Beyond Flesch-Kincaid: New Neural Metrics to Improve Difficulty Prediction for Educational Texts

Anonymous NAACL 2024 submission

Abstract

Using large language models (LLMs) for educational applications such as dialogue-based teaching is a hot topic. Effective teaching, however, requires adapting 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 reliably measure adaptation success. 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 *neural* metrics for text difficulty. Based on a user study, we create Neural metrics as LLM prompts that leverage the general language understanding capabilities of LLMs to capture more abstract and complex text features than Static metrics. Through regression experiments, we show that our Neural metrics improve text difficulty prediction over Static metrics alone. Our results demonstrate the promise of Neural metrics as a new class of features for evaluating text adaptation to different education levels.

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.

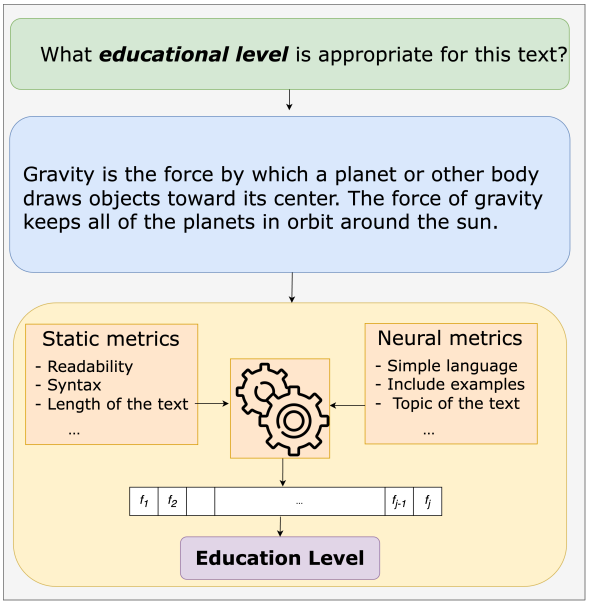

Figure 1: Schematic overview of our proposed approach to text difficulty classification. We calculate Static and Neural metrics relevant to text difficulty based on a given input text[1]. Either or both types of metrics are then fed into a regression classifier that makes a final prediction.

Consequently, even the best LLMs today struggle to provide appropriate answers for a given audience – they can adapt to different writing styles (casual or formal) and domains (emails, blog posts, and essays), but their ability to adapt to different age ranges or levels of education is much more limited (Kasneci et al., 2023). At best, they can differentiate between child and adult learners, but that distinction is rarely useful when adapting recurring material for different grades.

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.

However, whether a given output is appropriate for a given level of education is still very hard to

measure. Existing *Static* metrics, like the Flesch-Kincaid Reading Ease score, 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 reliability on counts, their estimates are unreliable in shorter formats. To make improvements on the adaptability of LLMs to education levels measurable, we need better metrics. Only when we can measure improvements can we make tangible progress in automation.[2]

Alternatively, we can use (Neural) classifiers to predict the educational level of a given text. They generalize better and can be applied to texts of varying length. However, these classifiers are expensive to train, and require more training data than we usually have for a niche domain like educational purposes.

The obvious goal is human assessment, 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 Neural metrics for text difficulty as complements to existing Static metrics. Neural metrics exploit the general language understanding capabilities of LLMs to capture more abstract features of educational texts than Static metrics. LLMs can, for example, easily and flexibly classify whether a given text uses a metaphor or not (which is one of the adaptation techniques used by teachers to adjust content to higher education levels). This would be difficult to do with Static approaches.

First, to motivate our selection of Neural metrics, we conduct 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 Neural metrics – like the metaphor example above.

Second, to evaluate the usefulness of our new Neural metrics for measuring text appropriateness for education levels, we run a series of experiments using a subset of the ScienceQA dataset (Auer et al., 2023), which contains question-answer pairs across several topics and education levels. Specifically, we run multinomial logistic regressions based on Static metrics, Neural metrics, and the combination of the two to evaluate the marginal benefits of our new Neural metrics. We also compare these regression approaches with using an LLM directly for zero-shot classification.

We find that Neural metrics by themselves are much less useful for text difficulty classification than Static metrics – even if both perform on par or better than a zero-shot LLM classifier. However, the combination of both Neural and Static metrics clearly performs best overall. This shows that Neural metrics do indeed capture relevant signals beyond what Static metrics can capture despite the large number of Static metrics that we include in our experiments. Therefore, the use of Neural metrics in addition to Static metrics appears to be a promising direction for improving text difficulty classification.

To summarise, our main contributions are:

1. We conduct a user study to motivate the creation of novel Neural metrics of text difficulty for educational texts.

2. We show in a series of regression experiments that these novel Neural metrics hold additional value for text difficulty classification beyond what Static metrics can capture.

3. By leveraging the interpretability of our regressions, we highlight the relative importance of individual Static and Neural metrics.

2 User Study

In this paper, we introduce novel Neural metrics for text difficulty. To provide an empirical foundation for these metrics, we ran a one-day user study with a group of university students in November 2023.

2.1 Study Design

Our user study consists 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 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

[2] As examplified by the cases of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BLANC (Recasens and Hovy, 2011), among others, which kickstarted and sustained the development of automated approaches to machine translation, summarization, and coreference resolution, respectively.

either from the social, natural, or language sciences. 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 [REDACTED UNIVERSITY]. There were seven participants aged between 21 and 31 years. Four participants were female, three male. All participants were students at [REDACTED UNIVERSITY] and were 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 were diverse, including 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, as well as prompts in Appendix A.

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

3 Metrics for Text Difficulty

3.1 Neural Metrics

Since the metrics we introduce are based on the prompts of Neural language models rather than discrete heuristics, we refer to them as 'Neural' to distinguish them. The goal of the Neural 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.

We derive our Neural metrics from the results of our user study. Figure 2 shows an illustrative example of our derivation process. We consider the explanations provided by users for why they consider a certain educational text to be of elementary, middle, or high school level difficulty. Then, we identify recurring attributes and other explanation features that are mentioned by several users to reflect them in Neural metrics. More specifically, 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, *explicit, explanation, middle*.

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

In total, we construct 41 Neural metrics using this process. Each Neural metric relates to either education level (24 metrics), lexical or syntactic complexity (7 metrics), or the topic of the text at hand (10 metrics). Table 1 shows an example of each type of metric. For the full list of all Neural metrics, see 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 or other textual features. This makes them simple to apply but also limits the conceptual complexity of what they can reasonably be expected to measure. In total, we include 38 Static metrics, selected from those compiled in prior work by Flekova et al. (2016). These

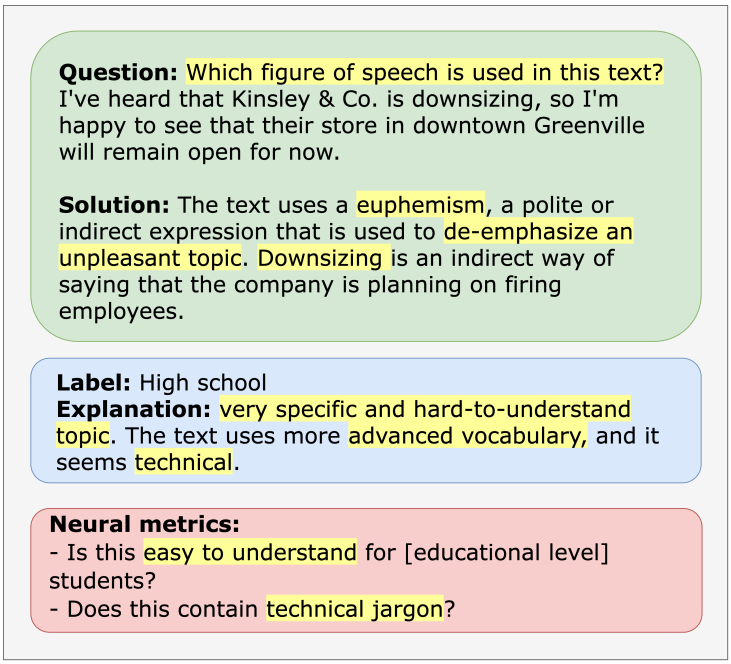

Figure 2: An illustrative example of the Neural metric process. The green box contains the education text from the ScienceQA dataset. The blue box shows the predicted educational level and the explanation. The red box contains the Neural metrics based on the sample.

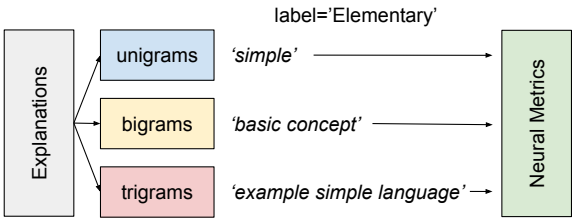

Figure 3: High-level view of the derivation process for the Neural metrics using n-gram frequencies. Function words are excluded.

Neural Metric	Category
Is this text suitable for an **elementary school** student?	Education level
Does this text contain **technical jargon**?	Lexical/syntactic complexity
Is this text about **math**?	Topic

Table 1: Three example Neural metrics across the three categories of Neural metrics. Each Neural metric is a prompt for an LLM.

range from text-level metrics that capture general linguistic properties such as vocabulary size and word frequency to sentence-level metrics such as sentence length and syntactic complexity. For the full list of 38 Static metrics see Appendix **??**.

4 Experiments

To evaluate the usefulness of our novel Neural metrics for measuring text difficulty we conduct a series of text difficulty classification experiments.

4.1 Dataset

All our experiments are based on the recently-released ScienceQA dataset (Auer et al., 2023). There are ca. 21k texts in ScienceQA. Each text consists of a question (e.g. "What is the mass of a dinner fork?") with answer choices ("70 grams or 70 kilograms") 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).[3] From the total 21k texts in ScienceQA, we sample only those that do not use images in questions or explanations. We then deduplicate and sample 1,500 texts for each education level to create a balanced dataset of 4,500 texts. Of these 4,548 texts, we use 3638 (80%) for training and 910 (20%) for evaluation. To our knowledge, ours

[3] https://usahello.org/education/children/grade-levels/

is the first use of the ScienceQA dataset for the purposes of training and evaluating text difficulty classifiers.

4.2 LLMs for Neural Metrics

We mainly use LLMs in our experiments to compute the 41 Neural metrics described in Section **??**. In principle, any LLM can be used for this purpose. We use three state-of-the-art LLMs: Llama2 (Touvron et al., 2023), Mistral (Jiang et al., 2023) and GPT-4 (OpenAI, 2023). Llama2, launched in July 2023, is a collection of both pre-trained and fine-tuned LLMs, with sizes ranging from 7 billion to 70 billion parameters. It has been reported to outperform other open-access LLMs and demonstrates capabilities comparable to ChatGPT across various tasks. In the main body of this paper, we focus on the chat-optimised 13 billion parameter version of Llama2, which we refer to as Llama2-13b. The Mistral-7B, released in September 2023, is another open-source language model that surpasses similar-sized open LLMs, including Llama2-13b. We test the instruction-tuned version, Mistral-7b-Instruct-v0.2, which was published in December 2023. Access to these models is provided with Hugging Face. Additionally, we examine GPT-4, a proprietary model from OpenAI released in March 2023. GPT-4 is accessed through its API. This study aims to provide a comprehensive comparison of these state-of-the-art LLMs.

In Appendix D, we report corresponding results for Mistral-7B and GPT-4. We set the model temperature to zero to make responses deterministic. The maximum response length is 256 tokens. Otherwise, we use standard generation parameters as provided by the Hugging Face transformers library. We collected all responses in November and December 2023.

4.3 Multinomial Logistic Regression

To classify the difficulty level of texts, we use simple multinomial logistic regression. Formally, 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/or Neural metrics $\mathbf{M}_i$ that are computed based on S_i. We regress $\mathbf{M}_i$ on C_i on the 3638 training texts and then evaluate on the 910 test education texts.

4.4 Baseline: Zero-Shot Classification

We exploit the general language capabilities of LLMs to compute Neural metrics, which we then use as inputs to a logistic classifier for text difficulty. A natural follow-up question is whether LLMs could just as well predict education level text difficulty directly. Therefore, we include a zero-shot text classification baseline. We use the same model for zero-shot classification as for computing our Neural metrics. For the experiments here, this is Llama2-13b. Note that while the logistic classifier is fitted to our training data, the zero-shot LLM has not seen any examples at inference time.

4.5 Experimental Setups

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) Neural metrics only, which we refer to as NEURAL, 2) Static metrics only, which we refer to as STATIC, and 3) the combination of the two, which we refer to as COMBO.

4.6 Results

Table 2 reports the overall results of our different logistic classifier setups along with the zero-shot LLM classification baseline.

Method	Macro-F1	Accuracy (%)
NEURAL Reg.	0.42	45.9
STATIC Reg.	0.76	78.4
COMBO Reg.	**0.80**	**80.7**
Zero-shot LLM	0.42	45.5

Table 2: Overall performance for difficulty level classification on our ScienceQA testset. NEURAL metrics and zero-shot LLM classification use the *Llama2-13b* model. Best performance is highlighted in **bold**.

NEURAL regression performs relatively poorly, at 0.42 macro-F1 and 45.9% accuracy. This is on-par with directly prompting the LLM (in this case Llama2-13b) to make a zero-shot classification. By comparison, STATIC regression performs substantially better, at 0.76 macro-F1 and 78.4% accuracy. Best overall is COMBO, the combination of the two sets of metrics, at 0.80 macro-F1 and 80.7% accuracy.

To investigate performance in more detail, we split out the results for each regression setup by label, i.e. education level, in Table 3.

	Level	Precision	Recall	F1-Score
NEURAL	Elem.	0.55	0.36	0.43
	Middle	0.50	0.21	0.30
	High	0.43	0.78	0.55
STATIC	Elem.	0.80	0.84	0.82
	Middle	0.71	0.59	0.64
	High	0.79	**0.87*****	0.83
COMBO	Elem.	**0.86*****	**0.87*****	**0.86*****
	Middle	**0.73*****	**0.67*****	**0.70*****
	High	**0.81*****	**0.87*****	**0.86*****

Table 3: Performance for difficulty level classification on our ScienceQA testset, split by level. *** indicates statistically significant improvements. NEURAL metrics use the Llama2-13b model. Best performance per level is highlighted in **bold**.

The overall picture remains unchanged: NEURAL regression still performs worst, while STATIC performs much better, and COMBO performs best, which indicates some marginal benefit to including the NEURAL metrics. The most challenging task appears to be identifying middle school-level texts, with no model scoring more than 0.70 macro-F1. This may, in part, be explained by the ordinal nature of our labels, making it easier to classify content at the extremes of the difficulty scale than in the middle. The NEURAL model struggles more with elementary school than high school texts, whereas the STATIC and COMBO models perform equally well on both.

we collect multiple (1000) 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. The results indicate a statistically significant improvement when using COMBO over both STATIC and NEURAL approaches individually.

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 4 shows the top-five most important features each among the Neural and the Static metrics, based on these F-tests.

Most noticeably, the Neural metrics are generally much less important than the Static metrics. The top Neural metric, which gives a binary assessment of whether a given text is appropriate for the skill level of elementary school students, is roughly 10% as relevant to difficulty level as the fifth-most-important Static metric counting the number of unique tokens (62.38 vs 629.00 F). However, while they may not hold the same importance, all of the top metrics are highly statistically significant.

5 Discussion

5.1 The Value of Neural Metrics

Neural metrics by themselves may not be a good-enough basis for classifying text difficulty (Table 2). Static metrics are much more effective by comparison. However, our results also show that Neural 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.

Beyond the demonstrated practical utility of the specific Neural metrics we introduced in this paper, the use of Neural metrics more generally appears to be a promising direction for assessing text difficulty. Better Neural metrics identified in future work may be even more effective complements to Static metrics.

5.2 Limitations

The user study we conducted provides a clear empirical motivation for the Neural 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 Neural metrics.

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

	Rank	Metric	F
Neural Metrics	1	Is this **appropriate for the skill level** of elementary school students?	62.38***
	2	Does this contain a **complex language** structure?	58.04***
	3	Is this **easy to understand** for elementary school students?	46.72***
	4	Is this **suitable** for an elementary school student?	34.43***
	5	Is this about **earth science**?	27.71***
Static Metrics	1	`Herdan's C` (measures **lexical diversity**)	916.40***
	2	`Entropy` (measures **the lexical diversity**)	875.73***
	3	`Flesch-Kincaid Reading Ease` (measures **readability**)	715.95***
	4	`Simpson's Diversity Index` (measures **lexical diversity**)	686.87***
	5	`# unique tokens` (measures **length** and **lexical diversity**)	629.00***

Table 4: Top five most important features among the Neural and Static metrics. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance. *** indicates significance at >99.999% confidence.

both Static and Neural metrics, which would give useful insights into which text features are most generally indicate of text difficulty.

6 Related Work

6.1 Automatic Evaluation of Educational Content

The difficulty level prediction 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. Imperial and Madabushi (2023) and Rooein et al. (2023) use language models for content generation over text simplification tasks and controlling readability scores for specific age and educational levels. They show the limitations of the LLMs in adaptation to the specific educational grades.

6.2 Question Answering Datasets in Education

The review study by AlKhuzaey et al. (2023) about the literature on item difficulty prediction 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 prediction 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, AQG tends to be more straightforward in structure and cognitive demand compared to 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 1657 programming problems from LeetCode[4], 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

[4] https://leetcode.com

in training students for language proficiency exams. The remaining sources were developed by domain experts 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) in 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 new 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 several metadata fields associated with each question, including extensive solutions and general explanations. In contrast to SQuAD, this dataset demands more advanced reasoning abilities to answer its questions.

7 Conclusion

Good teachers succeed in making the material understandable for their respective audiences. This adaptation is a complex process which goes well beyond replacing individual words and phrases. However, existing Static metrics for text difficulty, like the Flesh-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.

Large Language Models are increasingly used in educational domains and offer ways to go beyond individual word replacement due to their general language capabilities. However, at the same time, they still struggle to adapt to precise education levels. To effectively automate text adaption to education levels, we need to measure the success of that adaptation, which requires more flexible metrics than the ones we currently have.

We introduce a suite of prompt-based Neural metrics for text adaptation based on a user study. We empirically show that these metrics, in combination with traditional Static metrics, improve text difficulty prediction. Our work opens up new avenues for the use of LLMs in educational applications.

Ethical Considerations

The participants in the user study we used in our paper were student volunteers of a course on related topics. They were able to 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 their ethics assessment. The risk for harm to the participants in this setting was assessed as minimal.

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.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240.

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 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.

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*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

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).

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.

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*.

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 year 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 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 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 difficulty level with comprehensive examples.

B Parameter settings

The Static matrics are collected by Python packages such as pandas, `textstat`, `spacy`, `wordfreq`, and `wordfreq`. We use `_en_core_web_sm` parameter from the `spacy` model. For Regression model, we use sklearn package and SelectKBest with `f_classif` score function.

C List of Metrics

Static Metrics. `num_words`: Number of words in the text. `num_sentences`: Number of sentences in the text. `type_token_ratio`: Ratio of unique words to total words, indicating vocabulary richness. `avg_word_length`: The average length of words in the text. `academic_word_list_ratio`: Ratio of academic or domain-specific words to total words. `complex_words`: Number of complex words, often based on syllable count or other criteria. `mtld`: Measure of Textual Lexical Diversity, a metric for vocabulary richness. `num_unique_tokens`: Number of unique words or tokens in the text. `avg_sent_length`: The average length of sentences in the text. `std_sent_length`: The standard deviation of sentence lengths, indicating sentence length variation. `clauses_sentences_ratio`: Ratio of clauses to sentences, providing insight into sentence complexity. `pos_ratios`: Ratios of different parts of speech (e.g., nouns, verbs) in the text. `FKGL`: Flesch-Kincaid Grade Level, an estimate of the text's readability. `FKES`: Flesch Reading Ease score is a readability measure for US education systems. `ttr`: TypeToken Ratio. `brunet_index`: A metric for text diversity and richness. `d_measure`: A measure of lexical diversity. `yules_k`: Yule's K, a measure of text's distribution of word frequencies. `herdan_c`: Herdan's C, a metric for vocabulary richness. `simpsons_di`: Simpson's Diversity Index, indicating the diversity of words in the text. `entropy`: A measure of information entropy, indicating unpredictability or randomness of the text.

Neural Metrics. Is this readable for an elementary school student based on the Flesch-Kincaid grade scale?, Is this suitable for an elementary school student?, Is this easy to understand for elementary school students?, Is this relevant to curriculum topics for elementary school students?, Is this relevant to the knowledge and experiences of elementary school students?, Is this content suitable for meeting the expected knowledge level of elementary school students? Is this able to provide detailed feedback to help elementary school students learn?, Is this appropriate for the skill level of elementary school students?, Is the length of this suitable for elementary school students?, Is this readable for a middle school student based on the Flesch-Kincaid grade scale?, Is this suitable for a middle school student?, Is this easy to understand for middle school students?, Is this relevant to curriculum topics for middle school students?, Is this relevant to the knowledge and experiences of middle school students?, Is this content suitable for meeting the expected knowledge level of middle school students? Is this able to provide detailed feedback to help middle school students learn?, Is this appropriate for the skill level of middle school students?, Is the length of this suitable for middle school students?, Is this readable for a high school student based on the Flesch-Kincaid grade scale?, Is this suitable for a high school student?, Is this easy to understand for high school students?, Is this relevant to curriculum topics for high school students?, Is this relevant to the knowledge and experiences of high school students?, Is this content suitable for meeting the expected knowledge level of high school students? Is this able to provide detailed feedback to help high school students learn?, Is this appropriate for the skill level of high school students?, Is the length of this suitable for high school students?, Does this contain metaphors and/or figurative language?, Does this contain a complex language structure?, Does this contain technical jargon?, Is the language of this simple?, Is this about science?, Is this about language science?, Is this about natural science?, Is this about social science?, Is this about math?, Is this about physics?, Is this about chemistry?, Is this about earth science?, Is this about world history?, Is this about geography?

D Details over Mistral-7B and GPT-4 Models

We describe the performance of Mistral-7B over the subset of our dataset with 4500 samples (1500 samples for each educational level) and GPT-4 for

450 samples (150 for each educational level).

Method	Macro-F1	Accuracy (%)
NEURAL Reg.	0.37	41.66
STATIC Reg.	0.73	74.00
COMBO Reg.	**0.76**	**77.3**
Zero-shot LLM	0.37	36.90

Table 5: Overall performance for difficulty level classification on our ScienceQA testset. NEURAL metrics and zero-shot LLM classification use the *Mistral-7B* model. Best performance is highlighted in **bold**.

Method	Macro-F1	Accuracy (%)
NEURAL Reg.	0.55	54.44
STATIC Reg.	0.74	74.44
COMBO Reg.	**0.76**	**76.22**
Zero-shot LLM	0.56	54.88

Table 6: Overall performance for difficulty level classification on our ScienceQA testset. NEURAL metrics and zero-shot LLM classification use the *GPT-4* model. Best performance is highlighted in **bold**.

	Level	Precision	Recall	F1-Score
NEURAL	Elem.	0.59	0.34	0.43
	Middle	0.55	0.11	0.18
	High	0.36	0.91	0.51
STATIC	Elem.	0.76	0.88	0.82
	Middle	0.81	0.45	0.58
	High	0.68	**0.94**	0.79
COMBO	Elem.	**0.80**	**0.90**	**0.85**
	Middle	**0.82**	**0.51**	**0.63**
	High	**0.71**	**0.94**	**0.81**

Table 7: Performance for difficulty level classification on ScienceQA testset, split by level. NEURAL metrics use the *Mistral-7B* model. Best performance per level is highlighted in **bold**.

	Level	Precision	Recall	F1-Score
NEURAL	Elem.	0.65	0.57	0.61
	Middle	0.41	0.38	0.39
	High	0.57	0.74	0.64
STATIC	Elem.	**0.87**	**0.74**	**0.80**
	Middle	**0.68**	**0.72**	**0.70**
	High	0.69	**0.78**	**0.73**
COMBO	Elem.	0.81	**0.74**	0.78
	Middle	0.65	0.69	0.67
	High	**0.71**	0.74	**0.73**

Table 8: Performance for difficulty level classification on our ScienceQA testset, split by level. NEURAL metrics use the *GPT-4* model. Best performance per level is highlighted in **bold**.

	Rank	Metric	F
Neural Metrics	1	Is this **easy to understand** for elementary school students?	79.75***
	2	Does this contain **metaphors and/or figurative language**?	42.20***
	3	Is this **readable** for elementary school students	40.80***
	4	Is this **appropriate for the skill level** of elementary school students?	35.65***
	5	Is this **relevant to curriculum topics** for elementary school students?	16.57***
Static Metrics	1	Herdan's C (measures **lexical diversity**)	348.82***
	2	Entropy (measures **variability or complexity**)	346.36***
	3	Flesch-Kincaid Reading Ease (measures **readability**)	284.60***
	4	Simpson's Diversity Index (measures **lexical diversity**)	255.80***
	5	Flesch-Kincaid Grade Level (measures the school **grade level**)	235.71***

Table 9: Top five most important features among the Neural and Static metrics. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance. NEURAL metrics use the *Mistral-7B* model. *** indicates significance at >99.999% confidence.

	Rank	Metric	F
Neural Metrics	1	Is this **appropriate for the skill level** of elementary school students?	62.38***
	2	Does this contain a **complex language** structure?	58.04***
	3	Is this **easy to understand** for elementary school students?	46.72***
	4	Is this **suitable** for an elementary school student?	34.43***
	5	Is this about **earth science**?	27.71***
Static Metrics	1	Herdan's C (measures **lexical diversity**)	916.40***
	2	Entropy (measures **variability or complexity**)	875.73***
	3	Flesch-Kincaid Reading Ease (measures **readability**)	715.95***
	4	Simpson's Diversity Index (measures **lexical diversity**)	686.87***
	5	Number of unique tokens (measures **length** and **lexical diversity**)	629.00***

Table 10: Top five most important features among the Neural and Static metrics. Feature importance is measured using univariate F-tests. Larger F indicates higher feature importance. NEURAL metrics use the *GPT-4* model. *** indicates significance at >99.999% confidence.