Beyond Flesch-Kincaid: Prompt-based Metrics Improve Difficulty Classification of Educational Texts

Donya Rooein[1], Paul Röttger[1], Anastassia Shaitarova[2], Dirk Hovy[1]

[1]Bocconi University, [2]University of Zurich
{donya.rooein, paul.rottger, dirk.hovy}@unibocconi.it,
anastassia.shaitarova@uzh.ch

Abstract

Using large language models (LLMs) for educational applications like dialogue-based teaching is a hot topic. Effective teaching, however, requires teachers to adapt the difficulty of content and explanations to the education level of their students. Even the best LLMs today struggle to do this well. If we want to improve LLMs on this adaptation task, we need to be able to measure adaptation success reliably. However, current STATIC metrics for text difficulty, like the Flesch-Kincaid Reading Ease score, are known to be crude and brittle. We, therefore, introduce and evaluate a new set of PROMPT-BASED metrics for text difficulty. Based on a user study, we create PROMPT-BASED metrics as inputs for LLMs. They leverage LLM's general language understanding capabilities to capture more abstract and complex features than STATIC metrics. Regression experiments show that adding our PROMPT-BASED metrics significantly improves text difficulty classification over STATIC metrics alone. Our results demonstrate the promise of using LLMs to evaluate text adaptation to different education levels.
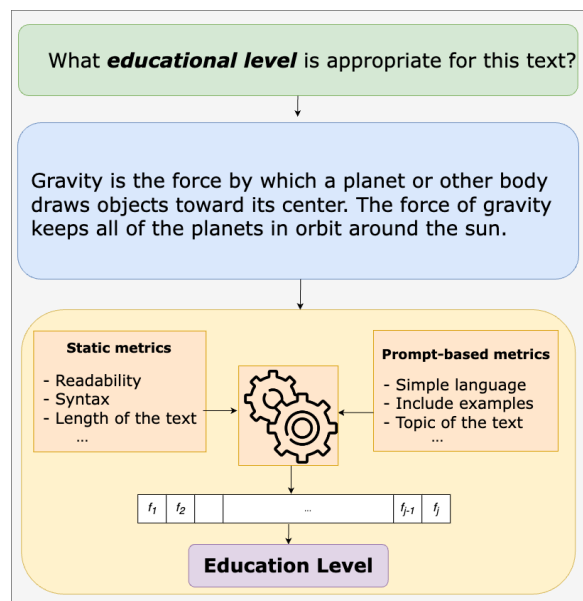
Figure 1: Schematic overview of our approach to text difficulty classification. We calculate relevant STATIC and PROMPT-BASED metrics for a given input text. Either or both metrics are then fed into a regression classifier that makes a final classification.

1 Introduction

Large language models (LLMs) today can answer wide-ranging questions and explain complex concepts with high accuracy (Chung et al., 2022; OpenAI, 2023). This development has motivated explorations into their uses for education, ranging from automated student assessment and personalised content to dialogue-based teaching (Upadhyay et al., 2023; Sallam, 2023; Yan et al., 2023; Hosseini et al., 2023).

Effective teaching requires that the difficulty of content and explanations is tailored to the education level of the students. Human teachers are trained to do this, and adjust their material and style without much prompting. However, this adaptation is not just the adjustment of one variable. It is a complex undertaking, touching upon lexicon, syntax,

pragmatics, and semantics. Improving the ability of LLMs to adapt their outputs to different levels of education is therefore crucial to unlocking their usefulness for education. One of the most basic requirements to achieve this goal is a way to measure adaptation success.

Measuring whether a given output is appropriate for a given level of education, however, is a very difficult task. Existing STATIC metrics, like the Flesch-Kincaid Reading Ease score (Flesch, 1948), are based on simple formulas, heuristics, and word counts. They share the brittleness of all heuristic approaches and are known to be noisy measures of text difficulty at best. Also, these metrics were developed for longer-form explanations, like those found in textbooks, rather than dialogue-style teaching. Due to their reliance on counts, their estimates

are unreliable in shorter formats. We need better metrics to make improvements on the adaptability of LLMs to education levels measurable. Only when we can measure improvements can we make tangible progress in leveraging LLMs for educational applications.[1]

As an alternative to STATIC metrics, we can use classifiers to predict the educational level of a given text. They generalize better and can be applied to texts of varying lengths. However, these classifiers are expensive to train and require more training data than we usually have for a niche domain like educational purposes. Similarly, human assessment of difficulty may provide a gold standard, but it is expensive to collect and, like all annotation tasks, suffers from disagreement.

In this paper, we introduce and evaluate a new set of PROMPT-BASED metrics for text difficulty as complements to existing STATIC metrics. PROMPT-BASED metrics are LLM prompts that exploit the general language understanding capabilities of LLMs to capture more abstract features of educational texts than STATIC metrics. For example, LLMs can flexibly classify the topic of a text, which is one adaptation technique used by teachers to adjust the content which called curriculum compacting in pedagogy (Stamps, 2004). This would be difficult to do with STATIC approaches.

We develop our selection of PROMPT-BASED metrics based on a user study, where we ask a group of university students to 1) assess the difficulty of educational texts and explain their reasoning, and 2) come up with prompts for an LLM to change the difficulty of a given text. We then translate the qualitative findings from both parts of the study into concrete LLM prompts that serve as PROMPT-BASED metrics. We incorporate prompts from other studies to manage text readability with LLMs (Imperial and Madabushi, 2023; Gobara et al., 2024). We evaluate the ability of our new PROMPT-BASED metrics to measure text appropriateness for different education levels with a series of regression experiments.

While PROMPT-BASED metrics perform on par or better than zero-shot and few-shot LLM classifiers, they are less useful for text difficulty classification by themselves than STATIC metrics. However, combining PROMPT-BASED and STATIC metrics significantly improves performance. This suggests that PROMPT-BASED metrics capture relevant signals beyond those captured by the large number of STATIC metrics.

A combination of STATIC and PROMPT-BASED metrics also provides a deeper understanding of the key metrics or features that influence complexity than classifiers could. Additionally, the factors that contribute to complexity in a scientific text differ from those in a medical or a legal document. By considering a range of metrics, we can develop more accurate domain-specific measures. Our multifaceted approach allows us to break down complexity into its basic components, such as its appropriateness for different education levels, lexical or syntactic complexity, thematic topics, and text readability.

Overall, PROMPT-BASED metrics empower educators to develop more effective content development strategies with LLMs to engage learners of all levels and backgrounds. We could have directly trained classifiers; however, this approach would not have enabled us to identify the most relevant metrics.

Contributions

1. We conduct a user study to motivate the creation of novel PROMPT-BASED metrics of text difficulty for educational texts (§2).

2. We show in a series of regression experiments that these PROMPT-BASED metrics hold additional value for text difficulty classification beyond what STATIC metrics can capture (§4.3).

3. By leveraging the interpretability of our regressions, we highlight the relative importance of individual STATIC and PROMPT-BASED metrics (§4.5).

2 User Study

Our PROMPT-BASED metrics for text difficulty are prompts based on the results of a one-day user study we ran with a group of university students in November 2023.

2.1 Study Design

The user study consisted of two main parts.

In the first part of our study, we asked participants to review 60 educational texts randomly sampled from the ScienceQA dataset (Lu et al., 2022). Each text consists of a question (e.g., "What is

---

[1]Similarly, metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BLANC (Recasens and Hovy, 2011), among others, kickstarted and sustained the development of automated approaches to machine translation, summarization, and coreference resolution, respectively.

the mass of a dinner fork?") with answer choices ("70 grams or 70 kilograms") and a longer-form explanation of the solution. All texts we select here are authentic educational materials from the social, natural, or language sciences in schools. Participants were tasked with a) labeling the education level of each text as appropriate for either elementary school, middle school, or high school and b) explaining the reasoning behind their choice in a short, free-text answer.

In the second part of our study, we asked participants to rewrite scientific text explanations, also sampled from ScienceQA, to be appropriate for different education levels, with the help of an LLM – in this case, ChatGPT. For example, participants were asked to rewrite a middle school explanation of thermal energy at the elementary and high school levels with the help of prompts. We recorded the prompts they used to get ChatGPT to accomplish the adaptation for them. Thus, we collected prompts that are used both for text *simplification* and for text *complexification*.

2.2 Study Participants

We ran our study as part of a hackathon at the University of Zurich. There were seven participants aged between 21 and 31 years. Four participants were female, three male. All participants were students at Department of Computational Linguistics from University of Zurich, enrolled at the time in programs specializing in computational linguistics, computer science, and AI. Five were studying for a bachelor's degree and two for a master's degree. The participants held prior educational degrees from school systems across five different countries. Their native languages include English, Italian, German, Greek, and Ukrainian. They self-reported their English language proficiency at C1 and C2 levels. Participants were compensated in study credits that could be counted towards completing their program.

2.3 Study Results

The first task of our study yielded 276 classification labels together with their corresponding descriptive justifications. These include 120 label-explanation pairs for middle school texts, 89 for high school, and 67 for elementary school texts. In the second task of our study, we collected 103 prompts for text simplification and complexification. We share illustrative examples of classifications, explanations, and prompts in Appendix A.

In the next section, we use the qualitative results from our study to motivate the construction of novel PROMPT-BASED metrics for text appropriateness for various education levels.

3 Metrics for Text Difficulty

3.1 Prompt-based Metrics

Since the metrics we introduce are based on the prompts of language models rather than discrete heuristics, we refer to them as 'PROMPT-BASED' to distinguish them. The goal of the PROMPT-BASED metrics we develop is to capture more abstract features of educational texts than would be possible with STATIC metrics, which typically focus on individual words and their statistics.
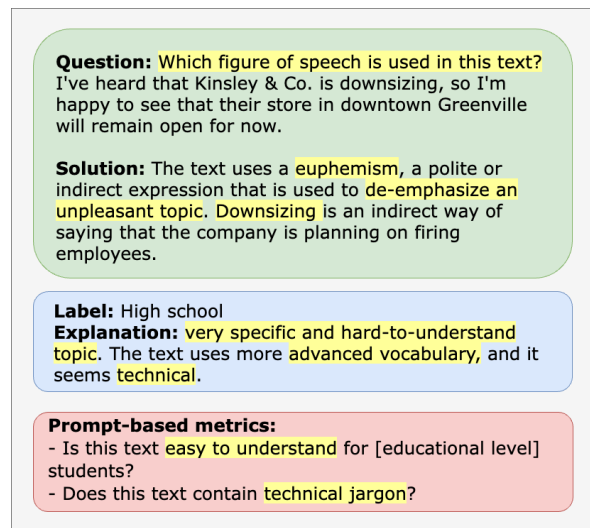

Figure 2: An illustrative example of the PROMPT-BASED metric process. The green box contains the education text from the ScienceQA dataset. The blue box shows the predicted educational level and the explanation. The red box contains the PROMPT-BASED metrics based on the sample.

We derive our PROMPT-BASED metrics from the results of our user study. Figure 2 shows an illustrative example of our derivation process. We
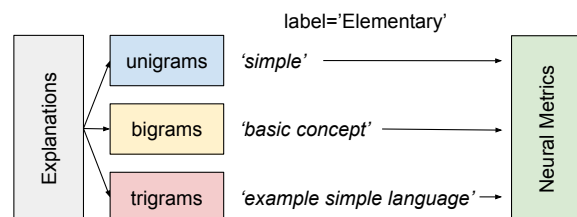

Figure 3: High-level view of the derivation process for the PROMPT-BASED metrics using n-gram frequencies. Function words are excluded.

consider users' explanations for why they consider a specific educational text to be of elementary, middle, or high school level difficulty. Then, we identify recurring attributes and other explanation features that several users mention to reflect them in PROMPT-BASED metrics. We examine the distributions of unigrams, bigrams, and trigrams across all three labels, excluding function words (see Figure 3). Some of the most frequent unigrams for the elementary level include *simple, basic, elementary*; for the high school level, *high, complex, concepts*; and for the middle school level, *explicit, explanation, middle*.

We qualitatively assessed the n-gram distributions, considering both frequencies and topic appropriateness, before finalizing the query construction. Each PROMPT-BASED metric is a simple yes-no question, which we use to prompt the LLMs. These metrics encompass the most frequent unigrams and less common bigrams and trigrams derived from the findings of our study.

While, Gobara et al. (2024) demonstrate a correlation between readability scores of LLM-generated texts in education and human assessments, Imperial and Madabushi (2023) indicate challenges in LLMs effectively adjusting the readability of text. We construct 63 PROMPT-BASED metrics using this process. Each PROMPT-BASED metric relates to either education level (30 metrics), lexical or syntactic complexity (8 metrics), and the topic of the text at hand (10 metrics). In addition, we include metrics about the text's readability score (15 metrics) based on the work by Imperial and Madabushi (2023). The complete list of all our PROMPT-BASED metrics is in Appendix C.

3.2 Existing Static Metrics

STATIC metrics are the baseline we want to improve on. All STATIC metrics are based on simple formulas, heuristics, or counts of words and other textual features. These properties make them simple to apply but limit the conceptual complexity of what they can reasonably measure. In total, we include 46 STATIC metrics, selected from those compiled in prior work (Flekova et al., 2016; Yaneva et al., 2019; Xue et al., 2020; Baldwin et al., 2021).

These metrics encompass a variety of linguistic characteristics, spanning from basic text-level measures like vocabulary size and word frequency to sentence-level attributes such as sentence length and syntactic complexity. Additionally, they take into account the question-answering structure within the input text. In the ScienceQA dataset, each question is paired with its respective solution and corresponding lecture. This segmentation of information across educational levels facilitates the computation of STATIC features for each section of the question-answer solution and lecture independently. For the complete list of 46 STATIC metrics, see Appendix C.

4 Experiments

We conduct a series of classification experiments to evaluate the usefulness of our novel PROMPT-BASED metrics for measuring text difficulty. We use a subset of the ScienceQA dataset, which contains question-answer pairs across several topics and education levels. Specifically, we run multinomial logistic regressions based on STATIC metrics, PROMPT-BASED metrics, and the combination of the two to evaluate the marginal benefits of our new PROMPT-BASED metrics. We also compare these regression approaches to using an LLM for zero-shot and few-shot classification.

4.1 Dataset

All our experiments are based on the ScienceQA dataset (Lu et al., 2022). There are 21,208 texts in ScienceQA. Each text consists of a question with answer choices, and a longer-form explanation of the solution. Texts in ScienceQA are classified according to their grade level using the K12 system from the US education system. We simplify this classification by collapsing the 12-grade levels into just three: elementary school (grades 1 to 5), middle school (grades 6 to 8), and high school (grades 9 to 12).[2] From the 21,208 texts in ScienceQA, we sample only those that do not use images in questions or explanations. We then deduplicate and sample 1,516 texts for each education level to create a balanced dataset of 4,548 texts. Of these 4,548 texts, we use 3,638 (80%) for training and 910 (20%) for evaluation. To our knowledge, ours is the first use of the ScienceQA dataset for training and evaluating text difficulty classifiers.

4.2 LLMs for Prompt-based Metrics

We use LLMs to compute the 63 PROMPT-BASED metrics described in Section 3.1. In principle, any LLM can serve this purpose. With 63 metrics for 4,548 texts, we get 286,524 prompts from each LLM. This amount is prohibitively expensive for

---

paid services like GPT4. Hence, we concentrate on state-of-the-art open LLMs, which we can execute at a low cost: Llama2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), and Gemma (Google, 2024). Llama2, launched in July 2023, comprises both pre-trained and fine-tuned LLMs, ranging in size from 7 billion to 70 billion parameters. It has been reported to outperform other open-access LLMs and exhibits capabilities comparable to Chat-GPT across various tasks. In this paper, we use Llama2-7b and Llama2-13b. The next model is Mistral-7B, released in September 2023, another open LLM surpassing similar-sized open LLMs. We use Mistral-7b-Instruct-v0.2, which was published in December 2023.

The last model we use is Gemma7b-it, based on the Gemma base model and trained on open-source mathematics datasets.

We set the model temperature to zero to make responses deterministic. The maximum response length is 256 tokens. Otherwise, we use standard generation parameters from the Hugging Face transformers library. We collected all responses in February 2024.

4.3 Multinomial Logistic Regression

We use simple multinomial logistic regression to classify the difficulty level of texts. The task is to predict the difficulty level C_i of a given educational text S_i. C_i can take three ordinal values: elementary, middle, or high school difficulty. Instead of including S_i directly, we include sets of STATIC and PROMPT-BASED metrics $\mathbf{M}_i$ that are computed based on S_i. We regress $\mathbf{M}_i$ on C_i on the 3,638 training texts and then evaluate on the 910 test education texts.

We vary which metrics we include across experimental setups to evaluate the marginal benefits of different metrics. There are three main setups of interest: 1) PROMPT-BASED metrics only, 2) STATIC metrics only, 3) the combination of the two, which we refer to as COMBO.

4.4 Baseline: Zero- and Few-Shot Classification

We exploit the general language capabilities of LLMs to compute PROMPT-BASED metrics, which we then use as inputs to a logistic classifier for text difficulty. A natural follow-up question is whether LLMs could directly predict text difficulty related to education levels. Therefore, we incorporate a baseline for zero-shot and few-shot text classifica-

tion. We test zero-shot and few-shot classification with the same LLMs that we use for calculating our PROMPT-BASED metrics. As an additional comparison point, we test GPT-4 Turbo.

Note that while the logistic classifier is fitted to our training data, the zero-shot LLM has not seen any examples at inference time. In the few-shot setting, we provide two examples for each education level and prompt the model to assign one of the desired labels without explanations.

To investigate the effect of different prompting styles, we test five distinct prompt templates in our zero-shot setup, each consisting of 25-30 words. Additionally, each prompt contains a textual segment describing the text of the science question answering for educational-level classification. We compare performance across the five prompt templates to determine the most effective prompt, i.e., the strongest baseline for our experiments. We evaluate the models' responses on a subset of randomly selected samples (n=100). The lowest performance stands at 29%, while the highest achievement reaches 42%. We proceed with our experiments under zero-shot and few-shot setups, using the best performance style as our baselines. The selected prompt for zero-shot experiments is: "Your task is to predict the education level corresponding to a given text. You are provided with three labels to choose from: 1) elementary school 2) middle school 3) high school. Text: [text] Educational level: "

We instructed LLMs to return one of the education levels. Due to the difficulty of LLMs in directly predicting the levels and complexity of the text, we have responses without the desired educational level. In this case, we assigned a default level to this invalid response, which is the "elementary level". For example, Llama2-13b has 2.86% invalid in zero-shot and 4.07% in few-shot. The most-predicted class is elementary school level, with 75.93% in zero-shot and 80% in few-shot. The number of invalid responses for other models is available in the Appendix D.

4.5 Results

Overall Performance   Table 1 reports the overall results of our different logistic classifier setups along with the ZERO-SHOT and FEW-SHOT LLM classification baselines. We use Gemma-7b, Mistral-7b, Llama2-7b, and Llama2-13b across all referenced classification methods. GPT-4 is exclusively used in the baseline due to the high cost of

experiments.

The findings highlight the consistent superiority of the COMBO approach in achieving the highest macro-F1 score, surpassing all other models. Specifically, while the Llama2-7b model exhibits comparatively lower performance when employing the Prompt-based method, the Llama2-13b model demonstrates the best performance across PROMPT-BASED metrics. Notably, the Gemma-7b model stands out as the best-performing model when using the COMBO metric. In terms of Prompt-based regression, the average macro-F1 score across all models stands at 0.62, with all PROMPT-BASED metrics obtained directly through LLMs' binary classification prompts. The best performance overall is achieved by COMBO, which combines both sets of metrics, resulting in a macro-F1 score of 0.86.

Nearly all models encounter difficulty in predicting the educational level across both ZERO-SHOT and FEW-SHOT methodologies. However, in these experiments, the FEW-SHOT approach notably enhances the macro-F1 score. Additionally, Table 1 highlights that the best performance among baseline approaches is achieved by GPT-4, attaining a macro-F1 score of 0.63 in the FEW-SHOT setting.

Performance by Education Level To delve into the performance more comprehensively, we split out the results for each regression setup by label, i.e., education level, in Table 2. Here, we display only the top-performing model based on the PROMPT-BASED metric and provide the details of the other models in Appendix D.

The overall picture of PROMPT-BASED regression shows that it faces difficulty in the classification of educational level, while STATIC performs much better, and COMBO performs best, which indicates that there is an additional benefit to including the PROMPT-BASED metrics.

We collect 1,000 bootstrap samples to train and test the logistic regression models for each approach. This method helps in understanding the variability and reliability of the model performance. We use t-tests to determine if the observed differences in accuracies are statistically significant over COMBO vs. STATIC. Results in Table 2 indicate a statistically significant improvement.

Feature Importance One big benefit of our regression approach over, for example, classification with an LLM, is that we can easily measure the feature importance of each metric that goes into

the classification result. For this purpose, we calculate univariate F-tests between each metric and the difficulty level variable. Table 3 shows the top-five most important features, each among the PROMPT-BASED and the STATIC metrics, based on these F-tests for *Llama2-13b* model.

Most notably, the PROMPT-BASED metrics are generally less important than the STATIC metrics. On average, the top five most important STATIC metrics are at least twice as significant as the top five PROMPT-BASED metrics. The STATIC metrics mainly focus on readability and lexical diversity, while PROMPT-BASED metrics capture topic relevancy and the inclusion of simple examples. Although they may not carry the same weight, all of the top metrics are highly statistically significant.

5 Discussion

5.1 The Value of Prompt-based Metrics

PROMPT-BASED metrics by themselves may not be a good-enough basis for classifying text difficulty (Table 1). STATIC metrics are much more effective by comparison. However, our results also show that PROMPT-BASED metrics do indeed capture relevant features of the text that are not captured by STATIC metrics since models that combine both kinds of metrics clearly perform best overall. This is despite the fact that the STATIC metrics we include are many and highly diverse.

The practical usefulness of the particular PROMPT-BASED metrics outlined in this paper is evident. Moreover, the broader application of PROMPT-BASED metrics holds promise for evaluating text complexity. Our experiments indicate that the COMBO approach outperforms other models consistently. Notably, most models exhibit superior macro-F1 scores in predicting elementary-level texts, suggesting that distinguishing science questions at the elementary level is more discernible compared to other educational levels.

Furthermore, we present the feature importance of PROMPT-BASED metrics, noting that the primary PROMPT-BASED metrics pertain to readability, understandability, and suitability of text for particular educational levels. Additionally, topic relevance (e.g., math or natural science) emerges as a significant feature. In top 5 best features of STATIC metrics are summarized through readability scores ranging from the Gunning Fog Index to the Flesch-Kincaid Index, along with a metric evaluating the lexical diversity of the text.

Method	Gemma-7b	Mistral-7b	Llama2-7b	Llama2-13b	GPT-4
PROMPT-BASED Reg.	0.73	0.54	0.45	0.77	-
STATIC Reg.	0.81	**0.81**	**0.81**	0.81	-
COMBO Reg.	**0.95**	**0.82**	**0.81**	**0.88**	-
ZERO-SHOT LLM	0.35	0.34	0.35	0.35	0.51
FEW-SHOT LLM	0.37	0.37	0.45	0.47	**0.65**

Table 1: Macro-F1 for difficulty classification on test. PROMPT-BASED metrics, zero-shot, and few-shot (two examples) performance are specific to each LLM. STATIC metrics are the same across models. Zero-shot and few-shot classification use GPT4. Best performance per model in **bold**.

	Level	Precision	Recall	F1-Score
PROMPT	Elem.	0.84	0.82	0.83
	Middle	0.84	0.64	0.73
	High	0.68	0.84	0.75
STATIC	Elem.	0.86	0.85	0.86
	Middle	0.75	0.71	0.73
	High	**0.84**	0.88	0.84
COMBO	Elem.	**0.95**$*$	**0.93**$*$	**0.94**$*$
	Middle	**0.89**$*$	**0.77**$*$	**0.83**$*$
	High	0.82	**0.93**$*$	**0.87**$*$

Table 2: Difficulty classification performance on test. $*$ = statistically significant improvements of COMBO over STATIC at $p = 0.05$ (bootstrap). PROMPT-BASED metrics use *Llama2-13b*. Best performance per level in **bold**.

Better PROMPT-BASED metrics identified in future work may be even more effective complements to Static metrics.

5.2 Limitations

Limited Scope of User Study   The user study we conducted provides a clear empirical motivation for the PROMPT-BASED metrics we selected. This in itself is a core contribution of our work. However, due to resource and time constraints, the sample of participants in the study is fairly small and of limited diversity. Future work could improve on our approach by conducting larger studies or recruiting participants from even more relevant professions (e.g. teachers) to motivate the selection of even better PROMPT-BASED metrics.

Limited Availability of Relevant Data   Our experiments are mostly constrained by the availability of relevant data for text difficulty classification. The ScienceQA dataset that we use is, to our knowledge, the only dataset that fits our experimental

setup in terms of size and detail on education level. Therefore, we cannot make any strong claims about the generalisability of our results. Future work could invest into building new datasets and testing cross-domain performance of both Static and PROMPT-BASED metrics, which would give useful insights into which text features are most generally indicate of text difficulty.

6   Related Work

6.1   Question Answering Datasets in Education

The review study by AlKhuzaey et al. (2023) about the literature on item difficulty classification reveals a significant shortage of publicly accessible datasets with items that are labeled according to their difficulty levels. For example, Hsu et al. (2018) gathered their dataset from national standardized entrance tests that often concentrate on the medical and language fields, annotated with the performance data of 270,000 examinees. This study includes the necessity for a publicly accessible collection of standardized datasets and the need for further exploration into alternative methods for feature elicitation and classification modeling. The lack of publicly available datasets for measuring difficulty has led researchers toward the domain of Automatic Question Generation (AQG) in recent years. Typically, questions generated by AQG tend to be more straightforward in structure and cognitive demand than questions written by humans.

Most of these automatically generated questions are basic, primarily addressing only the first level of Bloom's taxonomy, which is focused on recall (Leo et al., 2019). Another source of educational datasets is retrieved from online learning platforms or websites specific to the study's domain. An example includes the collection of 1,657

	Rank	Metric	F
Prompt Metrics	1	Based on the **ARI**, is this text suitable for ES readers?	251.77*
	2	Is this text **relevant to curriculum** topics for ES students?	249.07*
	3	Is this text about **math**?	248.17*
	4	Is this text about **natural science**?	240.07*
	5	Does this text contain **simple examples**?	235.96*
Static Metrics	1	Gunning Fog (measures **readability**)	817.86*
	2	Coleman-Liau index (measures **readability**)	785.60*
	3	Flesch-Kincaid Reading Ease (measures **readability**)	725.15*
	4	Automated Readability Index (measures **Readability**)	686.87*
	5	Number of unique Words (measures **lexical diversity**)	613.89*

Table 3: Five most important features for PROMPT-BASED and STATIC metrics in *Llama2-13b*. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance. (ES: Elementary School, ARI: Automated Readability Index) * indicates significance at >99.999% confidence.

programming problems from LeetCode[3], labeled with the number of solutions submitted and the pass rate for each problem. Additionally, fewer datasets are from domain-specific textbooks and preparation books, particularly prevalent in the language domain for their role in training students for language proficiency exams. Domain experts developed the remaining sources to meet specific study goals, and according to AlKhuzaey et al. (2023), only 7% from school or university-level assessments.

The Stanford Question Answering Dataset (SQuAD), developed by Rajpurkar et al. (2016), features 150,000 questions in the form of paragraph-answer pairs sourced from Wikipedia articles. This dataset was utilized by Bi et al. (2021) to develop and test their models for predicting the difficulty of reading comprehension questions. Lu et al. (2022) created a multimodal science question-answering datasets, which includes 21,000 English passages from school reading exams, each accompanied by four multiple-choice questions. The ScienceQA dataset provides metadata fields for each question, including extensive solutions and general explanations which made it suitable for this study (Lu et al., 2022).

6.2 Automatic Evaluation of Educational Content

The difficulty level classification of questions presented to students is crucial for facilitating more effective and efficient learning. Pérez et al. (2012) shows teachers usually fail to identify the correct difficulty level of the questions according to their

students' answers and final scores. The student's perception of the difficulty also changes across grades and subjects. AlKhuzaey et al. (2023) discovers that linguistic features significantly influence the determination of question difficulty levels in educational assessments. They have explored various syntactic and semantic aspects to understand the complexity of these questions. Crossley et al. (2019) shows the value of using crowdsourcing methods to gather human assessments of text comprehension, coupled with linguistic attributes derived from advanced readability metrics. This approach aids in creating models that explain how humans understand and process text, as well as factors influencing reading speed. Crossley et al. (2023) examined the effectiveness of new readability formulas developed on the CommonLit Ease of Readability (CLEAR) corpus using more efficient sentence-embedding models and comparing them to traditional readability formulas. They did not tru LLMs directly for difficulty classification task. In their respective studies, Imperial and Madabushi (2023), Rooein et al. (2023), and Gobara et al. (2024) leverage Large Language Models (LLMs) for content generation, focusing specifically on controlling readability scores. Their research illuminates the inherent challenges and limitations encountered when attempting to effectively adapt LLMs for this purpose.

7 Conclusion

Good teachers succeed in making the material understandable for their respective audiences. This adaptation is a complex process that goes well beyond replacing individual words and phrases. How-

---

[3]https://leetcode.com

ever, existing STATIC metrics for text difficulty, like the Flesch-Kincaid Reading Ease score, still focus on precisely those elements. As a result, these metrics are crude and brittle, failing to adapt to new domains and working mainly on long-form documents.

Our experiments reveal the promising potential of LLMs in predicting educational difficulty through using the PROMPT-BASED metrics rather than prompting the model directly. These metrics were derived from a small-scale user study involving students. Empirically, we demonstrate that when combined with traditional static metrics, these PROMPT-BASED metrics enhance text difficulty classification.

Our study paves the way for novel applications of LLMs in educational contexts. By involving more educational stakeholders, such as teachers, we can gather more representative PROMPT-BASED metrics, facilitating future advancements in difficulty classification.

Ethical Considerations

The participants in the user study we used in our paper were student volunteers for a course on related topics. They could leave the study at any point and were compensated in course credits that could be counted towards their study program. The study was conducted in accordance with the rules of the host university and passed its ethics assessment. The risk for harm to the participants in this setting was assessed as minimal.

Acknowledgements

References

Samah AlKhuzaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2023. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, pages 1–53.

Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E Clauser, and Le An Ha. 2021. Using natural language processing to predict item response times and improve test construction. *Journal of Educational Measurement*, 58(1):4–30.

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Lizhen Qu, Shirong Shen, Guilin Qi, Lu Pan, and Yinlin Jiang. 2021. Simple or complex? complexity-controllable question generation with soft templates and deep mixture of experts model. *arXiv preprint arXiv:2110.06560*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Scott Crossley, Joon Suh Choi, Yanisa Scherber, and Mathis Lucka. 2023. Using large language models to develop readability formulas for educational settings. In *International Conference on Artificial Intelligence in Education*, pages 422–427. Springer.

Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319.

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

Seiji Gobara, Hidetaka Kamigaito, and Taro Watanabe. 2024. Do llms implicitly determine the suitable text difficulty for users? *arXiv preprint arXiv:2402.14453*.

Google. 2024. Responsible Generative AI Toolk, GemmaTechnical Report. https://ai.google.dev/gemma/docs. Accessed: March 6, 2024.

Mohammad Hosseini, Catherine A Gao, David M Liebovitz, Alexandre M Carvalho, Faraz S Ahmad, Yuan Luo, Ngan MacDonald, Kristi L Holmes, and Abel Kho. 2023. An exploratory survey about using chatgpt in education, healthcare, and research. *medRxiv*, pages 2023–03.

Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. *arXiv preprint arXiv:2309.05454*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jared Leo, Ghader Kurdi, Nicolas Matentzoglu, Bijan Parsia, Ulrike Sattler, Sophie Forge, Gina Donato, and Will Dowling. 2019. Ontology-based generation of medical, multi-term mcqs. *International Journal of Artificial Intelligence in Education*, 29:145–188.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

OpenAI. 2023. GPT-4 Technical Report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Elena Verdú Pérez, Luisa M Regueras Santos, María Jesús Verdú Pérez, Juan Pablo de Castro Fernández, and Ricardo García Martín. 2012. Automatic classification of question difficulty level: Teachers' estimation vs. students' perception. In *2012 Frontiers in Education Conference Proceedings*, pages 1–5. IEEE.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural language engineering*, 17(4):485–510.

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*.

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11(6).

Lisa S Stamps. 2004. The effectiveness of curriculum compacting in first grade classrooms. *Roeper Review*, 27(1):31–41.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shriyash Upadhyay, Etan Ginsberg, and Chris Callison-Burch. 2023. Improving mathematics tutoring with a code scratchpad. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 20–28.

Kang Xue, Victoria Yaneva, Christopher Runyon, and Peter Baldwin. 2020. Predicting the difficulty and response time of multiple choice questions using transfer learning. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 193–197.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2023. Practical and ethical challenges of large language models in education: A systematic literature review. *arXiv preprint arXiv:2303.13379*.

Victoria Yaneva, Peter Baldwin, Janet Mee, et al. 2019. Predicting the difficulty of multiple choice questions in a high-stakes medical exam. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–20.

A Selected Prompts from the User Study

We collect the top prompts of the students from the chat history with analytical, manual, and AI Assistant (ChatGPT).

A.1 Elementary School:

- Simplify a text for elementary school, using simple language for 6-12 years olds. - Create an elementary version of a high school lecture text. - Simplify a high school text for elementary school. - Explain in a way an 8-year-old would understand. - This is a text meant for high school students. Can you help me make an appropriate version for elementary school students with very simple language and comprehensive, easy-to-understand examples?

A.2 Middle School:

- Give examples from middle school lectures. - Adapt a high school text for middle school, using less advanced language. - Be more textbook-like and more to the point for the level of middle school. - Adapt content for a student in middle school. - Simplify a lecture text for middle school using illustrative examples.

A.3 High School:

- Enhance scientific accuracy and add comprehensive examples for the high school level. - Adapt a middle school text for high school, using advanced language. - Increase difficulty for high school, with advanced vocabulary and scientific concepts. - Can you make it more scientific and less story-telling-like? - Increase the difficulty level with comprehensive examples.

B Parameter settings

The Static metrics are collected by Python packages such as `nltk(3.8.1)`, `pandas(2.2.0)`, `textstat(0.7.3`, `spacy(3.7.4)`. We use `nltk.download` to get data for 'stopwords', 'cmudict', 'wordnet', and 'averaged_perceptron_tagger'.

We performed 8-bit quantization for collecting prompt-based metrics and baselines, with a maximum input length of 2048 tokens and a maximum output length of 256 tokens. This process was restricted to a single run due to our utilization of pre-trained models readily accessible in Hugging-Face Transformers.

For Regression model, we use scikit-learn package and SelectKBest with `f_classif` score function.

C List of Metrics

C.1 Static Metrics

Table 4 shows all static metrics.

C.2 Prompt-based Metrics

Is this text readable for an elementary school student?, Is this text suitable for an elementary school student?, Is this text easy to understand for elementary school students?, Is this text relevant to curriculum topics for elementary school students?, Is this text relevant to the knowledge and experiences of elementary school students?, Could an average elementary school student engage with the content of this task?, Could most elementary school students complete this task without significant difficulty?, Is this text appropriate for the skills and knowledges of elementary school students?, Is the length of this text suitable for elementary school students?, Would the vocabulary in this text be comprehensible to elementary school students?, Is this text readable for a middle school student?, Is this text suitable for a middle school student?, Is this text easy to understand for middle school students?, Is this text relevant to curriculum topics for middle school students?, Is this text relevant to the knowledge and experiences of middle school students?, Could an average middle school student engage with the content of this task?, Could most middle school students complete this task without significant difficulty?, Is this text appropriate for the skills and knowledges of middle school students?, Is the length of this text suitable for middle school students?, Would the vocabulary in this text be comprehensible to middle school students?, Is this text readable for a high school student?, Is this text suitable for a high school student?, Is this text easy to understand for high school students?, Is this text relevant to curriculum topics for high school students?, Is this text relevant to the knowledge and experiences of high school students?, Could an average high school school student engage with the content of this task?, Could most high school students complete this task without significant difficulty?, Is this text appropriate for the skills and knowledges of high school students?, Is the length of this text suitable for high school students?, Would the vocabulary in this text be comprehensible to high school students?, Does this text contain metaphors and/or figurative language?, Does this text use complex language?, Does this text use simple language?, Does this text contain technical jargon?, Is this text about science?, Is this text about language science?, Is this text about natural science?, Is this text about social science?, Is this text about math?, Is this text about physics?, Is this text about chemistry?, Is this text about earth science?, Is this text about world history?, Is this text about geography?, Based on the Flesch-Kincaid reading-ease score, is this text suitable for elementary school readers?, Based on the Flesch-Kincaid reading-ease score, is this text suitable for middle school readers?, Based on the Flesch-Kincaid reading-ease score, is this text suitable for high school readers?, Based on the Gunning Fog Index, is this text suitable for elementary school readers?, Based on the Gunning Fog Index, is this text suitable for middle school readers?, Based on the Gunning Fog Index, is this text suitable for high school readers?, Based on the Coleman-Liau Index, is this text suitable for elementary school readers?, Based on the Coleman-Liau Index, is this text suitable for middle school readers?, Based on the Coleman-Liau Index, is this text suitable for high school readers?, Based on the Automated Readability Index (ARI), is this text suitable for elementary school readers?, Based on the Automated Readability Index (ARI), is this text

Table 4: List of Static metrics

Feature	Description
n_words_q	Number of words in the question
n_words_a_solution	Number of words in the solution of an answer
n_words_a_lecture	Number of words in the lecture
Text_Length	Length of the text
Word_Count	Total word count
Nouns	Number of nouns
Verbs	Number of verbs
Adjectives	Number of adjectives
Adverbs	Number of adverbs
Num_Numbers	Number of numeric characters
Num_Commas	Number of commas
Num_Complex_Words	Number of complex words
Num_Unique_Words	Number of unique words
Num_Content_Words	Number of content words
Num_Content_Words_No_Stopwords	Number of content words excluding stopwords
Word_Length_Syllables	Average word length in syllables
Avg_Sentence_Length	Average sentence length
Num_Prepositional_Phrases	Number of prepositional phrases
Num_Negated_Words_Stem	Number of negated words stemmed
Num_Negated_Words_Lead_In	Number of negated words leading in
Num_Main_Noun_Phrases	Number of main noun phrases
Avg_Main_NP_Length	Average length of main noun phrases
Num_Verb_Phrases	Number of verb phrases
Prop_Active_Voice_Verbs	Proportion of active voice verbs
Prop_Passive_Voice_Verbs	Proportion of passive voice verbs
Ratio_Active_to_Passive_Verbs	Ratio of active to passive voice verbs
Num_Words_Before_Main_Verb	Number of words before the main verb
Num_Agentless_Passive_Constructions	Number of agentless passive constructions
Word_Length_Std_Dev	Standard deviation of word lengths
Num_Polysemic_Words	Number of polysemic words
Num_Word_Senses	Number of word senses
Num_Word_Senses_For_Content_Words	Number of word senses for content words
Num_Word_Senses_For_Nouns	Number of word senses for nouns
Num_Word_Senses_For_Verbs	Number of word senses for verbs
Num_Word_Senses_For_Non_Auxiliary_Verbs	Number of word senses for non-auxiliary verbs
Num_Word_Senses_For_Adjectives	Number of word senses for adjectives
Num_Word_Senses_For_Adverbs	Number of word senses for adverbs
Distance_To_Root_Nouns	Distance to root for nouns
Distance_To_Root_Verbs	Distance to root for verbs
flesch_kincaid_grade	Flesch-Kincaid grade level
flesch_kincaid_ease	Flesch-Kincaid ease score
coleman_liau_index	Coleman-Liau index
automated_readability_index	Automated Readability Index
smog_index	SMOG index
gunning_fog	Gunning Fog index
traenkle_bailer_index	Traenkle-Bailer index

suitable for middle school readers?, Based on the Automated Readability Index (ARI), is this text suitable for high school readers?, Based on the SMOG Index, is this text suitable for elementary school readers?, Based on the SMOG Index, is this text suitable for middle school readers?, Based on the SMOG Index, is this text suitable for high school readers?, Does this text contain basic concepts that are easy to comprehend?, Does this text cover multiple concepts?, Does this text provide a very explicit explanation?, Does this text contain simple examples?

D Details over Gemma-7B, Mistral-7B, and Llama2-7B

We describe the performance of these models in detail. Gemma7b has 10.33% invalid response in zero-shot and 9.56% over few-shot. The majority of the predicted class is high school level 73.41% in zero-shot and 72.75% in few-shot. Mistral7b has 15.49% invalid response in zero-shot and 6.37% invalid in few-shot and with majority of classification for high school level with 66.04% in zero-shot and 42.31% for elemetary school in few-shot. Llama2-7b has 13.08% invalid in zero-shot and 5.49% in few-shot and the majority of elementary school classification with 66.26% in zero-shot and also 76.04% in few-shot. Gpt-4 has only 5.93% invalid in zero-shot and 0.77% in few-shot. Gpt-4 predicted also the high school level as the highest classification with 41.54% in zero-shot and 40.22% in few-shot.

	Level	Precision	Recall	F1-Score
PROMPT-BASED	Elem.	0.83	0.81	0.82
	Middle	0.75	0.57	0.65
	High	0.66	0.81	0.65
STATIC	Elem.	0.86	0.85	0.86
	Middle	0.75	0.71	0.73
	High	0.84	0.88	0.86
COMBO	Elem.	**0.98**∗	**0.98**∗	**0.98**∗
	Middle	**0.98**∗	**0.91**∗	**0.95**∗
	High	**0.91**∗	**0.97**∗	**0.94**∗

Table 5: Difficulty classification performance on test. ∗ = statistically significant improvements of COMBO over STATIC at $p = 0.05$ (bootstrap). PROMPT-BASED metrics use *Gemma-7b*. Best performance per level in **bold**.

	Level	Precision	Recall	F1-Score
PROMPT	Elem.	0.46	0.86	0.60
	Middle	**0.92**	0.83	**0.88**
	High	0.34	0.10	0.16
STATIC	Elem.	**0.86**	0.85	0.86
	Middle	0.75	0.71	0.73
	High	0.84	**0.88**	**0.86**
COMBO	Elem.	0.76	**0.95**∗	**0.84**∗
	Middle	0.85	**0.90**∗	**0.88**∗
	High	**0.89** ∗	0.64	0.75

Table 6: Difficulty classification performance on test. ∗ = statistically significant improvements of COMBO over STATIC at $p = 0.05$ (bootstrap). PROMPT-BASED metrics use *Mistral-7b*. Best performance per level in **bold**.

	Level	Precision	Recall	F1-Score
PROMPT	Elem.	0.44	0.47	0.45
	Middle	0.62	0.61	0.62
	High	0.29	0.28	0.28
STATIC	Elem.	0.86	0.85	0.86
	Middle	**0.75**	0.71	**0.73**
	High	**0.84**	**0.88**	**0.86**
COMBO	Elem.	**0.88**∗	**0.97**∗	**0.93**∗
	Middle	0.72	**0.74**∗	**0.73**∗
	High	0.83	0.73	0.78

Table 7: Difficulty classification performance on test. ∗ = statistically significant improvements of COMBO over STATIC at $p = 0.05$ (bootstrap). PROMPT-BASED metrics use *Llama2-7b*. Best performance per level in **bold**.

	Rank	Metric	F
Prompt	1	Based on the **Coleman-Liau Index**, is the text suitable for MS readers?	105.09*
	2	Is this text **readable** for a MS student?	104.42*
	3	Based on the **SMOG Index**, is this text suitable for MS readers?	103.53*
	4	Is this text **suitable** for a MS student?	94.21*
	5	Based on the **Gunning Fog Index**, is this text suitable for MS readers?	92.35*
Static Metrics	1	Gunning Fog (measures text **readability**)	817.86*
	2	Coleman-Liau index (measures text **readability**)	785.60*
	3	Flesch-Kincaid Reading Ease (measures **readability**)	725.15*
	4	Automated Readability Index (measures **lexical diversity**)	686.87*
	5	Number of unique Words (measures **lexical diversity**)	613.89*

Table 8: Top five most important features among the PROMPT-BASED and STATIC metrics. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance. (MS: Middle School) PROMPT-BASED metrics use the *Gemma-7B* model. * indicates significance at >99.999% confidence.

	Rank	Metric	F
Prompt	1	Based on the **Gunning Fog Index**, is this text suitable for ES readers?	209.84*
	2	Is this text **easy to understand** for ES students??	193.22*
	3	Is this text **Suitable** for ES students	190.61*
	4	Is this text about **math**?	175.72*
	5	Is this text **relevant to curriculum** topics for ES students?	175.08*
Static Metrics	1	Gunning Fog (measures text **readability**)	817.86*
	2	Coleman-Liau index (measures text **readability**)	785.60*
	3	Flesch-Kincaid Reading Ease (measures **readability**)	725.15*
	4	Automated Readability Index (measures **Readability**)	686.87*
	5	Number of unique Words (measures **lexical diversity**)	613.89*

Table 9: Top five most important features among the PROMPT-BASED and STATIC metrics. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance. (ES: Elementary School) PROMPT-BASED metrics use the *Mistral-7B* model. * indicates significance at >99.999% confidence.

	Rank	Metric	F
Prompt-based Metrics	1	Is this text **relevant to curriculum** topics for ES students?	139.66*
	2	Is this text suitable for an ES student?	136.97*
	3	Is this text **readable** for an ES student	132.89*
	4	Based on the **Gunning Fog Index**, is this text **suitable** for MS readers?"	125.51*
	5	Is this text about **natural science**?	124.52*
Static Metrics	1	Gunning Fog (measures text **readability**)	817.86*
	2	Coleman-Liau index (measures text **readability**)	785.60*
	3	Flesch-Kincaid Reading Ease (measures **readability**)	725.15*
	4	Automated Readability Index (measures **Readability**)	686.87*
	5	Number of unique Words (measures **lexical diversity**)	613.89*

Table 10: Top five most important features among the PROMPT-BASED and STATIC metrics. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance.(ES: Elementary School, MS: Middle School) PROMPT-BASED metrics use the *Llamma2-7B* model. * indicates significance at >99.999% confidence.