Evaluating Reader-Dependent Text Complexity Predictions in Large Language Models

Boris Thome
Heinrich Heine University Düsseldorf
Düsseldorf, Germany
boris.thome@hhu.de

Stefan Conrad
Heinrich Heine University Düsseldorf
Düsseldorf, Germany
stefan.conrad@hhu.de

Abstract

Perceived text complexity depends not only on linguistic properties of a text, but also on characteristics of the reader such as language proficiency or educational background. While large language models (LLMs) are increasingly used in educational and readability-related applications, it remains unclear whether they capture such reader-dependent variation. We evaluate multiple state-of-the-art LLMs on German sentence-level text complexity annotations enriched with reader-specific metadata. Beyond predictive performance, we analyze how linguistic features relate to human judgments and model predictions across reader groups. Our results show that LLMs adjust overall difficulty predictions based on reader profiles, but remain only moderately aligned with human annotations. Moreover, predictions from all evaluated models show stable associations with surface-level and structural features, such as sentence length and syntactic complexity, with only minimal variation across reader profiles. This suggests that current LLMs simulate reader adaptation mainly through global shifts in predicted difficulty rather than nuanced reader-dependent interpretations of linguistic complexity.

1 Introduction

Text complexity assessment plays an important role in educational technology, readability research, and adaptive learning systems. Automatically estimating how difficult a text is to understand is relevant for applications such as personalized learning support, readability-controlled text generation, educational recommendation systems, and text simplification. Traditional approaches to readability assessment typically model text complexity as an inherent property of the text itself, relying on surface-level indicators such as sentence length, word length, or predefined readability formulas (Flesch, 1948; Kincaid et al., 1975).

However, text complexity is multidimensional and cannot be fully captured by surface-level features such as word and sentence length alone (Graesser et al., 2004; McNamara et al., 2014; Pitler and Nenkova, 2008). The same text may be considered easy for one reader but difficult for another depending on factors such as language proficiency, background knowledge, and domain expertise (Beinborn et al., 2012; Seiffe et al., 2022). Recent work has therefore emphasized the importance of incorporating reader-specific information into complexity prediction models and has shown that person-related features can improve the modeling of perceived text difficulty (Thome et al., 2025).

As large language models are increasingly explored in educational applications (Kasneci et al., 2023; Wang et al., 2025), it is important to examine whether they can similarly account for reader-specific variation. Due to their strong instruction-following capabilities and broad linguistic knowledge, LLMs are often assumed to adapt flexibly to different audiences and user profiles. In educational settings, these capabilities have motivated growing interest in personalized learning support and in adapting LLM-generated content to readers of different ages and grade levels (Kasneci et al., 2023; Rooein et al., 2023; Naeem et al., 2025). Despite these developments, it remains unclear whether LLMs actually model reader-dependent text complexity in a human-like way.

Prior work has investigated controllable sentence simplification and its evaluation (Martin et al., 2020; Maddela et al., 2021; Alva-Manchego et al., 2020), while comparatively little attention has been paid to the decision patterns underlying LLM-based complexity assessments. A more detailed analysis of these patterns can reveal whether reader-profile information affects only the overall level of predicted difficulty, or also the linguistic cues associated with difficulty. Whether current LLMs exhibit such profile-dependent behavior remains an open question.

In this work, we investigate how LLMs model perceived text complexity across different reader profiles. Using a dataset of German sentence-level complexity annotations enriched with person-related metadata, we compare multiple state-of-the-art LLMs under both generic and profile-conditioned prompting settings. We analyze not only predictive performance, but also the linguistic signals underlying model predictions.

More specifically, we address the following research questions:

- **RQ1:** To what extent do LLM predictions align with human judgments of perceived text complexity?
- **RQ2:** How do reader-specific profiles influence LLM predictions of perceived text difficulty?
- **RQ3:** How strongly are LLM predictions associated with linguistic features compared with human ratings, and do these associations vary across reader profiles?

Our results show that current LLMs only partially align with human judgments of perceived text complexity. While models adapt their overall predictions depending on the provided reader profile, these changes do not consistently improve agreement with human annotations. Furthermore, linguistic feature analyses reveal strong and stable associations between model predictions and structural and surface-level features such as sentence length and syntactic complexity, with only minimal variation in these associations across reader profiles. This suggests that current LLMs primarily simulate reader adaptation through global shifts in predicted difficulty rather than through nuanced reader-dependent interpretations of linguistic complexity.

Overall, our work contributes to a better understanding of how LLMs model perceived text complexity and highlights important limitations of prompt-based personalization for educational NLP applications.

2 Related Work

2.1 Text Complexity and Readability Assessment

Research on text complexity and readability has a long tradition in computational linguistics and educational technology. Early approaches relied on handcrafted readability formulas such as Flesch Reading Ease (Flesch, 1948) or the Flesch–Kincaid grade level (Kincaid et al., 1975), which estimate difficulty mainly from surface-level characteristics such as sentence and word length. Although these formulas are efficient and interpretable, they provide only a limited view of text complexity, as they do not account for semantic, discourse-level, and cognitive aspects of text processing (Pitler and Nenkova, 2008; Graesser et al., 2004; McNamara et al., 2014).

To address these limitations, more recent approaches have incorporated broad linguistic feature sets into readability assessment (Weiss and Meurers, 2022). Frameworks such as Coh-Metrix assess multiple dimensions of text complexity using indicators related to cohesion, syntax, lexical sophistication, and discourse organization (Graesser et al., 2004; McNamara et al., 2014). These studies demonstrate that text difficulty cannot be fully explained by surface-level measures alone.

2.2 Reader-dependent Complexity Modeling

A key limitation of text-only approaches is that difficulty is not an inherent property of a text alone. Rather, it emerges from the interaction between textual properties and reader characteristics. Previous studies have shown that perceived text difficulty can vary across readers and is associated with reader characteristics such as self-assessed language proficiency (Thome et al., 2025) and domain expertise (Seiffe et al., 2022). This is especially relevant in educational settings, where a text that is accessible to one learner may be challenging for another. Reader-aware approaches therefore incorporate person-related information into complexity prediction models.

2.3 LLMs for Educational Personalization

The growing use of LLMs in educational contexts has increased interest in personalization and audience adaptation. LLMs are now widely explored for educational applications such as adaptive tutoring, question answering, grade-adapted content generation, and personalized learning support (Oh et al., 2026; Naeem et al., 2025; Wang et al., 2025; Kasneci et al., 2023). Related work on controllable text simplification has shown that neural models can be guided toward simpler outputs by controlling properties such as length or lexical complexity (Martin et al., 2020), as well as on the degree of splitting, deletion, and paraphrasing applied to the

input (Maddela et al., 2021). This suggests that modern models can, at least to some extent, adjust linguistic properties of generated text.

However, adapting model outputs to specific audiences remains challenging. Recent studies suggest that LLMs do not reliably adjust their language to different learner groups through prompting alone. For instance, Rooein et al. (2023) find that generated texts often remain within relatively fixed readability ranges, even when models are prompted for different age groups and education levels. Similarly, Naeem et al. (2025) show that LLMs still struggle to generate responses appropriate for early-grade students. Complementing these findings, Oh et al. (2026) propose Classroom AI, a grade-specific framework for educational assistance, and report that fine-tuning improves alignment compared to prompt-based adaptation.

In contrast to prior work on audience-adapted generation, we examine how LLMs incorporate reader-profile information in text complexity assessment. Specifically, we investigate whether such information merely affects the overall difficulty level predicted by the model, or whether it also leads to profile-dependent changes in the linguistic cues associated with difficulty.

3 Data

We use the German sentence-level text complexity dataset introduced in our previous work (Thome et al., 2025). The dataset combines perceived difficulty ratings with person-related annotator information, allowing us to analyze reader-dependent variation in perceived text complexity.

The annotations were collected from 193 secondary school students, who rated sentence difficulty on a 5-point Likert scale ranging from 1 (very easy) to 5 (very difficult). In addition to the ratings, the dataset contains reader-specific metadata, including age, gender, language background, self-assessed German proficiency, and German grades. These variables provide the basis for constructing the reader profiles used in our experiments.

The dataset contains 3,954 sentence-level annotations covering 1,697 unique sentences. The annotations are based on two text sources: 3,160 annotations refer to sentences from official German study guidance books published between 1971 and 2021 by the German Federal Employment Agency, while 794 annotations refer to Wikipedia sentences drawn from the TextComplexityDE dataset (Naderi et al., 2019). Intended for secondary school students, the study guidance books provide factual information about degree programs, fields of study, and post-secondary educational pathways. By contrast, the Wikipedia sentences cover a broader range of topics and were not specifically written for this target group. Despite these differences, both sources predominantly consist of informational and expository texts. Combining them introduces variation in topic, vocabulary, communicative purpose, and assumed background knowledge while preserving a largely factual text type. Such differences may affect perceived difficulty, as students' judgments can depend not only on general linguistic complexity but also on topic familiarity, domain-specific terminology, and the amount of prior knowledge presupposed by a sentence.

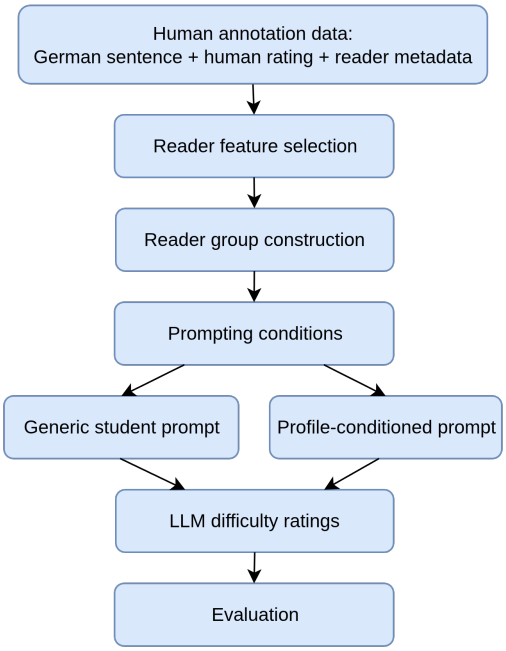

Figure 1: Overview of the reader-profile prompting and evaluation pipeline.

The number of annotations per sentence varies: 842 sentences were annotated by multiple students, while 855 received only a single annotation, resulting in an average of 2.33 annotations per sentence. Inter-annotator agreement is low (Krippendorff's $\alpha = 0.23$ for ordinal data), indicating substantial variation in perceived difficulty across readers (Thome et al., 2025). This variability reflects the reader-dependent nature of perceived text complexity and provides the basis for analyzing whether LLM predictions capture similar differences across reader profiles.

Feature	Type	Effect size
Language	ordinal	0.0223
Self-estimation	ordinal	0.0157
German grade	ordinal	0.0144
Gender	categorical	0.0102
Maths grade	ordinal	0.0044
Field of study	categorical	0.0032
Career path	categorical	0.0030
School type	categorical	0.0024
Age	ordinal	0.0019

Table 1: Associations between person-related features and human difficulty ratings.

4 Method

Figure 1 summarizes our experimental pipeline. We first identify reader-related variables associated with human difficulty ratings and use them to construct stronger and weaker reader profiles. We then compare LLM predictions under a generic baseline prompt and profile-conditioned prompts. Finally, we analyze whether linguistic features relate differently to human judgments and LLM predictions across reader groups.

4.1 Reader Feature Selection

Before constructing reader profiles for the LLM experiments, we conduct a preliminary analysis of the students' annotations to identify suitable reader-related variables. We select profile variables based on three criteria: they should show a measurable association with human difficulty ratings, be interpretable with respect to language experience or proficiency, and allow for meaningful comparisons between reader groups.

For ordinal variables, we compute Spearman rank correlations (Spearman, 1904) with the annotated difficulty scores and report squared correlations (ρ^2) as a rank-based measure of association. We treat *language background* as ordinal. The feature is based on students' self-reported home language ('German', 'German and another language', 'another language'), which we interpret as reflecting decreasing exposure to German in the home environment. For nominal categorical variables, we apply the Kruskal-Wallis test (Kruskal and Wallis, 1952) and report η^2 as an effect size. Since the reported effect sizes are derived from different statistical procedures, we use them descriptively to support the selection of suitable profile variables rather than as strictly equivalent quantities.

Table 1 summarizes the associations between reader-related variables and human difficulty ratings. Overall, the observed effect sizes are small, suggesting that perceived difficulty is influenced by multiple factors rather than by any single reader characteristic. Nevertheless, *language background*, *self-estimation of German proficiency*, and *German grade* show the strongest associations among the available reader metadata. Moreover, all three variables are directly related to language experience or language proficiency and can be mapped to interpretable reader groups. Group-specific agreement analyses provide additional descriptive support for this selection: within-group agreement is higher than agreement in the corresponding pooled subset for five of the six reader groups (see Appendix A.1). We therefore use these variables for the profile-conditioned experiments.

4.2 Model Selection and Prompting

To evaluate whether LLMs are able to capture reader-specific variation in perceived text complexity, we compare a diverse set of state-of-the-art models that differ in size, architecture, and training objectives. All models were accessed via the Azure-hosted model endpoints[1]. Specifically, we include two models from the `GPT-5.4` family (`GPT-5.4 nano`[2] and `GPT-5.4 mini`[3]), which allow us to analyze the effect of model scale within the same model family. In addition, we include `DeepSeek-V3.1`[4], which builds on `DeepSeek-V3` (Liu et al., 2024), and `Mistral Large 3`[5] to provide a broader comparison across independently developed LLMs. This selection enables us to examine whether observed patterns generalize across models with different design choices and training data. For each model, we treat model predictions as point estimates and analyze them directly in comparison to aggregated human annotations.

Feature-based profile groups. As a reference setting, we first define a baseline condition, in which the LLM is applied to all available texts (1,697 sentences) without providing any reader-

[1] https://ai.azure.com/catalog/models

[2] https://developers.openai.com/api/docs/models/gpt-5.4-nano

[3] https://developers.openai.com/api/docs/models/gpt-5.4-mini

[4] https://huggingface.co/deepseek-ai/DeepSeek-V3.1

[5] https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12

specific information beyond a generic student profile (see Appendix B).

We construct separate comparison subsets based on *language background*, *German grade*, and *self-estimated German proficiency*. Within each feature, students are divided into two groups reflecting relatively stronger and weaker language-related profiles. The stronger-profile group is denoted as *Group A*, while the weaker-profile group is denoted as *Group B*.

- **Language background:** German vs. non-German,
- **Self-estimated German proficiency:** high (values 1–2) vs. low (values 3–4),
- **German grade:** higher achievement (grades 1–3) vs. lower achievement (grades 4–6).

Each comparison includes only sentences that received at least one rating from a student in group A and at least one rating from a student in group B. As a result, both groups are evaluated on the same set of sentences within a given feature-based comparison. Sentence coverage may nevertheless vary across features, meaning that the resulting subsets can differ in both composition and size (see Appendix A.2). Each subset is analyzed independently, and the corresponding LLM predictions are obtained separately for both groups within the same set of sentences.

For language background, we exclude the intermediate group of students who report speaking both German and another language at home. This allows for a clearer contrast between students with German as the only home language and students with another home language, since bilingual home-language profiles cannot be unambiguously assigned to either group. For self-estimation, the original survey used a scale from 1 to 5, but the available annotations only cover values from 1 to 4. Since no students rated their German proficiency as 'very low' (5), we split the observed range at the midpoint, defining values 1–2 as *Group A* and values 3–4 as *Group B*.

Prompt design. All prompts follow a consistent structure in which the model is instructed to assume the role of a student and to rate the difficulty of a given sentence on a 5-point scale from 1 (very easy) to 5 (very difficult). The model is asked to assess how difficult each sentence is for the student personally. Each prompt contains 20 sentences to resemble the actual survey that the students faced during the annotation process.

To simulate reader-dependent variation, we introduce short profile descriptions that specify the three selected individual characteristics: *language background*, *German grade*, or *self-estimation of German proficiency*. Each profile is expressed as a single sentence (e.g., 'You grew up speaking German at home.'), which is inserted into the prompt.

Importantly, the prompt structure remains identical across all conditions, and only the profile sentence varies (see Appendix B). This ensures that any differences in model predictions can be attributed to the provided reader characteristics. In the baseline condition, no additional profile information is provided beyond the general instruction that the model should act as a student.

4.3 Linguistic Feature Analysis

In addition to reader-specific variables, we analyze how linguistic properties of the texts relate to perceived difficulty in both human annotations and LLM predictions.

Feature extraction. For each text, we extract a set of surface-level and lexical features using the spaCy library (Honnibal et al., 2020) with a German language model. The selected features capture different aspects of linguistic complexity:

- **Sentence length:** number of words per sentence, excluding punctuation and numerical expressions,
- **Average word length:** average number of characters per word,
- **Proportion of rare words:** share of words not contained in the German spaCy model[6] vocabulary
- **Dependency distance:** average linear distance between each word and its syntactic head in the dependency tree,
- **Dependency depth:** maximum depth of the syntactic dependency tree, reflecting hierarchical complexity,
- **Parenthesis count:** number of parenthetical elements (e.g., brackets or inserted expressions),

[6] https://spacy.io/models/de#de_core_news_md

- **Comma count:** number of commas, used as a proxy for clause boundaries and local syntactic segmentation,
- **Enumeration marker count:** number of coordination markers (e.g., commas, "und", "oder", "sowie") indicating list-like structures,
- **Abbreviation count:** number of abbreviated forms (e.g., "z.B.", uppercase abbreviations).

Together, these features provide a compact representation of lexical and syntactic characteristics commonly associated with text difficulty and allow us to compare which textual cues are reflected in human judgments and LLM predictions.

Correlation analysis. To assess how these features relate to perceived difficulty, we compute Spearman rank correlations (Spearman, 1904) between each linguistic feature and both human difficulty ratings and LLM predictions. All correlations are computed at the sentence level and are calculated separately for each reader group (Group A and Group B) within each subset (language background, German grade, and self-estimation of German proficiency).

5 Results

We evaluate the models from three complementary perspectives. First, we assess baseline performance in a setting without explicit reader-specific information. Second, we examine how predictions change when reader profiles are provided and whether this improves agreement with human annotations. Third, we analyze the relationship between linguistic features and predicted difficulty to better understand which textual cues the models rely on.

Together, these analyses allow us to distinguish between overall predictive performance, sensitivity to reader profiles, and the linguistic patterns underlying model predictions.

5.1 Baseline Setting

We first evaluate model performance in a baseline setting without providing any reader-specific information. In this condition, models are instructed to assess the difficulty of each sentence from the perspective of a generic secondary school student. This setup serves as a reference point for analyzing how models behave in the absence of explicit reader profiles. For evaluation, we compare each model's single prediction for a sentence with the mean of all available human ratings for that sentence.

Table 2 summarizes the results across all evaluated models. Overall, performance is comparable, with RMSE values ranging from 1.35 (`GPT-5.4 nano`) to 1.61 (`Mistral Large 3`), and MAE between 1.06 and 1.30. Correlation scores remain relatively low for all models (Spearman $\rho \approx 0.29$–0.32), indicating limited agreement with human judgments both in terms of absolute predictions and relative ranking of text difficulty.

Despite these differences, all models exhibit a consistent pattern: predictions are systematically shifted towards higher difficulty values. While the mean human rating is 2.00, the predicted means range from 2.72 to 3.02. This suggests that models tend to assign systematically higher difficulty scores even when no reader-specific information is provided.

5.2 Effect of Reader Profiles

We next examine whether incorporating reader-specific information improves the alignment between LLM predictions and human annotations. To this end, we condition the models on reader profiles reflecting the features *language background*, *German grade*, and *self-estimation of German proficiency*, and evaluate the resulting predictions against the corresponding human ratings.

Table 3 summarizes the predictive performance across all profile-based subsets. Overall, incorporating reader-specific information leads only to limited improvements compared to the baseline setting. Across all models and subsets, Spearman correlations slightly decrease compared to the baseline setting ($\rho = 0.21$–0.27), indicating that the models only partially recover the ranking of perceived sentence difficulty observed in the human annotations.

The results further suggest that profile conditioning does not consistently improve predictive performance. While some models achieve slightly lower RMSE and MAE values for individual subsets, these gains remain small and inconsistent across reader characteristics. In several cases, performance even deteriorates compared to the baseline condition, despite the additional profile information.

Across models, the *language background* and *self-estimation* subsets generally yield slightly stronger correlations than the German grade subset. However, the differences between subsets remain

Model	RMSE	MAE	Spearman ρ	Mean (LLM)
GPT-5.4 nano	1.35	1.06	0.32	2.72
GPT-5.4 mini	1.52	1.18	0.30	2.76
DeepSeek-V3.1	1.59	1.27	0.29	2.99
Mistral Large 3	1.61	1.30	0.30	3.02

Table 2: Performance of LLM predictions in the baseline condition without reader-specific information ($n = 1697$ unique sentences).

Model	Lang.			Grade			Self-est.		
	RMSE	MAE	Sp. ρ	RMSE	MAE	Sp. ρ	RMSE	MAE	Sp. ρ
GPT-5.4 nano	1.50	1.20	0.24	1.52	1.22	0.26	1.48	1.18	0.23
GPT-5.4 mini	1.72	1.37	0.26	1.73	1.38	0.21	1.63	1.27	0.27
DeepSeek-V3.1	1.49	1.16	0.24	1.63	1.29	0.22	1.45	1.15	0.26
Mistral Large 3	1.51	1.19	0.27	1.67	1.35	0.22	1.55	1.25	0.27

Table 3: RMSE, MAE and Spearman correlation of LLM predictions with human labels across profile-based subsets.

modest overall, suggesting that current LLMs struggle to make effective use of reader-specific information when predicting perceived text complexity.

At the model level, no single system consistently outperforms the others across all subsets. GPT-5.4 nano and DeepSeek V3.1 achieve the lowest RMSE values overall, while GPT-5.4 mini and Mistral Large 3 obtain slightly higher rank correlations in some conditions. Nevertheless, the differences between models remain relatively small, indicating broadly similar behavior across architectures and model families.

These results indicate that reader-profile information changes model predictions, but does not consistently improve their alignment with human judgments. While the models are sensitive to the provided profiles, their overall agreement with human annotations remains limited.

5.3 Linguistic Feature Analysis

To better understand the factors driving model predictions, we analyze how linguistic features correlate with perceived text difficulty for both human annotations and LLM predictions.

Across all evaluated models, we observe highly similar correlations between linguistic features and predicted difficulty scores (see Appendix C.1, Table 6). In particular, all models rely strongly on surface-level and structural cues such as sentence length and syntactic complexity, while differences between reader groups remain comparatively small. Since the observed patterns are highly consistent across models, we focus the following analysis on GPT-5.4 mini as a representative example.

Figure 2 shows the difference in Spearman rank correlations between GPT-5.4 mini and human annotations across a range of surface-level and structural linguistic features.

Overall, the model predictions show stronger correlations with all examined linguistic features than the human ratings do. The strongest deviations occur for sentence length, with consistently positive differences across all conditions and a maximum of +0.74 for Group B with the weaker language background. Enumeration markers also exhibit consistently positive deviations, suggesting that the model treats explicit enumeration and coordination cues as stronger indicators of difficulty than human annotators do.

Beyond surface-level cues, the model also assigns substantially greater importance to syntactic structure. Both dependency distance and dependency depth show consistently strong positive differences across all subsets, indicating that the model associates more complex syntactic structures with higher difficulty to a greater extent than humans do. These effects tend to be more pronounced for the lower-proficiency groups (Group B), although the pattern varies across individual linguistic features.

In contrast, features related to lexical complexity beyond surface form exhibit only minor or inconsistent effects. The proportion of rare words shows relatively small deviations, suggesting that the model does not strongly rely on less frequent or domain-specific vocabulary when estimating difficulty.

To further investigate whether these patterns de-

Figure 2: Difference in Spearman rank correlations between `GPT-5.4 mini` predictions and human annotations across linguistic features. Positive values indicate stronger feature-difficulty associations in the model than in the human data.

pend on the reader profile, we analyze the feature–difficulty correlations within the LLM predictions themselves (see Appendix C.2, Figure 3). Interestingly, we observe only minimal variation across reader groups: correlations between linguistic features and predicted difficulty remain largely stable for both stronger (Group A) and weaker (Group B) profiles.

This pattern suggests that profile conditioning primarily shifts the overall difficulty level rather than substantially changing the observed feature–difficulty relationships.

This differs markedly from the human annotations (see Appendix C.2 Figure 4), where feature–difficulty correlations are generally weaker and vary more strongly across reader groups. In particular, some human subgroups show only weak or near-zero associations with the selected linguistic features, suggesting that perceived difficulty is less uniformly tied to these textual cues. By contrast, the LLM predictions exhibit stable and consistently stronger feature correlations across profiles.

Our results suggest that `GPT-5.4 mini` does not capture reader-specific differences in a nuanced way. Rather than adapting its interpretation of linguistic complexity, the model relies on stable structural and surface-level cues and adjusts predictions primarily through global shifts in difficulty. Importantly, this pattern is not specific to `GPT-5.4 mini`: the other evaluated models show highly similar feature–difficulty correlation patterns across reader groups (see Appendix Table 6). This suggests that the observed reliance on stable linguistic heuristics is a broader tendency across the evaluated LLMs rather than an isolated behavior of a single model.

6 Conclusion

In this work, we investigated whether large language models capture reader-dependent variation in perceived text complexity. Using a German sentence-level dataset with human difficulty ratings and reader-specific metadata, we compared multiple LLMs under generic and profile-conditioned prompting settings. Beyond predictive performance, we analyzed how linguistic features relate to both human annotations and model predictions across reader groups.

Our results show that current LLMs only partially align with human judgments of perceived text difficulty. In the baseline setting, all evaluated mod-

els show modest correlations with human annotations and systematically overestimate text difficulty. Providing reader profiles changes model predictions, but does not consistently improve agreement with human ratings.

The linguistic feature analysis suggests that this limited improvement is not merely a performance issue, but reflects how models use reader-specific information. Across models, predicted difficulty is strongly associated with stable surface-level and structural cues such as sentence length and syntactic complexity. These feature–difficulty relationships remain largely consistent across reader profiles, indicating that models do not substantially adapt their interpretation of linguistic complexity to different reader characteristics.

Overall, our findings suggest that prompt-based reader profiles primarily lead to global shifts in predicted difficulty rather than to nuanced reader-dependent feature–difficulty associations. This highlights an important limitation of current LLMs for personalized readability assessment and educational applications: while models are sensitive to profile information, this sensitivity does not necessarily correspond to human-like reader adaptation.

Future work should investigate richer forms of reader modeling, including more detailed cognitive and educational profiles as well as interaction-based approaches. In addition, extending the analysis beyond sentence-level judgments to discourse-level comprehension could provide a more complete picture of how LLMs represent reader-dependent text difficulty.

Limitations

The findings of this study should be interpreted in light of the following limitations. First, our analysis focuses on German sentences rated by secondary school students and predominantly covers informational text sources. Language-specific properties, educational contexts, and reader populations may influence both human judgments and model predictions. Future research across additional languages, text genres, and reader populations could help determine the broader generalizability of the observed patterns.

Second, readers are represented through concise, single-attribute profile descriptions. This design enables controlled comparisons and allows us to isolate the influence of individual reader characteristics. Richer representations that combine multiple attributes, proficiency information, or interaction histories may reveal additional forms of reader adaptation. Examining such representations and alternative prompt formulations would therefore be a useful direction for future work.

Acknowledgements

The authors used ChatGPT for language editing, including improvements to grammar, wording, and writing style. The authors reviewed and approved all resulting changes.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2012. Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, pages 11–19. Linköping University Electronic Press.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, and 4 others. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel. Technical report, Defence Technical Information Center (DTIC) Document.

William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, and 180 others. 2024. Deepseek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language. *arXiv preprint arXiv:1904.07733*.

Numaan Naeem, Abdellah El Mekki, and Muhammad Abdul-Mageed. 2025. EduAdapt: A question answer benchmark dataset for evaluating grade-level adaptability in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34236–34263, Suzhou, China. Association for Computational Linguistics.

Jio Oh, Steven Euijong Whang, James Evans, and Jindong Wang. 2026. Classroom AI: large language models as grade-specific teachers. *npj Artificial Intelligence*, 2(1):28.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii. Association for Computational Linguistics.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do LLMs adapt to different age and education levels? *arXiv preprint arXiv:2312.02065*.

Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective text complexity assessment for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 707–714, Marseille, France. European Language Resources Association.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Boris Thome, Friederike Hertweck, and Stefan Conrad. 2025. Predicting perceived text complexity: The role of person-related features in profile-based models. *Journal of Educational Data Mining*, 17(1):276–307.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2025. Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6):51–63.

Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 141–153, Seattle, Washington. Association for Computational Linguistics.

A Profile-Based Feature Subsets

A.1 Subset Inter-Annotator Agreement

To examine whether ratings are more consistent within reader groups, we calculate ordinal Krippendorff's α for each pooled profile subset and separately for Groups A and B. Table 4 reports the number of annotators (n_{ann}), the number of sentences with at least two ratings within the respective group (n_{sent}), and the corresponding agreement value. Overall refers to the ratings of Groups A and B pooled within the respective profile subset.

Profile	Group	n_{ann}	n_{sent}	α_{ord}
Lang.	Overall	107	185	0.18
	Group A	79	90	0.21
	Group B	28	52	0.25
Grade	Overall	171	288	0.16
	Group A	132	152	0.17
	Group B	39	101	0.14
Self-est.	Overall	163	229	0.16
	Group A	142	146	0.18
	Group B	21	31	0.35

Table 4: Ordinal Krippendorff's α across profile subsets and reader groups.

Group-specific agreement exceeds that of the corresponding pooled subset in five of six groups. These results are descriptive, as the estimates are

based on partly different sets of multiply annotated sentences and Group B generally includes fewer annotators. Nevertheless, the observed pattern provides some support for examining profile-conditioned judgments separately.

A.2 Composition of the Profile-Based Subsets

The resulting subset sizes are summarized in Table 5. Differences across subsets arise because a sentence is retained only if it was rated by at least one student from each of the two respective groups.

Ratings	Language	Grade	Self-est.
Group A	590	1026	1131
Group B	307	528	266
Total	897	1554	1397

Table 5: Number of ratings per feature subset and per group.

B Model Prompts

Prompt: You are a secondary school student. *[PROFILE SENTENCE]* Your task is to rate how difficult the following texts are for you to understand. Rate each text on a scale from 1 (very easy) to 5 (very difficult).

Text 1:
Es sind zwei Fächer zu wählen.

Text 2:
Der allgemeine Stundenweltrekord für ausschließlich mit eigener Muskelkraft betriebene Fahrzeuge liegt deutlich über dem Rekord mit einem klassischen Rennrad.

...

Return only valid JSON in exactly this format, with one entry per text in order and no additional text:

```
[
{{"index": 1, "difficulty": <rating 1-5>}},
{{"index": 2, "difficulty": <rating 1-5>}},
...
]
```

Profile sentences. The profile sentence is inserted into the prompt depending on the respective profile dimension and reader group:

- **Language background**
  - *Group A:* You grew up speaking German at home.
  - *Group B:* You grew up speaking a language other than German at home.
- **German grade**
  - *Group A:* You usually receive good grades in German.
  - *Group B:* You usually receive lower grades in German.
- **Self-estimation of German proficiency**
  - *Group A:* You consider your German language abilities to be strong.
  - *Group B:* You consider your German language abilities to be more limited.

C Linguistic Feature Correlations

C.1 Overview across models

Table 6 reports the Spearman correlations between linguistic features and predicted difficulty scores for all evaluated LLMs, separately for each profile dimension and reader group. The table complements the main analysis by showing that the feature–difficulty patterns observed for `GPT-5.4 mini` are not specific to a single model. Across models, sentence length, dependency-based features, and enumeration markers show consistently strong positive associations with predicted difficulty, while correlations with rare words and abbreviations are weaker. This supports the observation that the evaluated LLMs rely on similar linguistic cues and that these correlations remain largely stable across reader groups.

C.2 Correlation Heatmaps for Model and Human Ratings

Figures 3 and 4 show the Spearman correlations between linguistic features and difficulty ratings for `GPT-5.4 mini` and the human annotations. They complement the main analysis by showing the absolute correlation patterns rather than the model-human differences.

For `GPT-5.4 mini`, correlations are consistently positive across reader groups, especially for sentence length, dependency-based features, and enumeration markers. In contrast, the human annotations show weaker and more variable correlations.

Model	Group	Sent. len.	Avg. word len.	Rare word %	Dep. dist.	Dep. depth	Paren. count	Comma count	Enum. markers	Abbrev. count
GPT-5.4 nano	Lang A	0.69	0.42	0.16	0.42	0.57	0.39	0.42	0.57	0.25
	Lang B	0.67	0.38	0.16	0.45	0.54	0.36	0.41	0.50	0.24
	Grade A	0.66	0.31	0.20	0.49	0.59	0.30	0.44	0.52	0.15
	Grade B	0.70	0.32	0.14	0.52	0.62	0.33	0.44	0.52	0.12
	Self A	0.62	0.16	0.18	0.45	0.50	0.19	0.42	0.48	0.13
	Self B	0.65	0.19	0.16	0.44	0.54	0.19	0.41	0.51	0.07
GPT-5.4 mini	Lang A	0.76	0.36	0.24	0.53	0.61	0.47	0.50	0.61	0.24
	Lang B	0.75	0.38	0.19	0.52	0.61	0.38	0.43	0.56	0.26
	Grade A	0.68	0.37	0.25	0.52	0.61	0.33	0.49	0.54	0.17
	Grade B	0.72	0.39	0.23	0.58	0.62	0.37	0.47	0.52	0.16
	Self A	0.73	0.31	0.27	0.54	0.62	0.28	0.50	0.59	0.14
	Self B	0.73	0.31	0.26	0.52	0.63	0.30	0.50	0.57	0.11
DeepSeek-V3.1	Lang A	0.65	0.40	0.20	0.42	0.53	0.32	0.37	0.53	0.22
	Lang B	0.72	0.42	0.24	0.47	0.58	0.39	0.45	0.58	0.24
	Grade A	0.62	0.42	0.26	0.45	0.56	0.33	0.41	0.51	0.12
	Grade B	0.67	0.42	0.23	0.54	0.59	0.34	0.42	0.52	0.13
	Self A	0.66	0.33	0.30	0.47	0.52	0.25	0.46	0.52	0.14
	Self B	0.68	0.31	0.30	0.50	0.56	0.27	0.45	0.53	0.15
Mistral Large 3	Lang A	0.61	0.42	0.19	0.37	0.50	0.39	0.34	0.48	0.23
	Lang B	0.69	0.33	0.20	0.48	0.56	0.41	0.42	0.54	0.28
	Grade A	0.57	0.37	0.19	0.40	0.50	0.33	0.36	0.43	0.10
	Grade B	0.60	0.35	0.20	0.44	0.52	0.36	0.37	0.43	0.12
	Self A	0.53	0.27	0.22	0.41	0.45	0.26	0.42	0.46	0.12
	Self B	0.58	0.28	0.26	0.47	0.47	0.29	0.43	0.49	0.13

Table 6: Spearman correlations between linguistic features and LLM complexity ratings per model, dataset, and group.

Figure 3: Spearman rank correlations between linguistic features and predicted difficulty scores for GPT-5.4 mini across stronger and weaker reader-profile groups.

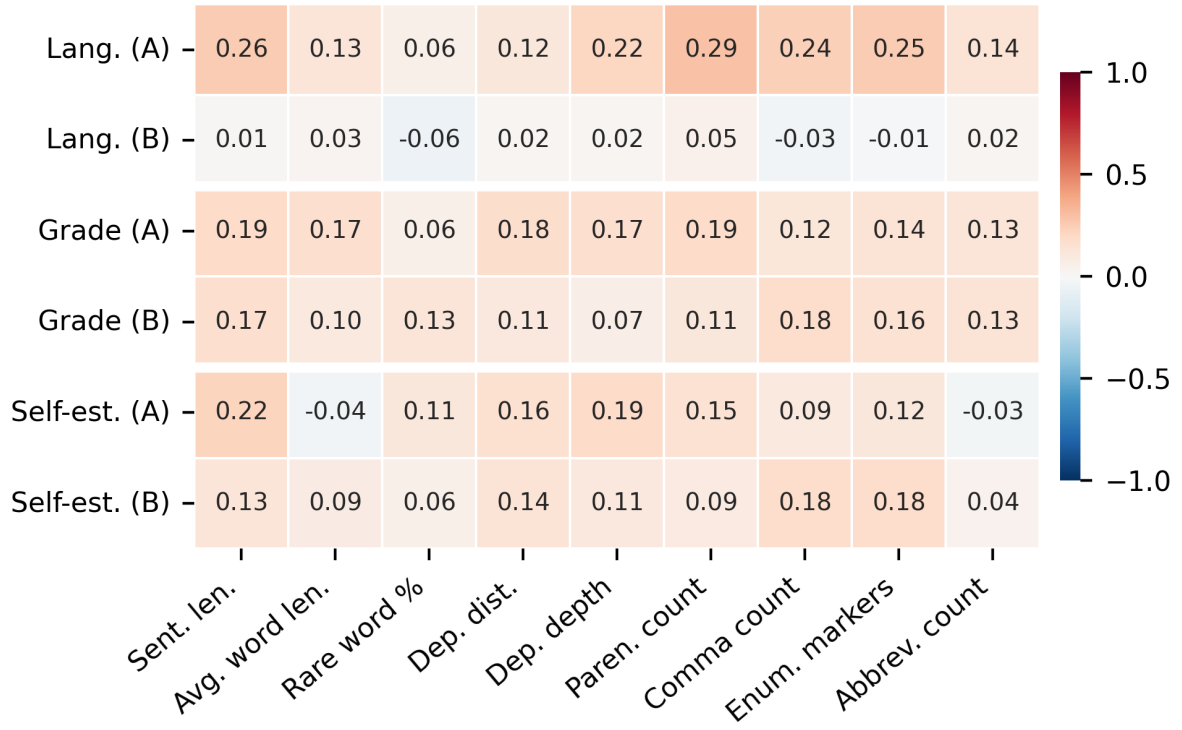

Figure 4: Spearman rank correlations between linguistic features and student-annotated difficulty scores across stronger and weaker reader-profile groups.