Parsing Bibliography Descriptions using Few-Shot Prompting with LLM

Kenan Kassab, Nikolay Teslya
SPC RAS
Saint-Petersburg, Russia
kassab.k@iias.spb.su; teslya@iias.spb.su

***Abstract*—In this work, we propose a large language model (LLM)-based approach for parsing Russian bibliographic references into structured components. We thoroughly assess current approaches, from rule-based techniques to machine learning and deep learning parsers, and point out their drawbacks when used with the Russian bibliography. These drawbacks are caused by issues including non-standard formatting, transliteration inconsistency, and Cyrillic script, in addition to the lack of annotated Russian datasets. On the other hand, LLMs are highly suited for low-resource tasks, since they exhibit remarkable flexibility to multilingual contexts and attain excellent accuracy even in zero-shot and few-shot scenarios. We evaluated the performance of multiple LLMs on a custom Russian bibliography dataset, achieving good performance with F1-scores above 80%. The results of our study demonstrate that LLMs are a viable and efficient approach to bibliographic parsing in languages with limited resources.**

I. Introduction

Extracting bibliographic data is a fundamental task in digital libraries, academic knowledge management, and information retrieval [1], [2]. In order to facilitate automated literature analysis, metadata enrichment, and citation indexing, references are parsed into structured elements including authors, titles, journals, publication years, and volumes. In limited environments, conventional methods—from manually constructed regular expressions to machine learning models trained on annotated datasets [3]—have proven successful. However, their general usefulness is limited because they often have trouble generalizing across languages, topics, and different citation styles.

In particular, processing bibliographic entries in Russian presents unique challenges due to a combination of linguistic, formatting, and resource-related factors [4]. Existing parsers are largely developed for English-language references and are heavily tailored to standardized citation formats (e.g., APA, MLA, Chicago). Citation formats in Russian publications deviate from commonly accepted international approaches. This can include differences in punctuation, author name ordering, journal title abbreviations, and the presence or exclusion of specific information fields. Moreover, the majority of open-source bibliographic parsing tools were created for English-language data and are ineffective at handling Cyrillic input, which results in incorrect classifications and parsing errors. The lack of large annotated datasets for Russian bibliographies limits the possibility of training supervised models, making the task considerably more complex than its English counterpart.

Recent developments in large language models (LLMs) offer a new approach to solve this task. In natural language comprehension, LLMs demonstrate good zero-shot and few-shot capabilities, which makes them well suited for problems with high format variation and little labeled input. The contribution of the paper can be summarized as:

- Propose a method for interpreting Russian bibliographic references based on LLMs.
- Analyze the drawbacks of previous parsing tools and highlight the impact of LLMs in low-resource and multilingual bibliographic parsing tasks

The rest of the paper is divided as follows: Section II explores previous tools used to parse bibliography descriptions. Section III describes the custom dataset used in this research. Section IV introduces the LLM-based approach followed to parse bibliographic descriptions. Section V presents the experiments conducted on the data. Section VI shows the results achieved by the LLM models with comprehensive comparison. Section VII concludes the paper.

II. Related Work

For many years, researchers have been studying the automatic parsing of bibliographic references, and their work has developed in parallel with developments in machine learning and natural language processing. This section examines the main reference parsing paradigms, emphasizing the enduring difficulties that drive our LLM-based methodology, especially in a multilingual setting.

Initially, manual rules and heuristics were used in automated bibliographic parsing attempts. To find structural clues in a citation string, such as punctuation, keywords (such ”Vol.” and ”pp.”), and numerical patterns for years and page numbers, these systems usually employed regular expressions and pattern-matching algorithms. This method served as the basis for processing structured formats such as MARC data and for early digital library systems [5]. These techniques were made possible by the creation of specific programming languages for text processing [6]. Nevertheless, there are serious problems with the rule-based paradigm. It is extremely fragile; even a slight departure from the intended format might lead to the failure of the parsing process as a whole. The process of developing and maintaining the complex rule sets needed to accommodate a wide variety of citation styles is time-consuming and prone to mistakes. Additionally, these methods

are not appropriate for diverse and changing scholarly data since they do not generalize effectively across languages or new, unseen citation forms.

Researchers used machine learning to reframe reference parsing as a sequence labeling task in order to overcome the limits of rule-based systems. This framework assigns a label to each token (word or character) in the citation string that corresponds to a bibliographic field (e.g., B-AUTHOR, I-AUTHOR, B-TITLE). Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs), which were more effective and went on to become the standard for many years because of their capacity to absorb a large number of features, were among the first models in this area [7] [8]. Additional advancements were made with the introduction of deep learning. Long-range dependencies in the text were better captured by Recurrent Neural Networks (RNNs), especially those with Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) architectures. By learning deep contextualized representations of text, Transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) have lately raised the bar [9]. On benchmark datasets, these models can be optimized for sequence tagging and achieve excellent accuracy. In order to improve performance, recent research has continued to improve these techniques, for example by introducing prompt-based learning and contrastive learning [10]. These supervised machine learning techniques, in spite of their achievements, are all dependent on the requirement for large, higher-quality, manually annotated training datasets. The development of such a resource is a major constraint, particularly for languages with limited resources and no extensive bibliographic databases. This ”data scarcity” issue has been the main obstacle to creating efficient Russian parsing systems.

A paradigm change in natural language processing has been brought about by the development of LLMs such as GPT models and its successors [11]. These models, which have been trained on web-scale text corpora, gain a general comprehension of syntax, semantics, reasoning, and world knowledge that can be applied to new tasks with little task-specific information [12]. This has been proven in a number of complex fields, such as scientific research [13], education [14], and medical [15].

The main benefit of LLMs for tasks like bibliographic parsing is their capacity to function in a few-shot or even zero-shot scenario. One can direct the LLM to generate structured output from an unstructured input string by giving it a well-crafted prompt that contains a task description and a few examples. This method directly addresses the primary constraint for low-resource languages by avoiding the requirement for a sizable labeled dataset. Because of their pre-trained understanding of literary patterns and structures, which they can apply to the particular format of a bibliographic reference, LLMs are effective at such information extraction tasks.

Even though NLP has advanced significantly, much of this improvement has been focused on English. When faced with the linguistic and stylistic diversity of other languages, the majority of parsing tools and datasets are ineffective because they are English-centric. The Russian language presents a particularly challenging case due to its unique citation styles, rich morphology (such as case endings that alter the form of words and names), and Cyrillic script. The design of such systems should also take into account at least the standards GOST 7.1–2003 [16], GOST 7.0.5–2008 [17] and GOST 7.0.100–2018 [18] as reference models for the structure of bibliographic elements.

Research on extracting fields from Russian-language bibliographic descriptions relies on regular expressions and grammars, including the use of the Tomita parser for metadata extraction from full-text publications [19]. Here up to 86.7% of records are reported as correctly extracted and highly formalized fields (e.g., ISBN) are recognized most reliably . However, such methods are highly sensitive to variability in GOST formatting.

Previous Natural Language Processing (NLP) studies on Russian have often emphasized the necessity for resources and models customized to the language. New datasets frequently have to be created from scratch for studies in named entity recognition and sentiment analysis [20] [21]. Fortunately, the most recent generation of LLMs has been trained on large, varied corpora that contain significant quantities of Russian material, making them more and more multilingual. These models have performed well on a variety of Russian natural language processing tasks [22], indicating that they may be able to handle the more difficult task of bibliographic parsing.

A study [23] presents a systematic comparison of open-source systems for field extraction from bibliographic references, evaluating ten parsers (including GROBID, CERMINE, ParsCit, etc.) in both out-of-the-box and retrained settings. The results indicate that machine-learning approaches substantially outperform rule-based methods in recall, with GROBID showing the strongest baseline performance, and that retraining on domain-specific data can further boost accuracy (e.g., CERMINE improving from F1 0.83 to 0.92).

In this section, we highlighted the methods used for bibliography parsing from rule-based to machine learning models. Although these techniques have advanced for English, there is still no reliable, low-cost, and precise way to parse Russian bibliographic references. The absence of extensive annotated data needed to train specialized models has been the main challenge. This paper aims to fill this gap by utilizing the capabilities of modern LLMs for this task. We propose that LLMs' few-shot learning capability and inherent multilingual knowledge offer a straightforward and practical answer to the bibliography parsing of Russian references.

III. Dataset Description

Bibliographic description is a key tool in scholarly communication and library-information activities. International and national standards regulate the rules for formatting references and source descriptions.

The dataset used in this study contains bibliographic descriptions of various types of publications. It includes books, book chapters, periodicals and their individual articles, and

collections of articles. The dataset is based on a more general dataset entitled *"Pushkiniana: Bibliography of Scholarly and Critical Works Dedicated to A. S. Pushkin"* [24]. It provides a complete summary of materials devoted to the works of Alexander Pushkin. Since the main body of sources consists of Russian-language publications issued in Russia, the dataset is presented in a format close to the Russian bibliographic description standard GOST 7.0.100–2018 [18]. After minor adaptation, the results can also be applied to other bibliographic description standards, since all of them are built around common principles. An overview of the standards is presented below.

The following sections provide a detailed analysis of the standards: APA [25], MLA [26], ISO 690:2010 [27], ISBD [28], BIBFRAME [28], and GOST 7.0.100–2018 [18].

A. Citation Styles Description

1) APA (American Psychological Association) style: is a citation style widely used in the social and behavioral sciences. The latest edition (7th) was published in 2019 [25]. It is characterized by a strict citation structure and focus on readability. In-text citations follow the author–year format. The reference list is arranged alphabetically. Mandatory elements for APA are author, year, title, publisher. For electronic sources DOI or URL should be provided.

2) MLA (Modern Language Association) style: is a citation style applied in the humanities, especially in literary and cultural studies. The latest edition (9th) was issued in 2021 [26]. It emphasizes transparency and flexibility: references contain all elements necessary to identify the source, regardless of medium. In-text references follow the author–page principle. Mandatory elements: author, title, container (journal, book, website), publisher, publication date, page numbers (if applicable).

3) Harvard referencing style: also known as the author–date system, is one of the most widely used referencing methods in academia, particularly in the UK and Commonwealth countries. It requires in-text citations with the author's surname and year of publication, and a corresponding reference list arranged alphabetically [29]. While not governed by a single official manual, it follows consistent principles similar to ISO 690. Mandatory elements in references are author, year, title, source (journal, book, or website), publisher, place of publication.

4) ISO 690:2021: is an international standard regulating references and bibliographic descriptions [27]. It supports three methods: author–year, numeric, and notes. It establishes order of elements, punctuation, and sequence. It applies to print, electronic, and audiovisual resources. Mandatory elements in ISO 690:2010 are author, year, title, place of publication, publisher. For electronic resources: access date and URL.

5) ISBD (International Standard Bibliographic Description): international rules developed by IFLA (International Federation of Library Associations and Institutions). The consolidated edition dates from 2011 with updates in 2021 [28]. Its purpose is to unify bibliographic descriptions in library catalogs, ensuring interoperability and data exchange. ISBD defines nine areas:

1) Content form and media type
2) Title and statement of responsibility
3) Edition
4) Material or resource-specific details
5) Production, publication, distribution
6) Physical description
7) Series and multipart resources
8) Notes
9) Resource identifier and terms of availability

Mandatory elements in ISBD are title, statement of responsibility, resource identifier (e.g., ISBN).

6) BIBFRAME (Bibliographic Framework): is a metadata model developed by the Library of Congress to replace the MARC format [30]. It is based on Semantic Web and Linked Data principles. The main entities are: *Work* (intellectual creation), *Instance* (publication), and *Item* (physical copy). Additional entities include *Agent*, *Subject*, and *Event*. The standard enables integration of bibliographic information into global information networks. Mandatory elements are classes Work (author, title), Instance (publisher, date), Item (identifier of the copy - DOI or URL).

7) GOST 7.0.100–2018: is the latest version of Russian national standard harmonized with ISBD [18]. It replaced the previously valid GOST 7.1–2003 and has been applied since July 1, 2019 for all types of bibliographic lists, including scientific works. It regulates bibliographic descriptions of all types of documents. Three levels of completeness are defined: short, extended, and full descriptions. It specifies strict punctuation and order of elements. Mandatory elements are title, statement of responsibility, edition, physical description including year, book title, volume, issue, page numbers, resource identifier (ISBN, DOI).

All the reviewed standards aim at ensuring correct description of sources, yet their scope and level of detail differ (see Table I). APA and MLA target scholarly writing, ISO 690 provides international unification, ISBD and GOST are used in library cataloging, while BIBFRAME represents a new paradigm of bibliographic metadata integration into the Semantic Web (see Table II). Despite the strictness of the standards due to their complexity these standards are usually soften in real usage.

From the complete set of 55,000 records in [24] dataset, about two thousand were selected, representing various types of publications. Each bibliographic record was manually divided by two experts in the field of bibliography into individual fields that constitute the entry, such as: authors, title of the work, book or series title, publisher, place of publication (or more cities), year of issue, edition number, information about, editors and pages.

IV. Approach

Our approach parses Russian bibliographic references into structured fields by utilizing large language models (LLMs).

TABLE I. MANDATORY FIELDS OF BIBLIOGRAPHIC DESCRIPTION STANDARDS

Standard	Mandatory fields / elements
APA	Author, year, title, publisher.
MLA	Author, title, container, publisher, publication date, pages.
Harvard	Author, year, title, source, publisher, place of publication.
ISO 690:2021	Author, year, title, place of publication, publisher.
ISBD	Title, statement of responsibility, resource identifier (ISBN, etc.).
BIBFRAME	Work (author, title), Instance (date, publisher), Item (copy).
GOST 7.0.100–2018	Title, statement of responsibility, edition, year, book title, volume, issue, page numbers, identifier.

When given a clear task description, LLMs can complete this work with little modification, in contrast to rule-based systems like supervised parsers that need a large number of hand-made rules or annotated training data.

A. Prompt Design

We frame bibliography parsing as a structured extraction problem and use prompts to instruct the model about the correct approach to identify entities. Each prompt included:

1) **Task Description** – A detailed explanation of the parsing objective, emphasizing that the input is an unstructured bibliographic entry (raw text representing the bibliography reference) and the output should be a structured JSON-like format (describing the bibliography parts).
2) **Entity Schema** – A predetermined list of labels that could be used to indicate different parts of a bibliographic reference: AUTHOR, TITLE, BOOK_TITLE, ORGANIZATION, VOLUME, EDITOR, ADDITION, PUBLISHER, PLACE, YEAR, PAGES, SERIES
3) **Examples** – Few-shot examples of the desired behavior of parsing. These examples included both the expected structured output and the unformatted bibliographic text. These examples explain the task to the model in an informative way and formulate what we expected to do.

Figure 1 shows an example of the prompt we used. It should be mentioned that this example is translated into English but the original prompt feed to the model was in Russian. The prompt was also followed by a set of examples (20 samples of raw bibliography text with the corresponding annotated tags in a dictionary format) following few-shot prompt technique. The output of the model was post-processed to make sure it has a dictionary format so it can be compared with the ground-truth to calculate the evaluation metrics.

Through testing, we found that in order to obtain reliable results, it was essential to include both the complete task description and explanatory examples. By trailing, we found that example-guided prompts consistently increased extraction accuracy, whereas minimal prompts or schema-only prompts resulted in higher error rates.

```
You are an intelligent assistant for
    extracting information from bibliographic
    records.
Analyze this bibliographic record and
    highlight the following elements:

AUTHOR
TITLE (title of the article or work)
BOOK_TITLE (if the work is part of a book or
    collection)
ORGANIZATION (if specified)
VOLUME (if specified)
EDITOR (if available)
ADDITION (for example, notes, additional
    information)
PUBLISHER
PLACE (city or place of publication)
YEAR (year of publication)
PAGES (if specified)
SERIES (if specified)

The output should be in a dictionary format.
    Don't forget to close all parentheses.
```

Fig. 1. Example of the prompt used for the LLM approach

B. Model Selection and Tuning

We evaluated several LLMs hosted on a local server with Nvidia RTX 3090 TI 24Gb VRAM, 128 Gb DDR5 RAM using the Ollama web server. Due to the limited resources the following criteria were applied to model selection: i) model should be available for free; ii) it should fit to 24Gb VRAM; iii) The model should support Russian language and not to replace Cyrillic text with Latin or Chinese in output, iv) it should have high score in tests. Following these criteria, three models were initially selected for evaluation:

- gpt-oss (OpenAI open source model), gpt-oss:20b model
- Llama 3.2 (Meta), llama3.2:3b model
- Gemma 3 (Google DeepMind), gemma3:27b model

Before moving on to the whole evaluation, we tried several prompt versions for each model on tiny sections of the dataset to determine the best configuration. All findings were achieved in a zero-shot or few-shot prompting scenario without any fine-tuning.

C. Evaluation Metrics

Precision, recall, and F1-score for each type of entity, along with overall averages across all fields, were used to evaluate each model's performance. These metrics allow us to capture not only correctness but also completeness of extraction, which is especially crucial when working with optional variables (e.g., editor, series, addition).

V. EXPERIMENTS

We experimented with current citation parsing methods to determine baseline performance prior to evaluating large language models. Three popular systems that represent various methods of bibliographic parsing were chosen by us:

TABLE II. COMPARISON OF BIBLIOGRAPHIC DESCRIPTION STANDARDS

Characteristic	APA	MLA	Harvard	ISO 690:2021	ISBD	BIBFRAME	GOST 7.0.100–2018
Format type	Academic style	Academic style	Academic style	International standard	International description rules	RDF model	National standard
Mandatory fields	Author, year, title, publisher	Author, title, container, date	Author, year, title, source, publisher, place	Author, year, title, publisher	Title, responsibility, identifier	Work, Instance, Item	Title, responsibility, edition, physical description (year, book title, volume, issue, page numbers), identifier
Level of detail	Medium	Medium	Medium	Medium	High: 9 areas	High, semantic	High, three levels
Purpose	Social/behavioral sciences	Humanities	Broad academic referencing	Reference unification	Interoperability of catalogs	Linked Data integration	Adaptation of ISBD to Russian sources

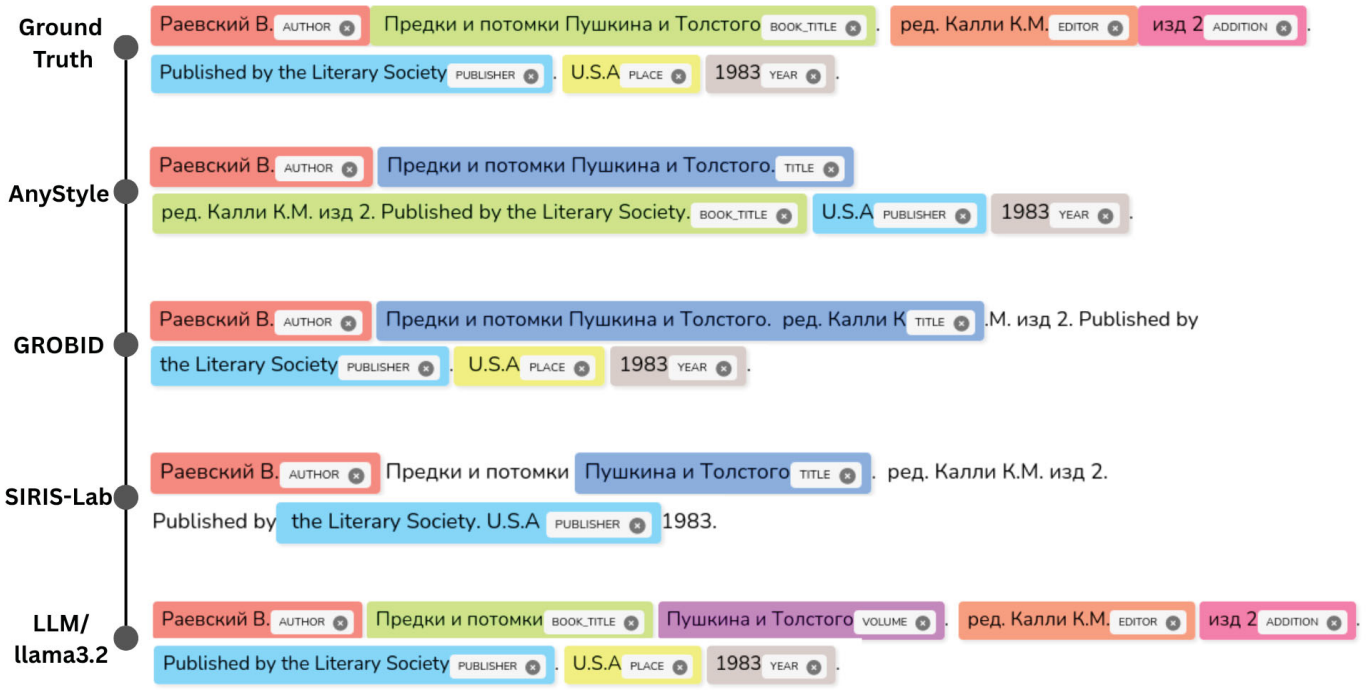

Fig. 2. Comparison between the parsing tools on a data sample

- **AnyStyle** [31]: A parser that breaks down citation strings into structured parts using machine learning algorithms.
- **GROBID** [32]: A machine learning library widely used for parsing PDF documents and extracting bibliographic metadata, primarily optimized for English and scientific publications.
- **SIRIS-Lab** [33]: A multilingual named entity recognition (NER) model trained with the DistilBERT-base-multilingual-cased architecture, designed to handle multilingual citation data.

The comparison of selected systems (except SIRIS-Lab) from [23], showing that the best out-of-the-box performance is achieved by GROBID (F1 = 0.89), while machine learning–based methods generally provide comparable precision but about three times higher recall compared to rule-based systems.

Each tool was additionally tested on our dataset, and the predicted outputs were compared with ground-truth annotations. Table III shows the overall metrics achieved by these method in parsing the bibliographic descriptions with out-of-the-box configuration. We can notice low values of metrics which reflect that the conventional approaches do not generalize well

to our task. Fig. 2 shows a visualization of the comparison results on a sample data example compared to the ground truth.

TABLE III. COMPARISON BETWEEN CONVENTIONAL APPROACHES ON OVERALL PERFORMANCE

Model	Metrics	Overall
AnyStyle	precision	0.2862
	recall	0.1662
	f1-score	0.2040
GROBID	precision	**0.3369**
	recall	**0.1979**
	f1-score	**0.2425**
SIRIS-Lab	precision	0.2527
	recall	0.1651
	f1-score	0.1963

AnyStyle frequently misclassified Russian entries, despite being useful for English-style references. Key data like publisher and place were either left out or assigned incorrectly, and cyclic characters were frequently handled wrongly. Compared to AnyStyle, GROBID demonstrated better structure recognition; nonetheless, it had trouble with Cyrillic script and deviance from international standards (APA, MLA, etc.). Titles were sometimes confused with book titles, and it frequently split fields incorrectly. SIRIS-Lab was tested by developers [33] and provide overall performance F1 = 0.94 on test dataset. Despite being multilingual, its performance on Russian bibliography was weaker than expected. The model handled optional fields incorrect and frequently failed to recognize parts such as addition and place.

The limits of traditional bibliographic parsers on Russian data are illustrated by these trials. Because rule-based and pretrained machine learning algorithms rely heavily on English-language datasets and standardized formats, they are not appropriate for the complexity and variety of Russian references. On the other hand, when given explicit task descriptions and examples, large language models (LLMs) demonstrated the ability to handle transliterations, parse irregular forms, and adapt to Cyrillic script. Figure 2, also shows the results obtained by Llama 3.2 model on the sample data compared to other tools.

VI. RESULTS

We compiled the results of LLM testing into a comparative table, reporting entity-level and overall performance on the dataset. Table IV presents a summary of our evaluation's findings, including overall averages for the three tested models (gpt-oss, Llama 3.2, and Gemma 3) as well as precision, recall, and F1-scores for each entity type. With an average F1-score that was higher than the other models, Llama 3.2 performed the best overall. While Gemma 3 fared inconsistently across all categories, gpt-oss performed competitively, especially on essential elements like AUTHOR and YEAR.

Table V and Table VI show the results of each model by entity type. Breaking down results by entity type reveals several patterns:

TABLE IV. COMPARISON BETWEEN LLM MODELS ON OVERALL PERFORMANCE

Model	Metrics	Overall
Gemma 3	precision	0.6469
	recall	0.6050
	f1-score	0.6229
Llama 3.2	precision	**0.8293**
	recall	**0.7907**
	f1-score	**0.8052**
gpt-oss	precision	0.7120
	recall	0.6766
	f1-score	0.6810

- **AUTHOR**: All models consistently produced the greatest ratings in this sector. Their significant linguistic cues (capitalization, ordering, and reference location) are reflected in this, as they are present in almost all entries.
- **TITLE** and **BOOK_TITLE**: The models are challenged by these fields to accurately identify and differentiate between them. On these tags, Llama 3.2 performed better than the other models.
- **YEAR** and **PAGES**: Extraction was generally reliable, as these fields often follow numerical patterns easily recognizable by LLMs.
- **PUBLISHER** and **ORGANIZATION**: With decreased recall and precision, these categories were more difficult, especially for Gemma 3. The variety of abbreviations and irregular formatting in Russian references are the source of the challenge.
- **EDITOR**, **SERIES**, and **ADDITION**: Due in significant part to their rarity in the dataset, these optional fields performed the worst overall. These fields were frequently left out of models (poor recall) or mistaken for related items (precision error).

In a nutshell, LLM performed well in solving the task of parsing Russian bibliography descriptions. The most balanced performance across entities was shown by Llama 3.2, which had high F1-scores overall. Additionally, it handled rare entities more robustly than the other models. On the other hand, gpt-oss shows competitive results in fields that were frequently used, but had trouble with entities that were optional or less common. At the end, Gemma 3 showed the lowest overall performance, especially in handling optional fields and publisher-related metadata.

A possible explanation for these differences may lie in the composition of the training corpora for each model. All models were updated just before evaluation from the Ollama repository. The most recent model is gpt-oss, since Llama 3.2 is the oldest one. Llama 3.2 likely had greater exposure to Cyrillic and Russian-language materials, which could explain its stronger handling of bibliographic conventions in Russian. In contrast, Gemma 3's performance may result from limited training coverage of Russian sources. The competitive scores of gpt-oss on AUTHOR and YEAR suggest that it benefits from robust handling of high-frequency and structurally simple tokens (e.g., capitalization and numeric patterns), but it lacks

TABLE V. COMPARISON BETWEEN THE LLM MODELS ON THE ENTITY LEVEL

Model	Metrics	AUTHOR	TITLE	BOOK_TITLE	ORGANIZATION	VOLUME	EDITOR
Gemma 3	precision	0.987239	0.222597	0.112882	0.956522	0.512500	0.384615
	recall	0.994159	0.381503	0.090426	0.105769	0.318653	0.336879
	f1-score	**0.990687**	0.281150	0.100413	0.190476	0.392971	0.359168
Llama 3.2	precision	0.882940	0.701657	0.753406	0.746032	0.843373	0.851064
	recall	0.908497	0.682796	0.724771	0.681159	0.793451	0.761246
	f1-score	0.895536	**0.692098**	**0.738811**	**0.712121**	**0.817651**	**0.803653**
gpt-oss	precision	0.828947	0.336449	0.610889	0.746914	0.695502	0.455253
	recall	0.948627	0.555270	0.476577	0.573460	0.477435	0.400000
	f1-score	0.884758	0.419011	0.535439	0.648794	0.566197	0.425842

robustness for low-frequency, linguistically varied fields. Additionally, sensitivity to prompting strategies may have amplified performance disparities, as models differ in their tolerance to instruction variation since common prompt was used for all models.

VII. CONCLUSION

This study shows that large language models provide an efficient solution for interpreting Russian bibliographic references. Without requiring large annotated datasets, LLMs are able to handle formatting variability, transcription, and Cyrillic script by utilizing zero-shot and few-shot prompting. Our tests demonstrate that LLMs provide the flexibility needed for low-resource and multilingual environments in addition to producing excellent quantitative outcomes. Our LLM-based approach was able to achieve high F1-score (more than 80%) in parsing the custom bibliography description dataset. These results demonstrate LLMs' potential as a workable substitute for bibliographic parsing, with further implications for metadata extraction in other languages and fields with limited resources.

Although our study shows that LLMs are a useful tool for parsing Russian bibliographic references, more research is required to validate and expand on these results. A next step is to carry out a thorough benchmarking research that directly contrasts the LLM-based approach with more conventional techniques like rule-based methods, supervised machine learning models, and deep learning architectures such as transformer-based citation parsers or sequence tagging. This will show the trade-offs between different paradigms, particularly in terms of accuracy, robustness, and computational efficiency.

ACKNOWLEDGMENT

This work was supported by the Russian State Research FFZF-2023-0001.

REFERENCES

[1] A. Eldallal and E. Barbu, "Bibrank: Automatic keyphrase extraction platform using metadata," *Information*, vol. 14, no. 10, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/10/549

[2] K. Williams, J. Wu, Z. Wu, and C. L. Giles, "Information extraction for scholarly digital libraries," in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, ser. JCDL '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 287–288. [Online]. Available: https://doi.org/10.1145/2910896.2925430

[3] Q. Zhang, Y.-G. Cao, and H. Yu, "Parsing citations in biomedical articles using conditional random fields," *Computers in Biology and Medicine*, vol. 41, no. 4, pp. 190–194, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482511000291

[4] D. Maltseva, V. Vashchenko, and K. Lika, "Methodology of processing bibliographic data in russian language to construct collaboration networks (using the example of the elibrary database)," *Sociology: methodology, methods, mathematical modeling (Sociology: 4M)*, no. 54-55, pp. 45–78, 2022. [Online]. Available: DOI:https://doi.org/10.19181/4m.2022.31.1-2.2

[5] J. Riley, C. Mullin, and C. Hunter, "Automatically batch loading metadata from marc into a work-based metadata model for music," *Cataloging & Classification Quarterly*, vol. 47, no. 6, pp. 519–543, 2009. [Online]. Available: https://doi.org/10.1080/01639370902936446

[6] H. Fosdick, "Programming languages for library and textual processing," *Bulletin of the American Society for Information Science and Technology*, vol. 31, no. 6, pp. 21–26, 2005. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/bult.2005.1720310607

[7] D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, "Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, ser. JCDL '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 99–108. [Online]. Available: https://doi.org/10.1145/3197026.3197048

[8] P. Lopez, "Automatic extraction and resolution of bibliographical references in patent documents," in *Advances in Multidisciplinary Retrieval*, H. Cunningham, A. Hanbury, and S. Rüger, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 120–135.

[9] S. Nakayama, T. Kanazawa, F. Uwano, and M. Ohta, "Error detection of bert-based bibliographic information extraction from reference strings," in *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2025, pp. 1–8.

[10] Z. Yin and S. Wang, "Enhancing bibliographic reference parsing with contrastive learning and prompt learning," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108548, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624007061

[11] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge, "Summary of chatgpt-related research and perspective towards the future of large language models," *Meta-Radiology*, vol. 1, no. 2, p. 100017, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2950162823000176

[12] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[13] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, vol. 624, no. 7992, pp. 570–578, 2023.

[14] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1041608023000195

TABLE VI. COMPARISON BETWEEN THE LLM MODELS ON THE ENTITY LEVEL

Model	Metrics	ADDITION	PUBLISHER	PLACE	YEAR	PAGES	SERIES
Gemma 3	precision	0.0	0.599602	0.944030	0.996875	0.100000	0.846154
	recall	0.0	0.419220	0.938195	1.000000	0.231884	0.835443
	f1-score	0.0	0.493443	**0.941104**	**0.998435**	0.139738	**0.840764**
Llama 3.2	precision	0.738739	0.758535	0.920934	0.996474	0.732456	0.791667
	recall	0.645669	0.731140	0.885590	0.916342	0.607273	0.459677
	f1-score	**0.689076**	**0.744585**	0.902916	0.954730	**0.664016**	0.581633
gpt-oss	precision	0.119048	0.733010	0.959470	0.994595	0.345679	0.398010
	recall	0.038760	0.576776	0.853675	0.909765	0.496454	0.634921
	f1-score	0.058480	0.645575	0.903486	0.950291	0.407569	0.489297

[15] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, and D. Chartash, "How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment," *JMIR Med Educ*, vol. 9, p. e45312, Feb 2023. [Online]. Available: https://mededu.jmir.org/2023/1/e45312

[16] "Russian state standard gost 7.1-2003. bibliographic record. bibliographic description. general requirements and rules of compiling," Standard, 2003, in Russian.

[17] "Russian state standard gost r 7.0.5-2008. bibliographic reference. general requirements and rules of compiling," Standard, 2008, in Russian.

[18] *GOST R 7.0.100-2018 Bibliographic record. Bibliographic description. General requirements and rules*, Federal Agency for Technical Regulation and Metrology (Russia) Std., 2018, in Russ.

[19] R. Suleymanov, "Extraction of metadata from the full-text electronic materials written in russian using tomita-parser," *Software Systems*, vol. 14, pp. 58–62, 2016, in Russian.

[20] V. Karyukin, G. Mutanov, Z. Mamykova, G. Nassimova, S. Torekul, Z. Sundetova, and M. Negri, "On the development of an information system for monitoring user opinion and its role for the public," *Journal of Big Data*, vol. 9, no. 1, p. 110, 2022.

[21] A. Sboev, A. Selivanov, I. Moloshnikov, R. Rybka, A. Gryaznov, S. Sboeva, and G. Rylkov, "Extraction of the relations among significant pharmacological entities in russian-language reviews of internet users on medications," *Big Data and Cognitive Computing*, vol. 6, no. 1, 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/1/10

[22] S. Chumakov, A. Kovantsev, and A. Surikov, "Generative approach to aspect based sentiment analysis with gpt language models," *Procedia Computer Science*, vol. 229, pp. 284–293, 2023, 12th International Young Scientists Conference in Computational Science, YSC2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923020203

[23] D. Tkaczyk, A. Collins, P. Sheridan, and J. Beel, "Evaluation and comparison of open source bibliographic reference parsers: a business use case," *arXiv preprint arXiv:1802.01168*, 2018.

[24] E. I. Vozhik, A. A. Dimyanenko, E. O. Kazakova, A. S. Kolotovkina, R. A. Lisyukov, and K. A. Maslinsky, "Pushkiniana: Bibliography of scholarly and critical works dedicated to a. s. pushkin," 2024. [Online]. Available: https://doi.org/10.31860/openlit-2024.9-B017

[25] *Publication Manual of the American Psychological Association*, 7th ed. Washington, DC: American Psychological Association, 2020, authoritative guide to APA Style, Seventh Edition.

[26] *MLA Handbook*, 9th ed. Modern Language Association of America, 2021, official, authorized guide on MLA style.

[27] *ISO 690:2021 — Information and documentation: Guidelines for bibliographic references and citations to information resources*, International Organization for Standardization Std., Jun. 2021, reference number ISO 690:2021, published June 2021. [Online]. Available: https://www.iso.org/standard/72642.html

[28] I. R. Group *et al.*, *ISBD: International Standard Bibliographic Description: Consolidated Edition*. Walter de Gruyter, 2011, vol. 44.

[29] *Harvard Referencing: A Guide Based on "Cite Them Right"*, https://librarydevelopment.group.shef.ac.uk/referencing/harvard.html, University of Sheffield Library, 2016, based on the "Cite Them Right" (10th revised and expanded edition).

[30] L. of Congress, *BIBFRAME: Model, Vocabulary, Guidelines, Examples, Notes, Analyses*, last visited September 8, 2025. [Online]. Available: https://www.loc.gov/bibframe/docs/index.html

[31] K. Sylvester, "Anystyle," https://github.com/inukshuk/anystyle?tab=readme-ov-file, 2022.

[32] "Grobid," https://github.com/kermitt2/grobid, 2008–2025.

[33] S. Lab and R. D. of SIRIS Academic, "Siris-lab/citation-parser-entity," https://huggingface.co/SIRIS-Lab/citation-parser-ENTITY, 2025.