Russian-Language Multimodal Dataset for Automatic Summarization of Scientific Papers

Tsanda Alena[1] and Bruches Elena[1,2]

[1] Novosibirsk State University, Russia
[2] A.P. Ershov Institute of Informatics Systems, Russia

Abstract. The paper discusses the creation of a multimodal dataset of Russian-language scientific papers and testing of existing language models for the task of automatic text summarization. A feature of the dataset is its multimodal data, which includes texts, tables and figures. The paper presents the results of experiments with two language models: Gigachat from SBER and YandexGPT from Yandex. The dataset consists of 420 papers and is publicly available on https://github.com/iis-research-team/summarization-dataset.

Keywords: Natural Language Processing · Automatic Text Summarization · Multimodal Dataset · Large Language Models.

1 Introduction

It is essential for any researcher to read scientific literature in order to be aware of the latest trends and state-of-the-art solutions in their field of interest. However, with the wealth of textual content growing dramatically on the Internet, especially scientific papers, the reading process can be inconvenient. Therefore, researchers need representative abstracts to understand the topics of the papers and focus on the relevant information. Manual text summarization consumes a lot of time, effort and cost and even becomes impractical with the upward trend towards automation. The main goal of any automatic summarization system is to produce a summary that includes the main ideas of the input document in less space without repetitions [1].

Existing solutions mainly work for the English language and for texts of the general domain. Besides, the methods of text summarization are not so developed for Russian, especially for domain-specific texts like scientific papers, hence the need to adapt such methods and develop new ones for this language. What is more, there are very few Russian-language datasets of scientific texts, especially for this task.

This paper is devoted to the development of a multimodal dataset of scientific papers written in Russian and testing existing language models on the collected data. The main specific feature of the dataset proposed in this work is its multimodality: it includes not only texts of papers and their abstracts, but also tables and figures with their descriptions. Singular focus on textual information fails to capture the full richness of multimodal data as most existing paper interpretation systems do [2]. Containing valuable information, tables and figures can noticeably improve the quality of abstracts. Moreover, the creation of such datasets is a long and painstaking process, but they are especially relevant due to the development of multimodal models. Multimodal models are the models that make predictions based on different modalities, i.e. different types of information [3].

In this work, we have two main contributions:

1. We present the carefully curated multimodal dataset for the scientific papers summarization for Russian, which covers 7 scientific domains and contains 420 scientific papers, including textual and visual information.
2. We evaluate the most widely used language models for Russian, namely GigaChat and YandexGPT, on this dataset. The study showed that handling scientific texts is ambiguous for these models as a number of texts were rejected due to ethical reasons.

2 Related Works

In this chapter, we review the literature about the task of summarization, existing datasets for this task and multimodal large language models.

2.1 The Task of Summarization

Large volumes of textual information on the Internet have complicated the task of searching and reading all the relevant documents. The field of automatic text summarization (ATS) has emerged as a solution for condensing extensive texts and creating accurate summaries [4]. ATS provides users with the opportunity to quickly grasp the main idea of a document without having to manually filter through information.

The main objective of any ATS system is to produce an effective summary, i.e. a summary that distills the most important information from a source (or sources) to produce an abridged version of the original information for a particular user(s) and task(s) [5].

There is a wide variety of different classifications and applications of ATS systems. Such systems may be classified based on their approach, input size, summarization algorithm, summary content, summary type, summarization domain, etc. They may be applied in many different ways as well: news, sentiment, books, email summarization, legal or biomedical documents summarization and others [1]. Particularly, text summarization is extremely functional for scientific papers. Indeed, scholars often face the challenge of sifting through vast amounts of literature to keep up with developments in their field.

In terms of approaches, automatic text summarization is categorized into extractive, abstractive and hybrid [4]. The first approach is based on extracting the most high-scored sentences from the original document. Extractive approach is fast and simple, yet it produces summaries that are far from those written by humans. On the contrary, abstractive summarization is generated by either rephrasing or using the new words instead of simply extracting sentences from the initial document [6], which makes it more like a human-written one.

Some challenges of ATS like finding the most informative segments of a text, summarizing long documents such as books, evaluating computer-generated summaries without the need for the human-produced summaries make this task one of the most arduous yet significant tasks of NLP.

2.2 Datasets for Summarization

Summarization aims to extract the most important information from the different types of texts. Thus, there are plenty of datasets serving the different tasks.The most popular summarization datasets are CNN/DailyMail [8] and XSum dataset [8], they are used for news summarization. Besides, there are task-specific datasets, such as for dialogue summarization [9,10], chat summarization [11], email summarization [12].

The most challenging aspect is long text summarization. Intuitively, long document summarization is harder than short document one due to the significant difference in the number of lexical tokens and breadth of content between short and long documents. As the length increases, the content that would be considered important increases as well, resulting in a more challenging task for an automatic summarization model to capture all salient information in the limited output length [13]. Such texts may be presented in books, long documents and scientific papers. The most widely used datasets for scientific paper summarization are arXiv and PubMed [14]. However, these datasets are intended for English. As for the Russian language, there are very few datasets. For text summarization there is Gazeta [15]. Besides, there are Russian-language parts of the datasets MLSUM [16] and XL-Sum [17].

Another important aspect of the current state-of-the-art is multimodal datasets, which contain not only text information, but also images, tables, etc. For example, in [18] SciMMIR benchmark was proposed to evaluate a model's abilities to use several types of information for information retrieval over the scientific texts. Constructing such datasets is a time-consuming process, but they may play an important role even in the model's evaluation.

2.3 Multimodal LLMs

Large Language Models, or LLMs, are basically AI systems that use deep learning techniques and massively large datasets to comprehend and generate human language text. These models have transformed natural language processing (NLP), branching their influence into various domains [19]. Undoubtedly, LLMs with their unprecedented capabilities have redefined the way we understand language and generate text.

Along with a great deal of modern research, a wide variety of different LLMs have emerged. Large language models can be classified based on their size - from small ($\geq$ 1B parameters) to very large ($>$ 100B parameters); on their type

- foundation, instruction or chat model; on their origin and availability [20]. Besides, there are some very popular LLM families such as the GPT family developed by OpenAI [21], the LLaMa family released by Meta [22], the PaLM from Google [23] and others.

Operating mainly with the pure text data, traditional large language models perform well at tasks like text generation and encoding but have limitations in understanding other data types [24]. Meanwhile, large vision foundation models make rapid progress in perception. Consequently, more and more attention focuses on combination with text, modality alignment and task unity [25]. This leads to the new field of multimodal LLMs, or MLLMs, which handle various modalities of information. What is more, multimodality is a key component of achieving general artificial intelligence, because it plays a pivotal role in interacting with the real world.

The use of multimodal input vastly expands the capabilities of existing language models. However, it still remains an active area of research as the majority of LLMs are trained only on textual data.

3 Dataset

In this section, we provide a description of the dataset, the procedure of its creation and the statistics.

3.1 Data Sources

Our goal is to create a dataset for the scientific papers summarization for Russian. This dataset should cover different domains, including technical, humanitarian and natural sciences. Such diversity is important as each domain has its own scientific traditions, patterns and different set of metadata such as formulas, tables, figures etc. The created dataset demonstrates this diversity, which can be seen in Table 1 and Table 2.

The following scientific domains were included in the dataset: Economics, History, Information Technologies (IT), Journalism, Law, Linguistics and Medicine.

The papers were collected from three scientific journals: Vestnik TSU[3] (Economics, History, Journalism, Law, Linguistics, IT), Vestnik NSU[4] (IT) and Medicinsky journal[5] (Medicine).

It should be noticed that even within the same journal and domain the structure of the papers can vary significantly. For example, there is no unified structure for the Journalism papers: some texts are divided by sections while others are continuous texts without any explicit structure.

We consider this feature as a challenge for the modern language models – in practice, we expect that they will extract the most important information without memorizing or relying on the section names, structure and other features that can be used as a hint for the models.

[3] https://journals.tsu.ru/

[4] https://vestnik.nsu.ru/

[5] https://sibmed.elpub.ru/jour/index

3.2 Dataset Creation

The dataset consists of the papers from 7 different scientific domains, each domain has 60 texts.

All texts were collected from the newest publications to the earlier ones. We chose only papers which contain the most part of the text in Russian. Some papers contain quite a lot of text inclusions in other languages such as Chinese, German, etc.

Each text has the following metadata:

1. Name of the paper;
2. Abstract;
3. Text of the paper;
4. Figures and their names;
5. Tables and their names.

It is worth noting that we clean text from such metadata as author names, keywords, footnotes, references and appendix. Additionally, we remove all links to the references from the text.

All figures and tables contained in the papers were screenshotted and saved as .png files. The descriptions for the visual data were extracted from their headings in the paper. In case there was no heading for the figure or table, this information was extracted from the context. For example, from the context "*We show the details of TIVE in Fig. 2*" one can extract the possible description for the figure "*The details of TIVE*".

Some tables may be torn into two or more pages. In this case, we save the tables in several .png files, depending on the number of pages containing this table.

In terms of the dataset storage, each paper occupies one folder, which contains the following files: name.txt, abstract.txt, text.txt, image_number, table_number, figures.json, tables.json.

3.3 Dataset Statistics

As it was mentioned before, the current version of the dataset consists of 420 texts from 7 scientific domains.

Tables 1 and 2 show the general statistics such as text length, number of tokens, figures and tables for the abstracts and texts correspondingly.

One may notice that in humanitarian domains texts tend to be longer, but they rarely include figures and tables. Technical and natural sciences contain the most part of visual information and shorter texts. This insight shows that quite a lot of useful information is encoded in other modalities than textual one and may play a crucial role for such semantic tasks as summarization, information extraction, etc.

4 Benchmarks

In this section, we describe the used methods and the obtained results.

Table 1. Statistics for the abstracts.

Domain	Length in chars	Length in tokens
Economics	64 271	7 004
History	34 211	3 787
IT	36 277	3 822
Journalism	31 981	3 664
Law	33 288	3 423
Linguistics	35 190	3 806
Medicine	97 061	11 296
Total	332 279	36 802

Table 2. Statistics for the texts.

Domain	Length in chars	Length in tokens	Figures	Tables
Economics	1 316 995	151 284	32	25
History	1 540 251	184 407	2	17
IT	1 002 115	114 721	238	27
Journalism	1 377 087	174 064	45	12
Law	1 243 153	143 675	0	2
Linguistics	1 557 481	190 478	1	1
Medicine	963 178	107 449	19	45
Total	9 000 260	1 066 078	337	129

4.1 Models

As a baseline, we decided to check whether the final section of the papers in the dataset correlates with the abstracts. Due to the fact that the papers do not always have a strict structure, not every paper has the section "Conclusion". In total, 183 pairs of "abstract – conclusion" were formed and taken to evaluate the quality. The results are presented in Tables 3 and 4 in section 4.2.

Besides, we conducted some experiments with large language models. The first model whose performance was tested is Gigachat[6] from SBER. This model is based on a neural network ensemble which includes, in particular, models for text generation. The significant advantage of this language model is that it was trained for the Russian language. However, Gigachat strives to avoid controversial ethical issues, and for this reason only 37% of texts from the dataset were processed without being censored. The model is particularly reluctant to work with legal and medical texts, as well as topics related to journalism. Nevertheless, 157 abstracts to dataset papers were generated with help of the langchain framework[7] designed to create applications using large language models. We evaluated the quality of generated abstracts taking the abstracts from the dataset as the reference summaries just as we did in the baseline. The prompt used for addressing the model: "*Below is a scientific paper. Highlight the main facts and write a summary of this paper*".

[6] https://developers.sber.ru/portal/products/gigachat

[7] https://github.com/langchain-ai/langchain

The second model that has been utilized is the generative language model from Yandex called YandexGPT[8]. Like the previous one, this model was trained on Russian-language data. With help of the Yandex GPT Lite model, we generated 295 abstracts for the dataset papers using the exact same prompt as earlier. Along with the previous one, this model did not manage to process all 420 papers due to censorship and length of some texts. All Yandex GPT API generation models currently have a limit of 8 000 tokens for the input and output sequences together.

4.2 Results

To evaluate the quality of the generated abstracts, it is required to compare the candidate summaries with the reference ones, i.e. the ones from the dataset, in order to determine how equivalent they are. For this purpose, the following text generation metrics were selected: BERTScore, BLEURT, ROUGE-1, ROUGE-2, ROUGE-L, BLEU. It is worth mentioning that expert evaluation of the collected dataset was not carried out.

The first metric, BERTScore, is based on pre–trained contextual embeddings BERT and calculates the similarity of sentences as the sum of cosine similarities between the embeddings of the tokens they consist of [26]. This metric is especially effective for abstractive summarization since it takes into account such points as paraphrasing and changing the order of words.

The second metric, BLEURT, similarly to the first one uses BERT embeddings and takes into account the semantics of the text. A key component of this metric is a special pre-training scheme on a large volume of synthetic data [27].

The ROUGE and BLEU metrics are considered to be standard for text generation and are based on calculating overlapping n-grams. They are more suitable for extractive summarization, since such summarization is formed from the most important sentences of the source document. ROUGE-N is a metric related to recall, while BLEU is related to accuracy [28]. Apart from this difference, the BLEU metric uses a brevity penalty so that the candidate text matches the reference one not only in the choice and order of words, but also in length [29]. Both metrics calculate the percentage of n-grams in the candidate summary that match the n-grams in the reference summary, i.e. abstracts from our dataset.

Except BLEURT, the values of the metrics range from 0 to 1, where 1 means the best match. The BLEURT score may go beyond the specified range and be negative, but the perfect value is also considered to be 1.

During experiments with the large language models, the first step was to calculate the values of these metrics on those abstracts that the models managed to generate (Table 3).

Unfortunately, it is quite problematic to compare the values of the metrics with one another in this case, since the number of generated texts, their lengths and domains differ significantly. To take into account the number of generated summaries and plausibly compare the work of the models, the same metrics were

[8] https://cloud.yandex.ru/ru/services/yandexgpt

calculated for all 420 abstracts, including empty lines if the model did not manage to generate a summary (Table 4). Although there was a substantial decline in values of the metrics in this case, it became possible to compare the models within this particular task. Additionally, we evaluated the models performance across different domains using the metrics: BERTScore, BLEURT, ROUGE-1, ROUGE-2, ROUGE-L (Table 5).

The first LLM, Gigachat, performed the worst according to the metrics. The reason for this might be that the model managed to generate the fewest number of summaries due to censorship, almost half as many as the YandexGPT model. For instance, the model generated only 9 abstracts for Journalism out of 60 possible due to these restrictions, yet it processed almost all IT papers. Unlike the model of Yandex, its considerable advantage is that it works with longer papers.

On the contrary, YandexGPT has the best scores for almost all the metrics. While this model does have some thematic limitations, it has processed 40 papers for Journalism. However, due to the length of the texts, it generated the fewest number of summaries for scientific fields like Linguistics and History.

In addition, the abstracts generated by the models differ in length. The average summary from Gigachat contains about 93 words or 5 sentences, while the average summary generated by YandexGPT has approximately 218 words or 15 sentences. This makes the latter significantly longer and may affect the metrics.

Some examples of the generated texts are provided here: https://github.com/iis-research-team/summarization-dataset/blob/main/generated_abstracts.md.

Table 3. Metrics on the generated abstracts.

Model	BERTScore	BLEURT	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Baseine	0.708	0.187	**0.173**	**0.099**	**0.167**	1.4e-155
Gigachat	**0.719**	**0.189**	0.148	0.071	0.142	1.3e-155
YandexGPT	0.692	0.204	0.118	0.053	0.113	5.1e-80

Table 4. Metrics on all the abstracts.

Model	BERTScore	BLEURT	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Baseine	0.310	-0.295	0.075	**0.044**	0.073	1.3e-155
Gigachat	0.270	-0.349	0.055	0.027	0.053	0.9e-155
YandexGPT	**0.486**	**-0.057**	**0.083**	0.037	**0.080**	5.1e-80

5 Limitations

Current version of the dataset covers only 7 scientific domains, but it should be increased to include more diverse areas.

Table 5. Metrics by domains.

Model	Texts	BERTScore	BLEURT	ROUGE-1	ROUGE-2	ROUGE-L
Baseine$_{Medicine}$	52	**0.725**	0.143	0.263	**0.191**	**0.251**
Gigachat$_{Medicine}$	39	0.718	0.158	0.203	0.116	0.194
YandexGPT$_{Medicine}$	**53**	0.718	**0.172**	**0.268**	0.128	0.248
Baseline$_{Economics}$	26	0.699	0.210	**0.030**	**0.005**	**0.026**
Gigachat$_{Economics}$	13	**0.740**	**0.251**	0.0	0.0	0.0
YandexGPT$_{Economics}$	**45**	0.686	0.226	0.017	0.0	0.017
Baseline$_{Linguistics}$	11	0.687	**0.207**	**0.123**	**0.109**	**0.123**
Gigachat$_{Linguistics}$	22	**0.718**	0.182	0.034	0.0	0.034
YandexGPT$_{Linguistics}$	**37**	0.685	0.199	0.067	0.022	0.064
Baseline$_{Law}$	9	0.678	0.220	0.0	0.0	0.0
Gigachat$_{Law}$	14	**0.718**	**0.253**	0.029	0.0	0.029
YandexGPT$_{Law}$	**39**	0.684	0.226	**0.033**	**0.009**	**0.033**
Baseline$_{IT}$	55	0.720	0.205	**0.210**	**0.105**	**0.204**
Gigachat$_{IT}$	50	**0.734**	**0.209**	0.178	0.098	0.169
YandexGPT$_{IT}$	**57**	0.691	0.203	0.114	0.047	0.111
Baseline$_{Journalism}$	22	0.687	**0.212**	0.149	0.029	0.149
Gigachat$_{Journalism}$	9	**0.702**	0.084	**0.262**	**0.089**	**0.262**
YandexGPT$_{Journalism}$	**40**	0.683	0.205	0.134	0.044	0.134
Baseline$_{Chemistry}$	**56**	**0.755**	**0.190**	**0.411**	**0.244**	**0.381**
Gigachat$_{Chemistry}$	41	0.728	0.172	0.252	0.121	0.242
YandexGPT$_{Chemistry}$	**56**	0.713	0.142	0.243	0.096	0.219
Baseline$_{History}$	9	0.674	0.158	0.108	0.069	0.108
Gigachat$_{History}$	10	**0.704**	0.206	**0.281**	0.100	**0.281**
YandexGPT$_{History}$	**24**	0.689	**0.210**	0.179	**0.136**	0.179

The most challenging part is to create a subset with technical papers such as Maths or Physics as they contain plenty of formulas that are important data. It is still an open question how one should store such information: as a raw text or as a LateX text. We are going to solve this problem in our future work.

Moreover, current evaluation does not support all modalities (both text and images) as input for the LLMs as only few models support such API. This is also a very important direction of our ongoing research. Nevertheless, this dataset may already be used to evaluate the systems that support multimodal inputs.

Another nuance is table representations as images. However, it seems that one can make use of processing tabular data (for example, in csv format) as such modality is included in some modern models.

6 Conclusion

In conclusion, we proposed a russian-language multimodal dataset for the task of automatic text summarization and conducted some experiments that include testing large language models. The dataset is publicly available on GitHub, and we plan to expand it with other scientific fields.

According to the traditional text generation metrics, the final sections in the scientific papers of the corpus turned out to be syntactically closer to the abstracts in the corpus than those generated by the models. On the contrary, in terms of semantics large language models performed better according to the neural network metrics, which take into account the content of the text. Despite the fact that LLMs today are capable of generating high-quality text in natural language, they have some limitations concerning the length and the content of texts.

Currently, we plan to test the performance of other language models on the resulting dataset and conduct experiments with different extractive and abstractive approaches to text summarization. Through combining methods and analyzing the collected figures and tables, we aim to enhance the quality of the generated abstracts.

References

1. El-Kassas W., Salama Ch., Rafea A., Mohamed Hoda K.: Automatic Text Summarization: A Comprehensive Survey. In: Expert Systems with Applications, vol. 165, pp. 1–46 (2021). doi: 10.1016/j.eswa.2020.113679
2. Jiang F., Wang K., Li H. Bridging Research and Readers: A Multi-Modal Automated Academic Papers Interpretation System. arXiv:2401.09150, 13 p. (2024)
3. Suzuki M., Matsuo Y. A survey of multimodal deep generative models. In: Advanced Robotics, vol. 36, pp. 261-278 (2022). doi: 10.1080/01691864.2022.2035253
4. Jin H., Zhang Y., Meng D., Wang J., Tan J. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. arXiv:2403.02901, 20 p. (2024)
5. Maybury, M. T. Generating summaries from event data. In: Information Processing & Management, vol. 31, pp. 735-751 (1995). doi: 10.1016/0306-4573(95)00025-C

6. Gupta S., Gupta K S. Abstractive summarization: An overview of the state of the art. In: Expert Systems with Applications, vol. 121, pp. 49-65 (2019). doi: 10.1016/j.eswa.2018.12.011
7. Nallapati R., Zhou B., Nogueira dos santos C., Gulcehre C., Xiang B. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In: The SIGNLL Conference on Computational Natural Language Learning (CoNLL), pp. 280-290. Berlin, Germany (2016). doi: 10.18653/v1/K16-1028
8. Narayan Sh., Cohen B. Sh., Lapata M. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In: Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1797–1807. Brussels, Belgium (2018). doi:10.18653/v1/D18-1206
9. Zhong M., Yin D., Yu T., Zaidi A., Mutuma M., Jha R. et al. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5905-5921 (2021). doi: 10.18653/v1/2021.naacl-main.472
10. Zhu Ch., Liu Y., Mei J., Zeng M. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization. In: North American Chapter of the Association for Computational Linguistics (NAACL). Mexico City, Mexico (2021). doi: 10.18653/v1/2021.naacl-main.474
11. Gliwa B., Mochol I., Biesek M., Wawer A. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics, pp. 70-79. Hong Kong, China (2019). doi: 10.18653/v1/D19-5409
12. Mukherjee S., Mukherjee S. Smart To-Do: Automatic Generation of To-Do Items from Emails. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8680–8689 (2020). doi: 10.18653/v1/2020.acl-main.767
13. Gidiotis A., Tsoumakas G. A divide-and-conquer approach to the summarization of long documents. In: IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 3029–3040 (2020). doi: 10.1109/TASLP.2020.3037401
14. Cohan A., Dernoncourt F., Soon Kim D., Bui T., Kim S., Chang W., Goharian N. A discourse-aware attention model for abstractive summarization of long documents. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 2, pp. 615–621. New Orleans, Louisiana (2018). doi: 10.48550/arXiv.1804.05685
15. Gusev I. Dataset for Automatic Summarization of Russian News. In: AINL 2020: Artificial Intelligence and Natural Language, pp. 122–134 (2020). doi: 10.1007/978-3-030-59082-6_9
16. Scialom T., Dray P. A., Lamprier S., Piwowarski B., Staiano J. MLSUM: The Multilingual Summarization Corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8051–8067 (2020). doi: 10.18653/v1/2020.emnlp-main.647
17. Hasan T., Bhattacharjee A., Islam Md. S., Mubasshir K., Li Y. F., Kang Y. B. et al. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 4693-4703 (2021). doi: 10.18653/v1/2021.findings-acl.413
18. Wu S., Li Y., Zhu K., Zhang G., Liang Y., Ma K. et al. SciMMIR: Benchmarking Scientific Multi-modal Information Retrieval. arXiv:2401.13478, 14 p. (2024)
19. Wei Sh., Xu X., Qi X., Yin X., Xia J., Ren J. et al. AcademicGPT: Empowering Academic Research. arXiv:2311.12315, 28 p. (2023)

20. Minaee Sh., Mikolov T., Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J. Large Language Models: A Survey. arXiv:2402.06196, 43 p. (2024)
21. GPT: OpenAI. Introducing ChatGPT (2022), https://openai.com/blog/chatgpt, last accessed 2024/03/17
22. Touvron H., Lavril Th., Izacard G., Martinet X., Lachaux M. A., Lacroix T. et al. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, p. 14 (2023)
23. Chowdhery A., Narang Sh., Devlin J., Bosma M., Mishra G., Roberts A. et al. PaLM: Scaling Language Modeling with Pathways. In: The Journal of Machine Learning Research, vol. 24, pp. 11324–11436 (2022). doi: 10.48550/arXiv.2204.02311
24. Wu J., Gan W., Chen Z., Wan Sh., Yu P. S. Multimodal Large Language Models: A Survey. arXiv:2311.13165, 10 p. (2023)
25. Yin Sh., Fu Ch., Zhao S., Li K., Sun X., Xu T., Chen E. A Survey on Multimodal Large Language Models. arXiv:2306.13549, 15 p. (2023)
26. Zhang T., Kishore V., Wu F., Weinberger Q. K., Artzi Y. BERTScore: Evaluating Text Generation with BERT, arXiv:1904.09675, 43 p. (2020)
27. Sellam T., Das D., Parikh A. P. BLEURT: Learning Robust Metrics for Text Generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892 (2020). doi: 10.48550/arXiv.2004.04696
28. Lin Ch. Y.. ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of the ACL Workshop: Text Summarization Braches Out, pp. 74-81 (2004)
29. Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318 (2002) doi: 10.48550/arXiv.2004.04696