Записки научных
семинаров ПОМИ
Том 540, 2024 г.

**A. Akhmetgareeva, A. Abramov, I. Kuleshov, V. Leschuk, A. Fenogenova**

# TOWARDS RUSSIAN SUMMARIZATION: CAN ARCHITECTURE SOLVE DATA LIMITATIONS PROBLEMS?

Abstract. In this work, we investigate the automatic summarization problem, focusing on its significance, challenges, and methodologies, particularly in the context of the Russian language. We highlight the limitations of current evaluation metrics and datasets, representing diverse summarization scenarios. We study various approaches, including the formats of supervised fine-tuning, a comparison of models designed for Russian and those with cross-lingual capabilities, and the influence of reinforcement learning alignment on the final results. Contributions of this work include an examination of the summarization task for the Russian language, publication of a new instruction-based dataset and the best open-source model, and insights for further advances in the field.

## 1. Introduction

Automatic summarization is a standard task in the field of natural language processing (NLP) that aims to extract the most important information from a document or set of documents. This not only saves time but also enhances comprehension for readers across diverse domains. The rise of language models in recent years has notably enhanced summarization capabilities: producing fluent and coherent text resembling human language is fundamental in constructing modern summarization systems. While automatic summarization has made significant strides in recent years, several challenges remain. Summarization scenarios vary in definitions, domains of application, and use cases of the systems.

Assessing the quality of summaries is a significant challenge due to limitations in standard evaluation datasets, which restrict the representation of scenarios characterized by variations in task definitions and domains of application. Automatic metrics are not always representative and cannot

*Key words and phrases*: abstractive summarization, Russian language, language models, RLHF.

evaluate the creative abilities of the models [27, 12]. For the Russian language, the situation is even more challenging. There is a lack of benchmarks for generative tasks for new models and no open-source models comparable with new ChatGPT-like models.

In this work, we investigate modern fundamental models (FM)[1] for the task of summarization in different setups. First, we explore the methods of Supervised Fine-Tuning (SFT) of the models, examining how the format of summarization datasets and training on general-purpose data improve the final quality and generalization abilities of the models. Second, we compare various FMs explicitly created for the Russian language with models exhibiting cross-lingual capabilities. Additionally, we hypothesize whether the architecture of the model influences the results. The third aspect of our research studies the capabilities of reinforcement learning (RL) alignment to improve the summarization abilities of the model. Restrictions of summarization models are based on the requirement that they understand the entire range correlated with the instruction summary.

The primary contributions of this work are as follows:

- we investigate the summarization task in Russian and compare different experimental setups, examining the influence of SFT and RL alignments, as well as model architectures and characteristics on the performance of the models;
- we release an instruction-based dataset, golden testset[2], and the best model[3] in open source under MIT license.

## 2. Related Work

The main idea of the summarization task is described as the creation of a model that takes a single document, a news article, a dialogue, or a review as input and produces compressed text (a summary) with essential information from the original content [1]. Thus, automatic text summarization can be considered as an unstructured sequence-to-sequence problem where the LM takes a text as input and generates a summary resembling the reference text.

[1]According to the definition proposed in [42], "FMs are models trained on broad data at a scale and adaptable to various downstream tasks".

[2]https://huggingface.co/datasets/RussianNLP/Mixed-Summarization-Dataset

[3]https://huggingface.co/RussianNLP/FRED-T5-Summarizer

Summarization systems can be roughly divided into two categories: extractive and abstractive. Extractive summarization such as the BertSum [15] approach outputs concatenated essential segments from the original text. Abstractive methods generate a summary from the original text's internal semantic representation that captures the text's core information. Abstractive summarization is computationally more challenging than extractive summarization and requires a deep understanding of the original content. In this work, we focus only on abstractive summarization methods.

**Models**. Standard approaches for text summarization have focused on developing neural network architectures. Works such as PEGASUS [6], mT5 [21], and BART [7] have demonstrated the effectiveness of pre-training sequence-to-sequence models with the encoder-decoder architecture for abstractive summarization downstream tasks. In particular, Russian summarization has been explored in mBART[4] and ruT5[5]. However, previous models trained on specific domain corpora (such as News or Dialogues domains) without instructive format tuning cannot be applied as a general summarization system and do not correspond to human preferences.

Recent studies indicate that Large Language Models (LLMs) can enhance the quality of summaries and exhibit more human-like performance by incorporating specific instruction-following examples into the training data, deliberately altering the models' behavior to align with human intents [5]. Model responses can be better tailored to human preferences through instruction tuning [3, 4] and reinforcement learning with human or model feedback (RLHF) [8, 9]. We note that a comprehensive investigation of all of these methodologies within the framework of the Russian language remains lacking.

**Datasets**. Numerous distinct summarization datasets have been designed for various tasks. The summarization task is challenging to pose in the general case; it is often divided into many subtasks, which results in a large variety of surveys on datasets [18, 19, 17] and diverse collections for evaluation of summarization systems [16]. We will discuss commonly used Russian summarization datasets without long context for the single-document summarization task.

[4]https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta

[5]https://huggingface.co/IlyaGusev/rut5_base_sum_gazeta

Multilingual collections CNN/DailyMail [23] and XLSum [21] have been the standard summarization datasets for many years and focus on abstractive summaries with extremely short targets. With regard to the Russian standard corpus of texts like RIA, Lenta, and Gazeta corpus [11], only headline generation is given for the target. These datasets lack a variety of summarization tasks, such as notes with main ideas or key thoughts. Moreover, their expansion to domains beyond news is crucial to address the growing demand for representative summarization evaluation in diverse contexts [12]. The recent large-scale multilingual dataset of Wikihow articles WikiLingua [24] expands domain diversity. The target summaries were collected as the first sentence of each step, which was assumed to be one paragraph of the Wikihow article. The source text was collected from the rest of the paragraph. This collection method can lead to a partial loss of information coherence between summaries and the text.

Translated from English by Google Translate SAMSum [25], DialogSum [26] dialogue collections for specific types of summarization available on HuggingFace[6,7] have been also used to extend the diversity domain and types of the abstractive summaries. These summaries do not incorporate any potential lead bias compared to previous news collections.

## 3. Experimental setup

In this work we investigate different hypotheses based on the foundation models for Russian on the summarization task. We analyze the impact of instruction datasets on the final performance and models of various sizes, architectures, and languages, and we additionally explore the influence of RL alignment on the result. We focus on the general Russian summarizer for a single document; studying longer contexts falls beyond the scope of the current research.

**SFT experiments**. During the supervised fine-tuning phase, we utilized large pre-trained models explicitly created for the Russian language and explored models with cross-lingual capabilities. We conduct experiments on the latter to determine the most effective model for the Russian language. We select models of varying sizes (see Section 3.1 for details) based on their performance on Russian benchmarks and fine-tune them using two distinct datasets: closed and open instructional data. For the closed dataset, we curate a syntactic-based proprietary set comprising

[6]https://huggingface.co/datasets/d0rj/dialogsum-ru

[7]https://huggingface.co/datasets/d0rj/samsum-ru

the most prevalent cases in summarization tasks. The second set is open and encompasses summarization datasets from diverse domains, supplemented with additional instructions. We fine-tune both decoder-based and encoder-decoder architecture models on both datasets.

**RL experiments**. Just as OpenAI enhances the architecture of ChatGPT[8] through reinforcement learning, we propose that summarization quality can also be improved using RLHF techniques. In a prior study [4] by the OpenAI team, the authors explored RLHF for the summarization task and proved that decoder-like architectures can learn from human feedback. In this study, we replicate these experiments using the models from the SFT step, comparing their performance with and without RL alignment (see Section 3.4 for details). Our investigation compares the Transformer's decoder-like architecture with the encoder-decoder commonly used in sequence-to-sequence tasks.

**Ablation study**. Additionally, we conduct two ablation studies.

1. The first experiment focuses on models pre-trained for English (or Chinese) but capable of generating Russian texts. We conducted an ablation study to investigate the impact of varying the amount of training data on the models' performance. In the experiment with fine-tuning multilingual models like LLaMA for the summarization task, we found that models, after fine-tuning in Russian, still switch to English while generating the text. We decided to add the perplexity metric of the target language, Russian, and the perplexity of the most popular language of pre-train English / Chinese (QWEN-7B focus on both Chinese and English) to the summarization SFT process.

We evaluate metrics for 0 SFT data (pre-trained model), 10,000 SFT data points in training, 100,000 SFT data points, and training on the maximum size of our dataset which is about 200k SFT data points (final step of the summarization training process). Thus, we can evaluate the quality of multilingual models with a lower perplexity of the target language. We also count the code-switching percentage [43] to another language relative to the target language of SFT data.

[8]https://chat.openai.com/

2. The second experiment investigates whether the models' abilities improve when we add general knowledge instructions to the summarization instructions set. We compare the pipelines with FRED-T5-1.7B fine-tuned on general-purpose instructions before training on closed summarization datasets with FRED-T5-1.7B trained only on closed summarization: FRED_1.7B_INSTR_$\text{SUM}_{\text{closed}}$ and FRED_1.7B_$\text{SUM}_{\text{closed}}$ respectively.

**3.1. Models.** The experimental setup requires FMs that can generate text in Russian. For the main pipeline of the experiments, we used two open-source models precisely for the Russian language and open-source English models with cross-lingual abilities for Russian. We have chosen the following models.

- **Mistral-7B-Instruct-v0.2**[9] [35] is an improved instruction fine-tuned version of the Mistral-7B pre-train generative text model with 7 billion parameters. The Mistral-7B model utilizes advanced architectural features such as grouped-query attention (GQA) for faster inference and sliding window attention (SWA) to effectively handle longer contexts. The model effectively copes with the Russian language and shows the high performance of both the pre-train version and the instructive versions on the new benchmark MERA[10] [38].
- **FRED-T5-1.7B model** [41] is an encoder-decoder model based on T5 and UL2. This model is specially designed for the Russian language, leveraging a substantial dataset to achieve high performance in text-to-text generation tasks. The model shows high performance[11] on the benchmark Russian SuperGLUE [40].
- **ruGPT-3.5 13B**[12] is an advanced language model with 13 billion parameters. It is the largest open-source foundation model for the Russian language, presented on the MERA benchmark.

For the ablation studies, in addition to Mistral-7B-Instruct-v0.2, we utilize several models known for their cross-lingual capabilities. These models were selected based on their performance on the Open LLM Leaderboard[13]

[9] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
[10] https://mera.a-ai.ru/ru/leaderboard
[11] https://russiansuperglue.com/leaderboard/2
[12] https://huggingface.co/ai-forever/ruGPT-3.5-13B
[13] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

at the time of the experiments. Specifically, we analyzed QWEN-7B[14], LLaMA-2[15] [39], and LLaMA-1[16].

**3.2. Data.** FMs interact with humans via instructions. Therefore, the restrictions to the summarization models are based on the requirement that they understand the entire range correlated with the instruction summary. Another crucial consideration is the diversity of summarization types that must be addressed. We analyzed open-source instruction-based generative datasets containing summarization tasks, as well as existing non-instructive datasets with various summary formats, comparing them with industry demands. Therefore, for all the datasets used in training and testing, we consider the nuances of different domains and various types of summarizations, such as bullet points, indirect speech, news specifications, and more.

**General-purpose SFT data.** We used the Orca dataset [22] for fine-tuning on general-purpose instructions. It was translated into Russian with DeepL API[17] and then filtered using the multilingual-e5-base[18] proposed in [20] and cosine similarity metric between original text and translation. All records with a metric lower than 0.87 were removed from the translated dataset, reducing its size to 3,579,872 instances. Second, we used an open summarization dataset for training our pre-train model and model after fine-tuning with general instructions tasks to fine-tune the downstream summarization task.

There are two training summarization instruction sets, which we will refer to in the paper as "open" and "closed". The closed set comprises manually verified synthetic data consisting of 18,241 items. It was automatically collected using the proprietary GPT-4 model, with requests for generating various types of summaries. We do not publish it due to the high cost of collection. The open training set is derived from summarization datasets sourced from open-access repositories.

- **XLSum**[19]: multilingual corpus of BBC news articles with titles given as a summary;

[14]https://huggingface.co/Qwen/Qwen-7B
[15]https://huggingface.co/meta-llama/Llama-2-7b-hf
[16]https://github.com/facebookresearch/llama
[17]https://www.deepl.com
[18]https://huggingface.co/intfloat/multilingual-e5-base
[19]https://huggingface.co/datasets/csebuetnlp/xlsum

- **Gazeta**[20]: Russian corpus of *Gazeta.ru* articles with the summaries retrieved from the contents of an HTML tag with the "description" property;
- **WikiLingua**[21]: collections of Wikihow articles, where the target summaries are the first sentence of each paragraph of the article;
- **MLSUM**[22]: multilingual summarization dataset of articles from CNN and Daily Mail with summaries as a short description of the text;
- **Reviews-russian**[23]: a small corpus of Russian reviews for hotels;
- **Curation-corpus (ru)**[24]: collections of news articles with professionally written summaries of news articles, translated into Russian;
- **Matreshka**[25]: collections of Russian life dialogues with summaries generated by ruGPT-3;
- **DialogSum (ru)**[26]: dialogue summarization dataset with manually labeled summaries, translated into Russian via Google Translate;
- **SAMSum (ru)**[27]: corpus of dialogues with human-written summaries, translated into Russian via Google Translate.

The number of examples in each open dataset and domain information are shown in Table 1.

Additionally, all open datasets were post-processed to accommodate model limitations. Instructions exceeding 1024 tokens in prompt length and 300 tokens in response length were removed from the training sets for the FRED-T5-1.7B model. Similarly, instructions exceeding a combined length of 1400 tokens were excluded from training sets for Mistral-7B, LLaMA-7B, and QWEN-7B. The limit for ruGPT-3.5 13B's is set at 1200 tokens due to memory constraints on graphics accelerators. Domains balance the resulting set based on open-sourced data and contain 197,561 items.

[20]https://huggingface.co/datasets/IlyaGusev/gazeta
[21]https://huggingface.co/datasets/GEM/wiki_lingua
[22]https://huggingface.co/datasets/mlsum
[23]https://huggingface.co/datasets/trixdade/reviews_russian
[24]https://huggingface.co/datasets/d0rj/curation-corpus-ru
[25]https://huggingface.co/datasets/zjkarina/matreshka
[26]https://huggingface.co/datasets/rcp-meetings/rudialogsum_v2
[27]https://huggingface.co/datasets/d0rj/samsum-ru

Table 1. Open source summarization datasets with the number of items in the training set ($N_{\text{train}}$) and validation set ($N_{\text{val}}$) splits, the domain of summarization (Domain), and the average percentage (%) of original texts compression.

| Dataset Name | $N_{\text{train}}$ | $N_{\text{val}}$ | Domain | Compression in % |
|---|---|---|---|---|
| **XLSum** | 62243 | 7780 | News | 4.5 |
| **Gazeta** | 74126 | 6369 | News | 6.7 |
| **WikiLingua** | 35313 | 4984 | Mixed WikiHow | 10.2 |
| **MLSUM** | 25556 | 750 | News | 1.6 |
| **Reviews-russian** | 95 | 15 | Reviews | 24.8 |
| **Curation-corpus (ru)** | 30454 | – | Curations | 14.8 |
| **Matreshka** | 6655 | 1664 | Dialogues | 32.3 |
| **DialogSum (ru)** | 12460 | 1500 | Dialogues | 17.6 |
| **SAMSum (ru)** | 14731 | 818 | Dialogues | 15.3 |

For instruction tuning, we randomly assign summarization dialogue prompts for dialog sets (Matreshka, DialogSum (ru), SAMSum (ru)) and general summarization prompts for the rest of the train sets. All instructions are presented in Appendix B.

**Golden testset**. The testset was created using a semi-automatic method. Initially, we manually collected multidomain texts and passed them through the GPT-4 API. We request the system to write different popular types of summaries with ten randomly assigned instructions. The GPT-4 outputs were checked and partially rewritten manually by a professional editor. This resulted in a minimum of 25 texts per prompt, each requiring a specific type of summary. The list of example instructions is shown in Appendix B. The final size of the golden set is 258 items.

**Reward data**. The training set for the reward model, in both open and closed versions, was taken from classical datasets, namely *summarize_from_feedback*[28] proposed in [4], *hh-rlhf*[29] presented in [14], and open datasets available on the HuggingFace platform: *webgpt_comparisons*[30]

[28]https://huggingface.co/datasets/openai/summarize_from_feedback
[29]https://huggingface.co/datasets/Anthropic/hh-rlhf
[30]https://huggingface.co/datasets/openai/webgpt_comparisons

and *synthetic-instruct-gptj-pairwise*[31]. Datasets were translated into Russian with google-translate-api[32] and filtered with cosine similarity $> 0.87$ of token embeddings between original and translation texts from multilingual-e5-base model[33]. The resulting training data comprised 11,000 unique examples, with good examples ("chosen") having summaries 2.5 times shorter than the original text. The golden testset described above was used for the automatic evaluation of the reward model. We expanded it with bad ("rejected") summaries by automatically generating short summaries using the open API tool[34], which was claimed by the authors to provide multi-domain support.

**3.3. Supervised fine-tuning details.** For all experiments, we employed 8xNvidia A100 80 Gb. For FRED-T5-1.7B training, we utilized the Hugging Face trainer. The model was trained for 7 epochs with the following parameters: optimizer – Adafactor [36] with a learning rate of 1e-4 and weight decay of 0.05. We employed a constant scheduler, and training was conducted in Bfloat16 format. For ruGPT-3.5 13B and Mistral-7B, we employed handwritten learning cycles using accelerate[35] and DeepSpeed[36]. These models were training for 7 epochs with the following parameters: optimizer – AdamW [37] with a learning rate of 1e-5, betas (0.9, 0.999), and a linear scheduler with warmup.

**3.4. Reinforcement learning alignment details.** The next step in the pipeline involves training a model to align summaries with human feedback. We adopt a solution based on the reward model described in [3]. For each SFT architecture mentioned in the previous section we add the head of BertEncoder[37] proposed in [41] as an embedding layer and a linear layer to get the scalar value. This model is trained to predict which summary $y \in \{y_0, y_1\}$ is better for given a text $x$ as judged by a human. Let the summary preferred by the human be $y_i$; the reward model (RM) loss is then calculated as follows:

$$\text{loss}(r_\theta) = -E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))], \qquad (1)$$

[31]https://huggingface.co/datasets/Dahoas/synthetic-instruct-gptj-pairwise
[32]https://cloud.google.com/translate?hl=ru
[33]https://huggingface.co/intfloat/multilingual-e5-base
[34]API summarizator
[35]https://github.com/huggingface/accelerate
[36]https://github.com/microsoft/DeepSpeed
[37]https://huggingface.co/ai-forever/ruBert-base

where $r_\theta(x, y)$ is the scalar output of the reward model for text $x$ and summary $y$ with parameters $\theta$, and $D$ is the dataset of human judgments. Subsequently, we employ the reward model with a best-of-N scheme to rerank generative summaries of texts based on the reward score, selecting the top-1 as the best generation.

## 4. Evaluation

**4.1. Metrics.** Our evaluation pipeline consists of two automatic assessment approaches. First, we review automatic metrics applied to generated text and reference summaries. The common metrics **BLEU** [29] and **ROUGE** [30] based on measuring the number of n-gram overlap between texts often correlate poorly with human judgments for many natural language generation (NLG) tasks and, in particular, for the summarization task [2]. To reduce the correlation gap, we add a cosine similarity score based on aligning contextualized token embeddings from **BERT** and **LaBSE** models [28]. We also add recent metrics based on the harmonic mean of precision and recall from calculated unigram mapping **METEOR** [32] and F-score statistic for character-based n-gram overlap **CHRF** [31].

Second, we examine models with **LM-as-an-Examiner** [33], an approach imitating human assessment by evaluating the correlations between models and human judgment [34]. The generated summaries are measured to be trustworthy (accurately reflecting the content of the original text for factual correctness), grammatically correct, coherent, and ensure no essential details are lost [27]. We ask GPT-4 to assess text with different abstracts generated by our models using the **Likert** scale from 1 (poor summary) to 10 (perfect summary), considering criteria like fluency, factuality, coherency, etc. The prompt is provided in Appendix A.

We evaluate the performance of the reward model by counting the **Accuracy** of the correct chosen summary as it agrees with human judgment. We also add the mean difference between the reward for chosen and reward for rejected summaries and the standard deviation values for the reward model for chosen summaries and rejected summaries: $\mu_{RM_{\text{chosen}}} - \mu_{RM_{\text{rejected}}}$, $\sigma_{RM_{\text{chosen}}}$, and $\sigma_{RM_{\text{rejected}}}$. The mean difference in rewards shows the ability of the model to distinguish between good and bad summaries.

**4.2. Results.** *Reinforcement learning alignment.* Reranking generations using RMs significantly increases auto metrics for decoder-based models.

However, the gain for seq2seq-like FRED-T5-1.7B models is a decrease in auto metrics.

The results are presented in Table 3. We find a noticeable correlation between the quality measurements of RM at the beginning of SFT and the final automatic metrics presented in Table 2. Decoder models exhibit notably high reward and pair accuracy scores, while FRED-T5-1.7B shows the lowest indicators. This trend persists in the automatic metrics reported in Table 2. These observations suggest a potential influence of the models' capacity level on both the quality of SFT and the RM when trained on the same datasets.

*SFT for summarization.* The results of the models described in experimental setup 3 are presented in Table 2. We compare the obtained models with the several open-source baselines:

- ruGPT-3 medium gazeta[38] is the ruGPT-3-medium [41] model fine-tuned on abstractive summarization standard Gazeta dataset;
- GPT 3.5 turbo is the open model API by OpenAI;
- mT5 multilingual XLSum[39] is the multilingual summarization baseline mentioned in [21] The authors fine-tuned the mT5 model on the 45 languages of the XLSum dataset;
- mbart gazeta[40] is the mBART base model fine-tuned on abstractive summarization standard Gazeta dataset.

Due to the high cost of GPT-4 evaluations, we do not provide the results for all models. The FRED-T5-1.7B Likert is measured without the RL alignment, as the reward models show worse performance by the auto metrics. Table 3 also shows that the FRED-T5-1.7B reward for closed and open datasets does not differ for the summaries. We, in contrast, report the reward Likert metrics for the ruGPT-3.5 13B and Mistral 7B models. Mistral 7B, also tuned on the closed set with RL alignment, performs virtually identically to the GPT 3.5 turbo model on the Likert scale.

We note a consistent trend where the performance of SFT models with RL significantly improves relative to open-source baselines. Table 2 shows that FRED-T5-1.7B trained on the closed dataset performs best according to both auto metrics and Likert evaluation, surpassing GPT 3.5 turbo but falling short relative to the gold summaries provided by humans.

[38] https://huggingface.co/IlyaGusev/rugpt3medium_sum_gazeta
[39] https://huggingface.co/csebuetnlp/mT5_multilingual_XLSum
[40] https://huggingface.co/IlyaGusev/mbart_ru_sum_gazeta

Table 2. Automatic metrics of model performance results on the golden testset. The best score is in bold, and the second-best one is underlined.

| **Model name** | **Rouge$_1$** | **Rouge$_2$** | **Rouge$_L$** | **BERT$_{score}$** | **BLEU** | **METEOR** | **LaBSE** | **ChrF** | **Likert** |
|---|---|---|---|---|---|---|---|---|---|
| Mistral 7B$_{\text{pretrain}}$ | 0.066 | 0.031 | 0.062 | 0.738 | 0.073 | **0.389** | 0.822 | 44.39 | - |
| Mistral 7B$_{\text{sft_closed}}$ | 0.303 | 0.181 | 0.292 | 0.763 | 0.154 | 0.321 | 0.824 | 42.794 | - |
| Mistral 7B$_{\text{reward_closed}}$ | 0.318 | 0.177 | 0.304 | 0.756 | 0.126 | 0.347 | 0.826 | <u>45.404</u> | <u>6.047</u> |
| Mistral 7B$_{\text{sft_open}}$ | 0.148 | 0.065 | 0.139 | 0.696 | 0.089 | 0.164 | 0.661 | 29.754 | - |
| Mistral 7B$_{\text{reward_open}}$ | 0.187 | 0.098 | 0.181 | 0.716 | 0.111 | 0.239 | 0.725 | 37.884 | 4.333 |
| ruGPT-3.5 13B$_{\text{pretrain}}$ | 0.274 | 0.154 | 0.264 | 0.719 | 0.067 | 0.370 | 0.792 | 44.34 | 1.409 |
| ruGPT-3.5 13B$_{\text{sft_closed}}$ | 0.307 | 0.178 | 0.295 | 0.721 | 0.090 | 0.360 | 0.811 | 45.236 | - |
| ruGPT-3.5 13B$_{\text{reward_closed}}$ | 0.310 | 0.186 | 0.299 | 0.714 | 0.095 | 0.366 | 0.811 | **45.694** | 4.363 |
| ruGPT-3.5 13B$_{\text{sft_open}}$ | 0.229 | 0.121 | 0.223 | 0.739 | 0.117 | 0.242 | 0.766 | 37.852 | - |
| ruGPT-3.5 13B$_{\text{reward_open}}$ | 0.274 | 0.133 | 0.268 | 0.752 | <u>0.157</u> | 0.321 | 0.802 | 43.552 | 5.954 |
| FRED-T5-1.7B$_{\text{pretrain}}$ | 0.252 | 0.128 | 0.238 | 0.726 | 0.112 | 0.329 | 0.788 | 43.37 | 2.244 |
| FRED-T5-1.7B$_{\text{sft_closed}}$ | **0.405** | **0.262** | **0.387** | <u>0.766</u> | **0.159** | 0.354 | **0.841** | 44.844 | **7.221** |
| FRED-T5-1.7B$_{\text{reward_closed}}$ | 0.307 | 0.158 | 0.295 | 0.754 | 0.120 | 0.301 | 0.818 | 40.468 | - |
| FRED-T5-1.7B$_{\text{sft_open}}$ | 0.175 | 0.087 | 0.172 | 0.715 | 0.059 | 0.204 | 0.736 | 26.648 | 4.331 |
| FRED-T5-1.7B$_{\text{reward_open}}$ | 0.156 | 0.054 | 0.151 | 0.669 | 0.054 | 0.170 | 0.690 | 22.488 | - |
| Gold human results | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 100 | **8.902** |
| ruGPT-3-medium gazeta | 0.229 | 0.114 | 0.218 | 0.720 | 0.074 | <u>0.383</u> | 0.812 | 45.18 | 1.238 |
| GPT 3.5 turbo | <u>0.361</u> | <u>0.227</u> | <u>0.351</u> | **0.781** | 0.129 | 0.338 | <u>0.833</u> | 40.164 | <u>6.763</u> |
| mT5 multilingual XLSum | 0.066 | 0.021 | 0.064 | 0.666 | 0.014 | 0.111 | 0.586 | 17.894 | 2.977 |
| mbart gazeta | 0.125 | 0.041 | 0.127 | 0.639 | 0.0616 | 0.160 | 0.679 | 21.989 | 1.977 |

Table 3. Automatic metrics for reward model performance on the golden testset; best score in bold, best score for each SFT architecture underlined; $\mu$ is the mean difference between reward values of "chosen" and "rejected" summaries, $\sigma$ is the standard deviation.

| **Model** | **Accuracy** | $\mu_{RM_{\text{chosen}}} - \mu_{RM_{\text{rejected}}}$ | $\sigma_{RM_{\text{chosen}}}$ | $\sigma_{RM_{\text{rejected}}}$ |
|---|---|---|---|---|
| Mistral 7B$_{\text{reward_closed}}$ | <u>0.8249</u> | 4.12 | 2.85 | 2.48 |
| Mistral 7B$_{\text{reward_open}}$ | 0.7939 | 3.86 | 3.34 | 3.09 |
| ruGPT-3.5 13B$_{\text{reward_closed}}$ | 0.865 | 1.67 | 2.35 | 1.66 |
| ruGPT-3.5 13B$_{\text{reward_open}}$ | **0.926** | 2.34 | 1.75 | 1.40 |
| FRED-T5-1.7B$_{\text{reward_closed}}$ | <u>0.712</u> | 0.01 | 0.04 | 0.06 |
| FRED-T5-1.7B$_{\text{reward_open}}$ | 0.649 | 0.036 | 0.08 | 0.07 |

We conclude that the encoder-decoder architecture is the best for the summarization task based on Table 2. However, in addition to a well-chosen

Table 4. Perplexity of Russian and English (Chinese) languages for multilingual models on different stages of iterative training with different numbers of samples $N$ in the training set (increasing $N$ leads to slightly forgetting the pre-training language while learning the SFT language).

| **Model and Language** | **SFT(0)** | **SFT(10k)** | **SFT(100k)** | **SFT(200k)** |
|---|---|---|---|---|
| QWEN-7B Russian | 27.328 | 24.987 | 20.897 | 18.753 |
| QWEN-7B Chinese | 6.622 | 7.258 | 12.158 | 14.857 |
| LLaMA-1-7B Russian | 34.762 | 29.367 | 25.486 | 22.628 |
| LLaMA-1-7B English | 7.179 | 9.658 | 14.648 | 19.654 |
| LLaMA-2-7B Russian | 27.884 | 25.234 | 21.321 | 18.995 |
| LLaMA-2-7B English | 6.798 | 7.893 | 11.384 | 15.415 |
| Mistral-7B-Instruct Russian | **25.641** | 23.487 | 19.239 | **17.788** |
| Mistral-7B-Instruct English | **6.214** | 7.782 | 10.897 | **13.214** |
| Mistral-7B Russian | 26.759 | 25.058 | 19.498 | 18.218 |
| Mistral-7B English | 6.589 | 8.056 | 11.358 | 14.079 |

Table 5. Automatic metrics of model performance on the golden testset with and without general-purpose tuning for FRED-T5-1.7B model; best score in bold.

| **Model name** | **Rouge$_1$** | **Rouge$_2$** | **Rouge$_L$** | **BERT$_{score}$** | **BLEU** | **METEOR** | **LaBSE** | **ChrF** |
|---|---|---|---|---|---|---|---|---|
| FRED_1.7B_INSTR_SUM$_{closed}$ | **0.399** | **0.260** | **0.385** | **0.769** | **0.157** | **0.355** | **0.840** | **44.826** |
| FRED_1.7B_SUM$_{closed}$ | 0.356 | 0.210 | 0.342 | 0.760 | 0.140 | 0.330 | 0.827 | 43.229 |

architecture, we still need a high-quality closed training dataset to obtain results comparable to those of GPT 3.5 turbo and humans.

**Ablation study**.

1. The results presented in Table 4 highlight that an increased amount of training data enhances the performance of multilingual models, leading to lower perplexity in specific languages. Mistral-7B demonstrates higher performance among the models discussed due to the distinct SFT step performed (the pre-trained Mistral exhibits slightly higher perplexity). Additionally, for the final SFT with approximately 200k samples in training, we calculate the percentage of code-switching to the SFT target language (Russian). For LLaMA-1-7B and LLaMA-2-7B, these percentages are 9% and 5%, respectively, while for QWEN-7B and Mistral, the percentages are 3% and less than 1%, respectively. Thus, training a model

with cross-lingual capabilities on a downstream task can yield a fully competitive model, and 200k examples are adequate for this purpose. This pipeline is more computationally efficient as we can skip the pre-learning part for a particular language. The best model based on percentages of code-switching for the Russian language is Mistral-7B-Instruct.

2. We conduct an automatic evaluation of two models: FRED-T5-1.7B trained on both general-purpose instructions and closed set with FRED-T5-1.7B trained only on closed summarization. Table 5 with automatic evaluations on the golden testset proves the hypothesis: fine-tuning first on the general-purpose instruction set, followed by tuning on the specific task data, enhances performance.

Finally, we recommend using the instructions' setup. Although it is more difficult to collect, it provides better quality and allows you to manage the length and type of a summary. Additionally, we propose the fine-tuned model FRED-T5-1.7B with encoder-decoder architecture for the Russian summarization. This model is more computationally efficient than classical FMs (1.7B vs 7B) regarding training time. Moreover, the model provides results comparable to those of humans on the presented evaluation.

## 5. Conclusion

Automatic summarization remains a crucial task in NLP. Despite recent advancements, challenges persist, particularly for the Russian language, due to limited data resources and open-source models. Our study explores various approaches to enhance summarization quality, such as supervised fine-tuning and RL alignment. We contribute by investigating these methods, providing a new instruction-based dataset, and releasing the best model. Notably, the FRED-T5-1.7B model trained on a closed semi-supervised instruction set demonstrates high quality across various auto metrics, including Likert-based evaluation via GPT-4. We also emphasize the lack of representative auto metrics for sequence-to-sequence tasks. We hope our work provides valuable insights for advancing automatic summarization in Russian and beyond.

## 6. Ethical consideration

**6.1. Possible Misuse and Biases.** One of the contributions of our work is an open-source model. Thus, it should not be used to create content that affects individual or communal well-being, such as manipulating information or spreading misinformation. Summarization models can involuntarily

reproduce biases present in the training data, leading to biased or unfair summaries. We highlight the importance of using diverse and representative datasets to train the models to mitigate bias and promote fairness in summarization outputs.

**6.2. Data.** The datasets for training and gold test sets include large segments representing the Internet domain, and therefore, they may possibly contain a mixture of stereotypes and biases. The lack of data in various domains is still a crucial problem for all the datasets, resulting in biases and poor performance on others. We filtered and verified our datasets manually; however, proper evaluation is still needed to explore possible model vulnerabilities in terms of generalizing on the new data and specific new data.

**6.3. Limitations.** The primary limitation of the proposed methodology is the model's context length. The best model, FRED-T5-1.7B, is limited to context size 1024 that cannot fully encompass the scope of a book. It is crucial to note the importance of trustworthiness and factual correctness of the generated text. Misleading or inaccurate summaries could have significant consequences, particularly in contexts such as news reporting or legal documents. The automatic evaluation needs to incorporate such metrics. However, only the human evaluation is reliable, as the LM-as-judge approach can introduce bias.

## Appendix A. Evaluation prompt

We provide the prompt for evaluation of the generated summary used for GPT-4 (both in Russian and translated English version of the prompt):

«Пожалуйста, оцените целым числом качество следующих суммаризаций текста (генераций), используя шкалу Ликерта от 1 до 10, где 1 означает "очень плохая суммаризация", а 10 означает "отличная суммаризация". Оценка должна учитывать такие аспекты, как сохранение смысла текста, объём сокращения, формат суммаризации (по пунктам / в одном предложении), точность переданных фактов, логическую связность и полноту изложения (достаточную релевантную информацию, имеющую практическую ценность). Также, пожалуйста, обоснуйте каждую оценку, избегая любой потенциальной предвзятости и гарантируя, что порядок, в котором были представлены ответы, не повлияет на ваше суждение.

На вход подается ”Исходный текст“ (текст + инструкция) и его сокращенные версии (суммаризации) — ”Генерация 1“, ”Генерация 2“ и тд, которые требуется оценить в поле ”Оценка“. Следуйте этому формату и дайте оценку каждой генерации.

Исходный текст:
Генерация 0: [Текст генерации от модели 0]
Оценка:
Комментарий:
Генерация 1: [Текст генерации от модели 1]
Оценка:
Комментарий:
Генерация 2: [Текст генерации от модели 2]
Оценка:
Комментарий:»

«Please rate the quality of the following text summarizations (generations) using the Likert scale from 1 to 10 (integer value), where 1 means ”very poor summarization“ and 10 means ”excellent summarization“. The assessment should take into account such criteria as the correctness of the information in the text, the amount of reduction, the format of summarization (point by point / in one sentence), the accuracy of the facts, logical coherence and completeness of the facts (sufficient relevant information of practical value). Also, please justify each assessment, avoiding potential bias and ensuring that the order in which the answers were presented will not affect judgment. The input is provided with the ”Source text“ (text + instructions) and its shorted versions (summarizations) — ”Generation 1“, ”Generation 2“, etc., which must be evaluated in the ”Evaluation“ field. Follow this format and give an estimate of each generation.

Source text:
Generation 0: [Text of generated summary from model 0]
Score:
Comment:
Generation 1: [Text of generated summary from model 1]
Score:
Comment:
Generation 2: [Text of generated summary from model 2]
Score:
Comment:»

## Appendix B. Summarization instructions

We provide examples of instructions for the **train set** both in Russian and translated English versions of the prompt. The prompts for the general summarization domain (not the dialogue domain):

- «Кратко суммаризируй текст:» / «Summarize the text briefly:»,
- «Расскажи основной смысл:» / «Tell the main idea:»,
- «Суммаризируй:» / «Summarize:»,
- «Сократи текст:» / «Shorten the text:»,
- «Напиши основные тезисы этого текста» / «Write the main thesis of this text:»,
- «Можешь коротко объяснить, о чем тут говорится?» / «Can you briefly explain what is being discussed here?»,
- «Дай краткое изложение этого текста» / «Give a brief summary of this text».

The prompts for the dialogue domain:

- «Расскажи основной смысл диалога:» / «Give a brief summary of the dialogue:»,
- «Суммаризируй диалог:» / «Summarize the dialogue:»,
- «Какая основная тема разговора?» / «What is the main topic of conversation?».

We provide examples of instructions for the **golden set** both in Russian and translated English version of the prompt:

- «Кратко суммаризируй текст:» / «Summarize the text briefly:»,
- «Расскажи основной смысл:» / «Tell the main idea:»,
- «Суммаризируй:» / «Summarize:»,
- «Сократи текст:» / «Shorten the text:»,
- «Напиши тезисы этого текста по пунктам» / «Write the theses of this text point by point:»,
- «Напиши основные тезисы этого текста» / «Write the main thesis of this text:»,
- «Перепиши этот текст так, чтобы он стал вдвое короче» / «Rewrite this text to make it half as short:»,
- «Можешь коротко объяснить, о чем тут говорится?» / «Can you briefly explain what is being discussed here?»,
- «Дай краткое изложением этого текста» / «Give a brief summary of this text».

## References


1. A. Nenkova and K. McKeown, *Automatic Summarization.* — Found. Trends Inf. Retr. **5** (2011).
2. W. Kryściński, B. McCann, C. Xiong, and R. Socher, *Evaluating the Factual Consistency of Abstractive Text Summarization*, ArXiv preprint arXiv:1910.12840 (2019).
3. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, *Training Language Models to Follow Instructions with Human Feedback*, ArXiv preprint arXiv:2203.02155 (2022).
4. N. Stiennon, L. Ouyang, J. Wu, D.M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano, *Learning to Summarize from Human Feedback*, ArXiv preprint arXiv:2009.01325 (2022).
5. M. Shu, J. Wang, C. Zhu, J. Geiping, C. Xiao, and T. Goldstein, *On the Exploitability of Instruction Tuning*, ArXiv preprint arXiv:2306.17194 (2023).
6. J. Zhang, Y. Zhao, M. Saleh, and P.J. Liu, *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*, ArXiv preprint arXiv:1912.08777 (2020).
7. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, ArXiv preprint arXiv:1910.13461 (2019).
8. P. Christiano, J. Leike, T.B. Brown, M. Martic, S. Legg, and D. Amodei, *Deep Reinforcement Learning from Human Preferences*, ArXiv preprint arXiv:1706.03741 (2023).
9. L. Gao, J. Schulman, and J. Hilton, *Scaling Laws for Reward Model Overoptimization*, ArXiv preprint arXiv:2210.10760 (2022).
10. I. Gusev, *Dataset for Automatic Summarization of Russian News*, Arxiv preprint 2006.11063 (2020).
11. I.O. Gusev, *Importance of Copying Mechanism for News Headline Generation.* — Komp. Lingv. Intell. Tekhnol. **18** (2019), 229–236.
12. P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. Manning, C. Ré, D. Acosta-Navas, D. Hudson, and Y. Koreeda, *Holistic Evaluation of Language Models*, ArXiv preprint arXiv:2211.09110 (2022).
13. O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov, and T. Shavrina, *mGPT: Few-Shot Learners Go Multilingual.* — Trans. Assoc. Comput. Linguist. **12** (2024), 58–79.
14. Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, and J. Kaplan, *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, ArXiv preprint arXiv:2204.05862 (2022).

15. Y. Liu and M. Lapata, *Text Summarization with Pretrained Encoders*, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3730–3740.
16. E. Clark, S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, and A. Parikh, *SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation*, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 9397–9413.
17. V. Rennard, G. Shang, J. Hunter, and M. Vazirgiannis, *Abstractive Meeting Summarization: A Survey.* — Trans. Assoc. Comput. Linguist. **11** (2023), 861–884.
18. H. Koh, J. Ju, M. Liu, and S. Pan, *An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics.* — ACM Comput. Surv. **55** (2022).
19. M. Cao, *A Survey on Neural Abstractive Summarization Methods and Factual Consistency of Summarization*, ArXiv preprint arXiv:2204.09519 (2022).
20. L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, *Multilingual E5 Text Embeddings: A Technical Report*, ArXiv preprint arXiv:2402.05672 (2024).
21. T. Hasan, A. Bhattacharjee, M.S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M.S. Rahman, and R. Shahriyar, *XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages*, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Aug. 2021, Association for Computational Linguistics, Online, pp. 4693–4703.
22. S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah, *Orca: Progressive Learning from Complex Explanation Traces of GPT-4*, ArXiv preprint arXiv:2306.02707 (2023).
23. R. Nallapati, B. Zhou, C. Dos Santos, C. Gulcehre, and B. Xiang, *Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond*, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Feb. 2016, pp. 280–290.
24. F. Ladhak, E. Durmus, C. Cardie, and K. McKeown, *WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization*, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Nov. 2020, Association for Computational Linguistics, Online, pp. 4034–4048.
25. B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, *SAMSum Corpus: A Human-Annotated Dialogue Dataset for Abstractive Summarization*, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Nov. 2019, Association for Computational Linguistics, Hong Kong, China, pp. 70–79.
26. Y. Chen, Y. Liu, L. Chen, and Y. Zhang, *DialogSum: A Real-Life Scenario Dialogue Summarization Dataset*, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Aug. 2021, Association for Computational Linguistics, Online, pp. 5062–5074.

27. A. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, *SummEval: Re-Evaluating Summarization Evaluation.* — Trans. Assoc. Comput. Linguist. **9** (2021), 391–409.
28. T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, and Y. Artzi, *BERTScore: Evaluating Text Generation with BERT*, in: International Conference on Learning Representations (ICLR), 2020. Available: `https://openreview.net/forum?id=SkeHuCVFDr`.
29. K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, *BLEU: a Method for Automatic Evaluation of Machine Translation*, Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2002.
30. C.-Y. Lin, *ROUGE: A Package for Automatic Evaluation of Summaries*, in: Text Summarization Branches Out, Jul. 2004, Association for Computational Linguistics, Barcelona, Spain, pp. 74–81. Available: `https://aclanthology.org/W04-1013`.
31. M. Popović, *chrF: Character N-Gram F-Score for Automatic MT Evaluation*, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, edited by O. Bojar et al., Sep. 2015, Association for Computational Linguistics, Lisbon, Portugal, pp. 392–395.
32. S. Banerjee and A. Lavie, *Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments*, in: Proceedings of ACL-WMT, 2004, pp. 65–72.
33. Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu, J. Zhang, J. Li, and L. Hou, *Benchmarking Foundation Models with Language-Model-as-an-Examiner*, ArXiv preprint arXiv:2306.04181 (2023).
34. Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, and C. Tao, *Leveraging Large Language Models for NLG Evaluation: A Survey*, ArXiv preprint arXiv:2401.07103 (2024).
35. A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L.R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, *Mistral 7B*, ArXiv preprint arXiv:2310.06825 (2023).
36. N. Shazeer and M. Stern, *Adafactor: Adaptive Learning Rates with Sublinear Memory Cost*, ArXiv preprint arXiv:1804.04235 (2018).
37. I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, ArXiv preprint arXiv:1711.05101 (2019).
38. A. Fenogenova et al., *MERA: A Comprehensive LLM Evaluation in Russian*, ArXiv preprint arXiv:2401.04531 (2024).
39. H. Touvron et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, ArXiv preprint arXiv:2307.09288 (2023).
40. T. Shavrina, A. Fenogenova, A. Emelyanov, D. Shevelev, E. Artemova, V. Malykh, V. Mikhailov, M. Tikhonova, A. Chertok, and A. Evlampiev, *RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark*, ArXiv preprint arXiv:2010.15925 (2020).
41. D. Zmitrovich et al., *A Family of Pretrained Transformer Language Models for Russian*, ArXiv preprint arXiv:2309.10931 (2023).
42. R. Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ArXiv preprint arXiv:2108.07258 (2022).

43. J. Zhao, Z. Zhang, L. Gao, Q. Zhang, T. Gui, and X. Huang, *LLaMA Beyond English: An Empirical Study on Language Capability Transfer*, ArXiv preprint arXiv:2401.01055 (2024).

SberDevices

*E-mail*: albina.akhmetgareeva@gmail.com

SberDevices

SberDevices
HSE University, St. Petersburg

SberDevices

SberDevices

Поступило 15 ноября 2024 г.