DOI: 10.14529/jsfi250310

Neurosymbolic Approach to Processing of Educational Texts for Educational Standard Compliance Analysis

Nikolai A. Prokopyev[1], *Marina I. Solnyshkina*[1], *Valery D. Solovyev*[1]

This article presents a neurosymbolic approach for analyzing the alignment between textbook content and educational standards. The study addresses the problem of assessing terminological coherence by evaluating a corpus of textbooks against the Russian Federal State Educational Standard. We employ a hybrid methodology combining classical symbolic NLP methods for topic modeling (keyword extraction and term alignment) with qualitative analysis and use of modern large language models for items not found algorithmically. The experimental results on a corpus of 5 textbooks on Physics for the 7th grade and corresponding educational standard indicate a mean coverage of standard topics of 71% across all textbooks with use of the symbolic methods. Application of large language model (ChatGPT 5) for the qualitative analysis recovered 51% keywords initially missed by the abovementioned methods. The findings are relevant for researchers in educational linguistics, computational linguistics, curriculum developers, and textbook authors. The proposed pipeline offers a scalable tool for automating analysis of educational content compliance, reducing the workload for manual expert assessment. This work contributes to the development of AI-assisted methodologies in educational standard alignment and textbook quality control.

Keywords: topic modeling, keyword extraction, symbolic NLP, large language model, textbook analysis.

Introduction

Alignment between educational standards and textbooks content as the degree of coherence between curricula and textbooks content has been an area of numerous disputes [26], and as such extensively studied around the world [4, 13]. Modern natural language processing (NLP) paradigm provides numerous approaches and powerful toolkits to measure this alignment: recent advances in large language models (LLMs) resulted in significant changes in the area, and LLMs enable considerate assistance in educational content assessment [1, 25] thus reducing the workload of academics and test developers.

Specifics and complexity of solving this problem is related to identifying the range of linguistic variability in texts, dynamism of modern discourse and active expansion of nonverbal signs into academic texts as well as growing number of nonlinear, polycode texts. All the above constitute the foundation of ongoing research in alignment between textbooks content and national standards.

Although the Russian textbook language quality and its compliance with national standards have been lately addressed by Russian and foreign researchers [16, 23], to the best of our knowledge, there is no research which explicitly uses both symbolic NLP methods and LLM evaluation tools to assess textbook language alignment with national educational standards. Thus the current research is aimed at evaluating terminological coherence between the educational content of five Physics textbooks and the Federal State Educational Standard.

The main goal of this research is to experimentally verify the algorithm for educational standard compliance analysis of educational texts. This algorithm can be used as a support tool for authors of textbooks and official experts by giving them the preliminary results for furhter investigation. Thus, the research questions we address are as follows:

[1]Kazan Federal University, Kazan, Russian Federation

1. Application of classical symbolic NLP methods to keyword extraction;
2. Application of NLP methods based on neural networks, particularly use of LLM for additional keyword extraction;
3. Combined use of these methods to solve the problem of educational standard compliance analysis for textbooks.

The article is structured as follows. Section 1 provides an overview of the pipeline for problem of educational standard compliance analysis. Section 2 presents the related works analysis of NLP technologies used in various relevant tasks. Section 3 describes the results of this study. Conclusion presents the main findings and directions for future developments.

1. Methodology

1.1. Problem of Educational Standard Compliance Analysis

Before we move on to clarified formulation of the problem of educational standard compliance analysis, it is necessary to describe the most important elements of the Federal State Educational Standard structure. These are the following elements:

- A set of topics on the school subject for one school class. Each topic has a description in free form, it determines which concepts from the topic should be disclosed in a textbook.
- A set of terms on the school subject for one school class. These terms should also be disclosed in a textbook.

The problem of educational standard compliance analysis then is defined as follows: to check whether a textbook aligns with topics and terms from the standard, also assessing to what extent it occurs.

To solve this problem, we propose to reduce it to the topic modeling problem partially, within which a topic description in the standard (standard topic) should be represented as a set of keywords. The set of terms in the standard (standard terms) can be considered a separate topic, isolated in its pragmatic significance for this problem.

Thus, we propose to use a hybrid neurosymbolic approach to this topic modeling problem. The symbolic part of the approach consists in usage of tokenization, lemmatization, n-gram retrieval and bag of words (BOW) methods. These methods are used for the primary algorithmic analysis of the textbook to search for keywords for each standard topic and to search for standard terms. The neural network part of the approach involves the use of LLM for subsequent qualitative analysis of keywords and terms not found algorithmically.

The general scheme of the solution pipeline is shown in Fig. 1. A detailed description of the pipeline stages is presented in the next subsections.

1.2. Preprocessing of Textbooks and Standards

Preprocessing of textbooks is an important pipeline stage, necessary for further analysis. This stage is fully automated, with the following algorithm:

1. Compilation of a textbook metadata file containing: a list of textbooks in the source dataset, their metadata (ID, title, school subject, school class), paths to files with different forms of textbook text representation.
2. Extraction of texts from the source textbook files and saving them in TXT format.

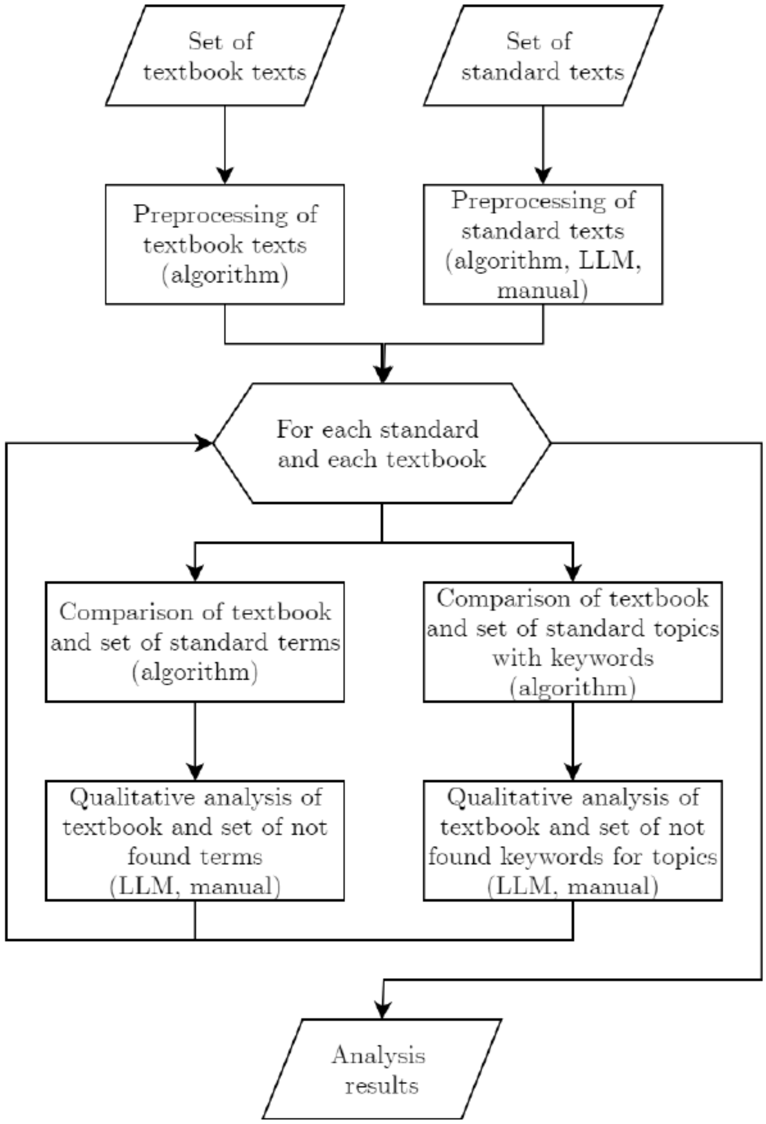

Figure 1. General scheme of textbook analysis stages

3. Tokenization of texts, which includes the removal of punctuation, stop words, and grammatical lemmatization (normalization) of tokens using a specific configuration. The resulting token set is saved in JSON format.

Preprocessing of standard texts is a more complex task, since the description of each standard topic is a free-form text, for automatic processing of which a semantic analyzer is required. Thus, while this stage is not fully automated, we propose to use LLM to extract keywords and terms as it can be an effective solution for semantic analysis as shown in Section 2. Sequence of actions:

1. Algorithmic compilation of a standards metadata file containing: a list of standards, their metadata (school subject, school class, level – basic or advanced), paths to files with different forms of standard representation.
2. Manual split of the standards text into topic descriptions and term descriptions.
3. Extraction of topic keywords and terms from the corresponding descriptions using LLM with a prepared prompt.
4. Algorithmic compilation of normalized n-grams of keywords and terms. Since keywords and terms can consist of several words in different word forms, n-grams should be compiled for

further analysis, and their normalization should be performed using the same configuration that was used to normalize textbook tokens. The resulting sets of keywords and a set of terms are saved in JSON format.

1.3. Primary Analysis: Comparison of Textbooks and Standards

For the primary analysis of textbooks for compliance with the educational standard, an automatic algorithm is used to align the tokenized textbook and sets of standard topic keywords or standard terms. The algorithm is as follows:

1. Compilation of ID-dictionary for the corpus of tokenized textbooks, taking into account the automatic collection of tokens into n-grams.
2. Compilation of a textbook BOW using the ID-dictionary, containing IDs of n-grams and their number of instances in the textbook.
3. For each standard topic, a BOW of the corresponding keywords set is compiled, after which a search for matching IDs of keywords in the textbook BOW is performed. The following are output: keyword, number of instances, frequency.

$$Frequency = \frac{number\ of\ instances}{textbook\ length\ in\ tokens}$$

4. For each standard topic, a separate search is performed among the not found keywords, those that had too small number of instances to be collected into an n-gram.
5. For each standard topic, the keywords found and not found by the algorithm are displayed, and the coverage is calculated.

$$Coverage = \frac{number\ of\ found\ keywords\ in\ textbook}{total\ number\ of\ keywords\ in\ topic}$$

It should be noted that classical metrics for keyword extractions, such as accuracy, precision and recall, are not suitable in this case, as they are metrics for supervised classification task demanding availability of a labeled data, while this algorithm is unsupervised.

6. Average coverage of the standard topics in the textbook is calculated.

The automatic algorithm for comparing the tokenized textbook and the previously obtained set of standard terms is performed in a similar way, while set of standard terms is accepted as a separate topic with a set of corresponding keywords.

1.4. Qualitative Analysis of Undetected Keywords and Terms

Since classical NLP methods do not take into account synonyms and semantically close descriptions of standard topics and terms, a qualitative analysis of the keywords and terms not found in each textbook is necessary.

Obviously, to solve this problem, manual expert assessment of each textbook can be used, but it is a labor-intensive task in the conditions of a large volume of textbooks. Thus, we propose to use LLM with a prepared prompt to check each set of keywords and terms not found in the primary analysis for each textbook.

However, the output of LLM also requires manual verification, so the set of terms additionally found using LLM is output separately. Recovery metric is calculated as:

$$Recovery = \frac{number\ of\ keywords\ found\ by\ LLM}{number\ of\ keywords\ notfound\ by\ algorithm}.$$

1.5. Source Data and Technologies

For the experimental verification of the proposed approach, a set of source data was used: the educational standard for Physics, 7th school class, basic level, and a set of textbooks. Table 1 presents the topics of this standard with English translation and their IDs used further. 5 textbooks on Physics for the 7th class of different publication years in WORD document format were considered. Table 2 presents the bibliographic data of these textbooks and their IDs used further.

Table 1. Standard topics

ID	Topic
Topic 1	Физика и ее роль в познании окружающего мира Physics and its role in understanding the world around us
Topic 2	Первоначальные сведения о строении вещества Initial information about structure of matter
Topic 3	Движение и взаимодействие тел Movement and interaction of bodies
Topic 4	Давление твердых тел, жидкостей и газов Pressure of solids, liquids and gases
Topic 5	Работа и мощность. Энергия Work and power. Energy

Table 2. Textbook bibliography

ID	Textbook
81412	Генденштейн Л. Э. Физика. 7 класс. В 2 ч. Ч. 1 : учебник для общеобразовательных учреждений / Л. Э. Генденштейн, А. В. Кайдалов ; под ред. В. А. Орлова, И. И. Ройзена. – 3-е изд., испр. – М.: Мнемозина, 2012. - 255 с. : ил. ISBN 978-5-346-02160-5
31329	Громов С. В. Физика: Учеб. для 7 кл. общеобразоват. учреждений/ С. В. Громов, Н. А. Родина.- 4-е изд.- М.: Просвещение, 2002.- 158 с.: ил.- ISBN 5-09-011495-1
28878	Физика. 7 класс : учеб. для общеобразоват. организаций / О.Ф. Кабардин. - 3-е изд. - М.: Просвещение, 2014. - 176 с.: с ил.
21915	Физика. 7 кл. : учеб. для общеобразоват. учреждений / А.В. Пёрышкин. - 2-е изд., стереотип. - М.: Дрофа, 2013. - 221 с.: с ил.
26802	Физика. 7 класс : учебник / Н.С. Пурышева, Н.Е. Важеевская. - 2-е изд., стереотип. - М.: Дрофа, 2013. - 224 с.: с ил.

For implementation, the Python programming language, PyCharm development environment, and Jupyter were used. `textract` library was used to extract textbook texts. `gensim` library was used to apply symbolic NLP methods: ID-dictionary compilation, bigram and trigram extraction (threshold = 1), BOW compilation. Additionally, less frequent n-gram candidates for $n > 3$ were extracted using custom algorithm. Tokenization was performed using `razdel` library, stop words were removed using `stopwords` library, and lemmatization was performed using `pymorphy2`.

ChatGPT 5 was used as the main LLM in experimental setup. The prompt for preprocessing the standard text is presented in Fig. 2 with English translation. The prompt for additional search of not found terms and keywords is presented in Fig. 3.

The best prompts obtained as a result of prompt engineering are shown. It was peculiar that small changes in the prompt in the conditions of processing an uploaded textbook file greatly affected the output of LLM, and whether it would take into account the attached textbook file or not.

2. Related Works

LLMs, being language models, are successfully applied in a wide range of problems related to information retrieval in texts, such as keyword extraction [11].

Application of the classical BERT model is described in [10], where KeyBERT is presented. The basic idea is that embeddings (vector representations) of the whole text and individual words are built, which are then compared by the degree of similarity. Its further improvement is proposed in [20].

Classical statistical methods also continue to be used. A general overview of statistical methods for keyword extraction can be found in [18].

Я извлекаю ключевые слова из стандарта ФГОС по {предмет} для среднего образования. Изначально они мне даны в виде описания на естественном языке. Преобразования, которые я применяю, следующие:
- Объединение нескольких слов в n-грамму ключевого слова;
- Извлечение нескольких n-грамм из их сокращенного описания;
- Написание собственных n-грамм из описания в виде предложения;
- Извлечение из n-грамм более обобщающих ключевых слов;
- Фильтрация слов, не являющихся ключевыми словами по предмету.

Мне необходимо автоматизировать ручную работу. Далее я буду давать тебе описания, а ты преобразуй их в списки n-грамм. Отвечай кратко, списком в формате Python, без пояснений.

I extract keywords from the State Educational Standard for {subject} for School Education. Initially, they are given to me as a description in natural language. The transformations I apply are following:
- Combining several words into an n-gram of a keyword;
- Extracting several n-grams from their abbreviated description;
- Writing own n-grams from a description in form of a sentence;
- Extracting more general keywords from n-grams;
- Filtering words that are not keywords of the subject.

I need to automate manual work. Next, I will give you the descriptions and you transform them into lists of n-grams. Answer briefly, with a list in Python format, without explanations.

Figure 2. Prompt for text processing of the standard

Тебе дан текст учебника по {предмет} для среднего образования в виде прикрепленного файла. Я делаю анализ наличия ключевых слов по темам. Для одной темы мой алгоритм выявил следующие не найденные ключевые слова (приведены в виде нормализованных n-грамм):
{ключевые_слова}
Попытайся найти эти ключевые слова в предоставленном учебнике в виде их синонимов или близких к ним описаний и тем. В ответе дай только список найденных ключевых слов в том виде, в котором они даны в изначальном списке.

You are given the text of a {subject} textbook for secondary education in the attached file. I do a keyword analysis by topic. For one topic, my algorithm identifies the following not found keywords (given as normalized n-grams):
{keyword_set}
Try to find these keywords in the provided textbook as their synonyms or close descriptions and topics. The answer only contains a list of found keywords in form in which they are given in the original list.

Figure 3. Prompt for search of not found terms and keywords

A combined approach with simultaneous application of several methods, including Transformer LLM, is proposed in [27]. Unfortunately, this work does not provide data allowing us to evaluate the effect of applying modern LLMs. Experimentally confirmed advantages of the combined approach are described in [30]. In this paper, LLM is combined with knowledge graphs, in which information is represented as triplets. Medical texts are processed and it is shown that the combined use of these two approaches provides a better result than they can give separately. The idea of combining LLM with knowledge graphs seems to be very promising.

In the task of extracting keywords from abstracts of Russian-language scientific articles using the BERTScore metric [33], the average result of three LLMs (Saiga, Mistral, and Vikhr) in zero-shot mode was 76.05, when soft fine-tuned on three random examples was 77.28. The average result of two classical statistical methods, YAKE and RuTermExtract, was 72.54. The average result of two fine-tuned neural networks, mT5 and mBART, was 77.37. Thus, LLMs demonstrate a better result than classical statistical methods, and are inferior to specialized fine-tuned neural networks, but not by much.

These results are further demonstrated in [8], fine-tuned mT5 model is compared to TopicRank, YAKE, RuTermExtract, KeyBERT using F-measure, ROUGE-1 and BERTScore metrics. According to this article, mT5 and RuTermExtract shows the highest performance in terms of the BERTScore metric (76.89 and 75.80), where mT5 demonstrates better results when generating keywords not presented in the source text.

Regarding the studies examining end-to-end solutions using only LLM for keyword extraction, given the different tasks, datasets, and metrics used, a direct comparison of results is not possible. It can only be noted that several studies [9, 24], like ours, have noted the high potential of LLM even without additional fine-tuning for keyword-related tasks.

A significant number of publications are devoted to the application of LLM for extracting critical terms in various subject areas to solve specific problems. In [7], extraction of keywords from electronic medical records to create a database of oncological diseases is described. In [2], application of the PhenoBERT system and LLM for extracting phenotypes from clinical records is considered and it is shown that LLM can find information missed by experts.

In [15], LLM-based framework for extracting terms from E-commerce platform texts is proposed. Specific applications are given and the effectiveness of this approach is experimentally demonstrated. In article [6], LLM is applied to the analysis of historical documents. In article [19], LLM is applied to extract objects and their properties from agricultural texts in order to obtain information about pests. It is shown that LLM can achieve better results than conventional methods.

Relatively few publications are available on application of LLM for term extraction in the social sciences and humanities. In [5], the lack of standardized datasets for applying LLM is noted and a dataset for political texts is presented.

It should be noted that not all works demonstrate the advantages of LLM in this problematics. Thus, in [32] it is shown that they successfully perform in extraction of terms only from simple sentences, while statistical methods are preferable in more complex situations. An important study is [3], which compares the results shown by 3 models: Llama2-7B, GPT-3.5, and Falcon-7B. This article discusses various factors: hallucinations, prompt quality, dependence on the subject area.

An important area of research closely related to keyword extraction is topic modeling. It has been extensively studied and applied in many subject areas, including in education for analyzing the structure of textbooks [21, 23]. A comprehensive overview of topic modeling can be found in [29].

In [17], LLM was proposed to be used to extract topics instead of traditional methods such as Latent Dirichlet Allocation (LDA). It was noted that LDA and similar methods do not take semantics into account, and in this regard, LLM has a significant advantage. The article shows that this is indeed the case and concludes that LLM can serve as an effective means of topic modeling. In [31], it was shown that LLM can be used to assess the relevance of topics instead of human experts.

LLMs have been used to extract topics in a number of subject areas. Thus, in [28], the BERTopic system, created on the basis of the early LLM BERT, was successfully applied to analyze psychotherapeutic texts. Its further development is presented in [12]. On a dataset of news texts, the combined use of modern LLM and standard topic modeling techniques improved topic coherence by about 10% compared to standard methods. In [14], it is proposed to combine LLM and cluster analysis methods. Thus, the general current trend is to combine LLM with standard techniques, and all studies demonstrate an improvement in quality of the extracted topics. LLM is becoming a new standard in topic modeling.

It is worth noting that the quality of textbooks is determined not only by their compliance with standards, but also by text complexity, which must be accessible to students. An overview of text complexity issues can be found in [22].

3. Results and Discussion

All textbooks were processed in the proposed pipeline. In terms of topic analysis, coverage of each topic for each textbook was calculated as well as mean coverage of all standard topics. Keywords not found by the primary analysis were processed with LLM, and were additionally found in all textbooks. Figure 5 shows an example of the result card for one textbook and one topic with all these data. The same steps were done for terms analysis. Figure 6 shows the corresponding result card for one book.

In summary, mean coverage of each topic and mean coverage of terms on all textbooks were calculated, as well as mean LLM recovery. Table 3 shows these statistics. Mean coverage among all topics is 71%. Mean LLM recovery is 51%. Figure 4 shows distribution of topic coverage in each textbook. This output data can be further used by textbook authors to revise their works and by official experts for comparative analysis of compliance to the educational standard.

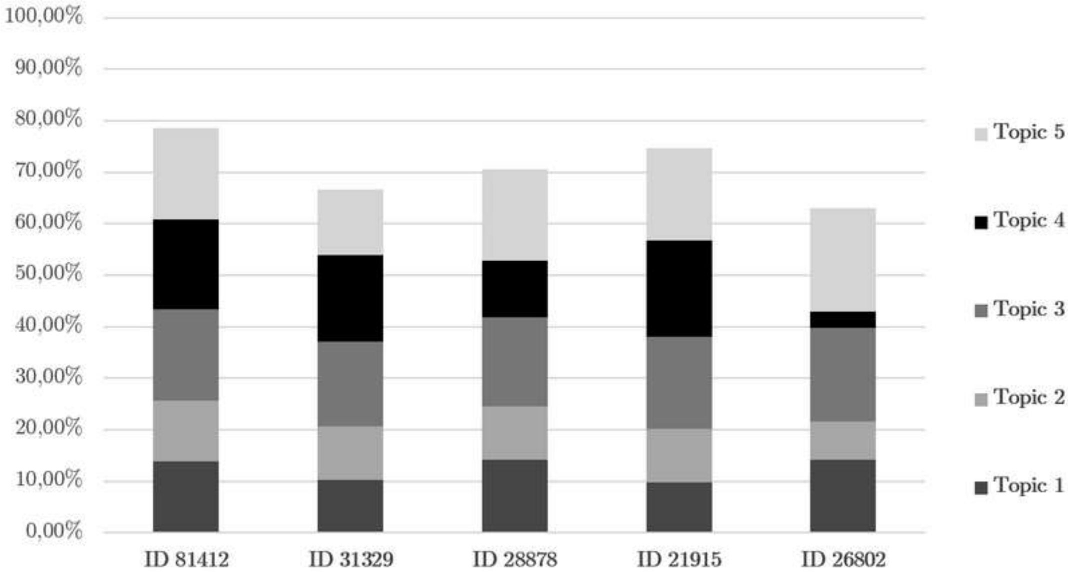

Figure 4. Topic coverage among all textbooks

Table 3. Results statistics

Topic	Number of keywords	Mean coverage	Mean LLM recovery
Topic 1	35	62%	56%
Topic 2	27	50%	52%
Topic 3	35	88%	60%
Topic 4	31	67%	24%
Topic 5	19	86%	64%
Terms	88	81%	70%

The first question for discussion is why the symbolic part of the algorithms does not find all the standard keywords and terms, even when they are present in textbook. The main weakness of the symbolic approach is its reliance on exact symbolic representation of words, without consideration of synonymy or semantics. This is why the preprocessing is an important step, where some nested keyword and term representations from the standard are expanded. For example, excerpt *деформация (упругая, пластическая) = deformation (elastic, plastic)* from the standard is expanded into terms *деформация, упругий_ деформация, пластический_ деформация = deformation, elastic_ deformation, plastic_ deformation.*

Furthermore, all of the keywords and terms not found by the symbolic approach, but recovered using LLM, are presented in textbooks in forms of synonyms or descriptions. First example, the term *правило_ равновесие = equilibrium_ rule* was not found, but its synonym *условие равновесия = equilibrium condition* is present in the textbook ID 81412, so it was recovered by

LLM (see Fig. 6). Second example, the term *агрегатный_ состояние* = *aggregate_ state* was not found, but the textbook ID 81412 contains an excerpt with the description of solid, liquid and gaseous states which correspond to this term, so it was recovered by LLM.

The second question for discussion concerns error analysis of LLM-based keyword recovery. Let us define our null hypothesis about every keyword not found by the symbolic part of the algorithm as not present in a given textbook. Then every error of LLM output is either type I error (keyword recovered, but not present in the textbook) or type II error (keyword not found by LLM, but present in the textbook in some form). Then type I errors are divided into type $\mathrm{I_a}$ errors (keyword recovered from false synonym or similar wording), type $\mathrm{I_b}$ errors (keyword recovered from similar description) and type $\mathrm{I_c}$ errors (keyword recovered as hallucination). Type II errors are divided into type $\mathrm{II_a}$ errors (keyword is present as synonym) and type $\mathrm{II_b}$ errors (description of keyword is present).

Table 4. Error statistics for terms recovery

Error	**ID 81412**	**ID 31329**	**ID 28878**	**ID 21915**	**ID 26802**	**Average**
Type I	0%	19%	0%	7%	22%	10%
Type $\mathrm{I_a}$	0%	13%	0%	7%	19%	8%
Type $\mathrm{I_b}$	0%	6%	0%	0%	4%	2%
Type $\mathrm{I_c}$	0%	0%	0%	0%	0%	0%
Type II	30%	19%	17%	7%	11%	17%
Type $\mathrm{II_a}$	10%	6%	6%	7%	0%	6%
Type $\mathrm{II_b}$	20%	13%	11%	0%	11%	11%
All types	30%	38%	17%	14%	33%	27%

Let us consider error analysis for terms recovery (Tab. 4), as they, by design of the educational standard, represent keywords from all of the topics. The table shows the ratio of LLM error number to number of terms not found by the symbolic part of the algorithm which should be processed by LLM.

As can be seen in these results, LLM produces less type I errors than type II errors on average, with the percentage of type I errors sufficiently low, that we can draw a considerable level of trust for experts and textbook authors to the output of LLM-based keyword recovery. No hallucinations occurred, for every false keyword recovery there can be given an explanation whether it was due to similar wording in textbook (dominant number of errors), or due to similar description. Additional analysis of hallucination occurrence was done in the form of introduction of trap words. We added to prompt some terms from physics topics of higher education level, such as *relativity theory*, *electrodynamics*, and some terms from completely different topics, namely literature and biology. No new hallucinations occurred after this addition.

As for type II errors, there are dominance of not found descriptions of keywords in textbook, which can be explained as attention mechanism in the LLM working in large context when the keyword description excerpts in the textbook are distributed sparsely. While the number of type II errors is higher, the full list of not found keywords is due for manual expert check in the educational standard compliance analysis regardless, therefore the level of trust in this case is out of question.

Another note on the results considers textbooks ID 31329 and ID 26802. These textbooks had a higher number of keywords not found by the symbolic part of the algorithm, and there are more errors than in other textbooks. We can conclude that the higher number of keywords the LLM have to find in the textbook, the less trust to its output should there be. This note raises a preliminary question about the effectiveness of end-to-end LLM-only solution to the problem of education standard compliance, without the symbolic preprocessing which can significantly reduce the number of keywords in the LLM input in an explainable and trusted manner.

It should be noted that in the LLM-based keyword recovery step we used the extracted text versions of the textbooks (.txt format) rather than source forms of Word documents (.docx) as the latter approach produced considerably more errors in the result.

Conclusion

The presented research tests application of classical symbolic NLP methods and LLM to keyword extraction. The practical task addressed and employed is the terminological alignment between educational standards and textbooks content. Within the neurosymbolic approach employed, we apply a hybrid methodology and combine symbolic methods for topic modeling of 5 Russian textbooks on Physics for 7th-graders. The pipeline is constituted of (1) keyword extraction with tokenization, lemmatization, n-gram retrieval and bag of words methods for topic modeling followed by (2) applying ChatGPT and qualitative analysis for items not found algorithmically. Having applied the symbolic methods we evaluated a 71% mean coverage of standard topics across all textbooks. Employment of LLMs resulted in recovery of 51% keywords initially missed by the abovementioned methods.

The research results show that hybrid neurosymbolic approach performs adequately good for the task of educational standard compliance analysis. The symbolic part of our algorithm finds most of the keywords and terms from the standard, with its output being explainable without manual post-check. Usage of LLM allows for additional recovery of keywords and terms as their synonyms and descriptions, with mostly reliable output.

As the textbooks were published under different Federal State Educational Standards over the period 2002–2019, the practical results are only partially applicable for educational and policy making purposes, although offer a scalable tool to automate analysis of content compliance of two sources, thus reducing the manual workload. Another limitation of the study is related to the application of ChatGPT 5 only. Given the rapid progress of LLMs and obsolescence of earlier models such as GPT2, in this study we used the most modern model, i.e., ChatGPT 5, and left comparison of the results to be achieved by other LLMs with those obtained in this work as a baseline for future studies. Another direction of future study is the experimental verification of an end-to-end solution for the educational standard compliance analysis based only on LLM with soft fine-tuning.

Acknowledgements

The research was supported by the RSF grant 24-28-01355 "Genre-discourse characteristics of the text as a function of lexical range".

Textbook: ID 81412
Standard: Physics 7th grade Basic
Mean coverage of standard topics: 79%

Standard topic: Topic 1
Keywords in topic: 35
Coverage: 69%
Found keywords (24, showing first 10):

	Keyword (Ru)	Keyword (En)	Instances	Frequency
1	гипотеза	hypothesis	11	0.000193
2	измерение	measurement	19	0.000333
3	модель	model	8	0.000140
4	наблюдение	observation	14	0.000245
5	наука	science	11	0.000193
6	природа	nature	27	0.000473
7	термометр	thermometer	4	0.000070
8	физика	physics	11	0.000193
9	физический_величина	physical_quantity	21	0.000368
10	физический_явление	physical_phenomenon	6	0.000105

Not found keywords (11):

	Keyword (Ru)	Keyword (En)
1	датчик_температура	temperature_sensor
2	физический_превращение	physical_transformation
3	постановка_научный_вопрос	formulation_scientific_question
4	жидкостный_термометр	liquid_thermometer
5	естественнонаучный_метод_познание	natural_scientific_method_cognition
6	научный_вопрос	scientific_question
7	физический_прибор	physical_device
8	объяснение_наблюдать_явление	explanation_observed_phenomenon
9	выдвижение_гипотеза	hypothesis_proposal
10	описание_физический_явление	description_physical_phenomenon
11	химический_превращение	chemical_transformation

LLM-found keywords (4):

	Keyword (Ru)	Keyword (En)
1	физический_прибор	physical_device
2	выдвижение_гипотеза	hypothesis_proposal
3	описание_физический_явление	description_physical_phenomenon
4	объяснение_наблюдать_явление	explanation_observed_phenomenon

Figure 5. Topic analysis results representation

Textbook: ID 81412
Standard: Physics 7th grade Basic
Terms in standard: 88
Coverage: 89%
Found terms (78, showing first 10):

	Term (Ru)	Term (En)	Instances	Frequency
1	атмосферный_давление	atmospheric_pressure	48	0.000841
2	атом	atom	32	0.000561
3	вес	weight	51	0.000894
4	весы	scales	8	0.000140
5	взаимодействие_тело	body_interaction	1	0.000018
6	время	time	43	0.000754
7	гипотеза	hypothesis	11	0.000193
8	давление	pressure	53	0.000929
9	движение	motion	52	0.000911
10	деформация	deformation	8	0.000140

Not found terms (10):

	Term (Ru)	Term (En)
1	агрегатный_состояние_вещество	aggregate_state_matter
2	правило_равновесие	equilibrium_rule
3	объём_вещество	substance_volume
4	правило_равновесие_рычаг	lever_equilibrium_rule
5	химический_явление	chemical_phenomenon
6	агрегатный_состояние	aggregate_state
7	высотомер	altimeter
8	пластический_деформация	plastic_deformation
9	превращение_механический_энергия	transformation_mechanical_energy
10	равновесие_твёрдый_тело	equilibrium_solid_body

LLM-found terms (6):

	Term (Ru)	Term (En)
1	правило_равновесие	equilibrium_rule
2	объём_вещество	substance_volume
3	правило_равновесие_рычаг	lever_equilibrium_rule
4	химический_явление	chemical_phenomenon
5	агрегатный_состояние	aggregate_state
6	равновесие_твёрдый_тело	equilibrium_solid_body

Figure 6. Term analysis results representation

References

1. Alier, M., Casañ, M.J., Filvà, D.A.: Smart Learning Applications: Leveraging LLMs for Contextualized and Ethical Educational Technology. In: Gonçalves, J.A.d.C., Lima, J.L.S.d.M., Coelho, J.P., *et al.* (eds.) Proceedings of TEEM 2023. pp. 190–199. Springer Nature Singapore, Singapore (2024). `https://doi.org/10.1007/978-981-97-1814-6_18`

2. Baddour, M., Paquelet, S., Rollier, P., *et al.*: Phenotypes Extraction from Text: Analysis and Perspective in the LLM Era. In: 2024 IEEE 12th International Conference on Intelligent Systems (IS). pp. 1–8 (2024). `https://doi.org/10.1109/IS61756.2024.10705235`

3. Chataut, S., Do, T., Gurung, B.D.S., *et al.*: Comparative Study of Domain Driven Terms Extraction Using Large Language Models (2024), `https://arxiv.org/abs/2404.02330`

4. Esfahani, M.N.: Content Analysis of Textbooks via Natural Language Processing. American Journal of Education and Practice 8(4), 36–54 (2024). `https://doi.org/10.47672/ajep.2252`

5. Foisy, L.O.M., Proulx, E., Cadieux, H., *et al.*: Prompting the Machine: Introducing an LLM Data Extraction Method for Social Scientists. Social Science Computer Review (2025). `https://doi.org/10.1177/08944393251344865`

6. Galushko, I.N.: Historical documents classification using BERT: LLM and historical domain. Perm University Herald. History 2(69), 147–158 (Jun 2025). `https://doi.org/10.17072/2219-3111-2025-2-147-158`

7. Gilbert, M., Crutchfield, A., Luo, B., *et al.*: Using a Large Language Model (LLM) for Automated Extraction of Discrete Elements from Clinical Notes for Creation of Cancer Databases. International Journal of Radiation Oncology*Biology*Physics 120(2, Supplement), e625 (2024). `https://doi.org/10.1016/j.ijrobp.2024.07.1375`

8. Glazkova, A.V., Morozov, D.A., Vorobeva, M.S., *et al.*: Keyword Generation for Russian-Language Scientific Texts Using the mT5 Model. Autom. Control Comput. Sci. 58(7), 995–1002 (Dec 2024). `https://doi.org/10.3103/S014641162470041X`

9. Glazkova, A., Morozov, D., Garipov, T.: Key Algorithms for Keyphrase Generation: Instruction-Based LLMs for Russian Scientific Keyphrases. In: Panchenko, A., Gubanov, D., Khachay, M., *et al.* (eds.) Analysis of Images, Social Networks and Texts. pp. 107–119. Springer Nature Switzerland, Cham (2025). `https://doi.org/10.1007/978-3-031-88036-0_5`

10. Grootendorst, M.: KeyBERT: Minimal Keyword Extraction with BERT. `https://www.maartengrootendorst.com/blog/keybert` (2020), accessed: 2025-08-29

11. Jo, T.: Keyword Extraction. In: Text Mining: Concepts, Implementation, and Big Data Challenge, pp. 421–443. Springer Nature Switzerland, Cham (2024). `https://doi.org/10.1007/978-3-031-75976-5_20`

12. Kapoor, S., Gil, A., Bhaduri, S., *et al.*: Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling (2024), `https://arxiv.org/abs/2409.15626`

13. Kuhn, J.: Computational text analysis within the Humanities: How to combine working practices from the contributing fields? Language Resources and Evaluation 53(4), 565–602 (Dec 2019). https://doi.org/10.1007/s10579-019-09459-3

14. Liu, J., Shang, Z., Ke, W., *et al.*: LLM-Guided Semantic-Aware Clustering for Topic Modeling. In: Che, W., Nabende, J., Shutova, E., *et al.* (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 18420–18435. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-long.902

15. Maragheh, R.Y., Fang, C., Irugu, C.C., *et al.*: LLM-TAKE: Theme-Aware Keyword Extraction Using Large Language Models. In: 2023 IEEE International Conference on Big Data (BigData). pp. 4318–4324 (2023). https://doi.org/10.1109/BigData59044.2023.10386476

16. Monakhov, S.I., Turchanenko, V.V., Cherdakov, D.N.: Terminology use in school textbooks: corpus analysis. RUDN Journal of Language Studies, Semiotics and Semantics 14(3), 437–456 (2023). https://doi.org/10.18413/2313-8912-2023-9-1-0-3

17. Mu, Y., Dong, C., Bontcheva, K., *et al.*: Large Language Models Offer an Alternative to the Traditional Approach of Topic Modelling. In: Calzolari, N., Kan, M.Y., Hoste, V., *et al.* (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 10160–10171. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.887/

18. Papagiannopoulou, E., Tsoumakas, G.: A review of keyphrase extraction. WIREs Data Mining and Knowledge Discovery 10(2), e1339 (2020). https://doi.org/10.1002/widm.1339

19. Peng, R., Liu, K., Yang, P., *et al.*: Embedding-based Retrieval with LLM for Effective Agriculture Information Extracting from Unstructured Data (2023), https://arxiv.org/abs/2308.03107

20. Priyanshu, A., Vijay, S.: AdaptKeyBERT: An Attention-Based approach towards Few-Shot & Zero-Shot Domain Adaptation of KeyBERT (2022), https://arxiv.org/abs/2211.07499

21. Sakhovskiy, A., Tutubalina, E., Solovyev, V., *et al.*: Topic Modeling as a Method of Educational Text Structuring. In: 2020 13th International Conference on Developments in eSystems Engineering (DeSE). pp. 399–405 (2020). https://doi.org/10.1109/DeSE51703.2020.9450232

22. Solnyshkina, M.I., Solovyev, V.D., Gafiyatova, E.V., *et al.*: Text complexity as interdisciplinary problem. Issues of Cognitive Linguistics (1), 18–39 (2022). https://doi.org/10.20916/1812-3228-2022-1-18-39

23. Solovyev, V., Solnyshkina, M., Tutubalina, E.: Topic Modeling for Text Structure Assessment: The case of Russian Academic Texts. Journal of Language and Education 9(3), 143–158 (Sep 2023). https://doi.org/10.17323/jle.2023.16604

24. Song, M., Jiang, H., Shi, S., *et al.*: Is ChatGPT A Good Keyphrase Generator? A Preliminary Study (2023), https://arxiv.org/abs/2303.13001

25. Steiss, J., Tate, T., Graham, S., *et al.*: Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction 91, 101894 (2024). https://doi.org/10.1016/j.learninstruc.2024.101894

26. Sukying, A., Barrot, J.S.: Friend or Foe? Investigating the Alignment of English Language Teaching (ELT) Textbooks with the National English Curriculum Standards. The Asia-Pacific Education Researcher 34(2), 793–801 (Apr 2025). https://doi.org/10.1007/s40299-024-00896-5

27. Tunyan, E.G., Sazikov, R.S., Kharlamov, S.A.: Automatic Extraction of Keywords and Summaries for Knowledge Base Population. Russian Journal of Cybernetics 6(2), 108–113 (Jun 2025), https://en.jcyb.ru/nisii_tech/article/view/413

28. Vanin, A., Bolshev, V., Panfilova, A.: Applying LLM and Topic Modelling in Psychotherapeutic Contexts (2024), https://arxiv.org/abs/2412.17449

29. Wu, X., Nguyen, T., Luu, A.T.: A survey on neural topic models: methods, applications, and challenges. Artificial Intelligence Review 57(2), 18 (Jan 2024). https://doi.org/10.1007/s10462-023-10661-7

30. Yang, J.: Integrated application of LLM model and knowledge graph in medical text mining and knowledge extraction. Social Medicine and Health Management 5(2), 56–62 (2024). https://doi.org/10.23977/socmhm.2024.050208

31. Yang, X., Zhao, H., Phung, D., *et al.*: LLM Reading Tea Leaves: Automatically Evaluating Topic Models with Large Language Models. Transactions of the Association for Computational Linguistics 13, 357–375 (2025). https://doi.org/10.1162/tacl_a_00744

32. Zagatti, F.R., Lucrédio, D., Caseli, H.d.M.: Unsupervised Statistical Keyword Extraction Pipeline: Is LLM All You Need? In: Paes, A., Verri, F.A.N. (eds.) Intelligent Systems. pp. 460–475. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-79032-4_32

33. Zhang, T., Kishore, V., Wu, F., *et al.*: BERTScore: Evaluating Text Generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net (2020), https://openreview.net/forum?id=SkeHuCVFDr