information

MDPI

Article

Assessing the Readability of Russian Textbooks Using Large Language Models

Andrei Paraschiv [1], Mihai Dascalu [1,2,*] and Marina Solnyshkina [3]

[1] Computer Science and Engineering Department, National University of Science and Technology Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania; andrei.paraschiv74@upb.ro
[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania
[3] Laboratory "MultiDisciplinary Investigations of Text", Institute of Philology and Intercultural Communication, Kazan (Volga Region) Federal University, 18 Kremlevskaya St., 420008 Kazan, Russia; mesoln@yandex.ru
* Correspondence: mihai.dascalu@upb.ro

Abstract

This study aims to assess the capability of Large Language Models (LLMs), particularly GPT-4o, to evaluate and modify the complexity level of Russian school textbooks. We lay the groundwork for developing scalable, context-aware methods for readability assessment and text simplification in Russian educational materials, areas where traditional formulas often fall short. Using a corpus of 154 textbooks covering various subjects and grade levels, we evaluate the extent to which LLMs accurately predict the appropriate comprehension level of a text and how well they simplify texts by targeted grade reduction. Our evaluation framework employs GPT-4o as a multi-role agent in three distinct experiments. First, we prompt the model to estimate the target comprehension age for each segment and identify five key linguistic or conceptual features underpinning its assessment. Second, we simulate student comprehension by instructing the model to reason step-by-step through whether the text is understandable for a hypothetical student of the given grade. Third, we examine the model's ability to simplify selected fragments by reducing their complexity by three grade levels. We further measure model perplexity and output token probabilities to probe the prediction confidence and coherence. Results indicate that while LLMs show considerable potential in complexity assessment (i.e., MAE of 1 grade level), they tend to overestimate text difficulty and face challenges in achieving precise simplification levels. Ease of understanding assessments generally align with human expectations, although texts with abstract, technical, or poetic content (e.g., Physics, History, and Literary Russian) pose challenges. Our study concludes that LLMs can substantially complement traditional readability metrics and assist teachers in developing suitable Russian educational materials.

Keywords: readability assessment; Large Language Models; Russian educational textbooks; text simplification

check for updates

Academic Editor: Katsuhide Fujita

Received: 31 October 2025
Revised: 10 November 2025
Accepted: 29 November 2025
Published: 4 December 2025

Citation: Paraschiv, A.; Dascalu, M.; Solnyshkina, M. Assessing the Readability of Russian Textbooks Using Large Language Models. *Information* **2025**, *16*, 1071. https://doi.org/10.3390/info16121071

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Textbooks serve as a cornerstone to educational systems by providing structure for learning at both early and later stages of schooling. As such, ensuring that these materials match the learners' cognitive and linguistic levels is essential. Consequently, researchers have long sought objective methods for assessing the difficulty of textbooks [1–3]. Developing automated, standardized measures would significantly benefit curriculum designers

and educators by providing scalable evaluation tools. However, accurately determining the difficulty of a textbook's content is a complex task that requires considerable expertise, as it involves not only linguistic and lexical analysis but also an evaluation of the underlying informational and conceptual density [4].

Automatically determining whether a textbook's content is appropriate for a specific educational level remains challenging. Although traditional readability metrics such as the Flesch–Kincaid Reading Ease score [5] have been widely used, they rely on simplified formulae based primarily on surface-level linguistic features, including sentence length, word length, and vocabulary frequency [6]. While useful, these measures capture only a narrow dimension of textual complexity and fail to account for the informational and conceptual demands embedded within educational content.

Large Language Models (LLMs) can address a broad spectrum of tasks that involve both linguistic and content-related challenges. They achieved high accuracy in explaining complex concepts and answering domain-specific questions [7,8]. The rapid advancement of LLMs has significantly accelerated research activity, with new applications emerging at an unprecedented pace. While their integration into education is still in its early stages, existing studies have explored diverse use cases, including automated essay evaluation [9], personalized content generation [10], and the deployment of conversational teaching assistants [11].

This paper explores and evaluates the potential of Large Language Models (LLMs) in assessing the complexity of Russian-language textbooks. Rather than replacing traditional static readability metrics, these models are proposed as complementary tools that leverage their broad linguistic capabilities to offer deeper, more context-aware insights. While our primary focus is on Russian educational materials, the findings are broadly applicable given the multilingual architecture of current LLMs. Our study centers on OpenAI's GPT-4o, although the methodologies presented are model-agnostic and can be adapted to other LLMs with comparable capabilities. To the best of our knowledge, this study is the first to systematically apply large-scale decoder LLMs rather than encoder models to assess the complexity and readability of Russian-language textbooks.

Research Objectives

Understanding whether LLMs can reliably assess the complexity of certain texts is crucial for their potential use in automating text selection or evaluation. This cannot be achieved without identifying the key influential features that lead the LLM to one conclusion or another. Evaluating its biases or shortcomings might reveal whether its assessment aligns with established linguistic or pedagogical principles of text complexity.

This study aims to conduct a comprehensive evaluation of LLMs' capability to evaluate the complexity and cognitive accessibility of Russian educational texts. Our analysis is structured around the following three key research questions:

- **RQ1**: To what extent do LLMs correctly assess the complexity of a Russian-language text? What are the key features that influence the decision?
- **RQ2**: Can LLMs be used as effective judges (or proxies) of the ease of understanding of Russian schoolbooks, while simulating how a student might understand the material? What are the key features that influence the LLM's decision regarding the ease of understanding?
- **RQ3**: Can LLMs successfully reduce the complexity of a schoolbook text to a certain level of comprehension (in our specific case, by 3 school years)? Can this be achieved across different school subjects?

Addressing these research questions provides insights into the capabilities and limitations of LLMs in analyzing and modifying Russian educational texts. The results offer valuable guidance to educators, researchers, and developers regarding the practical applications and constraints of these models in generating and assessing appropriately leveled learning materials within the Russian education framework.

2. Related Work

There has been a long line of research aimed at identifying and validating predictors of text complexity. Each of the five paradigms of discourse complexology, i.e., formative, classical, closed tests, constructive-cognitive, and Natural Language Processing, experimented with a new approach expanding the earlier models [12] In their systematic review, AlKhuzaey et al. [13] observed that traditional methods for assessing text difficulty predominantly rely on syntactic features, which are usually extracted using NLP tools—such as counts of complex words [14], sentence length [15], and overall word count [16]. In contrast, recent research has increasingly turned to semantic-level analysis, enabled by advances in semantically annotated data structures such as domain ontologies and the emergence of neural language models. Among the earliest and most influential surface-level approaches are the Flesch Reading Ease [5] and Flesch–Kincaid readability score [17], which estimate readability based on surface indicators such as orthographic (e.g., number of letters), syntactic (e.g., number of words), and phonological (e.g., number of syllables) features. In this framework, a higher score reflects greater difficulty. However, Yaneva et al. [18] highlighted the limitations of these traditional metrics, arguing that Flesch-based scores are weak predictors of actual difficulty, as they fail to distinguish between easy and difficult content based on lexical features alone.

The emergence of neural language models has transformed the landscape of automatic text complexity assessment. Trained on large-scale corpora, these models can capture not only syntactic and semantic patterns but also elements of world knowledge, allowing for deeper interpretation of textual content. Modern readability estimation methods now frequently use Transformer-based architectures such as BERT and RoBERTa, which have superseded earlier neural models, like recurrent networks. For instance, Imperial [19] proposed a hybrid model combining BERT embeddings with traditional linguistic features, achieving a notable 12.4% improvement in F1 score over classic readability metrics for English texts, and achieving robust performance on a low-resource language (i.e., Filipino). Similarly, Paraschiv et al. [20] applied a BERT-based hybrid model within the ReaderBench framework to classify the difficulty of Russian-language textbooks by grade level, achieving an F1 score of 54.06%. The use of multilingual transformer models such as mBERT, XLM-R, and multilingual T5 has further expanded the applicability of these methods across languages. The ReadMe++ benchmark introduced by Naous et al. [21] includes nearly 10,000 human-rated sentences across five languages—English, Arabic, French, Hindi, and Russian—spanning various domains. Their findings reveal that fine-tuned LLMs exhibit stronger correlations with human judgments than traditional scores such as Flesch–Kincaid and unsupervised LM-based baselines.

With the emergence of very large generative models like GPT-3 and GPT-4, another convenient method is to prompt the model to evaluate a text's readability or generate a difficulty rating. In this approach, we consider the LLM as an expert evaluator: we feed it a passage and ask (in natural language) for an analysis of how easy or hard that passage would be for a certain reader group [22,23].

In addition, LLMs have also been employed for generating or rewriting texts at specified reading levels, mostly in English—an inverse but closely related task to readability evaluation [24,25]. The capacity to simplify a complex passage requires an underlying

model of readability, as it requires the LLM to recognize what makes a text difficult and how to reduce that difficulty while preserving meaning. Huang et al. [26] introduced the task of leveled-text generation, in which a model is given a source text and a target readability level (e.g., converting a 12th-grade passage to an 8th-grade level) and asked to rewrite the content accordingly. Evaluations of models such as GPT-3.5 and LLaMA-2 on a dataset of 100 educational texts, measured using metrics like the Lexile score, revealed that while the models could generally approximate the target level, human inspection identified notable issues, including occasional misinformation and inconsistent simplification across passages. In parallel, Rooein et al. [27] introduced PROMPT-BASED metrics, an evaluation method that uses targeted yes/no questions to probe a model's judgment of text difficulty. When integrated with traditional static metrics, this approach improved overall classification performance.

Further, Gobara et al. [28] argued that instruction-tuned models show stronger alignment between user-specified difficulty levels and generated outputs, having a stronger influence than the model size. In contrast, Imperial and Tayyar Madabushi [29] highlighted the limitations of proprietary models like ChatGPT (gpt-3.5-turbo and gpt-4-32k) for text simplification, showing that, without carefully crafted prompts, their outputs lagged behind open-source alternatives such as BLOOMZ and FlanT5 in both consistency and effectiveness.

3. Method

3.1. Russian Schoolbook Corpus

We base our analysis on a corpus of 154 Russian textbooks curated by linguistic experts at the Multidisciplinary Investigations of Texts Laboratory, Institute of Philology and Intercultural Communication, Kazan, in 2023 [3]. This corpus spans 13 key subjects and covers grade levels 2 through 11. However, due to the structure of the Russian educational curriculum and the availability of digitized textbooks, certain subjects are not represented at all grade levels. For example, subjects like Mathematics, Science, and Technology are absent in this corpus beyond the 4th grade. In contrast, others (i.e., Biology, Geography, History, and Physics) are introduced only in higher grades. The textbooks selected for this study are approved by the Ministry of Education and Science of the Russian Federation and are part of the standard curriculum for secondary and high schools.

Table 1 provides a detailed breakdown of the topics and the number of textbooks available for each grade level. In it, we can observe the uneven distribution of subjects across textbooks. Notably, subjects such as Arts and Music are only offered in lower grades, while subjects such as Biology and Physics are predominantly offered in higher grades. In particular, Biology for grades 10 and 11 presents a unique case, as all three textbooks used in these grades were evenly distributed between both levels. Since a clear separation between the textbooks for these two grades was not possible, we grouped all three under 10th grade.

Furthermore, Table 2 provides a detailed analysis of the distribution of paragraphs, phrases, and words across textbooks by grade level. Phrase segmentation was determined using punctuation markers, while words were separated by either whitespace or punctuation. This data reveals the increasing complexity and length of the texts as students progress through the grades. For example, while the average number of paragraphs remains relatively consistent throughout the dataset, the number of phrases and words shows a noticeable progression, in step with the increasing linguistic and conceptual demands placed on students as they advance in their education.

Table 1. Class distribution for subjects after splitting the documents using the two approaches (# denotes counts per subject).

Subject	# Textbooks	Paragraph Split	"Optimal Split" with Textsplit
Art	4	2306	143
Biology	23	29,612	3247
Ecology	4	2248	173
Geography	8	8510	828
History	17	21,226	3379
Informatics	19	11,277	608
Maths	20	33,618	1493
Music	4	1212	79
Physics	19	3690	627
Russian	13	63,819	3200
Science	6	9906	685
Social studies	2	15,845	1885
Technology	24	5032	332

Table 2. Average number of paragraphs, phrases, and words per textbook and grade (# denotes counts per grade level).

Grade Level	# Books	# Paragraphs	# Phrases	# Words
2nd Grade	24	1203	1837	13,126
3rd Grade	22	1758	2562	19,951
4th Grade	27	1960	2840	24,500
5th Grade	15	1665	3214	31,015
6th Grade	11	1090	2509	29,807
7th Grade	15	1751	3847	42,815
8th Grade	13	1377	3558	43,766
9th Grade	10	1457	3252	43,768
10th Grade	10	1552	4630	68,930
11th Grade	7	1725	4941	72,857

3.2. Extractive Summarizing of Textbooks

With more than 15,000 paragraphs across the entire corpus, compute time and costs can quickly become prohibitive. To mitigate this issue, we applied extractive summarization, selecting the 10 most representative segments from each textbook. Unlike abstractive methods, the extractive approach preserves the original text complexity within the chosen segments.

The summarization process involves an initial pass through the textbook, segmenting the content into discrete paragraphs. Each paragraph was subsequently embedded using the *cointegrated/LaBSE-en-ru* (https://huggingface.co/cointegrated/LaBSE-en-ru, accessed on 28 November 2025) model, which is specifically fine-tuned for sentence embeddings in the Russian language. To address potential disruptions in text cohesion caused by simple paragraph-based splitting, we apply the *split_optimal* method from the textsplit library [30]. This approach ensured summarization coherence by grouping adjacent text fragments with high cosine similarity. Following this re-segmentation, we exclude segments with fewer than 100 tokens and further split segments exceeding 512 tokens.

After the segmentation phase, centrality scores for each segment were computed using the LexRank algorithm [31], a graph-based extractive summarization algorithm that relies on PageRank [32]. In extractive summarization, the most important content is often defined by centrality, reflecting the intuition that the sentences or segments most similar to the rest of the document embody its core meaning [33]. In particular, LexRank pinpoints the salient sentences that collectively represent the original text [34].

Using cosine similarity as edge weights, the model ranks all fragments based on their connections. A highly connected segment (similar to many other segments or to highly ranked segments) accumulates a high centrality score. In effect, the PageRank-like computation will elevate segments by how representative they are of the overall text. Segments that cover content common to multiple parts of the document naturally rise to the top, as they are strongly connected to the rest of the document through content similarity. After processing each document, the ten segments with the highest centrality for each textbook were chosen as the summary.

The selected key segments serve as reliable proxies for the textbook's overall complexity level. By using these central text fragments, our approach ensures that complexity assessments remain both representative of the document and computationally efficient. This enabled us to significantly reduce the volume of tokens required for further processing by LLMs, resulting in lower computational overhead and reduced operational costs.

To deliver a qualitative dimension to our analysis, we visualized the key segments within the embedding space (see Figure 1). We can observe that the selected key phrases are well distributed across the embedding space, capturing a broad range of thematic and semantic aspects of the textbook. This supports the claim that our methodology effectively identifies central passages, ensuring that the summary accurately reflects the text's complexity and thematic focus.

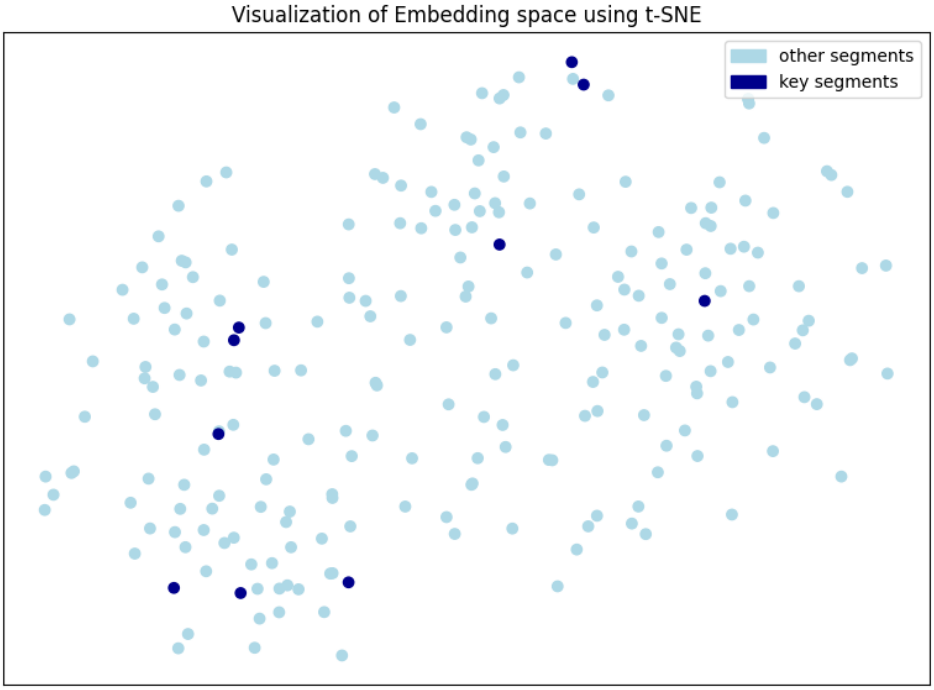


Figure 1. The graph displays the distribution of text segments in the embedding space for the 4th Grade Math schoolbook. Light blue points represent general segments, and dark blue points indicate the key segments selected through extractive summarization.

3.3. Assessing the Readability with LLM Prompting

LLMs have increasingly been employed to assess text readability and evaluate reading comprehension [27,29]. Unlike traditional readability metrics, such as Flesch–Kincaid Reading Ease score [17], LLMs can exhibit intrinsic linguistic understanding and can offer more nuanced assessments. These methodologies can broadly be categorized into three distinct approaches: LLMs as Readability Evaluators [21], Prompt-Based Comprehension Testing [35,36], and Automated Response Grading and Summarization [9].

The first approach involves using LLMs as evaluators to determine the reading level or complexity of a passage. For example, an LLM can be prompted to assess the complexity of a text, allowing for a nuanced evaluation of whether a topic or prior knowledge requirements might increase the difficulty, areas where traditional metrics may fall short [27].

In the second approach, LLMs are prompted to answer questions based on the text, simulating how a student's comprehension would be tested. This method uses the LLM as a stand-in for a student [37,38].

The final approach employs LLMs to evaluate short-answer or essay responses, assessing both correctness and comprehension. In this role, the LLM acts as a proxy for a teacher, providing a detailed evaluation in natural language that goes beyond assigning a simple numerical score. It can replicate the qualitative judgment a teacher might apply, checking that important information is not missing or distorted [9,39,40].

In this study, we employ a hybrid methodology that leverages all three strategies, first by prompting the LLM to determine the target grade level for each text fragment, producing a numeric prediction between grades 2 and 11. To gain a deeper understanding of the LLM's decision-making process, we request the extraction of 5 key phrases that highlight the text's complexity. Additionally, we simulate a comprehension assessment by prompting the LLM to adopt the perspective of an n-th grade student and evaluate whether the fragment would be accessible to students at that educational level. Finally, we instruct the model to simplify texts designed for grades 5 through 11 to assess its capability to reduce text complexity by 3 grade levels.

In addition to these methods, we also measure the model's perplexity during response generation. Perplexity is a quantitative measure of the model's capability to process long-form texts. Generally, a lower perplexity score indicates better language modeling performance, suggesting the model generates coherent, contextually appropriate responses. This metric has proven effective in evaluating LLMs' capability to manage and interpret lengthy text passages [41].

3.4. Experimental Setup

To address our research questions, we designed a series of experiments that investigate the effectiveness of LLMs in evaluating and understanding the complexity of Russian schoolbook texts. By leveraging the previously mentioned key sequences extracted from the textbook corpus, we devised 3 experiments to highlight the model's abilities. These experiments were conducted using GPT-4o with a temperature setting of 0.5, ensuring a balanced approach to creativity and response consistency. We accessed GPT-4o via the official OpenAI API endpoint, programmatically submitting prompts and retrieving completions using a predefined template for each experiment. This template ensured that the model consistently adhered to structured output expectations, delivering answers in a predefined JSON format.

The first experiment aimed to evaluate the LLM's capability to accurately assess the complexity of Russian educational texts. We prompted the model with the following query: *"For what age would this Russian-language school text be suitable, assuming that the reader is a native Russian speaker? Additionally, extract a list of five key phrases that best pinpoint the complexity level of the text. Write your answer as a JSON with the following structure: {'age': 0, 'phrases': ['..']}."* This task not only required the LLM to predict the grade level, ranging from 2 to 11, but also to provide insights into its decision-making process by highlighting the specific linguistic or thematic elements that influenced its judgment. By examining the extracted key phrases, we aim to reveal the features the model uses to assess text complexity and to uncover how closely the LLM's assessments align with established educational standards.

In the second experiment, we shifted the focus to the evaluation of text cognitive accessibility from a student's perspective. The LLM was prompted with the following instruction: *"This text is in your [Subject] class schoolbook. Is this text comprehensible for you? Answer in English. Let's think step by step. It's crucial that the final sentence contains your conclusion in the form 'Answer: yes/no'."* Here, the model was instructed to simulate the thought process of a student, critically analyzing the text's clarity and accessibility. By instructing the LLM to adopt a step-by-step approach, we encouraged it to articulate the rationale behind its conclusions [42], enabling us to assess not only the final answer but also the reasoning path leading to it. This approach aimed to validate whether the model could effectively serve as a proxy for human judgment in educational settings, identifying specific features that contribute to or hinder comprehension.

The third experiment extended our analysis to the domain of text simplification. We explored whether the LLM could actively reduce the complexity of schoolbook texts to match a lower grade level, aiming to reduce the complexity by approximately three school years. The prompt provided to the model was: *"Rephrase this Russian text in a way that would be comprehensible for a Russian student in the 6th grade (12 years old). Reply in Russian. No other comments."* This experiment required the model to transform the original text while preserving essential information, making sure that its content remained valuable yet accessible to younger students. In this experiment, we specifically targeted texts from Russian Language, Social Studies, and History, since these subjects rely more on lexical complexity, unlike subjects such as Mathematics or Informatics, where structural and formulaic complexity play a more dominant role. Additionally, these subjects span a broader range of grade levels than most other disciplines. By analyzing simplified outputs, we aimed to determine whether LLMs could be used to develop simplification strategies across different subjects while still maintaining the intended educational outcomes.

In conjunction, these experiments offer a multifaceted evaluation of LLMs in the context of Russian-language textbook evaluation and facilitate the analysis of LLMs' immediate effectiveness and the exploration of their broader potential to support teachers and educators in developing age-appropriate and pedagogically sound materials.

4. Results

In this section, we present a comprehensive analysis of the LLM's capability to assess the reading complexity of Russian educational texts across multiple dimensions. Our evaluation is structured around three core experiments designed to probe different facets of model performance: age prediction accuracy, subjective cognitive accessibility, and the ability to simplify a text. The aim is not only to evaluate the model's raw prediction accuracy, but also to explore the underlying cognitive and linguistic factors that shape these predictions—particularly in the context of how LLMs perceive and assess educational difficulty.

4.1. Experiment 1—Comprehension Age Prediction

In the first experiment, we evaluate the large language model's ability to predict the target comprehension age for Russian schoolbook texts across various academic subjects. This assessment serves as a baseline for understanding how well the model aligns with age-appropriate educational standards in the Russian curriculum.

To this end, we evaluate the performance of the LLM in predicting the age of comprehension for Russian schoolbook texts, based on their lexical complexity and information content (Table 3).

Table 3. Performance Metrics by Subject over the whole dataset.

Subject	MSE	MAE	Mean Prob (%)	Mean Joint Prob (%)	Mean Perplexity	Count
Art	3.46	1.42	61.07	55.02	1.09	50
Biology	2.08	1.17	59.47	52.26	1.09	230
Ecology	1.70	1.03	56.83	51.43	1.10	30
Geography	1.63	0.90	57.82	51.56	1.08	60
History	2.64	1.24	54.90	48.51	1.10	190
Informatics	4.22	1.73	59.26	53.23	1.10	120
Maths	1.25	0.72	61.21	55.17	1.10	160
Music	5.17	1.93	66.12	58.59	1.11	40
Physics	3.05	1.32	54.59	42.94	1.16	60
Russian	2.34	1.14	55.93	50.44	1.11	250
Science	2.31	1.14	68.47	61.53	1.09	70
Social Studies	2.25	1.19	57.40	50.87	1.10	160
Technology	1.61	0.96	60.25	51.81	1.09	120

In Table 3, we evaluate the performance of the LLM in predicting the age of comprehension for Russian schoolbook texts, based on their lexical complexity and information content. Overall, the LLM reasonably estimated the target age group for the presented texts, with predictions generally aligning with the intended grade levels. Our findings indicate that the mean absolute error (MAE) across most subjects is approximately one year. Considering the natural overlap of educational content between adjacent grade levels, this degree of error is within a reasonable and acceptable range.

Figure 2 illustrates the relationship between the predicted and actual comprehension ages for each text segment, with marker size proportional to the frequency of that (actual, predicted) pair. The red LOESS smoothing curve depicts the overall trend, while the dashed diagonal line (y = x) represents perfect predictions. The plot shows that GPT-4o's predictions broadly follow the expected grade levels but exhibit a systematic bias. It is overestimating the difficulty of texts for younger students and underestimating the target age for more complex fragments, consistent with the quantitative results reported above.

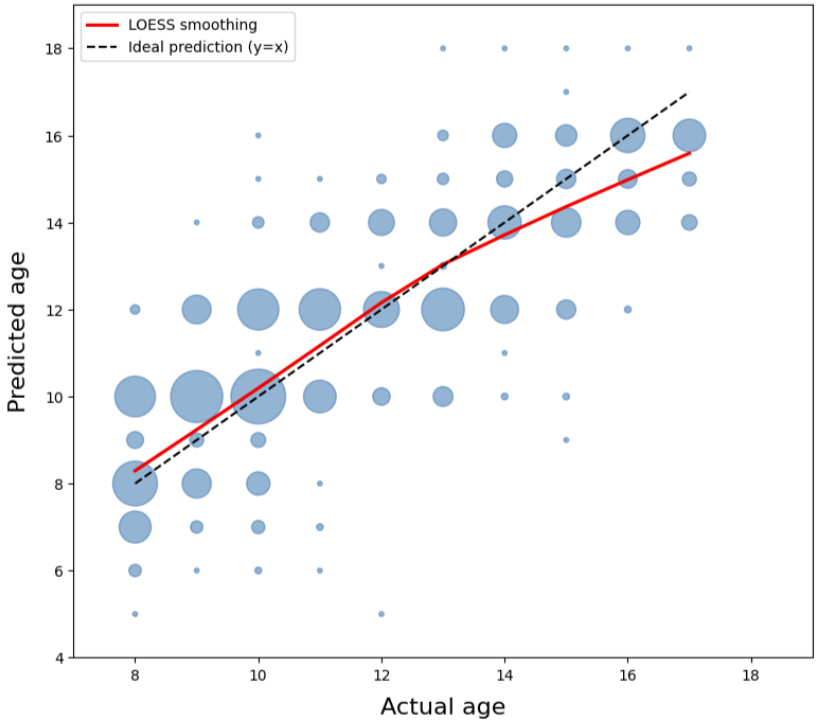


Figure 2. Relationship between predicted and actual comprehension ages across Russian school textbook fragments. The marker size is proportional to the frequency of the "actual age"–"predicted age" pair.

We analyzed the log probabilities of predicted output tokens to quantify the model's confidence, specifically examining the probability associated with the token that encodes the predicted age and the cumulative joint probability of the preceding tokens. The joint probability was computed as the exponential of the sum of all log probabilities from the beginning of the generated sequence up to and including the age token, effectively capturing the total likelihood of the model producing that specific prediction path. This measure provides a more expressive indicator of confidence, as it reflects not only the certainty of the final numeric token for the age but also the probability associated with the reasoning chain leading to it. The mean probability derived from this analysis reflects the LLM's confidence in its predictions. Interestingly, the table shows that while the Music subject shows an MAE approaching 2 years, the model maintains a high degree of confidence in its predictions; nevertheless, these textbooks have the shortest history and tradition and, as such, lack consistency. Moreover, this observation suggests that the model's confidence does not always align with actual prediction accuracy, particularly in subjects with less structured content.

The relatively high error rate in Informatics is understandable, given the advanced technical concepts that may not align well with the cognitive level of younger students. However, the significantly higher error rates for Art and Music are less intuitive and require further investigation. This phenomenon may be partially explained by Solovyev et al. [3], Solnyshkina et al. [43], who noted that 5th-grade textbooks sometimes contain a higher density of specialized terms than those for higher grades. Future research could examine whether this discrepancy arises from difficulties in evaluating creative, subjective content or from biases in the training data.

One notable trend was the LLM's systematic overestimation of text complexity. On average, errors were predominantly skewed towards assigning a higher age group than expected when the model's predictions had an MAE over 1 (see Figure 3). This tendency suggests a conservative bias in the model's complexity assessments, potentially reflecting a cautious interpretation of challenging vocabulary or abstract concepts within the texts. In the case of higher error subjects, such as Art, Informatics, Music, and Physics, this skew is more pronounced (see Figure 4).

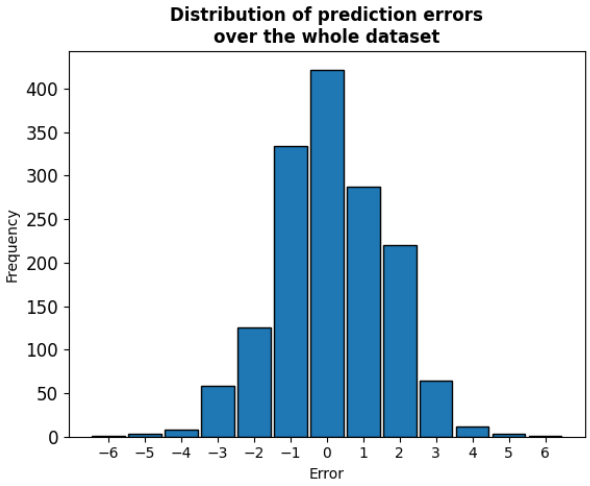
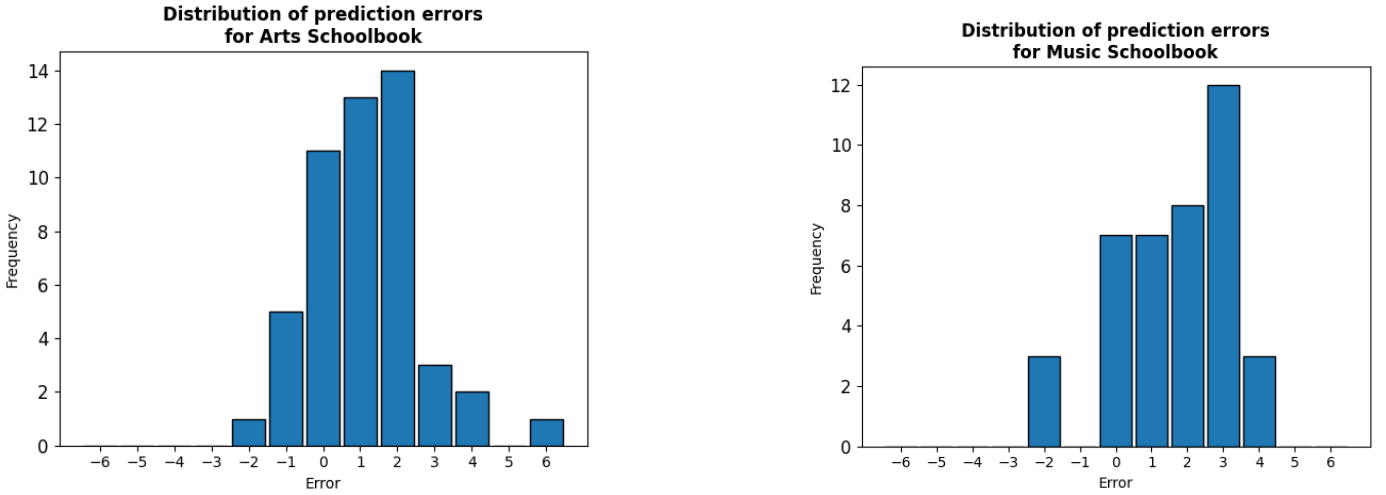


Figure 3. Overall distribution of prediction errors for comprehension ages over all school textbooks.



(**a**) (**b**)

Figure 4. Error distributions of comprehension age predictions for (**a**) Art and (**b**) Music textbooks.

Focusing on the relationship between the model's confidence levels and its prediction accuracy, we extract only the segments with an MAE of 2 years or higher in Table 4. Here, we also notice no strong correlation between confidence and correctness. Even when the model produced erroneous age predictions, its confidence scores remained comparable to or even higher than the average. Physics was the only subject in which a notable drop in confidence was consistently observed for incorrect predictions, suggesting a degree of uncertainty in the model's processing of physics-related material that is less evident in other subjects.

Table 4. Mean Probability for erroneous predictions with an age difference higher than 2 years.

Subject	Mean Prob (%)	Mean Joint Prob (%)	Mean Perplexity	Count	Percentage
Art	59.22	53.41	1.09	21	42.00
Biology	55.90	48.02	1.11	59	25.65
Ecology	70.05	65.15	1.09	8	26.67
Geography	44.87	40.38	1.07	17	28.33
History	46.14	41.36	1.10	71	37.37
Informatics	59.91	54.26	1.10	68	56.67
Maths	61.38	54.69	1.10	34	21.25
Music	66.63	60.89	1.10	26	65.00
Physics	45.72	27.76	1.24	15	25.00
Russian	54.07	47.90	1.11	62	24.80
Science	70.61	63.84	1.09	28	40.00
Social Studies	54.37	48.17	1.09	53	33.13
Technology	57.83	48.66	1.09	35	29.17

The mean perplexity scores were relatively uniform across all subjects. However, an increase in perplexity was observed alongside erroneous predictions in Physics texts. This elevated perplexity suggests that the model found these texts particularly challenging to process, potentially contributing to its reduced accuracy and confidence in this subject area.

Table 5 analyzes extreme outliers identified within the prediction dataset. Observations indicate that these outliers predominantly belong to the subjects of Russian, Physics, and Informatics. Notably, the elevated Mean Absolute Error (MAE) values for Physics (1.32) and Informatics (1.73) suggest difficulties in accurately assessing the appropriate comprehension age for these subjects. Contrastingly, Russian exhibits a lower MAE, despite its frequent appearance among the outliers.

Table 5. Prediction results for different subjects, including predicted age, probability scores, perplexity, and error.

Subject	Grade	Age	Predicted Age	Pred. Probability	Joint Probability	Perplexity	Error
Art	04	10	16	10.24	9.45	1.05	6
Informatics	04	10	15	19.36	17.88	1.08	5
Informatics	03	9	14	50.04	45.24	1.06	5
Physics	09	15	9	9.08	0.00	1.50	−6
Physics	09	15	10	92.24	0.00	1.67	−5
Russian	09	15	10	55.04	53.41	1.09	−5
Russian	05	11	6	36.32	34.55	1.41	−5
Russian	06	12	5	58.86	54.38	1.06	−7
Social Studies	07	13	18	82.08	72.25	1.05	5

To better understand the underlying problem, we focus on the 3 Russian-language outliers (see Figure 5). All 3 texts are Literary and, in isolation, might appear of much lower complexity than the intended age target. The first example is a narrative excerpt featuring a dialogue between a forester and the narrator about spring. The vocabulary is mostly accessible, and the dialogue structure aids comprehension. The second text is a famous example of grammatically structured nonsense, created by the writer Ludmila Petrushevskaya to illustrate linguistic principles. The vocabulary is almost entirely composed of nonsense words, but the syntax follows standard Russian patterns. Due to its playfulness, the text might be regarded as targeting a lower age than the textbook was designed for. We also observe a similar case for the third text, a lullaby poem by Vasily Lebedev-Kumach. If we take it by itself, without the literary analysis in the textbook, the poem is indeed aimed at a much younger audience. These instances underscore a limitation inherent in text segmentation approaches: the partial removal of surrounding context, particularly the pedagogical framework (e.g., analytical tasks), can lead to inaccurate complexity assessments and produce such outliers.

	Original Russian Text	English Translation
Grade: 09 Age: 15 Predicted Age: 10	Родник! — сказал лесник, глядя, как из оконца всплыл и тотчас пошёл на дно неистово барахтавшийся жук. — Должно, Волга тоже начинается из такого оконца? Да, должно быть, — согласился я. Я большой любитель разбирать слова, — неожиданно сказал лесник и смущённо усмехнулся. — И вот скажи на милость! Бывает же так, что пристанет к тебе одно слово и не даёт покоя. А какое слово к вам привязалось сейчас? — спросил я. Да вот этот самый «родник».	"A spring!" said the forester, watching a beetle, struggling furiously, floating to the surface from the window and immediately sinking to the bottom. "The Volga must also start from a window like that? Yes, it must," I agreed. "I'm a big fan of analyzing words," the forester suddenly said and smiled sheepishly. "And tell me, please! It happens that one word sticks to you and doesn't give you peace. And what word has stuck to you now?" I asked. "Yes, this very 'spring.'
Grade: 05 Age: 11 Predicted Age: 6	Пуськи бятые. Сяпала Калуша по папушке и увозила бутявку. И волит: — Калушата, калушаточки! Бутявочка! Кал ушата присяпали и бутявку стрямкали. И подудонились. А Калуша волит: — Оее, оее! Бутявка-то некузявая! Калушата бутявку вычучили. Бутявка вздребезднулась, сопритюкнулась и усяпала с напушки. А Калуша волит: — Бутявок не трямкают. Бутявки дюбые и зюмо-зюмо некузявые. От бутявок дудонятся. А Бутявка волит за напушкой: — Калушата подудонились! Калушата подудонились! Зюмо некузявые? Пуськи бятые!	“Byuski Byata” is a fairy tale written in an invented language resembling children’s speech or modern youth slang. Although each individual word is nonsensical, the overall plot remains easily comprehensible.
Grade: 06 Age: 12 Predicted Age: 5	Сон приходит на порог, Крепко, крепко спи ты! Сто путей, сто дорог Для тебя открыты. Всё на свете отдыхает, Ветер затихает, Небо спит, солнце спит, И луна зевает. Спи, сокровище моё, Ты такой богатый; Всё твоё, всё твоё — Звёзды и закаты. Завтра солнышко проснётся, Снова к нам вернётся. Молодой, золотой, Новый день начнётся. Чтобы завтра раньше встать Солнышку навстречу, Надо спать, крепко спать, Милый человечек. Спит зайчонок и мартышка, Спит в берлоге мишка, Дяди спят, тёти спят, Спи и ты, малышка.	Dreams come to the door, Sleep soundly, soundly! A hundred paths, a hundred roads are open to you. Everything in the world is resting, The wind is dying down, The sky is sleeping, the sun is sleeping, And the moon is yawning. Sleep, my treasure, You are so rich; All is yours, all is yours Are the stars and the sunsets. Tomorrow the sun will wake up, Will return to us again. Young, golden, A new day will begin. To get up early tomorrow To meet the sun, You need to sleep, sleep soundly, Dear little man. The bunny and the monkey are sleeping, The bear is sleeping in the den, men are sleeping, women are sleeping, Sleep too, baby.

Figure 5. Outliers in the Russian-language subject with an error in prediction of over 4 years.

To gain deeper insights into the linguistic features that influence the model’s decisions, we analyzed the key phrases the LLM extracted during its assessments. As Viswanathan et al. [44] and Tarekegn [45] show, LLMs are highly effective at producing contextually relevant keyphrase expansions, which can enhance document clustering. To structure and interpret these key phrases, we applied the BERTopic framework [46], which

clusters short texts based on semantic similarity using transformer-based embeddings. This method yielded approximately 39 distinct clusters (see Figure 6), revealing patterns in the model's interpretive framework. Each cluster was automatically labeled by BERTopic using the most representative features—typically high-frequency or contextually significant n-grams within the cluster—providing interpretable topic names. Notably, many top clusters were closely tied to specific subjects, indicating that the model associates particular linguistic or conceptual motifs with complexity in those domains. Prominent examples of these clusters include "Interconnected Aspects of Life, Nature, and Human Culture", "Historical and Cultural Dynamics of Russia and Europe", "Intersections of Nature, Communication, and Society", "Challenges of Warfare, Disease, Substance Abuse, and Social Conflict", "Quantitative Descriptions and Numerical Comparisons Across Various Contexts".

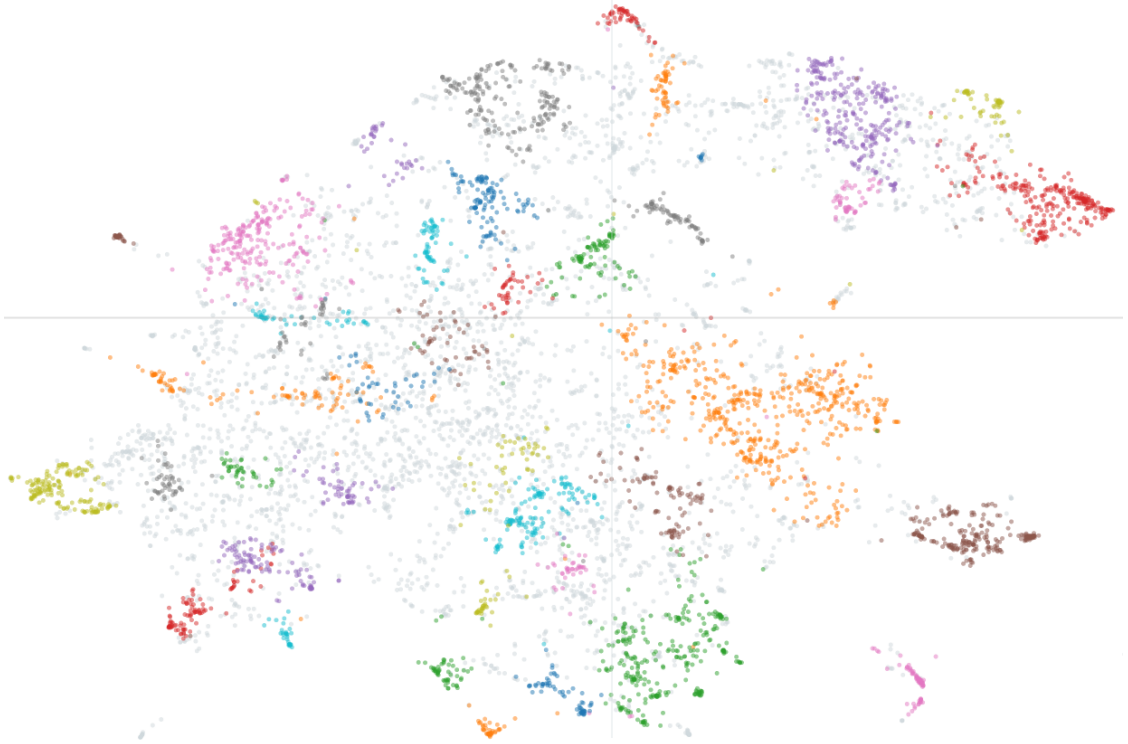


Figure 6. Key phrases clustering that underlines the complexity of the fragments, as extracted by the LLM. Each color represents a detected cluster.

These findings suggest that the model's complexity judgments are not based solely on surface-level metrics such as sentence length or vocabulary difficulty, but also involve a deeper analysis of thematic and contextual elements. The identified clusters indicate that the themes, topics, and underlying narrative influence the LLM's decision.

4.2. Experiment 2—Ease of Understanding Classification

In our second experiment, we prompted the LLM to assess the ease of understanding the selected fragments. Table 6 shows that, for most subjects, the fragments are evaluated as being comprehensible. This finding suggests that the model is generally aligned with a student's perspective when assessing the clarity and accessibility of educational content.

Table 6. Ease of understanding Assessment by Subject. High incomprehensible rates (> 20%) marked bold.

Subject	Comprehensible	Incomprehensible	Incomprehensible Rate
Art	46	4	8.70%
Biology	382	78	**20.42%**
Ecology	34	6	17.65%
Geography	106	14	13.21%
History	314	66	**21.02%**
Informatics	113	25	**22.12%**
Maths	129	13	10.08%
Music	32	6	18.75%
Physics	101	18	17.82%
Russian	243	84	**34.57%**
Science	66	3	4.55%
Social Studies	273	46	16.85%
Technology	113	7	6.19%

However, there were notable exceptions to this trend. Texts related to Russian, Informatics, Biology, and History exhibited higher comprehension difficulty scores. These subjects pose unique challenges, possibly due to more advanced language requirements or technical terminology that students at lower grade levels may find difficult to grasp.

Interestingly, the relationship between the model's prediction errors in Experiment 1 and the ease of understanding assessments in Experiment 2 was not straightforward. Not all subjects with high age prediction errors (e.g., Art, Music) showed similarly low intelligibility. This discrepancy indicates that while the model might overestimate the complexity of certain texts, it does not necessarily perceive them as incomprehensible, suggesting the presence of some biases in its training data.

To gain deeper insights into the model's decision-making process, we specifically analyzed the conclusion the LLM reached in its chain of thought for cases in which texts were deemed incomprehensible. By applying clustering techniques to these concluding statements, we identified seven distinct clusters, each reflecting specific factors that might hinder comprehension:

- Challenging Texts for Young Students: Complexity in Language, Concepts, and Contexts;
- Challenges in Comprehending Complex Instructions and Concepts: Across Educational Levels;
- Challenges of Understanding Complex Language and Concepts: At Elementary and Middle School Levels;
- Complexity of Educational Texts for Early Grade Students: Challenges in Comprehension and Vocabulary;
- Challenges of Complex Language and Concepts for Young Russian Students
- Complex Topics and Vocabulary: Beyond the Understanding of Young Children;
- Comprehension with Guidance and Instructional Support.

Similar to the first experiment, these clusters indicate that the model's judgments are influenced not only by linguistic difficulty but also by the concepts and contextual demands presented by the documents. Particularly, the clusters emphasize challenges related to abstract language, complex vocabulary, complex instructions, and the need for guidance and support to facilitate understanding. By identifying these themes, we provide valuable guidance for future research, particularly in developing educational tools to support students at different cognitive and developmental stages.

4.3. Experiment 3—Academic Text Simplification with LLMs

Using a previously trained BERT classification model [20], we assessed the generated texts to determine whether the targeted complexity reduction was achieved, namely a shift of 3 school years. The confusion matrix for the model's prediction (see Figure 7) indicates a significant performance gap, with the classification model achieving very low accuracy on the generated set. The overall accuracy of the BERT model on the generated texts was 14.88%, reflecting the difficulty in aligning the generated text complexity with the targeted grade levels. The very low Recall in almost all target grades further highlights the significant discrepancy between the intended and achieved complexity levels.

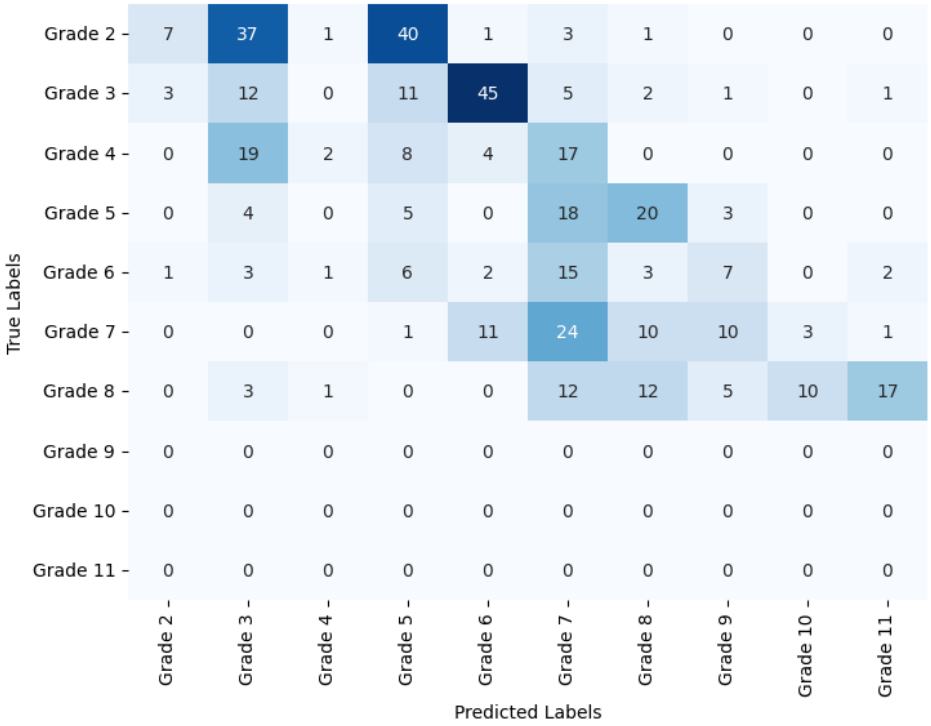


Figure 7. Confusion matrix of the BERT-based predictions on the simplified texts (higher color intensity for higher agreement). As the simplification aimed to reduce complexity by three grade levels, no true labels exist for grades 9–11. Nevertheless, BERT assigns a number of samples to these higher-grade categories.

One particular observation is that most prediction errors aligned precisely with a three-year gap (see Figure 8), the exact reduction target. This pattern implies that the LLM largely failed to effectively lower the complexity of the texts. Instead of adapting the material to the intended, simpler level, the model maintained complexity closer to the original grade level, suggesting a potential limitation in its simplification capabilities or in its interpretation of the prompt.

To further validate these findings, we employ the Textometr backend as described by [47], with added use of the internal *formula_pushkin_100* metric3—a composite measure of grammatical complexity (*structure_complex*), lexical coverage (*lexical_complex*), and *narrativity*. While the 2021 paper explains the ingredients of these features (average sentence length, lexical list coverage, Flesch-style readability, passive/participle counts), it does not formally publish the aggregation formula above. We nevertheless use this formula, as it maps to the well-known Flesch–Kincaid readability score (ranging from 0 to 100), has a straightforward open-source implementation, and serves as the foundation of the Russian Text Complexity Analyzer developed by the Pushkin State Russian Language Institute (https://textometr.ru, accessed on 28 November 2025). Interestingly, according to this

metric, the LLM performed bette, with an error distribution similar to that when applied to the original texts (see Figure 8). When examining the score distributions, we observed a notable leftward shift in the mean scores from 60.52 to 37.19, indicating that the model did achieve a reduction in complexity, albeit not to the full extent initially targeted (see Figure 9). The Wilcoxon signed-rank test revealed a statistically significant downward shift in the distribution of scores for the newly generated texts relative to the Pushkin 100 scores from the original text samples ($p < 0.001$). This shift suggests that while the model may not have precisely met the specified grade-level adjustment, it did contribute to a general simplification of the texts.

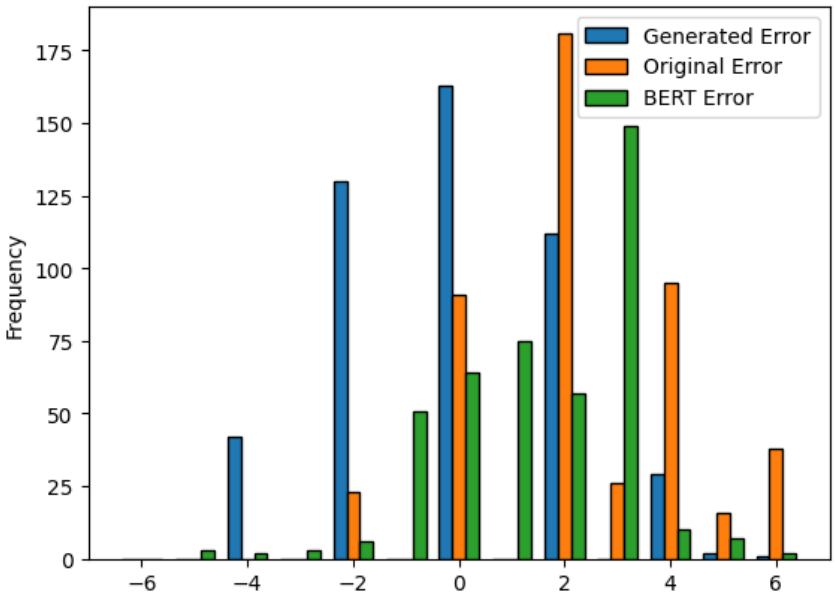
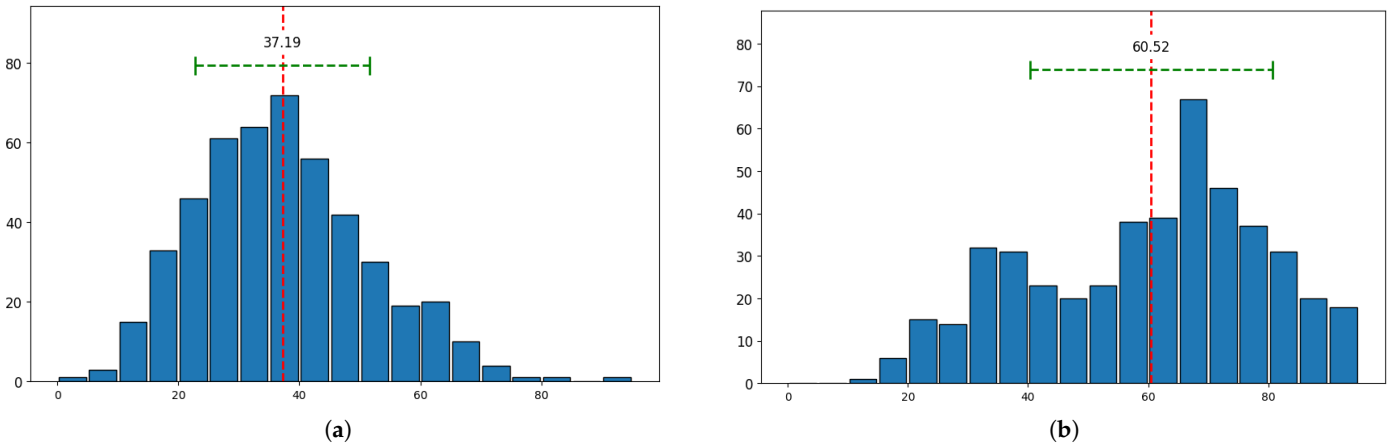


Figure 8. Error distribution for the Pushkin 100 score for the generated, simplified texts, BERT predicted grade for the simplified texts, compared to the error for the original texts.



(**a**) (**b**)

Figure 9. Distributions of Pushkin 100 scores (generated versus original texts). Mean value marked with a red vertical line and the standard deviation represented in green. (**a**) Distribution of the Pushkin 100 score for the generated, simplified texts. (**b**) Distribution of the Pushkin 100 score for the original texts.

A particularly surprising outcome was the variation in complexity reduction across subjects, as measured by this metric (Table 7). While all subjects exhibited a clear decrease in complexity, the Russian-language subject showed the smallest reduction, from a mean score of 38.34 to one of 25.45. This finding may reflect the inherent challenges of simplifying texts

in the literature domain, where maintaining grammatical structure, stylistic nuances, and educational appropriateness adds complexity to the simplification process. For example, theoretical texts often contain specialized terminology and dense explanations, instructional materials [48] that prioritize clarity and procedural accuracy, while literary texts emphasize narrative style and expressive language. Keeping this style intact while maintaining grammatical integrity and educational value increases the task's complexity.

Further, we analyze the outliers in these newly generated texts. namely, text where the complexity score increased through generation, or texts where the complexity scores remained unchanged. In Figure 10, we observe the two samples where the complexity score increased after the generation based on Pushkin 100 scores. Both textbook samples that showed an increase in complexity are for the 5th grade. The first text from the Russian subject textbook is a fill-in-the-blanks task with adjectives. In the generated text, the LLM filled in the blanks, thereby increasing the vocabulary and complexity of the resulting text. The other is from a social studies book. Here we can see the LLM trying to simplify the texts by shortening the phrases and using more colloquial language. Since the original score was relatively low, we can assume that small changes could increase it by 4–5 points. Overall, these outliers are exceptions, and we can infer that the LLM generally reduces the complexity of the processed texts.

	Original Russian Text	English Translation
Subject: Russian Grade: 05 Initial Score: 30	Спишите текст, вставляя вместо слов в скобках прилагательные с -н- и -нн-. Объясните их правописание. Всё вокруг застыло в крепком (осень) сне. Сквозь (туман) мглу виднеются вдали широкие луга. Солнце греет (пустыня) землю. Солнечный луч упал на (сон) воду сквозь заросли зелёных трав. Надо мною звенит (листва) лес, отряхивает с зелёных лап капли росы. На узорчатых листьях папоротника тоже сверкает (утро) роса. (Сон) трава немного примята вчерашним дождём. Просыпаются птицы. (Чин) синицы перескакивают с ветки на ветку, (огонь) клесты крошат крепкими носами сосновые шишки. Скоро весь лес огласится птичьими криками и песнями.Подчеркните прилагательные, определите их разряд. Какие прилагательные — качественные или относительные — употреблены в тексте? Почему? Найдите в тексте пять слов с безударными гласными в корне слова. Объясните их правописание. Спишите текст, заменяя глаголы в форме 1 -го лица множественного числа глаголами повелительного наклонения в единственном числе. Как изменится текст?	Write down the text, inserting instead of the words in parentheses adjectives with -н- and -нн-. Explain their spelling. Everything around froze in a deep (autumn) sleep. Wide meadows can be seen in the distance through the (fog) haze. The sun warms the (desert) earth. A sunbeam fell on the (sleep) water through the thickets of green grass. Above me rings the (foliage) forest, shaking drops of dew from its green branches. On the patterned leaves of the fern also sparkles the (morning) dew. The (sleep) grass is slightly pressed down by yesterday's rain. The birds are waking up. (Orderly) tits are jumping from branch to branch, (fire) crossbills are crumbling pine cones with their strong beaks. Soon the whole forest will resound with bird cries and songs. Underline the adjectives, determine their category. Which adjectives — qualitative or relative — are used in the text? Why? Find in the text five words with unstressed vowels in the root of the word. Explain their spelling. Write down the text again, replacing verbs in the form of the 1st person plural with verbs in the imperative mood, singular. How will the text change?
Generated text: New Score: 40	Перепишите текст, вставляя вместо слов в скобках прилагательные с -н- и -нн-. Объясните их написание. Всё вокруг застыло в крепком (осеннем) сне. Сквозь (туманную) мглу видны вдали широкие луга. Солнце греет (пустынную) землю. Солнечный луч упал на (сонную) воду сквозь заросли зелёных трав. Надо мной звенит (лиственная) роща, стряхивает с зелёных веток капли росы. На узорчатых листьях папоротника тоже сверкает (утренняя) роса. (Сонная) трава немного примята вчерашним дождём. Просыпаются птицы. (Чинные) синицы перескакивают с ветки на ветку, (огненные) клесты крошат крепкими клювами сосновые шишки. Скоро весь лес наполнится птичьими криками и песнями. Подчеркните прилагательные, определите их вид. Какие прилагательные — качественные или относительные — использованы в тексте? Почему? Найдите в тексте пять слов с безударными гласными в корне. Объясните их написание.Перепишите текст, заменяя глаголы в форме 1-го лица множественного числа глаголами повелительного наклонения в единственном числе. Как изменится текст?	Rewrite the text, inserting instead of the words in parentheses adjectives with -н- and -нн-. Explain their spelling. Everything around froze in a deep (autumn) sleep. Wide meadows are visible in the distance Through the (foggy) haze The sun warms the (desolate) earth. A sunbeam fell on the (sleepy) water through the thickets of green grass. Above me rings the (leafy) grove, shaking drops of dew from the green branches. On the patterned leaves of the fern also sparkles the (morning) dew. The (sleepy) grass is slightly pressed down by yesterday's rain. The birds are waking up. (Graceful) tits are jumping from branch to branch, (fiery) crossbills crumble pine cones with their strong beaks. Soon the whole forest will be filled with bird cries and songs. Underline the adjectives, determine their type. Which adjectives — qualitative or relative — are used in the text? Why? Find in the text five words with unstressed vowels in the root. Explain their spelling. Rewrite the text, replacing verbs in the form of the 1st person plural with verbs in the imperative mood, singular. How will the text change?
Subject: Social Studies Grade:05 Initial Score: 23	Это Родина. Это было в далёкой южной стране… Там никогда не бывает зимы и очень редко идёт дождь. На стройку в эту страну наше государство послало лучших инженеров, техников и рабочих. Поехал и отец Коли, инженер-строитель. С отцом поехали мама и Коля. За посёлком, в котором жил Коля, начиналась пустыня. А берега реки были зелёные. Над рекой росли стройные пальмы. Над ними парили диковинные птицы. Но всё это казалось Коле чужим. Его тянуло на родину. Он не мог забыть курских полей и садов, зелёных лугов и соловьиной песни. Однажды в школе для детей российских специалистов на уроке учительница прочитала стихотворение:	This is the Homeland. It was in a distant southern country… There, winter never comes and it rains very rarely. Our state sent its best engineers, technicians, and workers to build in that country. Kolya's father, a construction engineer, went there too. His mother and Kolya went with him. Beyond the settlement where Kolya lived began the desert. But the riverbanks were green. Slender palms grew above the river. Strange birds soared above them. Yet all this seemed foreign to Kolya. He longed for his homeland. He could not forget the fields and gardens of Kursk, the green meadows, and the song of the nightingale. One day, at school for the children of Russian specialists, the teacher read a poem aloud:
Generated text: New Score: 28	Это Родина. В далёкой южной стране никогда не бывает зимы и редко идёт дождь. Наше государство отправило туда лучших инженеров и рабочих. Папа Коли, который был инженером, тоже поехал. С ним поехали мама и Коля. За посёлком, где жил Коля, начиналась пустыня, а берега реки были зелёные. Там росли пальмы, и летали необычные птицы. Всё это казалось Коле чужим. Он скучал по родине, вспоминал родные поля, сады и соловьиную песню. Однажды в школе учительница прочитала стихотворе	This is the Homeland. In a distant southern country, there is never winter and it rarely rains. Our state sent its best engineers and workers there. Kolya's father, who was an engineer, went too. His mother and Kolya went with him. Beyond the settlement where Kolya lived began the desert, but the riverbanks were green. There grew palm trees, and unusual birds flew above them. All this seemed foreign to Kolya. He missed his homeland, remembering the familiar fields, gardens, and the nightingale's song. One day at school, the teacher read a poem aloud…

Figure 10. Samples of generations with increased complexity.

Table 7. Comparison of original and modified average metrics and predicted grades or ages across subjects.

Subject	Original Text			Generated Text		
	Pushkin 100 Scores	Predicted Age	Predicted Grade	Pushkin 100 Scores	Predicted Age	Predicted Grade
History	70.44	14.42	8.32	45.96	11.56	5.56
Social Studies	66.76	14.03	7.86	36.32	10.49	4.48
Russian	38.34	10.80	4.78	25.45	9.58	3.58

5. Discussion

This study aimed to provide an exploration and evaluation of a state-of-the-art LLM, specifically GPT-4o, in assessing the complexity, judging the ease of understanding, and performing targeted simplification of Russian educational textbook texts. By analyzing the LLM's performance across three distinct experimental setups using a curated corpus of Russian schoolbooks, we deepen the understanding of its potential utility and limitations within the Russian educational context. To accomplish these objectives, we examined three core research questions, each targeting a distinct aspect of LLM performance.

Our first research question investigated whether LLMs can correctly assess the complexity of Russian-language texts and, during this experiment, identify the key features that have influenced the decisions. The results from Experiment 1 (see Table 3) show that the LLM adequately estimated the target age group, with a Mean Absolute Error (MAE) generally within 1 school year for most subjects. This suggests a fairly good capability to evaluate the difficulty of a textbook for the average Russian student.

However, several important observations emerged. Firstly, we can notice a consistent trend across specific subjects, particularly Art, Informatics, Music, and Physics (see Figure 4). For these textbooks, the LLM tends to overestimate complexity, predicting a higher age level than the intended one. This bias might stem from the model overestimating factors such as challenging vocabulary, abstract concepts, or specific technical terminology encountered in these subjects. Analysis through the clustering of key phrases extracted by the LLM (see Figure 6) supports this, indicating that the model's decisions are influenced not merely by surface-level metrics (like sentence length or vocabulary used) but also by the addressed topics and domain-specific concepts that are discussed in the evaluated fragments.

In addition, the analysis of extreme outliers, especially within the Russian-language subject (Figure 5), highlighted an important limitation: the lack of context when evaluating a textbook fragment. Literary excerpts, when assessed in isolation as text segments, were often judged by the LLM to be far simpler than their intended target age. This occurred because the surrounding pedagogical context within the textbook (such as accompanying literary analysis tasks that significantly increased the effective cognitive load) was necessarily lost during the segmentation process required for analysis. This finding underscores that LLM assessments based on isolated text chunks may not accurately reflect the complexity a student encounters when interacting with the full material.

Furthermore, the model's confidence, as measured by output token probabilities (see Tables 4 and 8), did not always correlate with the accuracy. Even in cases of significant error (MAE > 2 years), the model often maintained high confidence, except notably in Geography, History, and Physics, where incorrect predictions were associated with lower confidence.

Table 8. Mean errors for predictions with a high confidence.

Subject	MSE	MAE	Mean Prob (%)	Mean Joint Prob (%)	Mean Perplexity	Count
Art	2.64	1.36	81.88	73.74	1.12	11
Biology	1.73	1.11	81.93	71.62	1.09	71
Ecology	2.88	1.62	78.14	72.99	1.12	8
Geography	0.23	0.23	83.44	73.03	1.08	13
History	0.56	0.44	81.49	70.32	1.09	34
Informatics	4.25	1.83	80.90	73.25	1.08	24
Maths	1.23	0.64	81.17	73.29	1.08	44
Music	7.73	2.64	80.16	72.22	1.09	11
Physics	3.67	1.53	82.68	61.93	1.14	15
Russian	1.75	0.95	80.86	71.32	1.10	40
Science	2.10	1.07	82.76	74.63	1.08	29
Social Studies	2.93	1.22	81.62	73.59	1.08	27
Technology	1.13	0.74	81.84	67.45	1.11	31

To conclude RQ1, while LLMs show considerable potential in assessing text complexity, their evaluations often display a bias toward overestimation. Notably, the model's judgments appear to depend not only on linguistic features but also on the alignment of content themes with the cognitive and developmental expectations of the target age group. This thematic sensitivity, if properly understood and controlled, could serve as a valuable complement to traditional readability metrics. LLMs thus hold promise as valuable tools in augmenting classical methods of comprehension assessment. A promising direction for future research would be the development of techniques to disentangle purely linguistic signals from content- and theme-aware influences in the model's predictions.

The second research question explored whether LLMs could serve as an effective proxy for student comprehension, simulating the extent to which the text is understandable. Results from Experiment 2 (see Table 6) suggest a general alignment, with the LLM deeming the majority of fragments across most subjects as "comprehensible." This indicates that the model can broadly distinguish between texts that are likely accessible and those that pose significant challenges.

However, subjects with lower comprehension ease included Russian, Informatics, Biology, and History. Analysis of the LLM's rationale, by clustering its chain-of-thought conclusions, indicated that judgments of lower comprehension ease were often linked to a high terminological density, as well as challenges with complex language, abstract concepts, technical vocabulary, intricate instructions, or the perceived need for external guidance—factors that are highly relevant to actual student comprehension difficulties.

Interestingly, there was no direct one-to-one mapping between the age-prediction errors observed in Experiment 1 and the cognitive accessibility judgments in Experiment 2, suggesting that the LLM uses distinct pathways for these two tasks. Subjects in which the LLM significantly overestimated age (e.g., Art and Music) did not necessarily show lower comprehension ease. This suggests that the LLM might employ different internal weighting or criteria when assessing abstract "complexity" versus simulating direct "comprehension ease," potentially separating linguistic difficulty from conceptual accessibility in its judgment process.

Our third research question assessed the LLM's capability to successfully reduce text complexity to a specific target level, i.e., 3 school years lower across different subjects. The findings here present a mixed picture depending on the evaluation metric used.

When evaluated using a BERT-based grade-level classifier (see Table 9), the LLM achieved very poor performance (14.88% accuracy), largely failing to align the generated text with the target grade. Many prediction errors matched the intended three-year gap

exactly, indicating that the model often maintained complexity close to the original level rather than sufficiently simplifying the text.

Table 9. Classification Performance Metrics of BERT Model by Grade Level.

Target Grade	Precision	Recall	F1-Score	Support
Grade 2	0.6364	0.0778	0.1386	90
Grade 3	0.1538	0.1500	0.1519	80
Grade 4	0.4000	0.0400	0.0727	50
Grade 5	0.0704	0.1000	0.0826	50
Grade 6	0.0317	0.0500	0.0388	40
Grade 7	0.2553	0.4000	0.3117	60
Grade 8	0.2500	0.2000	0.2222	60

Conversely, a statistically significant overall reduction in complexity was observed when evaluated using the Pushkin 100 readability score (see Figure 9) ($p < 0.001$), evidenced by a clear leftward shift in the mean score distribution, from 60.52 to 37.19 (see Figure 9). This reflects an average simplification of 23.33 points, corresponding to a decrease in estimated reading age from 14 to 10 years. Given that the Pushkin 100 score predicts comprehension across 2-year age intervals, the observed reduction closely aligns with our intended 3-grade simplification target.

Performance varied significantly across subjects, with Russian-language texts showing the least reduction according to the Pushkin 100 metric, potentially reflecting the inherent difficulty of simplifying literary texts while preserving essential meaning and style. Analysis of outliers where complexity increased (see Figures 5 and 10) revealed these were often artifacts of the original task—e.g., the task of filling blanks in the Russian subject example steers the LLM to perform this task rather than simplifying it, or possibly metric sensitivity to minor changes in already simple texts (e.g., social studies). These outliers were deemed exceptions to the general trend of simplification.

Therefore, as a conclusion to RQ3, LLMs can contribute to text simplification, but achieving precise, controlled reduction to a specific pedagogical level (especially one defined by grade levels) remains a challenge that researchers can address. In addition, subject-specific consistent simplification outcomes remain an open research question that merits further exploration.

Practical and Ethical Implications

LLMs inherit and may amplify biases present in their training data, raising fairness concerns when applied to educational content assessment. These models have been shown to encode both explicit and implicit stereotypes—for example, prompts using male-associated names can provide more confident evaluations than those using female-associated names [49]. Cultural and linguistic biases are also a concern. Many LLMs display Anglocentric biases that privilege Western or dominant-curriculum perspectives, which can marginalize content rooted in other languages, cultures, or educational systems [50–52]. Such misalignment risks producing unfair assessments of materials that deviate from these norms, potentially disadvantaging learners in diverse cultural contexts.

Without deliberate safeguards, these biases can be perpetuated or even intensified by LLM-based tools. Notably, the findings of the present study reflect these broader concerns: even within a specific national context (e.g., Russian-language textbook analysis), the model exhibited some interpretations of textual content that align with known patterns of systemic bias.

Beyond these biases, LLMs face key technical limitations when used to evaluate or simplify academic content. Their assessments often rely on surface-level features—such

as sentence length or vocabulary—without understanding the deeper conceptual or pedagogical context, leading to over- or underestimations of text complexity [50]. Moreover, LLM-generated feedback lacks the intentionality of human instruction. Unlike educators who adapt explanations based on curriculum goals or student needs, LLMs tend to produce generic simplifications that may overlook important learning objectives or do not align with instructional intent [53]. Thus, over-reliance on LLM-generated simplifications can diminish valuable learning opportunities [54].

To realize the benefits of LLMs in education while minimizing risks, their use must be guided by responsible oversight. Educators should treat AI-generated outputs, such as readability ratings or summaries, as starting points, not final judgments, always adapting them to learner needs. Policymakers can support equitable adoption by promoting culturally inclusive AI tools and funding models trained on diverse curricula. As this study and others show, LLMs offer both promise and pitfalls, making ongoing research, human oversight, and strong ethical safeguards essential to ensuring their alignment with sound educational practice.

6. Conclusions and Future Work

Our study conducted a comprehensive evaluation of a state-of-the-art LLM, namely GPT-4o, in the context of assessing Russian educational materials to investigate its ability to gauge the complexity and readability of school textbooks. Through a series of experiments using a curated corpus of Russian school textbooks, we identified several insights into the strengths and limitations of current LLM models in this domain.

LLMs adequately estimated text complexity and served as a good proxy for student comprehension. However, their complexity assessments frequently overestimated, predicting a higher text complexity. We showed that the model's judgments do not rely solely on linguistic features, such as sentence length or vocabulary; they also reflect thematic alignment with students' age and the conceptual density of the text. Nevertheless, evaluating text segments in isolation, without pedagogical scaffolding or contextual cues (e.g., accompanying tasks or full-unit structure), reduced the reliability of these complexity assessments.

In addition to refining the use of GPT-4o, future work will evaluate the performance of alternative LLMs, both open-source and via API, to benchmark their capabilities across educational domains and linguistic contexts. Comparative model analysis could reveal important differences in accuracy, bias, or simplification strategies. Furthermore, to better assess the quality and pedagogical utility of LLM-generated simplifications, we recommend conducting blind human evaluations in which educators rate simplified outputs without knowing whether they were AI- or human-authored. Such studies would offer critical insights into the real-world applicability of LLMs in instructional settings.

Overall, our findings indicate that large-scale LLMs, such as GPT-4o, have substantial potential as assistive tools in educational contexts, particularly for Russian-language materials. They can complement traditional complexity-scoring systems and support preliminary assessments and adaptations. However, despite acceptable performance, these models are not yet sufficiently reliable to replace expert human judgment. Future research aimed at aligning LLM outputs with established educational standards, such as through fine-tuning on domain-specific corpora or employing reinforcement learning with readability- and pedagogy-oriented objectives, could offer more reliable results. Nevertheless, current capabilities lay a solid foundation for enhancing conventional readability frameworks and fostering more responsive, personalized educational content.

Author Contributions: Conceptualization, A.P. and M.D.; methodology, A.P., M.D. and M.S.; software, A.P.; validation, M.S.; formal analysis, A.P. and M.D.; investigation, A.P.; resources, M.S.; data curation, M.S.; writing—original draft preparation, A.P.; writing—review and editing, M.D. and M.S.; visualization, A.P.; supervision, M.D.; project administration, M.D.; funding acquisition, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the project "Romanian Hub for Artificial Intelligence—HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 351416.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request due to copyright concerns.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Models
MAE	Mean Absolute Error
MSE	Mean Squared Error
NLP	Natural Language Processing

References

1. Benjamin, R.G. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educ. Psychol. Rev.* **2012**, *24*, 63–88. [CrossRef]
2. Solovyev, V.D.; Solnyshkina, M.I.; Ivanov, V.; Timoshenko, S. Complexity of Russian academic texts as the function of syntactic parameters. In *Computational Linguistics and Intelligent Text Processing, Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing, Hanoi, Vietnam, 18–24 March 2018* ; Springer: Cham, Switzerland , 2018; pp. 168–179.
3. Solovyev, V.D.; Ivanov, V.; Solnyshkina, M.I. Readability Formulas for Three Levels of Russian School Textbooks. *J. Math. Sci.* **2024**, *285*, 100–111. [CrossRef]
4. McNamara, D.S.; Graesser, A.C.; McCarthy, P.M.; Cai, Z. *Automated Evaluation of Text and Discourse with Coh-Metrix*; Cambridge University Press: Cambridge, UK, 2014.
5. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221. [CrossRef]
6. Solnyshkina, M.; Ivanov, V.; Solovyev, V.D. Readability Formula for Russian Texts: A Modified Version. In *Advances in Computational Intelligence, Proceedings of the 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, 22–27 October 2018*; Proceedings, Part II; Batyrshin, I.Z., de Lourdes Martínez-Villaseñor, M., Espinosa, H.E.P., Eds.; Springer: Cham, Switzerland, 2018; Volume 11289, pp. 132–145. [CrossRef]
7. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
8. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Mahdaouy, Y.E.; Lample, G.; Babaei, Y.; Bashlykov, N.; Batra, S.; et al. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2406.10552. [CrossRef]
9. Quah, B.; Zheng, L.; Sng, T.J.H.; Yong, C.W.; Islam, I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Med. Educ.* **2024**, *24*, 962. [CrossRef] [PubMed]
10. Wang, T.; Tao, M.; Fang, R.; Wang, H.; Wang, S.; Jiang, Y.E.; Zhou, W. AI PERSONA: Towards Life-long Personalization of LLMs. *arXiv* **2024**, arXiv:2412.13103. [CrossRef]
11. Khan, I.; Chohan, I.; Malik, S.I. Leveraging ChatGPT-4 for Enhanced Education: Personalized Problem-Solving and Consistent Learning. In Proceedings of the 2024 2nd International Conference on Computing and Data Analytics (ICCDA), Shinas, Oman, 12–13 November 2024; pp. 1–6.
12. Solnyshkina, M.I.; McNamara, D.S.; Zamaletdinov, R.R. Natural language processing and discourse complexity studies. *Russ. J. Linguist.* **2022**, *26*, 317–341. [CrossRef]

13. AlKhuzaey, S.; Grasso, F.; Payne, T.R.; Tamma, V. Text-based question difficulty prediction: A systematic review of automatic approaches. *Int. J. Artif. Intell. Educ.* **2024**, *34*, 862–914. [CrossRef]
14. Maddela, M.; Xu, W. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3749–3760. [CrossRef]
15. Schwarm, S.E.; Ostendorf, M. Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, MI, USA, 25–30 June 2005; pp. 523–530.
16. Pitler, E.; Nenkova, A. Revisiting readability: A unified framework for predicting text quality. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI, USA, 25–27 October 2008; pp. 186–195.
17. Kincaid, J.P.; Fishburne, R.P., Jr.; Rogers, R.L.; Chissom, B.S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical Report, Naval Technical Training Command Millington TN Research Branch. 1975. Available online: https://apps.dtic.mil/sti/tr/pdf/ADA006655.pdf (accessed on 28 November 2025).
18. Yaneva, V.; Ha, L.A.; Baldwin, P.; Mee, J. Predicting Item Survival for Multiple Choice Questions in a High-Stakes Medical Exam. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 6812–6818.
19. Imperial, J.M. BERT Embeddings for Automatic Readability Assessment. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, 1–3 September 2021; INCOMA Ltd.: Shoumen, Bulgaria, 2021; pp. 611–618.
20. Paraschiv, A.; Dascalu, M.; Solnyshkina, M. Classification of Russian textbooks by grade level and topic using ReaderBench. *Res. Result Theor. Appl. Linguist.* **2023**, *9*, 73–86. [CrossRef]
21. Naous, T.; Ryan, M.J.; Lavrouk, A.; Chandra, M.; Xu, W. ReadMe++: Benchmarking Multilingual Language Models for Multi-Domain Readability Assessment. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12230–12266. [CrossRef]
22. Herianah, H.; Setiawan, E.C.; Adri, A.; Tamrin, T.; Judijanto, L.; Supatmiwati, D.; Sutrisno, D.; Musfeptial, M.; Damayanti, W.; Martina, M. Automated Assessment of Text Complexity through the Fusion of AutoML and Psycholinguistic Models. *Forum Linguist. Stud.* **2025**, *7*, 46–62. [CrossRef]
23. Morozov, D.A.; Glazkova, A.V.; Iomdin, B.L. Text complexity and linguistic features: Their correlation in English and Russian. *Russ. J. Linguist.* **2022**, *26*, 426–448. [CrossRef]
24. Farajidizaji, A.; Raina, V.; Gales, M. Is It Possible to Modify Text to a Target Readability Level? An Initial Investigation Using Zero-Shot Large Language Models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, 20–25 May 2024; pp. 9325–9339.
25. Trott, S.; Rivière, P. Measuring and Modifying the Readability of English Texts with GPT-4. In Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Miami, FL, USA, 15 November 2024; pp. 126–134. [CrossRef]
26. Huang, C.Y.; Wei, J.; Huang, T.H.K. Generating educational materials with different levels of readability using LLMs. In Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants, Honolulu, HI, USA, 11–16 May 2024; pp. 16–22.
27. Rooein, D.; Röttger, P.; Shaitarova, A.; Hovy, D. Beyond Flesch-Kincaid: Prompt-based Metrics Improve Difficulty Classification of Educational Texts. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2024, Mexico City, Mexico, 20 June 2024; Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., Yuan, Z., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 54–67.
28. Gobara, S.; Kamigaito, H.; Watanabe, T. Do LLMs Implicitly Determine the Suitable Text Difficulty for Users? In Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, Tokyo, Japan, 7–9 December 2024; pp. 940–960.
29. Imperial, J.M.; Tayyar Madabushi, H. Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. In Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Singapore, 6 December 2023; pp. 205–223.
30. Alemi, A.A.; Ginsparg, P. Text segmentation based on semantic word embeddings. *arXiv* **2015**, arXiv:1503.05543. [CrossRef]
31. Erkan, G.; Radev, D.R. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* **2004**, *22*, 457–479. [CrossRef]
32. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford Infolab: Stanford, CA, USA, 1999.
33. Radev, D.R.; Jing, H.; Styś, M.; Tam, D. Centroid-based summarization of multiple documents. *Inf. Process. Manag.* **2004**, *40*, 919–938. [CrossRef]

34. Li, M.; Conrad, F.; Gagnon-Bartsch, J. FastLexRank: Bring Order into Social Media Posts Using Lexical Ranking Algorithm. In Proceedings of the ICWSM Workshops: R2CASS 2025—Social Science Meets Web Data, Copenhagen, Denmark, 23–26 June 2025; [CrossRef]
35. Olney, A.M. Assessing Readability by Filling Cloze Items with Transformers. In *Artificial Intelligence in Education, Proceedings of the 23rd Proceedings of the International Conference on Artificial Intelligence in Education, Durham, UK, 27–31 July 2022*; Springer: Cham, Switzerland, 2022; pp. 307–318.
36. Kamalloo, E.; Upadhyay, S.; Lin, J. Towards robust qa evaluation via open llms. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, USA, 14–18 July 2024; pp. 2811–2816.
37. Benedetto, L.; Aradelli, G.; Donvito, A.; Lucchetti, A.; Cappelli, A.; Buttery, P. Using LLMs to simulate students' responses to exam questions. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 11351–11368.
38. Jain, Y.; Hollander, J.; He, A.; Tang, S.; Zhang, L.; Sabatini, J. Exploring the Potential of Large Language Models for Estimating the Reading Comprehension Question Difficulty. In Proceedings of the Adaptive Instructional Systems—7th International Conference, AIS 2025, Held as Part of the 27th HCI International Conference, HCII 2025, Gothenburg, Sweden, 22–27 June 2025; Proceedings, Part I; Sottilare, R.A., Schwarz, J., Eds.; Springer: Cham, Switzerland, 2025; Volume 15812, pp. 202–213. [CrossRef]
39. Kortemeyer, G. Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discov. Artif. Intell.* **2024**, *4*, 47. [CrossRef]
40. Chiang, C.; Chen, W.; Kuan, C.; Yang, C.; Lee, H. Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, 12–16 November 2024; Al-Onaizan, Y., Bansal, M., Chen, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 2489–2513. [CrossRef]
41. Hu, Y.; Huang, Q.; Tao, M.; Zhang, C.; Feng, Y. Can Perplexity Reflect Large Language Model's Ability in Long Text Understanding? In Proceedings of the Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, 11 May 2024.
42. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
43. Solnyshkina, M.I.; Solovyev, V.D.; Ebzeeva, Y.N. Approaches and tools for Russian text linguistic profiling. *Russ. Lang. Stud.* **2024**, *22*, 501–517. [CrossRef]
44. Viswanathan, V.; Gashteovski, K.; Gashteovski, K.; Lawrence, C.; Wu, T.; Neubig, G. Large language models enable few-shot clustering. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 321–333. [CrossRef]
45. Tarekegn, A.N. Large Language Model Enhanced Clustering for News Event Detection. *arXiv* **2024**, arXiv:2406.10552. [CrossRef]
46. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794.
47. Laposhina, A.N.; Lebedeva, M.Y. Textometr: An online tool for automated complexity level assessment of texts for Russian language learners. *Russ. Lang. Stud.* **2021**, *19*, 331–345. [CrossRef]
48. Bulina, E.N.; Solnyshkina, M.I.; Ebzeeva, Y.N. Russian language textbook as agent of change: From USSR to the new century. *Russ. Lang. Stud.* **2024**, *22*, 540–554. [CrossRef]
49. Haim, A.; Salinas, A.; Nyarko, J. What's in a Name? Auditing Large Language Models for Race and Gender Bias. *arXiv* **2024**, arXiv:2402.14875. [CrossRef]
50. Delikoura, I.; Fung, Y.R.; Hui, P. From Superficial Outputs to Superficial Learning: Risks of Large Language Models in Education. *arXiv* **2025**, arXiv:2509.21972. [CrossRef]
51. Dokic, K.; Pisker, B.; Radisic, B. Mirroring Cultural Dominance: Disclosing Large Language Models Social Values, Attitudes and Stereotypes. *Societies* **2025**, *15*, 142. [CrossRef]
52. Dudy, S.; Ahmad, I.S.; Kitajima, R.; Lapedriza, À. Analyzing Cultural Representations of Emotions in LLMs Through Mixed Emotion Survey. In Proceedings of the 12th International Conference on Affective Computing and Intelligent Interaction, ACII 2024, Glasgow, UK, 15–18 September 2024; IEEE: New York City, NY, USA, 2024; pp. 346–354. [CrossRef]
53. Seßler, K.; Fürstenberg, M.; Bühler, B.; Kasneci, E. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In Proceedings of the 15th International Learning Analytics and Knowledge Conference, LAK 2025, Dublin, Ireland, 3–7 March 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 462–472. [CrossRef]
54. Gerlich, M. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* **2025**, *15*, 6. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.