

Word Sense Disambiguation in Russian: A Generative LLM Approach

Polina Gousyatskaya, Natalia Loukachevitch

Lomonosov Moscow State University
polinagousyatskaya@gmail.com, louk_nat@mail.ru

Abstract. Various statistical and neural approaches to Word Sense Disambiguation (WSD) for the Russian language have been extensively studied, yet little attention has been given to leveraging generative LLMs for this task. The ability of generative LLMs to draw insightful conclusions due to their black-box nature and the possibility of interacting with them in natural language make them a promising tool for linguistic research. The aim of this study is to test the ability of generative models, particularly those adapted for Russian language, to distinguish the senses of ambiguous Russian words. The experimental setup we employ involves prompting the model to select the correct sense of an ambiguous word from a predefined list. We conduct the experiment on a range of Russian-language decoder-based LLMs, from state-of-the-art models to smaller ones. The objective of this study is twofold: first, to establish a baseline for generative LLM performance in Russian WSD; second, to conduct a linguistic analysis examining how different models handle various types of lexical ambiguity, namely homonymy and polysemy, as well as their ability to process sparse data. As a result, we outline the performance range across various models, with F1 scores spanning from 0.45 to over 0.9. We also identify linguistic factors influencing disambiguation difficulty, namely the semantic relatedness and representation of the senses. The largest average F1 score of 0.75 is achieved on homonyms with well-represented senses. These findings highlight key areas for future improvement in generative LLMs’ handling of lexical ambiguity in Russian.

Keywords: Word Sense Disambiguation, generative LLM, classification, lexical ambiguity, decoder-based models, homonymy, polysemic words

1 Introduction

Word Sense Disambiguation (WSD) is an indispensable problem in natural language processing (NLP), forming the foundation for the analysis of all other linguistic levels: WSD is not considered an end goal but rather a component of other natural language processing tasks, such as machine translation, information retrieval, or named entity extraction and linking [1].

This paper focuses on generative, or decoder-based, large language models. The encoder-based models, vividly represented by the BERT family, are capable of ana-

lyzing text with high precision and depth and have been successfully applied across various NLP tasks, achieving notable results in all its subfields. Decoder-based models, which generate text based on a natural language prompt, have been explored less frequently in NLP research. Being able to interpret user’s instructions in natural language is one of the most prominent features of generative LLMs; in comparison, encoder-based models, while being extremely effective, require a classic machine learning experimental setup, which makes them much less accessible to the general user. In modern research generative LLMs are widely regarded as a disruptive innovation, driving advancements across multiple domains, including NLP. Their rapid development has significantly expanded LLM-based research, with computational linguistics being no exception.

This study aims to integrate generative LLM-based methodologies with Word Sense Disambiguation as a key challenge in NLP. Given that this research direction remains relatively unexplored, particularly for Russian-language data, our objective is to establish a baseline through a foundational experiment, as well as to investigate how Russian-adapted models generally respond to the WSD task: how the models handle different types of lexical ambiguity, namely polysemy and homonymy, and to what extent the known influential factors in automated WSD affect the performance of decoder-based models on this task.

2 Related work

Word Sense Disambiguation (WSD) is traditionally framed as a classification problem, in contrast to Word Sense Induction, which is approached through clustering. One of the most well-established approaches in the field is the knowledge-based approach that utilizes structured, machine-readable lexical resources, such as WordNet and its Russian counterparts, RussNet [3] and RuWordNet [13]. Resources of this kind prove to be highly effective in disambiguation tasks, as the structure of ontologies and thesauri reflects paradigmatic and syntagmatic relationships between concepts, their corresponding lexical representations, and their meanings [1] [16]. Nevertheless, the time and labor required to develop structured lexical resources hinder the active development of knowledge-based approaches in the modern WSD.

The most widely-used setting for WSD is supervised machine learning encompassing classic machine learning algorithms that typically demand complex feature representations of natural language data, and neural approaches with encoder-based LLMs methods as their subset: the latter managed to notably relax the strict requirements for the complexity of feature representations, making it possible to embed individual tokens without the need for additional text annotation. The latest evolution in this approach involves relying not on human-constructed feature representations, but on the reasoning capabilities of the models themselves, a shift exemplified by the emergence of generative, or decoder-based, LLMs.

Disambiguation experiments on generative LLMs have not been extensively conducted due to the relative novelty of generative LLMs as a phenomenon. Such studies are relatively scarce even on the English-language material, though they have pro-

gressed further than those conducted on Russian-language material: in English-language research the foundational experiments have been carried out, providing reference points for further studies, while the more nuanced and innovative approaches are already in active exploration [4] [22].

A significant shift in modern LLM-based Word Sense Disambiguation (WSD) research is marked by the increasing use of such models as primary disambiguation tools. Previously, generative models were employed in WSD tasks as supplementary tools to support supervised machine learning methods, such as for generating and labeling synthetic datasets, augmenting data [5], among other tasks.

Using generative LLMs as a main disambiguation tool has several notable advantages, the most prominent of which being the opportunity to perform inference without any training or fine-tuning with the high probability of achieving noteworthy results. Among the other advantages of this approach are the capabilities of such models to use advanced reasoning to draw meaningful conclusions, the relative easiness of interaction, due to the possibility to communicate with models in natural language via prompts.

Among the trailblazing English-language WSD studies it is crucial to name [22] and [20], introducing the basic experimental setup implemented in the present paper, as well as outlining baseline model performance, providing reference values for comparison.

Further studies basing on English models are focusing on implementing the known techniques of improving the LLM performance from the side of the user: prompt optimization, human in the loop approaches, Chain-of-Thought reasoning, etc. Additionally, the studies combine the LLM setting with some of the disambiguation methods known since the pioneering works on WSD: disambiguation through translation, leveraging wide and narrow context, etc. [11]. Some studies, following the tendency of allowing the LLM to self-regulate and perform the entire pipeline independently, with the minimum of human-made constraints, resort to more basic tests: for instance, prompting generative LLMs to first detect ambiguity in a sentence and then proceed with disambiguation [17].

3 Experimental setup

The analysis of contemporary research on LLM-based WSD methods led us to develop the following experimental setup:

Basic zero-shot LLM-based classification: given that our primary goal is to conduct foundational research and establish a baseline for Russian-language WSD, selecting the simplest setup was a logical choice.

Structured output and batch evaluation: most LLM-based experiments have involved human-evaluated, long-form responses. However, for the purposes of our study, we deemed structured output as a crucial condition for the experiment: to facilitate the calculation of metrics, we required the models to produce output in JSON format, i.e. a dictionary where keys correspond to context indices and values represent the sense indices predicted by the model. Precisely this requirement led us to reject

Chain-of-Thought prompt enhancement methods, as instructions for reasoning are known to hinder the model's ability to generate structured output, the latter being prioritized in this study.

Data handling: in the present experiment, involving the simultaneous disambiguation of multiple contexts, we do not aim to observe the relationship of the number of contexts presented in the single prompt and the quality of disambiguation. In other words, we do not test if it is easier for the models to disambiguate 10 contexts at once versus 50. Instead, batch construction followed a straightforward approach: preliminary trials indicated that providing the model with more than 40 contexts in a single prompt often led to hallucinations and increased error rates. Consequently, we structured the dataset such that the contexts for each ambiguous word were divided into multiple balanced groups, ensuring that no batch contained more than 40 contexts.

4 Data

In this section we describe the experiment material and outline the principles underlying its selection. For the experiment we selected nine ambiguous words from the three datasets created for RUSSE'18 shared task [18]. The datasets differed in sense structure and the number of available contexts for each sense: the wiki-wiki dataset contained only homonyms with exactly two senses, the bts-rnc dataset comprised words with the average ambiguity of 3.2, and the active-dict dataset manifested even higher average ambiguity scores, along with the broader range of senses.

Below we provide a table with all the target ambiguous words, the extensive description of the senses they have in the Russian language, as well as the number of contexts representing each sense in the corresponding dataset.

Table 1. Experiment material with metadata

Word	Senses	Contexts per sense
бор	1. chemical element	14
	2. pine forest	42
лук	1. weapon	65
	2. vegetable	45
замок	1. building	100
	2. device for closing the doors	38
клетка	1. cage	38
	2. pattern	7
	3. biological cell	96
	4. staircase	4
	5. chest	4
крыло	1. body part of a bird	50
	2. part of a plane	19
	3. part of a car	4
	4. part of a building	3
	5. flank of an army	5
	6. group within a political party	7

винт	1. screw	39
	2. propeller	67
	3. card game	16
жаба	1. amphibia	79
	2. ugly person	6
	3. greed	9
	4. illness	27
дар	1. present	13
	2. divine blessing	8
	3. innate talent	15
девятка	1. number	11
	2. group of nine people	7
	3. card	5
	4. type of billiard	9
	5. car	6
	6. corner of a football goal	4
	7. place on a target	5

5 Models

In the experiment we used a range of generative LLMs adapted for Russian language: GigaChat-2-Max (closed-source flagship model of Sber released in 2025), T-Lite-it-1.0 and T-Pro-it-1.0 (open-source models, 7B and 32B parameters respectively, released in 2024), YandexGPT-5-Pro (closed-source model of Yandex released in 2025) and RefalMachine/RuadaptQwen2.5-32B-Pro-Beta (Ruadapt further: open-source model developed by the Research Computing Center of Lomonosov Moscow State University in 2024 [21]). The range of models taken into account varied from flagship models to smaller ones, more typically used for improving inference speed or for less challenging tasks. All chosen models, except YandexGPT-5-Pro, that has not been evaluated, are among the leaders on the MERA-benchmark for Russian-language model evaluation: even the smaller ones rank within the top 30 positions on the leaderboard [7].

GigaChat-2-Max was accessed through its official GUI, in other cases we resorted to using the LLM Arena, the benchmark for Russian-language LLMs, that offers a wide selection of models for testing and evaluation. The links to both GUIs are available in the references section.

Additionally, the aforementioned requirements to structured output led us to discarding some of the models from the initial list of test subjects: their inference capabilities did not prove advanced enough to abide by this rule. Any reasoning and comments the models outputted was not taken into account: the evaluation was carried out automatically basing only on the predictions in machine-readable format.

6 Experiment: selecting the correct sense from a predefined list

The experiment consisted in prompting a model to determine in which sense from a predefined list the target word is used in the provided list of contexts. The sense inventories we used for each ambiguous word exactly corresponded to the sense inventories present in the respective datasets.

Below we provide a prompt used for all the models:

```
You are a specialist in word sense disambiguation. You
are given a list of contexts containing an ambiguous
word. Your task is to determine in which meaning the tar-
get word is used in each context.
Target word: {word}
Senses of the target word with INDICES: 1.(...)\n2. (...)\n3.
(...)
Provide the answer in a JSON format with the following
fields:
{
  "context number in the analyzed list": sense index
  (int)
}
Carefully analyze the format and do not deviate from it.
Make sure to analyze all contexts from the list and out-
put a sense index for each one.
The response should be a JSON object with {number}
fields, where the keys are the context numbers in order,
and the values are your answers.
Do not output anything other than the JSON response.
```

6.1 Metrics overview

Models' performance was evaluated by the accuracy and macro-averaged F1 scores, the latter being chosen to better deal with class imbalance, as some of the senses of the ambiguous words, especially from *bts-rnc* and *active-dict* datasets are represented by significantly fewer contexts than the others.

Table 2. F1 and accuracy scores on the first experiment

	Yandex GPT 5 Pro		T-pro		Ruadapt		Giga-2-Max		T-Lite	
	acc	f1	acc	f1	acc	f1	acc	f1	acc	f1
бор	1.00	1.00	0.95	0.93	0.91	0.88	0.86	0.81	0.46	0.44
лук	0.98	0.98	0.90	0.90	0.81	0.80	0.66	0.60	0.56	0.56
замок	0.99	0.98	0.84	0.80	0.82	0.76	0.74	0.59	0.73	0.67

девятка	0.93	0.93	0.74	0.73	0.72	0.73	0.15	0.07	0.72	0.72
клетка	0.91	0.82	0.81	0.76	0.72	0.58	0.74	0.55	0.53	0.18
крыло	0.75	0.61	0.61	0.4	0.73	0.63	0.55	0.36	0.61	0.38
винт	0.65	0.39	0.6	0.27	0.67	0.29	0.55	0.24	0.53	0.15
жаба	0.6	0.54	0.65	0.6	0.64	0.54	0.54	0.43	0.54	0.34
дар	0.53	0.53	0.47	0.46	0.52	0.53	0.58	0.52	0.33	0.32
AVG	0.83	0.79	0.73	0.72	0.70	0.66	0.57	0.46	0.57	0.42

6.2 Models' response to the experiment

The experiment proved to be manageable for the selected models. The top results produced by the leading models align with the outcomes obtained by [22] on GPT-4o in a similar basic test using English data. As shown in the table, the flagship models – YandexGPT-5-Pro and T-Pro – occupy the leadership positions in this task based on average scores, although their performance, along with that of the other models, is not entirely stable across the entire dataset. Other models performed worse and in some cases appeared disoriented, behaving almost at random. Despite the nature of the experiment being a one-class classification, where the model must choose from multiple senses, allowing for some error tolerance, less powerful models experienced issues with confidence, particularly in cases of greater ambiguity.

Overall, the models exhibited consistent performance, with each model's results not varying significantly, except in two instances of unusually low metric values – *бор* on T-Lite and *девятка* on GigaChat-2-Max. These outliers, which deviate from the average results both in terms of the word and the model, can likely be attributed to model hallucinations and are considered anomalous. LLMs, especially those of the mid-tier, are known to occasionally produce hallucinations due to the black-box nature of the models and the inability to fully explain their reasoning processes. As such, these two results were excluded from the metric calculations.

While the models' performance was not entirely stable, the obtained metrics allowed us to divide the dataset into two groups. One group demonstrated consistent accuracy and F1 scores, indicating relative stability in precision and recall. The other group showed notable drops in F1 scores, suggesting that for certain words, the models struggled to correctly identify positive cases – either defaulting to the majority class or failing to generalize across the dataset. In general, the drops in F1 were attributed to low recall, which seems justifiable given the poorly represented senses for certain words. YandexGPT was the only model that occasionally exhibited significant drops in precision, misclassifying too many negative cases.

6.3 Models' performance: linguistic interpretation

In this section a closer look is taken on some linguistic patterns found in the received results. The highest and the most stable results was achieved on *бор*, *лук*, *замок* taken

from the wiki-wiki dataset, and on *девятка* from active-dict. On these four words the models achieve the average accuracy of 0,78 and the average f1 of 0,75 and manifest no severe discrepancy between accuracy and f1 scores. Linguistically, *бор*, *лук*, *замок* are textbook homonyms, i.e. the ambiguous words the senses of which are not connected diachronically not through metaphor or metonymy (refer to Table 2 for the detailed sense description); therefore, the coincidence of their exponents is accidental. Moreover, their sense structure is binary, and the numbers of contexts representing each sense, although being imbalanced, never differ by an order of magnitude. The word *девятка* possesses a more complex sense structure: with seven senses represented by 11, 7, 5, 9, 6, 4 and 5 contexts respectively, it would have been close to impossible to disambiguate in a supervised WSD framework: generative LLMs, on the other hand, do not seem to struggle with data sparsity, which prompts us to suggest that the class balance in this case results more useful than the absolute numbers of contexts representing each one. What is more, by its nature *девятка* is not a homonym, which apparently does not hamper the disambiguation as much as it could have been expected; we will provide a more extended commentary on this phenomenon further.

The next group, represented by polysemous words *крыло*, *клетка*, *винт*, *жаба* showed middle-range performance across all LLMs: drops on recall become more apparent, and the average metrics decrease to 0,64 on accuracy and 0,44 on f1. All words from this group share similar sense structures, with central and marginal senses, the latter being represented by 1-7 contexts, which often entails the number of contexts varying by an order of magnitude. Zero-shot LLM-based experiments prove more suitable for processing such underrepresented senses, as there is no need to split the dataset into training and testing subsets. Nevertheless, given that the presence of rare senses in such setups do not always result in total failure in recognizing the class, it is obvious that some LLMs do not manage to extract sufficient contextual information from too few contexts to successfully disambiguate the word. In this case the models tend to behave conservatively or shift towards more frequent senses, which produces notable drops in recall.

Lastly, the polysemous word *дап* proved to be the most challenging for all the LLMs, achieving average accuracy of 0,49 and f1 of 0,47, the maximum values on both metrics grouping around 0,5. This word stands out among all the others, being both contextually underrepresented and possessing three quite semantically similar senses. It is safe to say that in some cases the sense distinction was not entirely clear even to the human annotator.

7 Discussion

Following the basic principles of distributional semantics and the related works on WSD, the results obtained in this study can be explained through two primary factors: quantitative and qualitative. The quantitative factor refers to the number of contexts representing each sense in relation to other senses – even if LLM-based setting does not imply traditional model training, the model makes its decisions basing on the

patterns recognized in the provided data: therefore, the sufficiency of this data remains a critical consideration. The qualitative factor concerns the semantic properties of senses and contexts: during the inference process, the model transforms the texts in natural language into vector embeddings, which must be counted and weighted in a manner that conserves as much semantic and syntactic information about a target word as possible. These two factors happen to be the only parameters of computational semantics that can be tracked without resorting to more complex feature representations involving more extensive tagging. Their interdependence, reflecting the particularities of the disambiguation process at least to some extent, appears to be a very complex phenomenon, the precise and coherent description of which is hampered by the natural language material itself.

It is essential to acknowledge several key limitations in this analysis. First, the differences in word and sense frequency, as well as the availability of data make it impossible to conduct any experiments on exactly the same quantities of data. Second, even if two senses are represented by the same number of contexts, evaluating the quality of these contexts or determining how informative they are for an LLM is inherently challenging, as the semantic fields to which words and senses pertain vary in size and frequency, often correspond to different registers of the language, etc. As a result, the most we can do at this stage is to outline general patterns and tendencies observed in the data.

7.1 Qualitative factor: semantics and ambiguity types

The hypothesis that homonyms tend to be easier for automatic disambiguation than polysemous words has been occasionally acknowledged by NLP studies [18], not to mention that it aligns well with the classic semantic theory that deals primarily with prototypical examples of language categories. In fact, the exponents of the senses of a homonym coincide either coincidentally or through borrowing, sometimes followed by the modification of the uncommonly sounding foreign word by associating it with a phonetically similar word from the target language. As a result, the homonym senses are totally distinct and frequently pertain to different domains, rarely appearing in the same contexts. On the contrary, the senses of the polysemous word, being metaphorically or metonymically connected by definition, are likely to appear in similar collocations, as well as form part of closer semantic fields. Semantic distinction between contexts plays a major role in Word Sense Disambiguation: as it essentially reflects the difference between "close" and "distant" senses. In the context of NLP, this is what directly influences the distribution of the context embeddings on a vector space, provided the model itself is able to construct embeddings that adequately reflect the semantic and syntactic features of the words.

The results of the models' inference support the hypothesis to some extent models indeed demonstrate higher and more stable metric values when disambiguating textbook homonyms compared to other types of ambiguous words. The example of *девятка*, nevertheless, implies that the proposed homonym/polysemous word distinction is not entirely suitable for explaining the phenomena at hand. *Девятка* constitutes a prototypical example of a polysemous word: the concept of nine as a number

or value serves as a point of derivation of all its senses – the number nine, the group of nine people, the playing card with the respective value, the variation of billiard played with nine balls, the Soviet car colloquially named by its model (VAZ-2109), the upper corner of a football goal (due to a specific zoning principle), and a target in darts that corresponds to nine points. Although the senses are diachronically related, they appear distinct in actual usage, with little to no overlap in their semantic fields. This shows that the diachronic connection between the senses does not necessarily imply the similarity of the contexts where each sense is used. Sense origins and evolution, provided they have no manifestation at the context level, are imperceptible to the embedding model. Additionally, numerals typically do not form their own semantic fields, as their meanings are ubiquitous and not confined to a specific domain.

The example of *девятка* suggests that considering polysemous words as a uniform class is an apparent mistake: the presence of metonymical connection between the senses of a word does not correlate with the semantic proximity of its senses whatsoever. This conclusion could be enforced by an example from the other extremity of the spectrum: the polysemous word *дар*, manifesting the senses of a present, a divine blessing and an innate talent. These contexts manifest a notable contextual similarity, especially on the level of collocations, which results in the fact that in some cases only the author's intent determines the precise meaning of the word.

To demonstrate the arbitrariness of the homonym/polysemous word distinction in NLP terms, we measured the pairwise cosine similarity between sense embeddings of each ambiguous word in order to observe if semantic proximity of the contexts correlate with the target word being homonymic or polysemic. The received cosine similarity values suggest that there is no statistically significant correlation between observed phenomena.

The sense embeddings were produced by the paraphrase-multilingual-MiniLM-L12-v2 model, available at HuggingFace. We provide the table with the results below.

Table 3. Context similarity in relation to homonymic/polysemous status

Word	Category	Average cosine similarity	Range
бор	Н	0.3225	-
лук	Н	0.57	-
замок	Н	0.78	-
винт	Р	0.51	0.3917
жаба	Р	0.78	0.1699
клетка	Р	0.42	0.5473
крыло	Р	0.52	0.5536
дар	Р	0.8	0.0324
девятка	Р	0.63	0.46

The obtained data suggests that in the context of polysemous words with more than two senses, the average context similarity does not impact the successful disambiguation.

tion as much as the range between each pairwise similarity values, i.e. the spread of sense embeddings in the vector space. The range parameter helps us define the semantic structure of the word in more detail, specifically, whether the senses are evenly spaced or, conversely, whether some form clusters while others are more isolated. This distinction could provide insight into how easily the senses can be disambiguated: for instance, a smaller similarity range in our data correlates with lower disambiguation quality.

Finally, the cosine similarity values produced on the three binary homonyms – *бор*, *лук*, *замок*, highlight the sheer importance of the correct interpretation of cosine similarity as a metric. The data shows that, despite the relatively high metric scores in disambiguation, two of the homonyms exhibit unexpectedly high cosine similarity between sense embeddings. Apparently, cosine similarity alone does not always capture the contextual markers crucial for disambiguation. Sense vectors may appear close in the vector space while still forming distinct clusters that the model can easily differentiate. These conclusions are consistent with findings from domain-labeling experiments as a WSD approach [19]. This study demonstrated that mapping a sense to a domain—generally aligning with the One Sense Per Discourse (OSPD) approach [8], which relies on measuring embedding similarity within a broad context window—does not necessarily guarantee high-quality disambiguation. While we adopt a more optimistic perspective and recognize OSPD’s continued relevance in WSD, these discoveries underscore the necessity of considering multiple parameters rather than relying on a single factor.

7.2 Quantitative factor: the number of senses and contexts

The discussion of the impact of quantitative factors in the present experiment should begin with the acknowledgment that our data is highly sparse. Regarding the number of contexts, previous research [12] [3] suggests that a minimum of 80 to 200 contexts per sense of an ambiguous word is necessary for effective supervised word sense disambiguation (WSD). As mentioned earlier, zero-shot large language model (LLM)-based experimental setups significantly mitigate this issue, as the absence of sufficient contexts no longer results in a technical failure of the experiment. Nevertheless, the negative impact of data sparsity on any natural language processing (NLP) setup, including WSD, cannot be overlooked.

While the absolute number of contexts is generally considered a positive factor in WSD, the results of the first experiment suggest that this may not hold in LLM-based settings. First, LLMs can infer meaning from much sparser data than traditional machine learning algorithms. Second, rather than the sheer number of contexts, a more critical factor for successful disambiguation appears to be the evenness of sense representation. It is possible that the most frequent sense disproportionately attracts the LLM’s attention, leading it to favor this sense while neglecting less common ones. This phenomenon explains the relatively high performance on words like *девятка*, where senses are represented in a balanced manner, in contrast to words with a more pronounced distinction between dominant and marginal senses.

The multiple-choice sense disambiguation experiment suggests that while the well-established factors influencing WSD remain relevant in LLM-based settings, they may operate differently than in supervised WSD. These differences primarily stem from the nature of generative LLMs in a zero-shot experimental setup. While semantic factors remain largely consistent, the advanced reasoning capabilities of generative models mitigate the traditionally critical issue of requiring sufficient data, making it less of a limiting factor.

8 Conclusion

In the present study we conducted a foundational experiment on generative LLM-based WSD using a set of common ambiguous Russian-language words. The experimental setup – a general multiple-choice sense classification task – allowed us to test a range of Russian-language LLMs and establish a baseline. The leading models performed similarly to OpenAI GPT on the same task in English [22].

We also made several linguistic observations regarding the WSD problem, focusing on the impact of two measurable linguistic factors – quantitative (the number of contexts representing each sense) and qualitative (the semantic proximity of senses) – on the quality of disambiguation. The results suggested that, while the general interdependency of these factors remains foundational in computational semantics, they behave differently in a generative LLM setting. This difference is largely attributed to the black-box nature of the model and the specific characteristics of its reasoning capabilities. It is clear that the tendencies and best practices for interacting with generative LLMs differ from those in supervised machine learning, and will need to be further elaborated in future research.

Acknowledgements

The work is supported by the RSF grant 24-18-00988.

References

1. Agirre E., Edmonds P. (ed.). Word sense disambiguation: Algorithms and applications. – Springer Science & Business Media, 2007. – T. 33.
2. Azarova I. V. et al. Razrabotka komp'yuternogo tezaurusa russkogo yazy'ka tipa WordNet (Development of a Russian-language computer thesaurus of the WordNet type)//Doklady` nauch. konf. «Korpusnaya lingvistika i lingvisticheskie bazy` danny'x» /pod red. A. C. Gerda. SPb.: Izd-vo Sankt-Peterburgskogo universiteta. 2002. S. 6-18.
3. Azarova I. V., Marina A. S. Avtomatizirovannaya klassifikaciya kontekstov pri podgotovke danny'x dlya komp'yuternogo tezaurusa RussNet //Komp'yuternaya lingvistika i intelektual'ny'e tekhnologii: Trudy` mezhdunarodnoj konferencii «Dialog» (Automated Context Classification for Data Preparation in the RussNet Computational Thesaurus). 2006. S. 13-17.
4. Basile P. et al. Exploring the Word Sense Disambiguation Capabilities of Large Language Models //arXiv preprint arXiv:2503.08662. – 2025.

5. Ding B. et al., Data augmentation using LLMs: Data perspectives, learning paradigms, and challenges, Findings of the Association for Computational Linguistics (ACL 2024), pp. 1679-1705
6. Embedding model paraphrase-multilingual-MiniLM-L12-v2 on HuggingFace, <https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2>, last accessed 2025/03/21
7. Fenogenova A. et al. Mera: A comprehensive llm evaluation in russian //arXiv preprint arXiv:2401.04531. – 2024.
8. Gale W. A., Church K., Yarowsky D. One sense per discourse //Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, 1992.
9. GigaChat-2-Max public GUI. <https://giga.chat/>, last accessed 2025/03/27
10. Ide N., Véronis J. Introduction to the special issue on word sense disambiguation: the state of the art //Computational linguistics. – 1998. – T. 24. – №. 1. – С. 1-40.
11. Kang H., Blevins T., Zettlemoyer L. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models //arXiv preprint arXiv:2304.13803. – 2023.
12. Kashkin E. V., Lyashevskaya O. N. Tipy` informacii o leksicheskix konstrukciyax v sisteme Frejmbank (Types of Information About Lexical Constructions in the FrameBank System) //Trudy` instituta russkogo yazy`ka im. VV Vinogradova. 2015. №. 6. S. 464-556.
13. Loukachevitch N. V., Lashevich G., Gerasimova A. A., Ivanov V. V., Dobrov B. V. Creating Russian WordNet by Conversion // In Proceedings of Conference on Computational linguistics and Intellectual technologies Dialog-2016, 2016. pp.405-415
14. LLM Arena benchmark, <https://llmarena.ru/>, last accessed 25/03/27
15. Moro A., Raganato A., Navigli R. Entity linking meets word sense disambiguation: a unified approach //Transactions of the Association for Computational Linguistics. 2014. V. 2. P. 231-244. Азарова И. В. и др. Разработка компьютерного тезауруса русского языка типа WordNet //Доклады науч. конф. «Корпусная лингвистика и лингвистические базы данных» /под ред. А. С. Герда. СПб.: Изд-во Санкт-Петербургского университета. 2002. С. 6-18.
16. Navigli, R. A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. // SOFSEM 2012: Theory and Practice of Computer Science. Lecture Notes in Computer Science. M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser, G. Turán. (Eds) V. 147. Berlin, Heidelberg: Springer, 2012. URL: https://doi.org/10.1007/978-3-642-27660-6_10
17. Ortega-Martín M. et al. Linguistic ambiguity analysis in ChatGPT //arXiv preprint arXiv:2302.06426. – 2023.].
18. Panchenko A. et al. RUSSE'2018: a shared task on word sense induction for the Russian language //arXiv preprint arXiv:1803.05795. – 2018.
19. Sainz O. et al. What do language models know about word senses? zero-shot wsd with language models and domain inventories //arXiv preprint arXiv:2302.03353. – 2023.].
20. Sumanathilaka T., Micallef N., Hough J. Can LLMs assist with Ambiguity? A Quantitative Evaluation of various Large Language Models on Word Sense Disambiguation //arXiv preprint arXiv:2411.18337. – 2024.
21. Tikhomirov M., Chernyshov D. Facilitating Large Language Model Russian Adaptation with Learned Embedding Propagation //Journal of Language and Education. – 2024. – T. 10. – №. 4. – С. 130-145.
22. Yae J. H. et al. Leveraging large language models for word sense disambiguation //Neural Computing and Applications. – 2025. – T. 37. – №. 6. – С. 4093-4110.