
Correct

Table 1. This table shows an example datapoint from the *SciEntsBank* dataset.

3.4. Models

We relied on the following open-weight models **Mistral-7B-Instruct** [Jiang et al. 2023], **Apertus-8B-Instruct** [Apertus et al. 2025] and **OpenChat-3.5** [Wang et al. 2023]. The reasoning behind focusing on open models instead of larger commercial ones here is that queries to large commercial models are not only comparably expensive (both in terms of accumulating monetary costs and environmental impact), but can also conflict with data protection rules established in many educational contexts. The cheaper, more privacy- and environmentally friendly alternative to commercial models such as GPT-5 is compact, open-weight models in the 7b- to 8b-parameter class (or smaller) [Jeanquartier et al. 2026].

3.5. Prompt Optimisation using ProTeGi

To optimise prompts, we used *ProTeGi* [Pryzant et al. 2023], a framework that treats prompt optimisation as a search problem over alternative prompt formulations. Starting from a human-written prompt, the method first evaluates model predictions on labelled examples and generates *textual gradients*, i.e., natural-language critiques describing why the current prompt may have led to incorrect predictions. These critiques are then used to propose revised prompt variants that address the identified weaknesses. To further diversify the search space, candidate prompts are additionally paraphrased.

The resulting prompt candidates are evaluated on subsets of the training data and ranked according to downstream task performance. Following a variant from [Pryzant et al. 2023], candidate selection was implemented using the Successive Rejects

Feature	Rationale
Token count	Measures overall prompt length. Longer prompts may provide more contextual guidance but can also introduce noise or redundant instructions.
Sentence count	Reflects the degree of segmentation of the instructions. More sentences may structure instructions more clearly.
Words per sentence	Approximates syntactic density. Longer sentences may encode more complex instructions, but can reduce readability and clarity.
Verb count	Indicates the amount of predicate structure in the prompt, reflecting how strongly instructions are framed as actions or procedures.
Content-word ratio	Measures the proportion of lexical content words relative to function words, capturing the informational density of the prompt.
Zipf mean (content lemmas)	Captures the average lexical log-frequency of content words in the prompt. Lower-frequency words indicate greater lexical sophistication, which may influence how precisely the model interprets task instructions. Frequencies are taken from the SUBTLEX-US corpus [Brysbaert and New 2009].
Reasoning verb count	Counts verbs associated with analytical or evaluative reasoning (e.g., <i>evaluate</i> , <i>justify</i> , <i>determine</i>). These may signal the type of reasoning expected from the model during scoring.
Clause proxy count	Approximates syntactic complexity through subordinate clauses and punctuation markers. Higher values indicate more complex instruction structures.
Instruction verb count	Measures the density of directive verbs (e.g., <i>classify</i> , <i>assign</i> , <i>determine</i>) that explicitly instruct the model how to perform the scoring task.
Imperative root ratio	Estimates how often instructions are phrased in imperative form. Imperatives may make task expectations more explicit but can also lead to overly directive prompts.
Discourse marker count	Counts explicit discourse markers (e.g., <i>because</i> , <i>therefore</i> , <i>if</i>) that signal logical relations and may clarify evaluation criteria.
Mean dependency depth	Reflects the average depth of syntactic dependency trees, serving as a proxy for hierarchical syntactic complexity.
Mean noun phrase length	Measures the average length of noun phrases. Longer noun phrases indicate more nominal compression and potentially denser conceptual instructions.

Table 2. Linguistic features used in the prompt analysis together with their theoretical motivation.

algorithm [Audibert and Bubeck 2010], which iteratively eliminates lower-performing candidates until only a single prompt remains. The final surviving prompt is used as optimised prompt for all subsequent analyses.

3.6. Hyperparameters and Setup

Prompt optimisation was run separately for each dataset and prompting configuration (i.e., once for *SciEntsBank* and once for *Beetle* per combination of task formulation and shot count). For *ProTeGi*, we extracted 4 textual gradients and generated a small candidate pool of 16 prompt variants (arms) in each round. Successive Rejects started with a sample size of 20 instances per candidate and increased by 2 per round. We ran three optimisation rounds (= full beam search iterations), after which the top-performing prompt was selected for the next round. For generating textual gradients and prompt candidates in *ProTeGi*, a temperature of 0.7 was used, whereas for evaluating model performance, a temperature of 0.0 was used to prevent fluctuations due to randomness. Each model–prompt configuration is evaluated on every SemEval-2013 test subset, yielding 270 total model–configuration–dataset combinations. This design enables us to disentangle the contributions of task framing, shot count, and prompt optimisation, and to observe interactions among these factors.

3.7. Statistical Analysis

To investigate the linguistic properties of optimised prompts, we adopted a two-stage analysis procedure. First, we identified systematic linguistic differences between human-

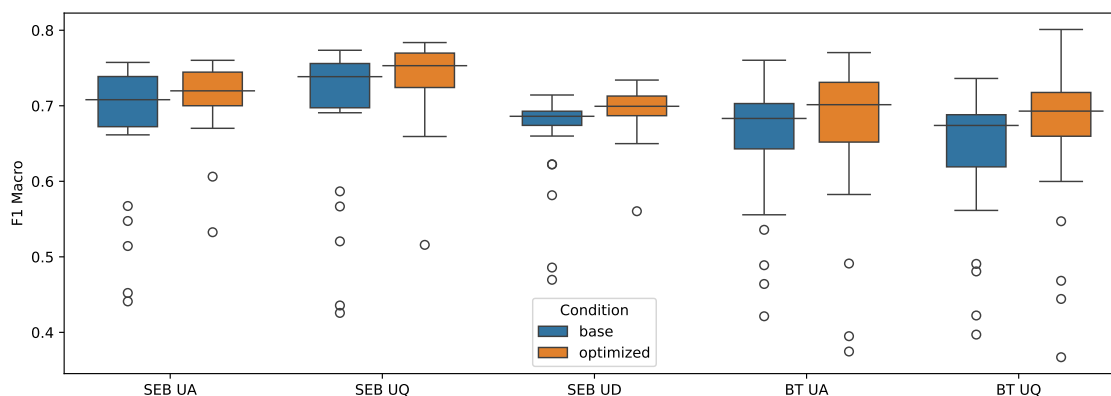


Figure 1. These box plots show the distribution of F1 Results for the base prompts and their optimised counterparts for all three models across all evaluation sets. SEB = SciEntsBank. BT = Beetle. UA = Unseen Answers. UQ = Unseen Questions. UD = Unseen Domains.

written base prompts and prompts obtained through optimisation. Second, we tested whether these linguistic properties predict downstream scoring performance.

In the first stage, we computed a set of linguistic prompt features (Table 2) for each base prompt and its corresponding optimised variant, using *SpaCy* [Honnibal et al. 2020] with its default English model. Because each optimised prompt originates from a specific human-written base prompt, the resulting observations form paired samples. To determine whether prompt optimisation leads to systematic changes in linguistic structure, we therefore compared base and optimised prompts using two-sided paired Wilcoxon signed-rank tests. The goal of this analysis was exploratory: rather than testing a specific theoretical model, we aimed to identify which linguistic properties consistently change when prompts are automatically improved for performance. These results provide empirical guidance for selecting a subset of linguistically meaningful predictors.

In the second stage, we used the features identified through this exploratory comparison to model prompt performance. Specifically, we estimated mixed-effects regression models predicting the raw scoring performance of prompts while controlling for task formulation, shot count, and prompting strategy. Because each prompt was evaluated across multiple test splits of the dataset (*unseen answers*, *unseen questions*, *unseen domains*), we included random intercepts to account for repeated observations within the same prompt configuration. This modelling step allowed us to determine whether the linguistic properties observed in the Wilcoxon analysis are associated with higher ASAS performance in practice.

4. Results

4.1. RQ1: Performance Improvement through Prompt Optimisation

Figure 1 compares the distribution of F1 Macro scores obtained with the human-written base prompts and the prompts produced through automated optimisation across all SemEval-2013 evaluation splits. Across all datasets, the distributions for the optimised prompts are consistently shifted upwards relative to their human-written counterparts, indicating that prompt optimisation generally improves scoring performance. This pattern

Feature	<i>N</i>	Base	Optimized	Diff.	<i>W</i>	<i>p</i>	<i>r</i>
Token count	54	54.00 [12.00, 79.00]	49.00 [40.25, 66.75]	2.00 [-16.75, 32.75]	511.50	= 0.047	0.27
Sentence count	54	3.00 [1.00, 4.00]	3.00 [2.00, 3.75]	0.00 [-1.00, 1.75]	316.00	= 0.083	0.26
Words per sentence	54	18.00 [12.00, 19.75]	18.83 [14.12, 22.94]	1.50 [-2.33, 8.88]	477.00	= 0.022	0.31
Verb count	54	10.00 [2.00, 17.00]	8.00 [6.00, 12.75]	0.00 [-5.00, 4.00]	671.00	= 0.870	0.02
Content-word ratio	54	0.58 [0.56, 0.58]	0.59 [0.56, 0.61]	0.02 [-0.02, 0.04]	548.00	= 0.094	0.23
Zipf mean (content lemmas)	54	4.68 [4.66, 4.97]	4.70 [4.61, 4.81]	-0.06 [-0.17, 0.03]	447.00	= 0.011	0.35
Reasoning verb count	54	4.00 [1.00, 5.00]	3.00 [2.00, 4.00]	0.00 [-2.00, 1.00]	374.50	= 0.102	0.24
Clause proxy count	54	7.00 [1.00, 11.00]	6.00 [4.00, 8.00]	0.00 [-4.00, 5.00]	613.00	= 0.813	0.03
Instruction verb count	54	3.00 [1.00, 4.00]	2.00 [2.00, 3.00]	0.00 [-1.00, 1.00]	384.50	= 0.274	0.16
Imperative root ratio	54	0.67 [0.50, 1.00]	1.00 [0.50, 1.00]	0.07 [-0.31, 0.33]	496.00	= 0.807	0.04
Discourse marker count	54	0.00 [0.00, 1.00]	1.00 [0.00, 2.00]	1.00 [0.00, 2.00]	55.00	< .001	0.74
Mean dependency depth	54	3.04 [2.92, 3.71]	3.20 [2.74, 3.68]	0.03 [-0.54, 0.62]	669.00	= 0.527	0.09
Mean noun phrase length	54	2.53 [2.50, 2.67]	2.32 [2.08, 2.46]	-0.23 [-0.46, -0.08]	62.50	< .001	0.79

Table 3. Paired two-sided Wilcoxon signed-rank tests comparing base and optimized prompts across unique prompt formulations.

is particularly visible for all test sets. To quantify this effect, we compared each optimised prompt against its corresponding human-written base prompt using a paired two-sided Wilcoxon signed-rank test. Across the 135 paired prompt configurations, prompt optimisation yielded a significant improvement in scoring performance, $W = 1757.00$, $p < .001$, with a large effect size ($r = .62$). The median F1 Macro score increased from $Mdn = 0.690$ for the human-written prompts to $Mdn = 0.712$ for the optimised prompts, corresponding to a median paired improvement of $\Delta Mdn = 0.019$. The mean paired improvement amounted to $\Delta M = 0.030$ F1 points.

4.2. RQ2: Linguistic Changes through Prompt Optimisation

Table 3 summarises the linguistic differences between the human-written base prompts and the prompts obtained through automated optimisation. Overall, prompt optimisation led to several systematic changes in prompt structure. Most notably, optimised prompts contained significantly more discourse markers than their human-written counterparts ($W = 55.00$, $p < .001$, $r = 0.74$), representing a large effect size. For example, the model turned the prompt *[s]core the "Student Answer" to the "Question" based on the "Sample Solution". To assign a correct score, you need to determine whether the "Student Answer" entails the "Sample Solution". Give a score from the provided list, depending on the degree of entailment. The "Reference Answer" is entailed in the "Student Answer" if a human reading of the "Reference Answer" would be justified in inferring the proposition expressed by the "Student Answer" from the proposition expressed by the "Reference Answer" into the prompt [a]nalyze the "Student Answer" in relation to the "Question" and "Sample Solution" by considering the human's interpretation of the "Reference Answer". Determine if the "Student Answer" includes the "Sample Solution" and if the "Reference Answer" would warrant the proposition expressed by the "Student Answer". Assign a score from the given list, representing the level of entailment.* At the same time, optimisation substantially reduced mean noun phrase length ($W = 62.50$, $p < .001$, $r = 0.79$), suggesting a shift away from dense nominal constructions towards simpler linguistic formulations. Smaller but statistically significant changes were also observed for words per sentence ($W = 477.00$, $p = .022$, $r = 0.31$), Zipf frequency of content lemmas ($W = 447.00$, $p = .011$, $r = 0.35$), and overall token count ($W = 511.50$, $p = .047$, $r = 0.27$). In contrast, no significant differences were found for sentence count, verb count, reasoning verb count, clause proxy count, instruction verb count, imperative root ratio, or mean dependency depth ($p > .05$ in all cases).

4.3. RQ3: Linguistic Predictors of Performance

Table 4 presents the results of the mixed-effects regression model predicting prompt performance from prompt characteristics while controlling for prompt type, task formulation, and shot count. Among the linguistic predictors, discourse marker count emerged as a significant positive predictor of performance ($\beta = 0.021$, $SE = 0.005$, $z = 4.343$, $p < .001$), indicating that prompts containing more explicit markers of logical relations achieved higher scoring accuracy. In contrast, both mean noun phrase length ($\beta = -0.010$, $SE = 0.005$, $z = -2.150$, $p = .032$) and instruction verb count ($\beta = -0.017$, $SE = 0.005$, $z = -3.490$, $p < .001$) were negatively associated with performance, suggesting that prompts relying on longer nominal constructions or large numbers of directive instructions tended to be less effective. Neither lexical sophistication, as measured by Zipf frequency, nor words per sentence significantly predicted performance ($p > .05$).

Term	Estimate	SE	z	p	
Intercept	0.686	0.017	39.369	< .001	***
C(Prompt Type)[T.optimized]	-0.006	0.011	-0.538	= 0.591	
C(Num Shots)[T.1]	-0.033	0.008	-3.862	< .001	***
C(Num Shots)[T.5]	0.024	0.008	2.865	= 0.004	**
C(strategy)[T.naive]	-0.001	0.010	-0.057	= 0.954	
C(strategy)[T.paraphrase]	0.004	0.009	0.493	= 0.622	
Zipf Mean (Content Lemmas) Z	-0.008	0.005	-1.584	= 0.113	
Words Per Sentence Z	0.002	0.004	0.504	= 0.614	
Discourse Marker Count Z	0.021	0.005	4.343	< .001	***
Mean Noun Phrase Length Z	-0.010	0.005	-2.150	= 0.032	*
Instruction Verb Count Z	-0.017	0.005	-3.490	< .001	***
Group Var	1.087	0.445	2.444	= 0.015	*

Table 4. Mixed-effects model predicting raw prompt performance from prompt type, control variables, and theory-driven linguistic features.

Among the control variables, five-shot prompting significantly outperformed zero-shot prompting ($\beta = 0.024$, $p = .004$), whereas one-shot prompting resulted in lower performance than the zero-shot baseline ($\beta = -0.033$, $p < .001$). An additional noteworthy finding is that whether a given prompt was optimised or not did not significantly predict performance once linguistic prompt characteristics were included in the model ($\beta = -0.006$, $p = .591$).

5. Discussion

Our findings provide clear evidence that prompt formulation substantially influences the performance of LLM-based automated short-answer scoring. With respect to **RQ1**, automatically optimised prompts consistently outperformed their human-written counterparts across datasets and experimental conditions. Although the median improvement of approximately two F1 points appears modest in absolute terms, the effect was highly consistent across 135 paired comparisons and resulted in a large overall effect size. This finding complements earlier studies on LLM-based ASAS [Chang and Ginter 2024, Kortemeyer 2023], which evaluated only a small number of manually designed prompts. Our results suggest that performance estimates obtained from individual hand-crafted prompts may underestimate the capabilities of prompting-based approaches and that prompt wording constitutes a meaningful source of variance in ASAS evaluations.

Regarding **RQ2**, prompt optimisation did not primarily alter prompts by making them longer, more syntactically complex, or more instruction-heavy. Instead, the strongest effects were observed for discourse marker usage and noun phrase complexity. Optimised prompts contained substantially more explicit discourse markers while simultaneously relying on shorter and less complex nominal constructions. Interestingly, these changes move in the opposite direction of a common intuition in prompt engineering that increasingly detailed and elaborate instructions necessarily improve model performance. Rather than producing longer prompts, optimisation appears to favour prompts that communicate task requirements more explicitly and coherently. This observation is consistent with the view that LLM prompting is fundamentally a communicative process, in which clarity of instruction may be more important than the sheer amount of information provided.

The findings for **RQ3** provide a possible explanation for why prompt optimisation improved scoring performance. The linguistic properties most strongly affected by optimisation were also among the strongest predictors of downstream performance. In particular, discourse marker usage emerged as a robust positive predictor, whereas longer noun phrases and larger numbers of instruction verbs were associated with lower performance. Moreover, once these linguistic characteristics were included in the regression model, prompt optimisation itself no longer explained additional variance. This suggests that optimisation acts primarily through modifying specific linguistic properties rather than through some opaque optimisation effect. From a practical perspective, these results indicate that effective prompts for ASAS should emphasise explicit logical relations and clear discourse structure while avoiding unnecessarily dense nominal formulations and excessive directive language.

Overall, our findings suggest that prompt optimisation is valuable not only as a performance-enhancement technique, but also as a methodological tool for studying effective prompt design. Optimised prompts consistently improved automated short-answer scoring performance, and the linguistic changes introduced by optimisation largely mirrored the linguistic properties that predicted downstream success. This indicates that prompt optimisation can reveal generalisable characteristics of effective prompts. More broadly, the results highlight that prompt formulation remains an important and often underexplored source of variation in LLM-based educational applications, warranting more systematic investigation in future work.

6. Limitations and Threats to Validity

The research in this paper used solely English language datasets. While it is likely that the general effectiveness of prompt optimisation transfers to other languages, the concrete linguistic properties behind this effect likely differ across languages. Future work therefore needs to test these hypotheses for different languages with available data sets such as Portuguese [Galhardi et al. 2018] or German [Gombert et al. 2026]. Moreover, this study focused on models of the 7b to 8b parameter class. Effects might differ for smaller or larger models. Future work therefore also needs to explore whether the results from this study hold up across models of different size. Prompt optimisation was limited to the task description. Other aspects of prompting, i.e., where to place slots for few-shot examples, were fixed. It is not unlikely that aspects such as the latter can also influence downstream performance.

AI Statement

Beyond the research topic, for this study, we used large language models for both programming and writing. Concretely, we generated parts of the code used for this study, which were manually checked for correctness thereafter, with *GPT* and *Claude*. *Grammarly* and *GPT* were used in the refinement of the text.

References

- Apertus, P., Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., et al. (2025). Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*.
- Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT - 23rd Conference on Learning Theory*, pages 13–p.
- Bai, X. and Stede, M. (2023). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, 33(4):992–1030.
- Bexte, M., Horbach, A., and Zesch, T. (2022). Similarity-based content scoring—how to make s-bert keep up with bert. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 118–123. Association for Computational Linguistics.
- Bexte, M. and Zesch, T. (2025). Is lunch free yet? overcoming the cold-start problem in supervised content scoring using zero-shot llm-generated training data. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 144–159. Association for Computational Linguistics.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.
- Burrows, S., Gurevych, I., and Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.
- Chang, L.-H. and Ginter, F. (2024). Automatic short answer grading for finnish with chatgpt. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23173–23181.
- Dada, I. D., Akinwale, A. T., Osinuga, I. A., Ogbu, H. N., and Tunde-Adeleke, T.-J. (2025). iAttention transformer: An inter-sentence attention mechanism for automated grading. *Mathematics*, 13(18):2991.

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Dzikovska, M. O., Nielsen, R. D., and Leacock, C. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of SemEval 2013*, pages 263–274.
- Fernando, C., Banarse, D., Michalewski, H., Osindero, S., and Rocktäschel, T. (2023). Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.
- Ferreira Mello, R., Pereira Junior, C., Rodrigues, L., Pereira, F. D., Cabral, L., Costa, N., Ramalho, G., and Gasevic, D. (2025). Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 93–103.
- Galhardi, L., Barbosa, C. R., de Souza, R. C. T., and Brancher, J. D. (2018). Portuguese automatic short answer grading. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 29, page 1373.
- Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J., Frey, A., and Drachsler, H. (2024). From the automated assessment of student essay content to highly informative feedback: a case study. *International Journal of Artificial Intelligence in Education*, 34(4):1378–1416.
- Gombert, S., Sun, Z., Zehner, F., Lossjew, J., Wyrwich, T., Czinczel, B. K., Bednorz, D., Bernholt, S., Neumann, K., Harms, U., et al. (2026). Report on the BEA 2026 shared task on rubric-based short answer scoring for german. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*. Association for Computational Linguistics.
- Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.
- Jeanquartier, F., Jean-Quartier, C., Rieder, P., Misirlić, V., Pasero, C., Hohensinner, R., Müller, H., and Holzinger, A. (2026). Assessing the carbon footprint of language models: Towards sustainability in AI. *Resources, Conservation and Recycling*, 226:108670.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7b.
- Jimenez, S., Becerra, C., and Gelbukh, A. (2013). SOFTCARDINALITY: Hierarchical text overlap for student response analysis. In Manandhar, S. and Yuret, D., editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (Se-*

- mEval 2013*), pages 280–284, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- Kaya, M. and Cicekli, I. (2024). A hybrid approach for automated short answer grading. *IEEE Access*, 12:96332–96341.
- Klein, R., Kyrilov, A., and Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of the 16th annual joint conference on Innovation and technology in computer science education*, page 158–162. ACM.
- Kortemeyer, G. (2023). Performance of the pre-trained large language model gpt-4 on automated short answer grading. *CoRR*, abs/2309.09338.
- Kreidler, C. (1998). *Introducing English semantics*. Routledge.
- Leacock, C. and Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.
- Lee, G.-G., Latif, E., Wu, X., Liu, N., and Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100213.
- Leidinger, A., van Rooij, R., and Shutova, E. (2023). The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.
- Levy, O., Zesch, T., Dagan, I., and Gurevych, I. (2013). UKP-BIU: Similarity and entailment metrics for student response analysis. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 285–289, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Li, Z., Tomar, Y., and Passonneau, R. J. (2021). A semantic feature-wise transformation relation network for automatic short answer grading. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6030–6040, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Livingston, S. (2009). Constructed-response test questions: Why we use them; how we score them. Technical Report 11, ETS.
- Padó, U., Eryilmaz, Y., and Kirschner, L. (2023). Short-answer grading for german: Addressing the challenges. *International Journal of Artificial Intelligence in Education*, 34(4):1321–1352.

- Poulton, A. and Eliens, S. (2021). Explaining transformer-based models for automatic short answer grading. In *Proceedings of ICDTE 2021*, pages 110–116.
- Pryzant, R., Iter, D., Li, J., Lee, Y., Zhu, C., and Zeng, M. (2023). Automatic prompt optimization with “gradient descent” and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.
- Ramnath, K., Zhou, K., Guan, S., Mishra, S. S., Qi, X., Shen, Z., Wang, S., Woo, S., Jeoung, S., Wang, Y., Wang, H., Ding, H., Lu, Y., Xu, Z., Zhou, Y., Srinivasan, B., Yan, Q., Chen, Y., Ding, H., Xu, P., and Cheong, L. L. (2025). A systematic survey of automatic prompt optimization techniques. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33078–33110, Suzhou, China. Association for Computational Linguistics.
- Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., and Sengupta, B. (2018). Sentence level or token level features for automatic short answer grading?: Use both. In *Artificial Intelligence in Education*, volume 10947 of *Lecture Notes in Computer Science*, pages 503–517. Springer.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 1.
- Sung, C., Dhamecha, T. I., and Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *Artificial Intelligence in Education*, volume 11625 of *Lecture Notes in Computer Science*, pages 469–481. Springer.
- Wang, G., Cheng, S., Zhan, X., Li, X., Song, S., and Liu, Y. (2023). Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Zesch, T., Horbach, A., and Zehner, F. (2023). To score or not to score: Factors influencing performance and feasibility of automatic content scoring of text responses. *Educational Measurement: Issues and Practice*, 42(1):44–58.