Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2025"

April 23–25, 2025

ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian

A. Vatolin
FRC CSC RAS
vatolinalex@gmail.com

N. Gerasimenko
MSU Institute for
Artificial Intelligence
nikgerasimenko@gmail.com

N. Loukachevitch
Lomonosov MSU
louk_nat@mail.ru

A. Ianina
Moscow Institute of
Physics and Technology
yanina.anastasia.mipt@gmail.com

K. Vorontsov
MSU Institute for
Artificial Intelligence
k.vorontsov@iai.msu.ru

Abstract

Against the backdrop of active LLM development, their tendency to hallucinate, as well as the growing volume of texts they generate, the validation of facts has become increasingly important and relevant. We propose ruSciFact[1], a new benchmark for fact-checking scientific claims in Russian. ruSciFact is structured as a NLI task, the goal is to verify whether a fact is confirmed by a given abstract. To generate facts, we used an 3-step pipeline based on LLaMA-405B, validating the resulting sentences with the help of assessors-terminologists. The ruSciFact dataset consists of 1128 pairs in the format <abstract, claim>, which we are releasing as open source together with the benchmark code. Additionally, we are open-sourcing the fact-generation pipeline[2], which facilitates the expansion of the dataset to specific scientific domains. We evaluated several popular language models on ruSciFact, including text embedders and generative models. The results show that this benchmark allows to effectively assess the fact-checking capabilities of LLMs in Russian.

Keywords: benchmark, large language model, embedder
DOI: 10.28995/2075-7182-2025-23-XX-XX

Аннотация

На фоне активного развития больших языковых моделей с их склонностью к галлюцинациям, а также роста количества сгенерированных с их помощью текстов, задача валидации фактов приобретает особую важность и актуальность. Мы представляем ruSciFact, новый бенчмарк для проверки научных утверждений на русском языке. ruSciFact имеет формат NLI: тестируется способность моделей проверить подтверждается ли утверждение аннотацией. Для генерации утверждений мы использовали 3-ступенчатый пайплайн на базе модели LLaMa-405B, затем валидируя полученные утверждения с помощью асессоров-терминологов. Всего ruSciFact содержит 1128 пар <аннотация, утверждение>, которые мы публикуем в открытом доступе вместе с кодом для оценки моделей на

[1] https://huggingface.co/collections/mlsa-iai-msu-lab/ruscifact-6803f71f2d3025988ebc489e
[2] https://github.com/mlsa-iai-msu-lab/ruscifact

бенчмарке. Также мы публикуем пайплайн генерации утверждений, который мы использовали, чтобы облегчить расширение датасета на дополнительные научные области. Мы протестировали на ruSciFact ряд современных языковых моделей, как текстовых эмбеддеров, так и генеративных LLM, и результаты показывают, что он позволяет эффективно оценить их способность к проверке научных фактов на русском языке.

Ключевые слова: бенчмарк, большие языковые модели, текстовый эмбеддер

1 Introduction

Assistance in evaluating the veracity of scientific claims becomes more and more important task in the modern research community due to widespread usage of Large Language Models in many spheres, including scientific investigation and paper writing processes. Such a task is called fact-checking – the verification of a statement against a corpus of documents that either support or refute the claim.

In this work we follow SciFact (Wadden et al., 2020), an expert-annotated dataset of $1,409$ scientific claims and related abstracts that either support or refute each claim. All the annotations in SciFact are provided with rationales (Lei et al., 2016), so each SUPPORTS / REFUTES decision has a reasoned justification. However, all the sentences presented there are written in English, making it impossible to use SciFact as a benchmark for fact-checking in other languages, thus limiting its usability. We propose a novel dataset and benchmark called ruSciFact, designed to evaluate the problem of scientific claim verification for texts written in Russian.

Another key distinction from SciFact is that, instead of relying on human annotators to reformulate naturally occurring claims in the papers, we used Large Language Models, making the data collection process entirely independent of human involvement. Specifically, we used a frontier-level open source language model LlaMa-405B (Touvron et al., 2023), which works across multiple languages including Russian. Using LLMs instead of human annotators enables the generation of larger datasets in less time.

We aimed to create a benchmark that is genuinely representative of the Russian scientific landscape. A direct translation of SciFact would inherently introduce biases toward international literature, thus potentially limiting the applicability and relevance of the benchmark to the Russian research community. Furthermore, translating specialized scientific claims and annotations involves complex terminology and domain-specific nuances. Automatic translation methods, such as using large language models (LLMs), do not yet reliably capture these nuances with perfect accuracy. Human translation by domain experts would mitigate this, but it requires substantial time and resources, making it impractical at scale.

Nevertheless, to support a more comprehensive evaluation of models, we still include a translated version of SciFact along with automated judgments by LLMs (prompts and evaluation results of wide range of LLMs can be found in Appendix A). This dual approach enables researchers to assess their systems across different dimensions: both in terms of language and scientific landscape, and also in terms of task formulation.

Our main contributions are as follows:

- a dataset of scientific claims with relevant abstracts;
- open-sourced code for the scientific claim generation pipeline, based on Llama-405B;
- open-sourced code for model evaluation using our proposed benchmark;
- the SciFact dataset translated to Russian.

2 Related Works

2.1 Scientific benchmarks

There are various benchmarks for assessing the quality of language models. Some are multi-purpose, covering a range of nearly unrelated tasks, while others focus on specific tasks, evaluating a model's ability to solve precise problems or work within a narrow domain. The first group of benchmarks include well-known GLUE (Wang, 2018) and SUPERGLUE (Wang et al., 2019) benchmarks. For evaluating embedders, there is the MTEB benchmark (Muennighoff et al., 2022), which spans 8 embedding tasks covering a total of 58 datasets and 112 languages.

A good example of benchmarks in the second group are those focused on scientific knowledge extraction and processing, such as SCIDOCS (Cohan et al., 2020). This benchmark consists of seven document-level tasks, ranging from citation prediction to document classification and recommendation of scientific articles. In our work, we follow its close relative, SciFact (Wadden et al., 2020), a dataset and benchmark for scientific claims and related abstracts that either support or refute the claims. SciFact consists of 1409 scientific claims verified against a corpus of 5183 abstracts. Each dataset entry includes a rationale providing an explanation for the final decision.

SciFact, in turn, is based on the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2019), a large corpus containing 81.1M English-language academic papers from `semanticscholar.org` spanning various academic disciplines. The original Semantic Scholar corpus (S2) consists of titles from scientific papers published in machine learning conferences and journals from 1985 to 2017, organized by year.

The SciFact public leaderboard is available online [3]. Based on SciFact, an AI fact-checking challenge called SciVER was developed. This shared task supports systems that should:

1. Take a scientific claim as input
2. Identify all relevant abstracts in a large corpus
3. Label them as Supporting or Refuting the claim
4. Select sentences as evidence for the label

SCIVER evaluation is performed at the abstract level (following (Thorne et al., 2018)) and sentence level (evaluating the correctness of the individual predicted evidence sentences).

All of the benchmarks discussed above lack non-English data, and we aim to bridge this gap for the Russian language. Despite being multilingual, even the MTEB benchmark (Muennighoff et al., 2022) lacks Russian-language tasks. A new benchmark for embedders designed specifically for the Russian language, called ruMTEB (Snegirev et al., 2024), has been introduced. Another example of a Russian benchmark is encodechka (Dale, 2022). However, neither of these contains a significant number of tasks focused specifically on scientific knowledge.

2.2 Automatic Generation of Claims for Fact-checking

Several approaches for the automatic generation of fact claims have been studied.

Pan et al. (Pan et al., 2021) proposed a method for automatically generating claims that can be supported, refuted, or unverifiable from evidence from Wikipedia. With this aim, the authors first employ a Question Generator to generate a question–answer pair for the evidence, then they

[3] `https://leaderboard.allenai.org/scifact/submissions/public`

convert the question-answer pair into a claim. The Question Generator and the QA-to-Claim model are off-the-shelf BART models trained on question-answering datasets.

The authors of (Wright et al., 2022) proposed both supervised and unsupervised methods for claim generation. For the supervised method, they trained the BART language model on a small set of <sentence, claim> pairs to directly generate claims (CLAIMGEN-BART). The unsupervised method, CLAIMGEN-ENTITY, involves identifying named entities in the target text fragment. For each named entity, the system generates a question about that entity, which can be answered from the sentence. The system then generates a claim from the question, as described in (Pan et al., 2021). Additionally, (Wright et al., 2022) proposed KBIN, a method for generating claims that contradict the text fragment. KBIN is designed for the biomedical domain. Entities from a given fragment are linked to concepts in the UMLS knowledge base. The initial concept is then replaced with a sibling concept (a sister concept) from UMLS. The closest concepts to the initial concept are selected based on pre-trained UMLS concept vectors. For each related concept, candidate claim variants are generated, with the best variants chosen according to the lowest perplexity and the highest probability of being a contradiction to the original claim.

Kozlova et al. (Kozlova et al., 2023) generate claims for a general fact-checking dataset in Russian. The authors used a trained paraphraser model to generate sentences for a text fragment. In addition, rule-based transformations were proposed as an alternative approach to synthetic data generation. For example, a randomly selected named entity in a statement was replaced with a different, randomly selected named entity.

Bussotti et al. (Bussotti et al., 2024) generate claims using both textual and tabular content. The authors experimented with the BART and LLaMa-2 models. To generate refuting claims, they used antonyms found by language models and resources such as WordNet and ConceptNet. The target entity in the initial claim is substituted with a related entity (e.g., antonym, sibling). Quantitative and human evaluations showed that the generated refuting examples demonstrate high variety, comparable to those created by humans.

3 Task Definition

Scientific fact-checking, or claim verification, focuses on finding evidence in the literature that SUPPORTS or REFUTES a given sentence (called a claim). According to (Wadden et al., 2020), "a scientific claim is an atomic verifiable statement expressing a finding". Validating such claims is a complex task that requires deep knowledge of specific scientific domains. This problem can be challenging even for human annotators, requiring specialists with a deep understanding of the topic. Therefore, the use of artificially accumulated knowledge systems, such as knowledge graphs or databases created using AI algorithms, can help address this issue. One of the most popular and effective examples of such approaches is Large Language Models (LLMs), whose prevalence in various applications is increasing daily. Using an LLM, one can automatically generate scientific claims and retrieve supporting abstracts without the involvement of human annotators.

4 ruSciFact dataset

4.1 Annotations dataset

We choose a publicly available dataset of Russian abstracts called ruSciBench as a starting point. It consists of 182000 pairs of titles and abstracts of scientific articles from Russian electronic library `elibrary.ru`. ruSciBench is the first benchmark for evaluating models on scientific texts in Russian. While all the papers in this dataset are written in Russian, they also include abstracts written in English.

4.2 Claims generation process

For the claim generation process, we used the largest model from the Llama family (Touvron et al., 2023). It has 405B parameters and is reported to perform on par with models like Gemini-1.5 (Team et al., 2024), GPT-4o (Achiam et al., 2023) and Claude-3.5 Sonnet for tasks such as MMLU Chat, instruction following, reasoning, code generation, long-context understanding (NIH/multi-needle task), and multilingual MGSM in a zero-shot scenario. However, Llama is an open-sourced model, unlike its proprietary counterparts from Google, OpenAI and Anthropic. We used the 8bit quantized variant of Llama-405B and ran it on 8 A100 with 80Gb Graphics RAM.

The generation pipeline can be described as follows (see Figure 1).

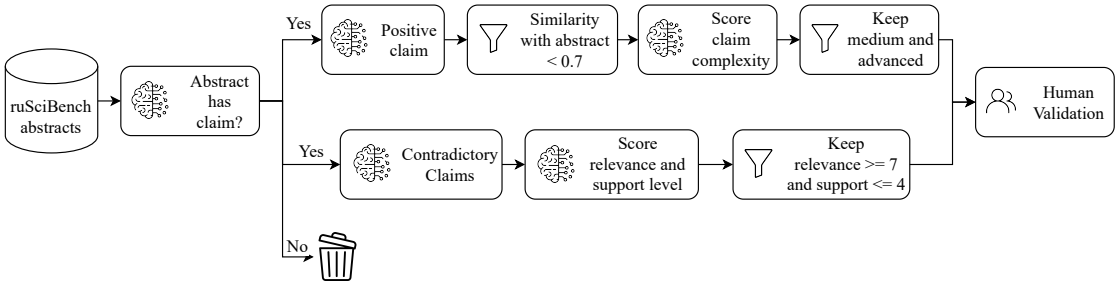

Figure 1: Overview of the data generation and validation pipeline.

4.2.1 Generation of positive examples (claims with evidence in the abstracts)

First, we filtered the abstracts in ruSciBench based on the following criteria. The eLibrary contains many abstracts that do not state meaningful claims. For instance, they may describe the conducted experiments or tell stories about the scientist, etc. A good example is the paper "COST-SENSITIVE LEARNING OF CLASSIFICATION KNOWLEDGE AND ITS APPLICATIONS IN ROBOTICS" by M.Tan[4].

Using a specific prompt (see Appendix B), we asked Llama-405B to generate a scientific claim based solely on the information in the abstract, or to specify that no claims were stated in the abstract. In the designed prompt, we listed the characteristics a potential claim should have and provided several examples of abstracts with corresponding claims. These characteristics include the following points:

[4] `https://www.elibrary.ru/item.asp?id=4996344`

- A fact must be related to the abstract and can be clearly supported by information from the abstract alone.
- A fact should not be too simple and must not contradict general knowledge or common sense.
- A fact should not contain uncertainty (e.g., a fact cannot include phrases like "in another study").

According to Llama-405B, only 42% of abstracts from eLibrary contain at least one scientific claim.

Second, after leaving out abstracts with no claims according to Llama, we continued the filtration process by eliminating facts that repeated information already present in the abstract. We performed this step using library **fuzzy** and its **partial_ratio** method to calculate how many words from the abstract and the fact overlapped. If the ratio was higher than 0.7, the claim was deleted.

Third, we filtered out simple facts. After manually annotating a small subset, we discovered that most of the generated facts were quite basic and could be derived from general knowledge (for example, "Russia is the biggest country territory-wise"). We created a prompt (see Appendix C) that asked the model to classify each fact into one of three classes: simple, medium, or advanced. Similar to the previous step, the prompt included a detailed description with examples. The distribution of fact complexity is the follows: most facts are of medium difficulty (61.4%), while simple and advanced facts are represented in roughly equal proportions (18.4% and 20.1% respectively).

4.2.2 Generation of negative examples (claims that contradict abstracts)

First, we selected abstracts that contained at least one fact (from the first step in the previous section). Then, we created a prompt (see Appendix E), that asks the model to generate a fact with the following features:

- A fact must contradict the abstract.
- A fact supporting the abstract cannot be obtained by removing the particle "not".
- A fact must be relevant to the abstract.
- A fact should not contain uncertainty (e.g., it cannot include phrases like "in another study").

Next, a fact should be derived from a chain of thought. It is a complex task for a language model. The model first changed truthful facts by adding particle "not" and then rephrased them so that the final claim would contain no negating expressions. The final fact was retrieved from such chains with another prompt.

The final step is to evaluate the fact's relevance and its contradiction to the abstract. We noticed that many facts neither relevant to the article nor contradict it. We designed an additional prompt (see Appendix F) that asks the model to grade each fact's relevance on a scale from 0 to 10 and indicate how much the abstract supports the fact. For ease of result interpretation, the model provides the response in JSON format. Only examples with a relevance score of 7 or higher and an abstract support level of 4 or lower were sent to the next step (labeling).

4.3 Human validation process

The generated facts were annotated by two terminologists with experience in analyzing scientific texts across various domains. The annotators obtained pairs: (generated claim, initial abstract) and

should provide the pairs with the following labels: "confirms", "contradicts", or "problematic". For the last two labels, the annotators wrote explanations. Problematic facts could contain non-existent words or be unclear. Each pair received two annotations.

In some cases, discrepancies arose between the annotations. For instance, one annotator might have deemed a fact too general for the abstract, while another annotator might have considered the same general fact to be confirmed. Currently, only facts with consistent annotations are included in the dataset. The final dataset contains 1128 pairs. Statistics on the length of the texts are presented in Table 1, the class distribution is shown in Table 2, and the distribution across scientific categories is detailed in Table 3.

Text Type	Mean Words	25% Words	50% Words	75% Words
Claim	15	12	14	18
Abstract	165	86	144	234

Table 1: Word count statistics for claims and abstracts in the ruSciFact dataset.

Label	Count
confirms	758
contradicts	369

Table 2: Class distribution in the final ruSciFact dataset after human validation.

Examples of claims and abstracts from the final dataset can be found in Appendix G.

5 Experiments

5.1 Supervised classification task

The task for evaluating embedders was formulated as a supervised classification problem. The input to the model was the concatenation of the claim and abstract embeddings. The model's objective was to predict a binary class: whether the abstract confirms or contradicts the claim. A linear classification layer was added on top of the embedder model to perform this classification. All layers of the embedder model were trained (unfrozen) for 10 epochs. The dataset was split equally into training and testing sets. This split presented a challenge due to the relatively small size of the resulting training set.

We tested several well-known embedders using the ruSciFact benchmark and summarised the results in Table 4. The list of models includes:

- Different RuBERT models (Kuratov and Arkhipov, 2019): rubert-tiny, rubert-tiny-turbo and rubert-tiny2
- SciRus-tiny (Gerasimenko et al., 2024; Vatolin et al., 2024)
- Multilingual E5 (base and large models)
- BGE-M3 embeddings (Chen et al., 2024)
- Universal Sentence Encoder for Russian (USER)[5]
- Language-agnostic BERT sentence embeddings (LaBSE) (Feng et al., 2020)

[5](https://huggingface.co/deepvk/USER-base)

Scientific Category (GRNTI)	**Percentage (%)**
Medicine and healthcare	30.85
Physics	7.09
Biology	6.83
Chemistry	6.12
Rural and forestry	5.85
Mechanical engineering	3.99
Mechanics	3.46
Geology	3.28
Polygraphy. Reprography. Photokinotechnology	3.10
Mathematics	2.75
Popular education. Pedagogy	2.04
Construction. Architecture	1.95
Mining	1.68
Automation. Computing technique	1.68
Linguistics	1.51
Psychology	1.51
State and law. Legal sciences	1.33
Electronics. Radio engineering	1.24
Electrical engineering	1.15
Informatics	1.15
Metallurgy	1.06

Table 3: Distribution of scientific categories in the ruSciFact dataset (categories with >1% representation).

- MPNet (Song et al., 2020)
- XLM-RoBERTa (Conneau, 2019)
- General Text Embedding family of models (gte-Qwen2-7B-instruct) (Bai et al., 2023)

The best result on ruSciFact among the models presented is achieved by BGE-m3 (Chen et al., 2024). Despite having only 359M parameters, compared to the larger E5 models (560M) and the significantly larger Qwen (7.61B), BGE still delivers the best performance in this set of experiments. Furthermore, the size of the embedder does not always correlate with the score for the other models as well: RuBERT-tiny2 is larger than SciRus-tiny3.1, yet their F1-scores are 0.66 and 0.68, respectively. On the other hand, mE5 is nearly 5 times bigger than RuBERT-tiny2 (118M vs. 29M) and its embeddings size is bigger(384 vs. 312), but it has the same score. The main reason for this discrepancy is context length: for mE5 it's 4 times (512 vs. 2048) compared to RuBERT-tiny2.

Model name	Model Config			F1-Score
	Model size	Embedding size	Max context length	
rubert-tiny	12M	312	512	0,66
sci-rus-tiny	23M	312	1K	0,67
sci-rus-tiny3.1	23M	312	1K	0,68
rubert-tiny-turbo	29M	312	2K	0,75
rubert-tiny2	29M	312	2K	0,66
multilingual-e5-small	118M	384	512	0,66
USER-base	124M	768	512	0,69
LaBSE-en-ru	129M	768	512	0,84
multilingual-e5-base	278M	768	512	0,78
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0,74
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0,78
USER-bge-m3	359M	1024	8K	**0,85**
multilingual-e5-large	560M	1024	512	0,8
multilingual-e5-large-instruct	560M	1024	512	0,77
gte-Qwen2-7B-instruct	7.61B	3584	131K	0,82

Table 4: Results of supervised evaluation on ruSciFact benchmark

5.2 Generative LLMs

We also tested 10 proprietary and 7 open-source generative language models on our benchmark, ruSciFact (see Appendix D for the prompt used in the evaluation). For the closed-source LLMs, we focused on well-known, powerful models. Since technical details for these models are not disclosed, we provide only brief notes about each of them.

- Claude 3.5 Sonnet (claude-3-5-sonnet-20241022) and Claude 3.5 Haiku (claude-3-5-haiku-20241022)
- GPT-4o (gpt-4o-2024-11-20) and GPT-4o mini (gpt-4o-mini-2024-07-18)
- Gemini 1.5 Pro (gemini-1.5-pro-002)
- YandexGPT 4 Pro and YandexGPT 4 Lite
- GigaChat: We tested following model versions: GigaChat, GigaChat Pro and GigaChat Max. As these models lack explicit version identifiers, we specify the access date (05.01.2025).

As for open-source LLMs, we conducted experiments with the following models:

- T-pro-it-1.0 (T-bank, 2024b) and T-lite-it-1.0 (T-bank, 2024a)
- Qwen2.5-32B-Instruct, Qwen2.5-7B-Instruct (Bai et al., 2023), RuadaptQwen2.5-32B-Instruct (Tikhomirov and Chernyshev, 2023; Tikhomirov and Chernyshev, 2024).
- sainemo-remix-12b-gptq-8bit (Gusev, 2024)
- Vikhr-Nemo-12B-Instruct (VikhrModels, 2024).

Most of the generative LLMs, both closed and open-source, achieve F1-scores above 90% (see Table 5). The best results are obtained by Claude 3.5 Sonnet and GPT-4o among models developed internationally, and by T-pro among Russian-developed LLMs. Additionally, even relatively small models, such as T-lite and Qwen2.5-7B, demonstrate strong performance, scoring

Model name	Open weights	Model size	F1-Score
claude-3-5-sonnet-20241022	No	-	0,99
gpt-4o-2024-11-20	No	-	0,98
T-pro-it-1.0	Yes	32B	0,98
gemini-1.5-pro-002	No	-	0,97
Qwen2.5-32B-Instruct	Yes	32B	0,97
RuadaptQwen2.5-32B-instruct	Yes	32B	0,96
YandexGPT 4 Pro	No	-	0,95
gpt-4o-mini-2024-07-18	No	-	0,94
T-lite-it-1.0	Yes	7B	0,94
Qwen2.5-7B-Instruct	Yes	7B	0,93
claude-3-5-haiku-20241022	No	-	0,93
GigaChat Max	No	-	0,87
sainemo_remix_12b_gptq_8bit	Yes	12B	0,80
YandexGPT 4 Lite	No	-	0,79
GigaChat	No	-	0,77
Vikhr-Nemo-12B-Instruct-R-21-09-24	Yes	12B	0,75
GigaChat Pro	No	-	0,75

Table 5: Results of zero shot LLM evaluation on ruSciFact benchmark

over 93% on our ruSciFact benchmark. Overall, the fact-checking task does not appear to be particularly challenging for powerful models with large context windows and well-trained reasoning abilities across multiple languages.

5.3 Retrieval evaluation

The retrieval task was formulated following the methodology described in MTEB (Muennighoff et al., 2022). Dataset entry consists of a claim (query) and the set of all abstracts in the dataset (corpus). The goal is to retrieve the relevant abstract for each given claim. To achieve this, we used the evaluated model to embed all claims and all abstracts. Cosine similarity scores were computed between each claim embedding and all abstract embeddings. Based on these scores, abstracts were ranked for each claim. The primary metric used for evaluation in this setup is Mean Reciprocal Rank at k (MRR@k), specifically MRR@1, which measures whether the correct abstract is ranked first for a given claim.

Several retrieval models were tested using our proposed benchmark ruSciBench. The list of models includes:

- GritLM-7B (Muennighoff et al., 2024)
- SFR-Embeddings by Salesforce Research (SFR- Embedding-2-R) (Meng* et al., 2024)
- Multilingual E5 (small, base, large and large-instruct) (Wang et al., 2024)
- Qwen2-7B-instruct (Bai et al., 2023)
- Stella 1.5B (Alibaba, 2024)

Model name	Model Config			MRR@1
	Model size	Embedding size	Max context length	
GritLM-7B	7.24B	4096	32K	0,91
SFR-Embedding-2_R	7.11B	4096	32K	0,90
multilingual-e5-large-instruct	560M	1024	512	0,88
multilingual-e5-base	278M	768	512	0,80
multilingual-e5-large	560M	1024	512	0,78
gte-Qwen2-7B-instruct	7.61B	3584	131K	0,81
stella_en_1.5B_v5	1.54B	1024	131K	0,80
multilingual-e5-small	118M	384	512	0,70
USER-bge-m3	359M	1024	8K	0,76
sci-rus-tiny	23M	312	1K	0,57
sci-rus-tiny3.1	23M	312	1K	0,59
USER-base	124M	768	512	0,62
rubert-tiny-turbo	29M	312	2K	0,55
LaBSE-en-ru	129M	768	512	0,42
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0,39
stella_en_400M_v5	435M	1024	8K	0,26
rubert-tiny2	29M	312	2K	0,15
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0,12
rubert-tiny	12M	312	512	0,06

Table 6: Results of retrieval evaluation on ruSciFact benchmark

- BGE-M3 (Chen et al., 2024)
- SciRus models (Gerasimenko et al., 2024)
- RuBERT models (Kuratov and Arkhipov, 2019)
- Language-agnostic BERT sentence embeddings (LaBSE) (Feng et al., 2020)
- Multilingual MPNet (Song et al., 2020)
- XLM-RoBERTa (Conneau, 2019)
- Universal Sentence Encoder for Russian (USER)

The results are presented in Table 6. As shown, GritLM provides the best result (it is also the largest model tested in this experiment). Multilingual E5 models also achieve high results being significantly smaller than GritLM or SFR Embeddings.

6 Conclusion and Future Work

In this work, we introduced a new benchmark, ruSciFact, for evaluating fact-checking capabilities within the scientific domain in the Russian language. We open-sourced a dataset of scientific facts with rationales, generated using a multistep pipeline based on LLaMa-405B and validated by human assessors. In addition to the benchmark, we open-sourced the fact-generation pipeline, aiming to support research in LLM skill validation, benchmarking, scientific data processing, and automatic scientific knowledge extraction. ruSciFact is intended to encourage the development of language models designed for research-oriented tasks while providing a platform for comparing

Russian-speaking models.

Acknowledgements

The research is part of project #23-Ш05-21 SES MSU "Development of mathematical methods of machine learning for processing large-volume textual scientific information". We would like to thank eLibrary for providing the datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alibaba. 2024. stella-en-1.5b-v5.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. // *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, P 12105–12122.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

David Dale. 2022. Рейтинг русскоязычных энкодеров предложений, June. [Online; posted 12-June-2022].

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Nikolai Gerasimenko, Aleksei Vatolin, Anastasia Ianina, and Konstantin Vorotsov. 2024. Scirus: Tiny and powerful multilingual encoder for scientific texts. *Doklady Mathematics, 2024, Vol. 110, No. 4, pp. 191–200*.

Ilya Gusev. 2024. sainemo-remix-12b-gptq-8bit.

Anastasia Kozlova, Denis Shevelev, and Alena Fenogenova. 2023. Fact-checking benchmark for the russian large language models. // *Proceedings of the International Conference "Dialogue*, volume 2023.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Rui Meng*, Ye Liu*, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. *// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, P 476–483.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2024. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *arXiv preprint arXiv:2408.12503*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

T-bank. 2024a. T-lite-it-1.0.

T-bank. 2024b. T-pro-it-1.0.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Mikhail Tikhomirov and Daniil Chernyshev. 2023. Impact of tokenization on llama russian adaptation. *// 2023 Ivannikov Ispras Open Conference (ISPRAS)*, P 163–168. IEEE.

Mikhail Tikhomirov and Daniil Chernyshev. 2024. Facilitating large language model russian adaptation with learned embedding propagation. *arXiv preprint arXiv:2412.21140*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: open and efficient foundation language models. arxiv. *arXiv preprint arXiv:2302.13971*.

Aleksei Vatolin, Nikolai Gerasimenko, Anastasia Ianina, and Konstantin Vorotsov. 2024. Ruscibench: Open benchmark for russian and english scientific document representations. *Doklady Mathematics, 2024, Vol. 110, No. 4, pp. 249–258*.

VikhrModels. 2024. Vikhr-nemo-12b-instruct-r.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. *// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 2448–2460.

A Scifact tranlation process and results

To adapt the original SciFact benchmark for Russian language evaluation, we undertook a multi-stage translation process for its English claims and abstracts. The initial translation was performed using the `gpt-4o-2024-11-20` large language model. Given the complexity of scientific terminology and the need for high fidelity, a rigorous quality control procedure was implemented based on the LLM-as-a-judge paradigm.

For quality assessment, we employed the `claude-3-5-haiku-20241022` model to act as an automated judge. This judge model evaluated each translated claim-abstract pair against the original English text, checking for semantic equivalence, grammatical correctness in Russian, and, crucially and the preservation of the core scientific meaning.

The translation and evaluation were conducted iteratively. After the initial translation pass by `gpt-4o`, all pairs were evaluated by the `claude-3-5-haiku` judge. Pairs flagged as containing translation errors (e.g., incorrect terminology, altered meaning, grammatical issues) were automatically collected. These problematic pairs were then fed back into the `gpt-4o` model for a second translation attempt. The re-translated pairs were subsequently re-evaluated by the judge model. This cycle of translation and evaluation was repeated for a total of three iterations.

Despite the iterative refinement, a small subset of examples remained flagged as incorrectly translated after the third iteration. These persistent cases, often involving particularly nuanced or complex scientific language, were manually reviewed and corrected by human experts to ensure the final dataset's accuracy and reliability. This combined approach of automated translation, iterative LLM-based judging, and final manual verification aimed to produce a high-quality Russian version of the SciFact dataset.

The evaluation of models on the translated SciFact dataset followed the same three setups as described for the ruSciFact benchmark: supervised classification, zero-shot LLM evaluation, and retrieval. The task formulations remain consistent with those outlined in the main body of the paper. The results for each setup are presented in Tables 7, 8, and 9.

Analysis of Results:

Comparing the results on the translated SciFact dataset (Tables 7, 8, 9) with those on the ruSciFact dataset (Tables 4, 5, 6), a noticeable decrease in performance is observed across all evaluation setups. This consistent trend suggests that the original SciFact data may be inherently more complex than ruSciFact, or it represents a different data distribution (e.g., varying scientific domains or writing styles) on which the evaluated models generally perform less effectively.

In the **supervised classification task** (Table 7), the highest F1-score achieved is 0.48 (by `multilingual-e5-large` and `gte-Qwen2-7B-instruct`), significantly lower than the 0.85 achieved on ruSciFact. This gap underscores the potential increased difficulty or distributional shift presented by the SciFact data, possibly amplified by translation nuances.

For the **zero-shot LLM evaluation** (Table 8), the performance drop is also substantial. While top models like `claude-3-5-sonnet-20241022` still perform relatively well (F1-score of 0.86), this is considerably lower than the near-perfect scores achieved by several models on ruSciFact. This difference highlights that even advanced LLMs find the translated SciFact version more challenging, likely due to the intrinsic properties of the data, although translation artifacts might also contribute.

The **retrieval task** (Table 9) shows the most dramatic performance decrease. The top MRR@1

Model name	Model Config			F1-Score
	Model size	Embedding size	Max context length	
rubert-tiny	12M	312	512	0,28
sci-rus-tiny	23M	312	1K	0,31
sci-rus-tiny3.1	23M	312	1K	0,33
rubert-tiny-turbo	29M	312	2K	0,23
rubert-tiny2	29M	312	2K	0,31
multilingual-e5-small	118M	384	512	0,28
USER-base	124M	768	512	0,27
LaBSE-en-ru	129M	768	512	0,40
multilingual-e5-base	278M	768	512	0,33
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0,27
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0,36
USER-bge-m3	359M	1024	8K	0,43
multilingual-e5-large	560M	1024	512	**0,48**
multilingual-e5-large-instruct	560M	1024	512	0,47
gte-Qwen2-7B-instruct	7.61B	3584	131K	**0,48**

Table 7: Results of supervised evaluation on translated SciFact benchmark

Model name	Open weights	Model size	F1-Score
claude-3-5-sonnet-20241022	No	-	0,86
T-pro-it-1.0	Yes	32B	0,81
gpt-4o-2024-11-20	No	-	0,81
gemini-1.5-pro-002	No	-	0,79
Qwen2.5-32B-Instruct	Yes	32B	0,77
GigaChat Max	No	-	0,75
gpt-4o-mini-2024-07-18	No	-	0,75
RuadaptQwen2.5-32B-instruct	Yes	32B	0,74
claude-3-5-haiku-20241022	No	-	0,73
YandexGPT 4 Pro	No	-	0,69
T-lite-it-1.0	Yes	7B	0,67
Qwen2.5-7B-Instruct	Yes	7B	0,64
GigaChat Pro	No	-	0,56
YandexGPT 4 Lite	No	-	0,53
Vikhr-Nemo-12B-Instruct-R-21-09-24	Yes	12B	0,49
sainemo_remix_12b_gptq_8bit	Yes	12B	0,48
GigaChat	No	-	0,47

Table 8: Results of zero shot LLM evaluation on translated SciFact benchmark

Model name	Model Config			MRR@1
	Model size	Embedding size	Max context length	
SFR-Embedding-2_R	7.11B	4096	32K	0.55
multilingual-e5-large-instruct	560M	1024	512	0.52
GritLM-7B	7.24B	4096	32K	0.52
multilingual-e5-large	560M	1024	512	0.52
gte-Qwen2-7B-instruct	7.61B	3584	131K	0.51
stella_en_1.5B_v5	1.54B	1024	131K	0.50
multilingual-e5-base	278M	768	512	0.49
USER-bge-m3	359M	1024	8K	0.49
multilingual-e5-small	118M	384	512	0.47
sci-rus-tiny3.1	23M	312	1K	0.41
sci-rus-tiny	23M	312	1K	0.39
USER-base	124M	768	512	0.37
LaBSE-en-ru	129M	768	512	0.37
rubert-tiny-turbo	29M	312	2K	0.36
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0.32
stella_en_400M_v5	435M	1024	8K	0.23
rubert-tiny2	29M	312	2K	0.21
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0.12
rubert-tiny	12M	312	512	0.09

Table 9: Results of retrieval evaluation on translated SciFact benchmark

score is 0.55 (`SFR-Embedding-2_R`), compared to 0.91 on ruSciFact. This indicates significant difficulty in accurately matching translated claims to translated abstracts based on semantic similarity, potentially stemming from the different nature of the SciFact content and how translation alters specific phrasing crucial for embedding models.

Overall, the translated SciFact benchmark serves as a valuable complementary evaluation tool. The observed lower scores across the board emphasize the challenges models face when dealing with potentially harder or differently distributed data, further complicated by cross-lingual transfer, underscoring the complexities of robust fact verification.

B Prompt for Positive Claim Generation

In the prompts presented in the appendices, the placeholder `{text}` is replaced with the actual text of the abstract during the generation process.

Вы ученый, который хорошо разбирается во всех областях науки. Ваша задача - записать один факт, который следует из аннотации к статье. Вы не можете скопировать текст из аннотации, вам нужно написать факт, который следует из аннотации, но прямо в ней не указан. Примечание: Убедитесь, что извлеченный факт точно выведен из содержания аннотации,

без добавления какой-либо дополнительной информации или интерпретации. При написании факта избегай ссылок на аннотацию (в приведенном текст, в данной работе, предложенный метод). Также при написании факта избегай неопределенности, например ”с определенным свойствами”, ”в некоторых состояниях”, ”определенной длины”. Если по аннотации невозможно написать точный факт, то напиши ”Аннотация не содержит факт”

Ниже приведены 2 примера фактов и аннотаций, на основе которых они были написаны.

Аннотация к статье: Ожидается, что снижение уровня гомоцистеина в сыворотке крови с помощью фолиевой кислоты снизит смертность от ишемической болезни сердца. Известно, что максимальное снижение уровня гомоцистеина достигается при приеме фолиевой кислоты в дозе 1 мг/сут, но эффект более низких доз (имеющих отношение к обогащению пищевых продуктов) неясен. МЕТОДЫ Мы рандомизировали 151 пациента с ишемической болезнью сердца на 1 из 5 доз фолиевой кислоты (0,2, 0,4, 0,6, 0,8 и 1,0 мг/сут) или плацебо. Первоначально, через 3 месяца приема добавок и через 3 месяца после прекращения приема фолиевой кислоты, были взяты образцы крови натощак для анализа на содержание гомоцистеина и фолиевой кислоты в сыворотке крови. РЕЗУЛЬТАТЫ: Средний уровень гомоцистеина в сыворотке крови снижался при увеличении дозы фолиевой кислоты до максимума при приеме 0,8 мг фолиевой кислоты в день, когда снижение уровня гомоцистеина (с поправкой на плацебо) составляло 2,7 мкмоль/л (23%), что аналогично известному эффекту приема фолиевой кислоты в дозах 1 мг/сут и выше. Чем выше был исходный уровень гомоцистеина в сыворотке крови человека, тем сильнее была реакция на фолиевую кислоту, но статистически значимое снижение наблюдалось независимо от исходного уровня. Уровень фолиевой кислоты в сыворотке крови повышался примерно линейно (5,5 нмоль/л на каждые 0,1 мг фолиевой кислоты). Индивидуальные колебания уровня гомоцистеина в сыворотке крови, измеренные в группе плацебо, были значительными по сравнению с эффектом приема фолиевой кислоты, что указывает на то, что мониторинг снижения уровня гомоцистеина у конкретного человека нецелесообразен. ВЫВОДЫ Для достижения максимального снижения уровня гомоцистеина в сыворотке крови во всем диапазоне уровней гомоцистеина в популяции, по-видимому, необходима доза фолиевой кислоты в размере 0,8 мг/сут. Нынешние уровни обогащения пищевых продуктов в США позволят достичь лишь небольшой доли достижимого снижения уровня гомоцистеина.

Факт из статьи: Дефицит витамина B9 снижает уровень гомоцистеина в крови.

Аннотация к статье: Для реакции выделения кислорода (OER) были разработаны известные на Земле катализаторы первого ряда (3d) на основе переходных металлов; однако они работают при потенциалах, значительно превышающих термодинамические требования. Теория функционала плотности предполагает, что не трехмерные металлы с высокой валентностью, такие как вольфрам, могут модулировать трехмерные оксиды металлов, обеспечивая почти оптимальную энергию адсорбции для предлагаемых промежуточных продуктов. Мы разработали метод синтеза при комнатной температуре для получения гелеобразных оксигидроксидных материалов с атомарно однородным распределением металлов. Эти гелеобразные оксигидроксиды FeCoW обладают самым низким перенапряжением (191 милливольт), зарегистрированным при 10 миллиамперах на квадратный сантиметр в щелочном электролите. Катализатор не проявляет признаков разложения после более чем 500 часов работы. Рентгеновское поглощение и компьютерные исследования показывают синергетическое взаимодействие между вольфрамом, железом и кобальтом в создании благоприятной локальной координационной среды и электронной структуры, которые повышают энергетику предложения.

Факт из статьи: Усовершенствованные катализаторы OER демонстрируют стабильную активность в течение нескольких сотен часов.
Если в аннотации не содержится фактов, например: В статье рассмотрено становление британо-японских отношений в период биполярности, отражена история формирования двусторонних отношений, а также сотрудничество в экономической, политической, научно-технической и социально-культурной сферах в указанный период. Главный акцент был сделан на рассмотрении зарождения отношений между Великобританией и Японией и анализа динамики их развития. то напиши ”Аннотация не содержит факт”.
Теперь ваша задача - записать этот факт в следующую аннотацию, точно следуя инструкциям
{text}
Не пиши никаких вводных слов (например ”из аннотации следует, что”), только факт. Напишите свой факт из этой аннотации:

C Prompt for Fact Complexity Classification

Инструкция по классификации фактов по научным статьям

Вы являетесь учёным, обладающим глубокими знаниями во всех областях науки. Ваша задача - определить сложность факта, изложенного в аннотации к научной статье. Факты могут быть классифицированы по трём уровням сложности: простой, средний и сложный. В отдельных случаях факт может

быть неопределённым.

Категории фактов:

1. Простой факт:

Факт очевиден большинству образованных людей и не требует дополнительных исследований или чтений для его подтверждения или опровержения. Такие факты известны из общих знаний.
Примеры простых фактов:

- Экономическая эффективность потребления может варьироваться в зависимости от личностных свойств потребителей.
- В России существует закон, регулирующий проведение медицинских научных исследований с участием человека и/или лабораторных животных.
- Нитраты могут вызывать токсические эффекты у животных.
- Российские вузы сокращают количество бюджетных мест.
- В Республике Алтай представлены многочисленные виды туризма и отдыха.
- У студентов вузов ценность внутреннего успеха выше, чем внешнего.
- Изменение жесткости элементов конструкции здания может быть вызвано разными факторами.
- Преступления, связанные с фальшивомонетничеством, совершались на территории Российской Федерации и Белгородской области.

2. Средний факт:

Факт достаточно сложен, большинству людей понадобится читать специализированные статьи или проводить запросы в интернете, чтобы понять, подтвердить или опровергнуть этот факт.
Примеры средних фактов:

- Поражение молочной железы эхинококком может быть излечено хирургическим путем.
- Татарстан стал более засушливым регионом за последние десятилетия.

3. Сложный факт:

Факт требует специфических знаний или экспертизы в данной научной области для его понимания.
Примеры сложных фактов:

- Прокатка СВС-продуктов в валках прокатного стана перед измельчением в шаровой мельнице увеличивает эффективность измельчения.
- У больных диабетическим макулярным отеком наблюдается повышенный уровень брадикинина в крови.
- Стентирование коронарных артерий не вызывает значимых изменений показателей глобальной и сегментарной продольной систолической деформации миокарда левого желудочка в первые 3 сут после процедуры.

4. Неопределённый факт:

Факт недостаточно ясен, неполон или содержит ссылки, требующие дополнительных разработок или исследований.
Примеры неопределённых фактов:

- Российские компании могут использовать разработанную шкалу для определения уровня развития ориентации на бренд.
- Предыдущие модели теплоусвоения вермикулита были неточными.

Ваша задача:

Внимательно изучите предоставленный факт и напишите одно из следующих определений сложности факта: "простой "средний "сложный"или "неопределенный". Напиши только одно слово, не объясняй причины такой классификации
Вот факт, для которого это нужно написать: {text}

D Prompt for zero shot LLM evaluation

Инструкция по разметке аннотаций и фактов
Цель
Цель этого задания - определить, насколько представленный факт соответствует содержанию аннотации научной статьи. Вы будете работать с парами "факт-аннотация"и назначать им один из трех предложенных классов.
Процедура разметки

1. **Внимательно прочитайте** указанный факт и соответствующую ему аннотацию из научной статьи.
2. **Определите соответствие** между фактом и содержанием аннотации, используя ниже приведенные классы.

Классы для разметки

1. **Аннотация подтверждает факт**: Выберите этот класс, если информация в аннотации подтверждает или прямо указывает на изложенный в ней факт. Выбирайте этот класс только если факт в точности соответствует аннотации, без необходимости "додумывать"что-либо.
2. **Аннотация не подтверждает или противоречит факту**: Аннотация опровергает факт: информация в аннотации прямо противоречит заявленному факту.

Важно

- Расценивайте смысл и контекст фактов и аннотаций, избегая слишком буквального истолкования.
- В ответе должно быть только название класса: ”Аннотация подтверждает факт”, ”Аннотация не подтверждает или противоречит факту”

E Prompt for Negative Claim Generation

Вы ученый, который хорошо разбирается во всех областях науки. Ваша задача - написать один факт, который был бы релевантен аннотации, но не следует из нее. Не используй отрицание, вместо этого напишите факт, который не подтверждается аннотацией, но релевантен ей! Вы не можете скопировать текст из аннотации, вам нужно написать факт, который не следует из аннотации. Примечание: Убедитесь, что извлеченный факт точно выведен из содержания аннотации, без добавления какой-либо дополнительной информации или интерпретации. При написании факта избегай ссылок на аннотацию (в приведенном текст, в данной работе, предложенный метод). Также при написании факта избегай неопределенности, например "с определенным свойствами "в некоторых состояниях "определенной длины". Если по аннотации невозможно написать точный факт, то напиши "Аннотация не содержит факт".
Ниже приведены 2 примера фактов и аннотаций, на основе которых они были написаны.

Пример 1

Аннотация к статье:
Статья посвящена актуальным проблемам установления уголовного запрета в сфере профессиональной медицинской деятельности. Проанализированы предложения Следственного комитета РФ о внесении изменений в действующий Уголовный Кодекс РФ, на основании социологического опроса и изучения уголовных дел, сделаны выводы о необходимости реформы уголовного закона. Основным является вывод о невозможности решения актуальных проблем в российском здравоохранении исключительно уголовно-правовыми средствами.
Факт из статьи: Актуальные проблемы в российском здравоохранении могут быть решены исключительно уголовно-правовыми средствами

Пример 2

Аннотация к статье:
Рассмотрены вопросы создания системы охраны территорий и объектов стратегического назначения. Предложены структура и способ построения такой системы, использующие методы теории решеток. Для обработки и анализа информации с датчиков физических величин и последующего принятия решений применяются решетки, построенные с помощью оператора замыкания.
Факт из статьи: Решетки могут быть построены без использования оператора замыкания

Пример аннотации без фактов

Если в аннотации не содержится фактов, например:
В статье рассмотрено становление британо-японских отношений в период биполярности, отражена история формирования двусторонних отношений, а

также сотрудничество в экономической, политической, научно-технической и социально-культурной сферах в указанный период. Главный акцент был сделан на рассмотрении зарождения отношений между Великобританией и Японией и анализа динамики их развития.
то напиши "Аннотация не содержит факт".

Пример неподходящего факта

Аннотация к статье: В статье обсуждаются вопросы взаимосвязи токсичности сточных вод и их химического состава. Для ряда гидрохимических показателей установлена достоверная связь между показателем токсичности, определявшимся с использованием методики, где в качестве тест-организма выступает P. Caudatum.
Факт из статьи: Для ряда гидрохимических показателей не существует достоверной связи между показателем токсичности, определявшимся с использованием методики, где в качестве тест-организма выступает P. Caudatum.
Факт не подходит, потому что можно удалить "не"и факт будет верным.

Не добавяй ничего к ответу, напиши только факт. Не пиши свои рассуждения, только факт! Теперь ваша задача - записать этот факт в следующую аннотацию, точно следуя инструкциям:

Аннотация к статье: {text}
Факт из статьи:

F Prompt for Relevance and Support Level Scoring

Описание Задачи

Вы ученый, который хорошо разбирается во всех областях науки.

Задача

Задача - оценить релевантность факта аннотации и насколько аннотация подтверждает факт.

Формат Вывода

На выходе нужно написать JSON.

Пример ответа:

```
{
   "relevance": "Релевантность факта аннотации",
   "support": "Насколько аннотация подтверждает факт"
}
```

Описание Полей

Поля relevance и support могут принимать значения от 0 до 10, где 0 — не релевантно (не подтверждает факт), 10 — максимально релевантно (подтверждает факт).

Входные Данные

Текст для анализа:
{text}

G Examples of Positive and Negative claims from ruSciFact

This section provides one positive (SUPPORTS) and one negative (REFUTES) example from the ruSciFact dataset.

Example 1

Claim: Пациенты с ишемической болезнью сердца, перенесшие чрескожные коронарные вмешательства, демонстрируют улучшение толерантности к физической нагрузке при использовании компьютеризированных систем поддержки врачебных решений.

Abstract: Цель. Изучить эффективность амбулаторных реабилитационно-профилактических программ у пациентов после чрескожных коронарных вмешательств (ЧКВ) с использованием компьютеризированной системы поддержки врачебных решений (СПКР), предназначенной для выбора режима контролируемых физических тренировок (КФТ) и предоставления полноценных рекомендаций по физической активности (ФА). Материал и методы. Исследование выполняли в течение 12 мес. с включением 194 пациентов (124 мужчины и 70 женщин, средний возраст 53,5) со стабильной формой ишемической болезни сердца (ИБС), перенесших ЧКВ (коронарную ангиопластику, коронарное стентирование). При выборе режима КФТ использовалась компьютеризированная СПКР. Традиционные врачебные решения анализировали по специально разработанной анкете. Результаты. Пациенты группы КФТ, продемонстрировали достоверное увеличение толерантности к физической нагрузке (ТФН) и средней продолжительности ФН, положительную динамику качества жизни (КЖ), высокий уровень приверженности лекарственной терапии на протяжении всего периода реабилитации. При формировании врачебных решений использовали, в среднем, 3 клинических признака. Наиболее типичные врачебные ошибки носили методологический характер. Заключение. Интегрирование реабилитационных программ с использованием СПВР в амбулаторных условиях у пациентов, перенесших ЧКВ, обеспечивает высокую эффективность реабилитационно-профилактических мероприятий и безопасность ФТ.

Label: SUPPORTS

Example 2

Claim: Порода карпа не влияет на содержание сухого вещества, жира, протеина у сеголетков карпа

Abstract: В статье приведены результаты сравнения биохимического состава тела сеголетков и годовиков некоторых коллекционных пород карпа, разводимых в спу «Изобелино»: немецкого, сарбоянского, отводок изобелинского карпа (столин XVIII, три прим, смесь чешуйчатая), выращенных одновременно, в одинаковых условиях и зимовавших совместно в одном пруду. Установлены породы, характеризующиеся повышенными уровнями содержания сухого вещества, жира, протеина у сеголетков карпа. В результате исследования биохимического состава тела сеголетков карпа разной породной принадлежности, выращенных в одинаковых условиях, проявляется тенденция к увеличению содержания сухого вещества, жира и протеина у коллекционных линий карпа белорусской селекции (изобелинский) по сравнению с породами зарубежной селекции (немецкий и сарбоянский), выращенными одновременно в одинаковых условиях. У годовиков коллекционных линий белорусской селекции отмечается тенденция к увеличению содержания сухого вещества, жира и протеина, снижению содержания влаги по сравнению с зимовавшими совместно коллекционными породами зарубежной селекции. В результате исследования изменения показателей, характеризующих биохимический состав тела, произошедших за зимний период, установлено, что отклонения биохимических показателей, особенно содержания сухого вещества и жира у пород зарубежной селекции значительно выше, чем у линий изобелинского карпа (белорусская селекция). Полученные данные свидетельствуют о большей приспособленности карпа коллекционных линий белорусской селекции к условиям зимовки в Беларуси, по сравнению с импортными породами (немецким и сарбоянским).

Label: REFUTES