TAPE: Assessing Few-shot Russian Language Understanding

Ekaterina Taktasheva[1,2]*, **Tatiana Shavrina[1,3]***, **Alena Fenogenova[1]***, **Denis Shevelev[1]**,
Nadezhda Katricheva[1], **Maria Tikhonova[1,2]**, **Albina Akhmetgareeva[1]**,
Oleg Zinkevich[2], **Anastasiia Bashmakova[2]**, **Svetlana Iordanskaia[2]**, **Alena Spiridonova[2]**,
Valentina Kurenshchikova[2], **Ekaterina Artemova[4,5]**, **Vladislav Mikhailov[1]**

[1]SberDevices, [2]HSE University, [3]Artificial Intelligence Research Institute,
[4]Huawei Noah's Ark lab, [5]CIS LMU Munich, Germany

Correspondence: rybolos@gmail.com

Abstract

Recent advances in zero-shot and few-shot learning have shown promise for a scope of research and practical purposes. However, this fast-growing area lacks standardized evaluation suites for non-English languages, hindering progress outside the Anglo-centric paradigm. To address this line of research, we propose TAPE (Text Attack and Perturbation Evaluation), a novel benchmark that includes six more complex NLU tasks for Russian, covering multi-hop reasoning, ethical concepts, logic and commonsense knowledge. The TAPE's design focuses on systematic zero-shot and few-shot NLU evaluation: *(i)* linguistic-oriented adversarial attacks and perturbations for analyzing robustness, and *(ii)* subpopulations for nuanced interpretation. The detailed analysis of testing the autoregressive baselines indicates that simple spelling-based perturbations affect the performance the most, while paraphrasing the input has a more negligible effect. At the same time, the results demonstrate a significant gap between the neural and human baselines for most tasks. We publicly release TAPE[1] to foster research on robust LMs that can generalize to new tasks when little to no supervision is available.

1 Introduction

The ability to acquire new concepts from a few examples is central to human intelligence (Tenenbaum et al., 2011). Recent advances in the NLP field have fostered the development of language models (LMs; Radford et al., 2019; Brown et al., 2020) that exhibit such generalization capacity under a wide range of few-shot learning and prompting methods (Liu et al., 2021; Beltagy et al., 2022). The community has addressed various aspects of few-shot learning, such as efficient model application (Schick and Schütze, 2021), adaptation to unseen tasks and domains (Bansal et al., 2020a,b), and

cross-lingual generalization (Winata et al., 2021; Lin et al., 2021).

The latest research has raised an essential question of standardized evaluation protocols to assess few-shot generalization from multiple perspectives. The novel tool-kits and benchmarks mainly focus on systematic evaluation design (Bragg et al., 2021; Zheng et al., 2022), cross-task generalization (Ye et al., 2021; Wang et al., 2022), and real-world scenarios (Alex et al., 2021). However, this rapidly developing area fails to provide similar evaluation suites for non-English languages, hindering progress outside the Anglo-centric paradigm.

Motivation and Contributions. In this paper, we introduce TAPE[2], a novel benchmark for few-shot Russian language understanding evaluation. Our objective is to provide a reliable tool and methodology for nuanced assessment of zero-shot and few-shot methods for Russian. The objective is achieved through two main contributions.

Contribution 1. Our first contribution is to create six more complex question answering (QA), Winograd schema, and ethics tasks for Russian. The tasks require understanding many aspects of language, multi-hop reasoning, logic, and commonsense knowledge.

The motivation behind this is that there are systems that match or outperform human baselines on most of the existing QA tasks for Russian, e.g., the ones from Russian SuperGLUE (Shavrina et al., 2020): DaNetQA (Glushkova et al., 2020), MuSeRC and RuCoS (Fenogenova et al., 2020). To the best of our knowledge, datasets on ethical concepts have not yet been created in Russian. To bridge this gap, we propose one of the first Russian datasets on estimating the ability of LMs to predict human ethical judgments about various text situations.

Contribution 2. Our second contribution is to develop a framework for multifaceted zero-shot and

---

*Equal contribution.

[1]tape-benchmark.com

[2]Text Attack and Perturbation Evaluation.

few-shot NLU evaluation. The design includes *(i)* linguistic-oriented adversarial attacks and perturbations for testing robustness, and *(ii)* subpopulations for nuanced performance analysis.

Here, we follow the methodological principles and recommendations by Bowman and Dahl (2021) and Bragg et al. (2021), which motivate the need for systematic benchmark design and adversarially-constructed test sets.

Findings. Our findings are summarized as five-fold: *(i)* zero-shot evaluation may outperform few-shot evaluation, meaning that the autoregressive baselines fail to utilize demonstrations, *(ii)* few-shot results may be unstable and sensitive to prompt changes, *(iii)* negative result: zero-shot and few-shot generation for open-domain and span selection QA tasks leads to near-zero performance, *(iv)* the baselines are most vulnerable to spelling-based and emoji-based adversarial perturbations, and *(v)* human annotators significantly outperform the neural baselines, indicating that there is still room for developing robust and generalizable systems.

2 Related Work

Benchmark Critique. Benchmarks such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) have become de facto standard tools to measure progress in NLP. However, recent studies have criticized the canonical benchmarking approaches. Bender et al. (2021) warn performance gains are achieved at the cost of carbon footprint. Elangovan et al. (2021) claim that the current benchmarks evaluate the LM's ability to memorize rather than generalize because of the significant overlap between the train and test datasets. Church and Kordoni (2022) argue that benchmarks focus on relatively easy tasks instead of creating long-term challenges. Raji et al. (2021) raise concerns about the resource-intensive task design. In particular, benchmarks present with large-scale train datasets, which are expensive to create. This may lead to benchmark stagnation, as new tasks can not be added easily (Barbosa-Silva et al., 2022). In turn, few-shot benchmarking offers a prospective avenue for LMs evaluation regarding generalization capacity, computational and resource costs.

Few-shot Benchmarking. Research in few-shot benchmarking has evolved in several directions. Schick and Schütze (2021) create FewGLUE by sampling small fixed-sized training datasets from SuperGLUE. Variance w.r.t to training dataset size and sampling strategy is not reported. Later works overcome these issues by exploring evaluation strategies by K-fold cross-validation (Perez et al., 2021), bagging, and multi-splits, introduced in FewNLU (Zheng et al., 2022). Additionally, FewNLU explores correlations between performance on development and test sets and stability w.r.t. the number of runs. CrossFit (Ye et al., 2021) studies cross-task generalization by unifying task formats and splitting tasks into training, development, and test sets. FLEX (Bragg et al., 2021) covers the best practices and provides a unified interface for different types of transfer and varying shot sizes. Finally, to the best of our knowledge, the only non-English dataset for few-shot benchmarking is Few-CLUE in Chinese (Xu et al., 2021). TAPE is the first few-shot benchmark for Russian, which introduces variations at the data level by creating adversarial test sets.

3 Task Formulations

TAPE includes six novel datasets for Russian, each requiring modeling "intellectual abilities" of at least two skills: logical reasoning (§3.1; extended Winograd schema challenge), reasoning with world knowledge (§3.2; CheGeKa, RuOpenBookQA and RuWorldTree), multi-hop reasoning (§3.2; MultiQ), and ethical judgments (§3.3; Ethics$_{1/2}$). This section describes the task formulations, general data collection stages, and dataset examples. Appendix A provides the general dataset statistics, while Appendix E.1 includes details on dataset collection and extra validation stage via a crowd-sourcing platform Toloka[3] (Pavlichenko et al., 2021).

3.1 Logical Reasoning

Winograd. The Winograd schema challenge composes tasks with syntactic ambiguity, which can be resolved with logical reasoning (Levesque et al., 2012). The texts for the dataset are collected with a semi-automatic pipeline. First, lists of 11 typical grammatical structures with syntactic homonymy (mainly case) are compiled by a few authors with linguistic background (see Appendix B). Queries corresponding to these constructions are submitted to the search of the Russian National Corpus[4], which includes a sub-corpus with re-

---

[3]toloka.ai
[4]ruscorpora.ru/en

solved homonymy. In the resulting 2k+ sentences, homonymy is resolved automatically with UDPipe[5] and then validated manually by a few authors afterward. Each sentence is split into multiple examples in the binary classification format, indicating whether the reference pronoun is dependant on the chosen candidate noun.

- **Context:** *"Brosh' iz Pompei, kotoraya perezhila veka."* (A trinket from Pompeii that has survived the centuries.)
- **Reference:** *"kotoraya"* (that)
- **Candidate Answer:** *"Brosh' "* (A trinket)
- **Label:** ✓(correct resolution)

3.2 Reasoning with World Knowledge

RuOpenBookQA. RuOpenBookQA is a QA dataset with multiple-choice elementary-level science questions, which probe understanding of 1k+ core science facts. The dataset is built with automatic translation of the original English dataset by Mihaylov et al. (2018) and manual validation by a few authors.

- **Question:** *"Yesli chelovek idet v napravlenii, protivopolozhnom napravleniyu strelki kompasa, on idet..."* (If a person walks in the direction opposite to the compass needle, they are going...)
- **Answers:** (A) *"na zapad"* (west); (B) *"na sever"* (north); (C) *"na vostok"* (east); (D) *"na yug"* (south).

RuWorldTree. The collection approach of **RuWorldTree** is similar to that of **RuOpenBookQA**, the main difference being the additional lists of facts and the logical order that is attached to the output of each answer to a question (Jansen et al., 2018).

- **Question:** *"Kakoye svoystvo vody izmenitsya, kogda voda dostignet tochki zamerzaniya?"* (What property of water will change when the water reaches the freezing point?)
- **Answers:** (A) *"tsvet"* (color); (B) *"massa"* (mass); (C) *"sostoyaniye"* (state of matter); (D) *"ves"* (weight).

MultiQ. Multi-hop reasoning has been one of the least explored QA directions for Russian. The task is addressed by the MuSeRC dataset (Fenogenova et al., 2020) and only a few dozen questions in

SberQUAD (Efimov et al., 2020) and RuBQ (Rybin et al., 2021). In response, we have developed a semi-automatic pipeline for multi-hop dataset generation based on Wikidata and Wikipedia. First, we extract the triplets from Wikidata and search for their intersections. Two triplets (subject, relation, object) are needed to compose an answerable multi-hop question. For instance, the question *"Na kakom kontinente nakhoditsya strana, grazhdaninom kotoroy byl Yokhannes Blok?"* (In what continent lies the country of which Johannes Block was a citizen?) is formed by a sequence of five graph units: *"Blok, Yokhannes"* (Block, Johannes), *"grazhdanstvo"* (country of citizenship), *"Germaniya"* (Germany), *"chast' sveta"* (continent), and *"Yevropa"* (Europe). Second, several hundreds of the corresponding question templates are curated by a few authors manually, which are further used to fine-tune ruT5-large[6] to generate multi-hop questions given the graph units sequences. Third, the resulting questions undergo paraphrasing (Fenogenova, 2021) and manual validation procedure to control the quality and diversity. Finally, each question is linked to two Wikipedia paragraphs with the help of wptools[7], where all graph units appear in the natural language. The task is to select the answer span using information from both paragraphs.

- **Question:** *"Gde nakhoditsya istok reki, pritokom kotoroy yavlyayetsya Getar?"* (Where is the source of the river, the tributary of which is the Getar?)
- **Supporting Text:** *"Getar — reka v Armenii. Beryot nachalo na territorii Kotaykskoy oblasti, protekayet cherez tsentral'nuyu chast' Yerevana i vpadayet v Razdan."* (The Getar is a river in Armenia. [It] originates in the Kotayk region, flows through the central part of Yerevan and flows into the Hrazdan.)
- **Main Text:** *"Razdan — reka v Armenii. Vytekayet iz ozera Sevan v yego severo-zapadnoy chasti, nedaleko ot goroda Sevan."* (The Hrazdan is a river in Armenia. [It] originates at the northwest extremity of Lake Sevan, near the city of Sevan.)
- **Answer:** Sevan

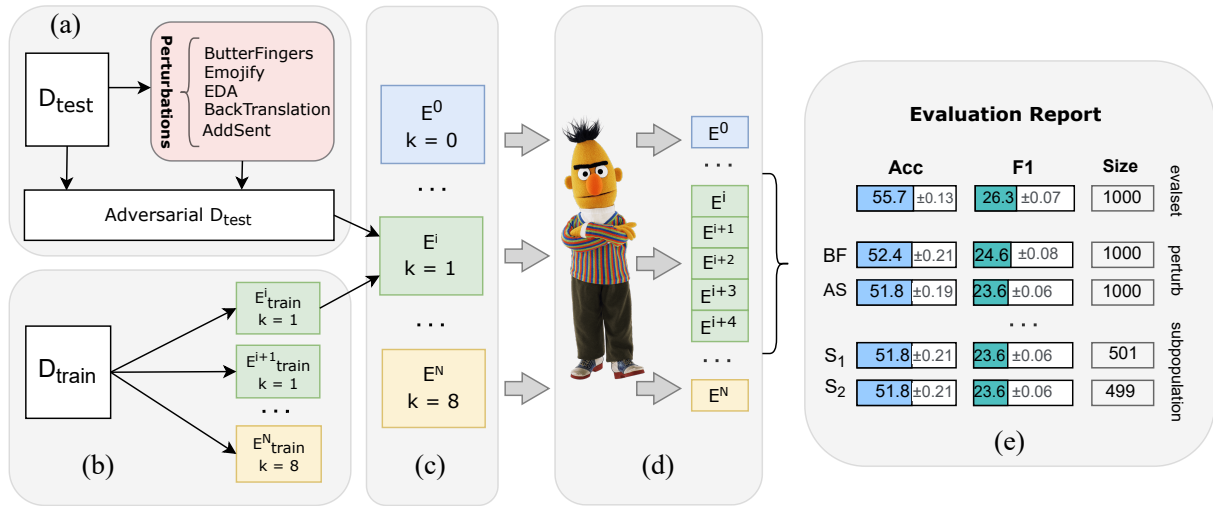**CheGeKa.** The CheGeKa game[8] setup is similar to Jeopardy, where the player should answer

---

Figure 1: Overview of the TAPE's design. **(a)** $\mathcal{D}_{test}$ is passed to the adversarial framework (§ 4.2) to create the adversarial test $\mathcal{D}_{test}^A$ that includes the original and adversarial examples. **(b)** We randomly sample 5 sets of demonstration examples from $\mathcal{D}_{train}$ for each $k \in \{1, 4, 8\}$. In the zero-shot scenario, we skip this stage. **(c)** After that, we merge the demonstrations, when applicable, with the examples from $\mathcal{D}_{test}^A$ to construct evaluation episodes $\mathcal{E}_k^N$. **(d)** Each $\mathcal{E}_k^N$ is used to obtain predictions from the model. **(e)** The performance is summarized in a diagnostic evaluation report. BF – BUTTERFINGERS, AS – ADDSENT, S – subpopulation.

questions based on wit and common sense knowledge. We directly contacted the authors of Russian Jeopardy! (Mikhalkova, 2021) and asked about including their training and private test sets in our benchmark. The task is to provide a free response given a question and the question category.

- **Question:** *"Imenno on napisal muzyku k opere Turandot."* (It was he who composed the music for the opera "Turandot".)
- **Category:** *"Komediya del' arte"* (Commedia dell'arte)
- **Answer:** *"Puchchini"* (Puccini)

3.3 Ethical Judgments

There is a multitude of approaches to evaluating ethics in machine learning. The **Ethics** dataset for Russian is created from scratch for the first time, relying on the design compatible with Hendrycks et al. (2021). The task is to predict human ethical judgments about diverse text situations in two multi-label classification settings. The first one is to identify the presence of concepts in normative ethics, such as *virtue, law, moral, justice, and utilitarianism* (**Ethics**$_1$). The second one is to evaluate the positive or negative implementation of these concepts with binary categories (**Ethics**$_2$).

The composition of the dataset is conducted in a semi-automatic mode. First, lists of keywords are formulated to identify the presence of ethical concepts (e.g., *"kill"*, *"give"*, *"create"*, etc.). The collection of keywords includes the automatic collection of synonyms using the semantic similarity tools of the RusVectores project (Kutuzov and Kuzmenko, 2017). After that, the news and fiction subcorpora of the Taiga corpus (Shavrina and Shapovalova, 2017) are filtered to extract short texts containing these keywords. Each text is annotated via Toloka as documented in Appendix E.1.

- **Text:** *"Pechen'kami sobstvennogo prigotovleniya nagradila 100-letnyaya Greta Plokh malysha, kotoryy pomog yey pereyti cherez ozhivlennoye shosse po peshekhodnomu perekhodu."* (100-year-old Greta Ploech gave handmade cookies to a toddler who helped her cross a busy highway at a pedestrian crossing.)
- **Labels$_1$:** ✓(Virtue) ✗(Law) ✗(Moral) ✓(Justice) ✓(Utilitarianism)
- **Labels$_2$:** ✓(Virtue) ✓(Law) ✓(Moral) ✓(Justice) ✓(Utilitarianism)

4 Design

4.1 Evaluation Principles

This section outlines our evaluation principles that are based on methodological recommendations and open research questions discussed by Bragg et al. (2021); Bowman and Dahl (2021); Beltagy et al. (2022), including sample size design, varying number of shots, reporting variability, diagnostic perfor-

mance analysis, and adversarial test sets. Figure 1 describes the TAPE's design.

Data Sampling. Each task in our benchmark consists of a training set $\mathcal{D}_{train}$ with labeled examples and a test set $\mathcal{D}_{test}$. The splits are randomly sampled, except for **RuOpenBookQA**, **RuWorldTree**, and **CheGeKa**, where we use the original splits. We keep the dataset size up to 1k and purposefully include imbalanced data for the text classification tasks.

No extra data. We do not provide validation sets nor any additional unlabeled data to test the zero-shot and few-shot generalization capabilities of LMs (Bao et al., 2019; Tam et al., 2021).

Number of shots. We consider $k \in \{1, 4, 8\}$ for few-shot evaluation to account for sensitivity to the number of shots. We also include zero-shot evaluation, which can be a strong baseline and simulate scenarios where no supervision is available.

Episode sampling. We provide 5 episodes in each k-shot setting $k \in \{1, 4, 8\}$ and report standard deviation over the episodes to estimate the variability due to the selection of demonstrations (Schick and Schütze, 2021). Each episode $\mathcal{E}^i = (\mathcal{E}^i_{train_k}, \mathcal{D}^A_{test})$ consists of k demonstration examples $\mathcal{E}^i_{train_k}$ randomly sampled from $\mathcal{D}_{train}$ with replacement, and a single test $\mathcal{D}^A_{test}$ acquired via the combination of original and adversarial test data.

Subpopulations. Subpopulations (Goel et al., 2021) are utilized for fine-grained performance analysis w.r.t. such properties of $\mathcal{D}_{test}$ as length, domain, and others.

Robustness. LMs are susceptible to adversarial examples, purposefully designed to force them output a wrong prediction given a modified input (Ebrahimi et al., 2018; Liang et al., 2018; Jia and Liang, 2017). We analyze the LMs' robustness to different types of adversarial data transformations. Here, each $\mathcal{E}^i_{train_k}$ corresponds to $T + 1$ test variations, including the original $\mathcal{D}_{test}$ and T adversarial test sets $\mathcal{D}^A_{test}$, acquired through the modification of $\mathcal{D}_{test}$. T depends on the dataset and can be adjusted based on the user's needs.

4.2 Adversarial Framework

4.2.1 Attacks and Perturbations

Table 1 summarizes the TAPE's adversarial attacks and perturbations based on the generally accepted typology (Zhang et al., 2020; Wang et al., 2021b).

Word-level Perturbations. Word-level perturbations utilize several strategies to perturb tokens, ranging from imitation of typos (Jin et al., 2020) to synonym replacement (Wei and Zou, 2019). We consider the following:

Spelling. BUTTERFINGERS is the typo-based perturbation that adds noise to data by mimicking spelling mistakes made by humans through character swaps based on their keyboard distance.

Modality. EMOJIFY replaces the input words with the corresponding emojis, preserving their original meaning. A few authors have manually validated translations of the English emoji dictionary.

Sentence-level Perturbations. In contrast to word-level perturbations, sentence-level perturbation techniques affect the syntactic structure:

Random. Easy Data Augmentation (EDA; Wei and Zou, 2019) have proved to be efficient in fooling LMs on text classification tasks. We use two EDA configurations: swapping words (EDA_SWAP) and deleting tokens (EDA_DELETE).

Paraphrasis. BACKTRANSLATION (Yaseen and Langer, 2021) allows to generate linguistic variations of the input without changing named entities. We use the OpusMT model[9] to translate the input text into English and back to Russian.

Distraction. ADDSENT is an adversarial attack that generates extra words or sentences with the help of a generative text model. We pass the input to the mGPT[10] LM and generate continuations with the sampling strategy. In the multiple-choice QA tasks, we replace one or more incorrect answers with their generated alternatives.

4.2.2 Data Curation

Adversarial perturbations and attacks are efficiently utilized to exploit weaknesses in LMs (Goel et al., 2021). At the same time, popular techniques may distort semantic meanings or generate invalid adversarial examples (Wang et al., 2021a). We aim to address this problem by using: *(i)* adversarial probability thresholds, *(ii)* task-specific constraints, and *(iii)* semantic filtering.

Probability thresholds. The degree of the input modification can be controlled with an adversarial probability threshold, which serves as the hyperparameter. The higher the probability, the more the input gets modified. The thresholds used in our experiments are specified in Table 1.

---

[9]hf.co/Helsinki-NLP/opus-mt
[10]hf.co/THUMT/mGPT

	Type	Name	Example	Adv. threshold	BERTScore
Word-l.	Spelling	BUTTERFINGERS	• This is a se m tence r o test t j e code	0.15	82.72
	Modality	EMOJIFY	• This is a sentence to test the 🏫	0.4	84.27
Sentence-level	Random	EDA$_{DELETE}$	• This a sentence to test the code	0.3	96.16
		EDA$_{SWAP}$	• code is a sentence to test This	0.3	93.99
	Paraphrasis	BACKTRANSLATION	• This sentence tests the code	0.5	95.38
	Distraction	ADDSENT	• This is a sentence to test the code , if you want to delete it	0.5	92.95

Table 1: Examples of the TAPE's adversarial attacks and perturbations. The examples are given for the English sentence "*This is a sentence used to test the code*" for illustration purposes. The similarity scores for each transformed sentence are given in percent. • – perturbations, • – adversarial attacks.

Constraints. The TAPE's attacks and perturbations do not drastically change the input's meaning. Despite this, we consider using rule-based constraints that keep the linguistic structure and task-specific aspects unchanged (see Table 4 in Appendix D). For instance, it is crucial to leave named entities in the QA tasks untouched or not modify the syntactic structure and anaphors when perturbing the **Winograd** examples.

Semantic filtering. We follow Wang et al. on filtering the adversarial examples with BERTScore[11] (Zhang et al., 2019), a BERT-based text similarity metric (Devlin et al., 2019). We measure the semantic similarity between the original input and adversarial output and keep examples with the highest similarity score. In cases when the score is lower than a specified threshold, we iteratively decrease the adversarial probability threshold and re-score the new adversarial examples.

5 Baselines

5.1 Non-neural Baselines

We use two models from the scikit-learn library (Pedregosa et al., 2011) as non-neural baselines for classification (**Ethics$_{1/2}$** and **Winograd**) and multiple-choice QA tasks (**RuOpenBookQA** and **RuWorldTree**). The baselines are fit on the corresponding $\mathcal{D}_{train}$ and evaluated on $\mathcal{D}_{test}$.

Random is a simple data-agnostic baseline that samples predictions uniformly from the set of target classes in a given task.

Linear is a logistic regression classifier over TF-IDF (Salton and Yang, 1973) N-grams within the range $N \in [1; 4]$. The classifier is trained on top-150k features with default L2-regularization hyperparameters.

5.2 Neural Baselines

We run zero-shot and few-shot evaluation of Russian GPT3[12] LMs available under the HuggingFace library (Wolf et al., 2020). We consider three model versions: ruGPT3$_S$[13] (125M), ruGPT3$_M$[14] (350M), and ruGPT3$_L$[15] (760M).

Perplexity-based evaluation. We consider the setting where the classification and multiple-choice tasks are formulated in natural language as a cloze-style prompt template: **Winograd**, **Ethics$_{1/2}$**, **RuWorldTree** and **RuOpenBookQA**. We provide examples of the prompt templates for each task in Appendix C. After filling in each possible target class or choice, we compute the per-token cross-entropy loss, which is reduced to negative log-probability due to one-hot encoding of the target tokens. The most probable string has the lowest sum of negative log probabilities of its tokens normalized over the total number of tokens in the input, as specified in Equation 1.

$$\text{PPL}(t) = \exp(-\frac{1}{|t|}\sum_{i=0}^{|t|} log_{p_\theta}(x_i|x_{<i})), \quad (1)$$

where t is the input prompt and $|t|$ is the length of the prompt in tokens. The choice relies on our preliminary experiments, where instead, the most probable string is chosen based on the lowest sum of negative log probabilities of the prompt's tokens. However, the latter approach has shown worse results on the subsets of the training sets.

Zero-shot and few-shot generation. Text generative baselines are of the greatest interest for tasks that can not be solved by the perplexity-based approach: **CheGeKA** and **MultiQ**. Here,

---

[11]hf.co/bert-base-multilingual-cased

[12]github.com/ai-forever/ru-gpts
[13]hf.co/sberbank-ai/rugpt3small
[14]hf.co/sberbank-ai/rugpt3medium
[15]hf.co/sberbank-ai/rugpt3large

	Dataset	BF	EMJ	EDA	BT	AS	Subpopulations	$|\mathcal{D}_{train}|$	$|\mathcal{D}_{test}|$	Metrics	Domain
Classification	Winograd	✓	✓	✓	✗	✓	● Gender, Number ● NumCandidates ● Class distribution ● Length	60	986	F1 / Acc.	Misc.
	Ethics$_{1/2}$	✓	✓	✓	✓	✓	● Domain ● Length ● Diversity, Readability	59/58	1k/1k	F1 / Acc.	News, fiction books
Question Answering	RuWorldTree	✓	✓	✓	✓	✓	● Answer Category ● ExamName,SchoolGrade ● Length ● Diversity, Readability	47	629	F1 / Acc.	Elementary science exams
	RuOpenBookQA	✓	✓	✓	✓	✓	● Answer Category ● Length ● Diversity, Readability	48	500	F1 / Acc.	Elementary science exams
	MultiQ	✓	✓	✓	✓	✓	● Answer Category ● Length ● Diversity, Readability	64	1k	F1 / EM	Wikipedia
	CheGeKa	✓	✓	✓	✓	✓	● Answer Category ● Length ● Diversity, Readability	49	520	F1 / EM	General domain

Table 2: Summary of the TAPE benchmark. Transformations: BF – BUTTERFINGERS, EMJ – EMOJIFY, AS – ADDSENT, EDA includes EDA$_{SWAP}$ and EDA$_{DELETE}$. Subpopulations: ● – Morphology, ● – Class Distribution, ● – Domain, ● – Answer Category, ● – Length, ● – Number of Candidates, ● – Text Statistics.

we generate the answer given the corresponding task prompt (see Appendix C) with nucleus sampling ($top_p = 0.8$). The choice of the strategy and hyperparameters is based on the grid search[16] on a subset of the corresponding $\mathcal{D}_{train}$. The output is limited by 100/200 (**CheGeKA**) and 400/800 (**MultiQ**) tokens in zero-shot/few-shot settings, respectively.

5.3 Human Baselines

The human evaluation is run via Toloka. Access to the annotation projects is granted to annotators certified as Russian native speakers. Each project consists of an unpaid training stage, control examples for monitoring annotation quality[17], and the main annotation stage. The annotator is given detailed instruction with a task description, annotation examples, and corresponding explanations. The instruction is linked to training and main annotation stages and available any time. Annotators who get less than 60% of the training examples cor-

rect on average do not qualify for the main stage. Each qualified annotator receives a page with a certain number of examples for annotation, one of which is a control one. The inter-annotator agreement (IAA) is based on the Dawid-Skene aggregation model[18] (Dawid and Skene, 1979), available directly from Toloka. Details on the annotation process, inter-annotator agreement rates, hourly pay rate, average response time, the overall project cost, annotation instructions, and examples of web interface are fully documented in Appendix E.

5.4 Metrics

We evaluate the baseline performance with macro-averaged F1 and accuracy scores for the classification (**Winograd**, **Ethics**$_{1/2}$) and multiple-choice QA tasks (**RuOpenBookQA**, **RuWorldTree**). F1-score and exact match (EM) are used for the open-domain (**CheGeKa**) and multi-hop QA tasks (**MultiQ**) – obtaining such metrics for generative and sequence-to-sequence tasks provides a comparable yet strict setup. The effectiveness of the perturbations and adversarial attacks is measured with the attack success rate (ASR; Wang et al., 2021a) computed as the percentage of the correct

---

[16]Beam search: number of beams $\{1, 2, 5, 10, 50\}$; Nucleus sampling: $top_p \in \{0.5, 0.8, 0.9, 0.95 - 0.99\}$. We also experimented with top-k sampling and greedy decoding.

[17]Control examples are commonly used on Toloka for filtering out results from bots, cheaters, and low-performing annotators. The examples are manually selected by a few authors from the $\mathcal{D}_{train}$ and guaranteed to be unambiguous.

[18]toloka.ai/docs/result-aggregation

Model	k-shot	Winograd	RuWorldTree	RuOpenBookQA	MultiQ	CheGeKa	Ethics$_1$	Ethics$_2$	Avg
ruGPT3$_S$	0	38.6 / 57.9	34.1 / 34.0	33.9 / 34.0	0.17 / 0.0	<u>1.0</u> / 0.0	34.1 / 55.5	48.5 / 60.9	27.24 / 34.64
	1	36.7 ±0.0 / 58.0 ±0.0	33.5 ±0.4 / 33.5 ±0.4	34.9 ±0.8 / 35.0 ±0.8	0.13 ±0.04 / 0.0 ±0.0	0.52 ±0.14 / 0.0 ±0.0	14.3 ±11.1 / 59.9 ±5.8	49.6 ±7.1 / 60.8 ±5.7	24.24 / 35.33
	4	41.0 ±5.3 / 54.8 ±3.9	34.3 ±0.9 / 34.3 ±0.9	34.6 ±0.6 / 34.6 ±0.6	0.09 ±0.06 / 0.0 ±0.0	0.62 ±0.34 / 0.65 ±0.29	23.5 ±13.8 / 54.3 ±12.3	56.7 ±10.8 / 61.3 ±5.0	27.44 / 34.31
	8	36.9 ±0.4 / 54.8 ±6.5	35.1 ±0.9 / 35.1 ±0.8	33.6 ±0.5 / 33.7 ±1.5	0.12 ±0.22 / 0.0 ±0.0	0.35 ±0.26 / 0.35 ±0.26	32.4 ±6.9 / 50.4 ±4.0	<u>63.7</u> ±9.5 / 61.7 ±4.2	<u>29.04</u> / 33.56
ruGPT3$_M$	0	39.5 / 57.2	38.0 / 38.0	34.8 / 34.8	<u>0.18</u> / 0.0	0.75 / 0.0	2.4 / 68.3	5.3 / 44.1	17.29 / 34.71
	1	36.7 ±0.0 / 58.0 ±0.0	36.3 ±0.6 / 36.3 ±0.6	36.7 ±0.8 / 36.8 ±0.8	0.11 ±0.03 / 0.0 ±0.0	0.47 ±0.11 / 0.0 ±0.0	14.3 ±11.1 / 60.8 ±5.7	49.6 ±7.1 / 60.8 ±5.7	24.9 / 35.99
	4	39.7 ±4.1 / 57.9 ±0.6	38.9 ±1.0 / 38.8 ±1.0	36.3 ±1.3 / 36.4 ±1.3	0.12 ±0.18 / 0.0 ±0.0	0.85 ±0.52 / <u>0.92</u> ±0.53	14.9 ±16.0 / 58.6 ±13.1	45.5 ±13.7 / 59.0 ±7.3	25.57 / 35.74
	8	39.0 ±5.9 / 54.3 ±6.5	40.1 ±0.5 / 40.1 ±0.5	35.7 ±0.9 / 35.8 ±1.0	0.1 ±0.15 / 0.0 ±0.0	0.42 ±0.19 / 0.42 ±0.19	19.6 ±10.6 / 60.1 ±5.1	57.7 ±9.0 / 64.2 ±2.1	27.78 / 36.83
ruGPT3$_L$	0	39.2 / 55.5	40.8 / 40.7	39.6 / 39.6	0.14 / 0.0	0.57 / 0.0	0.3 / <u>68.6</u>	8.0 / 44.9	18.15 / 35.73
	1	36.7 ±0.0 / 58.0 ±0.0	38.2 ±0.8 / 38.2 ±0.8	<u>39.8</u> ±1.0 / <u>39.8</u> ±1.1	0.12 ±0.04 / 0.0 ±0.0	0.47 ±0.10 / 0.0 ±0.0	14.3 ±11.1 / 59.9 ±5.8	49.6 ±7.1 / 60.8 ±5.7	25.6 / 36.69
	4	38.8 ±2.9 / <u>58.4</u> ±0.5	40.9 ±0.4 / 40.8 ±0.8	39.6 ±0.4 / 39.6 ±0.8	0.1 ±0.14 / 0.0 ±0.0	0.54 ±0.44 / 0.54 ±0.4	11.8 ±12.4 / 60.4 ±1.1	44.5 ±15.9 / 59.6 ±8.1	25.1 / 36.96
	8	39.2 ±3.2 / 57.6 ±1.5	<u>41.9</u> ±0.7 / <u>41.9</u> ±0.7	38.8 ±1.5 / 38.9 ±1.5	0.0 ±0.0 / 0.0 ±0.0	0.54 ±0.4 / 0.54 ±0.4	23.3 ±8.3 / 58.1 ±3.4	52.0 ±8.5 / 63.7 ±2.4	28.33 / <u>37.12</u>
Random	✗	<u>51.7</u> / 51.3	26.0 / 26.1	21.4 / 21.6	✗	✗	<u>36.5</u> / 47.9	54.1 / 52.0	✗
Linear	✗	36.7 / 58.0	14.3 / 25.8	11.4 / 25.2	✗	✗	0.0 / <u>68.6</u>	49.0 / <u>65.3</u>	✗
Human	✗	**86.6 / 87.0**	**83.8 / 83.7**	**87.5 / 86.5**	**92.8 / 91.0**	**71.9 / 64.5**	**51.0 / 52.9**	**73.0 / 67.6**	**78.09 / 76.17**

Table 3: Performance results of the non-neural, neural, and human baselines on the original test sets. Metrics: F1-score/accuracy (EM). The best score is put in bold, the second best is underlined.

predictions that are changed after the perturbation or attack is applied.

6 Results

6.1 Generalization Evaluation

Table 3 presents the zero-shot and few-shot performance results of the non-neural, neural, and human baselines on the original $\mathcal{D}_{test}$ sets.

Classification and multiple-choice tasks. The zero-shot evaluation provides a strong baseline, matching or exceeding the few-shot performance on **Winograd** ($k \in \{1, 4\}$) and **Ethics**$_1$ ($k \in \{1, 4, 8\}$). The zero-shot performance is similar among the models despite their size (**Winograd**), or it can steadily improve (**RuWorldTree**, **RuOpenBookQA**) and significantly drop when the model size increases (**Ethics**$_{1/2}$). We observe that introducing more examples increases variability on the imbalanced classification tasks (**Winograd**, **Ethics**$_{1/2}$) and leads to performance degradation, specifically for ruGPT$_S$. Furthermore, the performance degenerates into constant predictions, which is indicated by the significant difference in accuracy and F1 scores on the (**Winograd**) and (**Ethics**$_{1/2}$) tasks. In particular, the LMs predict the negative label for about 97% of the **Winograd** samples in the zero-shot setting. In the few-shot, however, the number of constant predictions is reduced to 80% ($k \in \{4, 8\}$). This result indicates that the demonstrations may help generalize to the task, but the predictions are still affected by the imbalanced classification setting. We also observe that **Ethics**$_{1/2}$ is the most challenging task for both human and neural baselines. The results are sensitive to prompt changing, and human annotators may receive low inter-annotator agreement on the examples due to subjectivity.

Zero-shot and few-shot generation results. However, approaching the **CheGeKa** and **MultiQ** tasks with zero-shot and few-shot generation results in near-zero generalization performance. Both generative tasks demonstrate the most significant difference between human evaluation and baseline results, which can be explained, on the one hand, by the lack of answer choices, on the other, by the limitations of standard QA metrics for assessing semantically correct but non-literal generated answers. To better understand this, we manually analyzed a sample of 100 predictions per task and found that the generated outputs rarely match the golden answers, e.g., the models generate irrelevant texts or texts related to the question.

Discussion. The neural baselines are capable to generalize to multiple-choice QA tasks well but perform worse than random baseline or blindly predict the target labels on the imbalanced classification tasks. Our results are consistent with Lin et al. (2021) in that: *(i)* the few-shot evaluation results may rely heavily on the input prompts, *(ii)* it is difficult for GPT-style LMs to perform judgments on social value tasks in the zero-shot and few-shot settings, and *(iii)* the few-shot results on some tasks are worse than zero-shot, meaning that LMs are not able to utilize given demonstrations for solving them. We also observe the negative result: zero-shot and few-shot generation baselines may receive near-zero performance on the open-domain and extractive QA tasks.

6.2 Robustness

Table 4 shows the ASR scores for each perturbation and k-shot setting averaged over the **RuWorldTree** and **RuOpenBookQA** tasks, where the model performance exceeds the random base-

Model	k-shot	BF	EMJ	EDA$_{\text{SWAP}}$	EDA$_{\text{DEL}}$	BT	AS	Avg
ruGPT3$_\text{S}$	0	53.0	37.3	<u>33.6</u>	37.9	22.3	18.5	33.77
	1	57.55	42.1	37.3	44.7	**19.4**	12.05	35.52
	4	58.8	35.85	39.8	47.5	<u>19.5</u>	8.85	35.05
	8	60.45	36.65	43.45	50.25	20.0	7.2	36.33
ruGPT3$_\text{M}$	0	**50.3**	38.9	**33.5**	**41.0**	20.75	14.25	**33.12**
	1	57.4	35.25	36.05	43.55	20.7	7.55	33.41
	4	60.1	34.55	37.8	46.0	20.35	5.45	34.04
	8	61.35	34.85	38.95	48.2	21.55	<u>3.55</u>	34.74
ruGPT3$_\text{L}$	0	<u>52.1</u>	36.2	35.8	<u>43.3</u>	25.25	12.65	34.21
	1	55.3	**33.1**	37.15	43.8	22.25	7.7	<u>33.22</u>
	4	59.15	34.95	39.05	46.3	21.85	7.9	34.87
	8	62.0	<u>34.4</u>	41.65	47.6	21.8	**2.7**	35.03

Table 4: The robustness evaluation results by adversarial perturbation and attack. The ASR values are averaged over the **RuOpenBookQA** and **RuWorldTree** tasks. The lower, the better. The best ASR value is put in bold and the second best is underlined.

line. We observe that the models are more susceptible to simple spelling-based perturbations (BUTTERFINGERS), token deletion (EDA$_{\text{DELETE}}$), and modality changes (EMOJIFY). Paraphrasing the input (BACKTRANSLATION) and, especially, addition of the distractors (ADDSENT), in turn, bear significantly less effect. Most of the perturbations either gain effectiveness as the number of demonstrations increases (BUTTERFINGERS, EDA$_{\text{DELETE}}$, EDA$_{\text{SWAP}}$) or does not seem to directly depend on the k value (EMJ, BACKTRANSLATION). However, the ASR scores for the distraction adversarial attack (ADDSENT) are inversely proportional to the number of shots. In other words, models are less likely to fall for the generated answer option as the number of demonstrations increases. The model size is not the main factor in model robustness, and the difference between the models is quite subtle. We notice, however, that larger models (ruGPT3$_\text{M}$, ruGPT3$_\text{L}$) are less affected by the ADDSENT perturbation and more vulnerable to BACKTRANSLATION, in comparison to ru-GPT3$_\text{S}$. Nonetheless, the latter proves to be less robust to adversarial perturbations, which is indicated by the higher ASR scores on average, closely followed by the largest model.

Discussion. The neural baselines are most vulnerable to word-level text perturbations: typo-based transformations and modality changes. In terms of the sentence-level perturbations, simple EDA techniques prove to be more effective than paraphrasis and the distraction-based adversarial attack, which can be due to the task-specificity. This is in line with Wang et al. (2021a), who observe typo-based adversarial perturbations as the most effective automatic attack methods.

6.3 Diagnostic Analysis

Subpopulation analysis reveals that the GPT-style models exhibit length bias, which is indicated by low performance on longer inputs (see Figure 1-Figure 4 in Appendix F). The effect is observed primarily on the **RuWorldTree** task, where the models' performance significantly drops on longer texts. As might be expected, for QA tasks, question complexity, determined by school grade or exam name, also affects the model performance. The models tend to better deal with easier questions, which becomes more prominent as the number of demonstrations increases. Seemingly, the readability and lexical diversity of a question, determined by Flesch Reading Ease scores and type-token ratio, respectively, do not affect model performance. Nevertheless, slight increases in performance on more readable and diverse texts are present.

Discussion. We reveal that the baseline performance depends on the input length. One of the reasons for such behavior can be the limited context window that the models have. Alex et al. (2021) have previously explored reasoning over long texts in a few-shot setting and their results are consistent with our findings.

7 Conclusion and Future Work

Zero-shot and few-shot methods have evolved as a new paradigm in NLP. Addressing the best practices, we introduced TAPE, a text attack and perturbation evaluation benchmark for Russian. TAPE combines the general language understanding evaluation techniques with the greener no-tuning approach, allowing the evaluation of LMs' robustness on complex intellectual tasks. We present six new datasets and a framework for generating adversarial attacks and perturbations, which can also be used as a standalone tool for practical purposes.

In future, we plan to incorporate more LMs with various architectures and prompting-based methods into the framework. Another direction is to evaluate the cross-lingual generalization capabilities of autoregressive LMs. We hope to encourage the community to foster evaluation of LMs' generalization capacity in non-English languages, leading to the development of more robust and reliable LMs.

8 Limitations

Performance aggregation. The well-established GLUE-style benchmarks evaluate systems using mean aggregation over heterogeneous task-specific

metrics (Wang et al., 2018, 2019, 2021a). Based on the criticism of this evaluation protocol by the research community (e.g., Waseem et al., 2021; Mishra and Arunkumar, 2021; Agarwal et al., 2021), we recognize that mean aggregation in our case does not account for the nature of the adversarial transformations and attacks and task specifications, such as the task type, domain, and the number of episodes in $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.

Baseline evaluation. First, our baseline model evaluation relies on using the same prompts for all language models unless mentioned otherwise. Second, we do not utilize related few-shot learning and prompt-tuning methods, which could serve as more solid baseline approaches. We recognize that it can lead to biased evaluation and spurious conclusions about the baseline performance. However, we aim to provide a scope of baseline solutions, ranging from perplexity-based to zero-shot open-ended generation approaches. At the same time, our training sets are publicly available, and it is not anticipated that the users will apply this data for fine-tuning.

Human performance. The comparison of our neural and human baselines is inconsistent regarding the number of demonstrations provided to understand a given task. The zero-shot and few-shot human performance can be comparable to neural LMs' performance when humans would receive $k \in \{0, 1, 4, 8\}$ examples in the annotation training stage (Mukherjee et al., 2021).

9 Ethics Statement

Subjectivity related to ethics. Ethics is a multi-dimensional subject, which remains a complicated problem for LMs and controversial for humans in a multitude of situations. Although our methodology spans general concepts in normative ethics, we acknowledge that it can be challenging to perform *objective* ethical judgments about some situations (Martineau, 2006). For instance, judgments about *law* are based on formal criteria (e.g., the criminal code), *morality* may rely on public sentiment, while *justice* may heavily rely on private sentiment and human worldview. At the same time, the real-life situations described in a given text are imbalanced concerning the number of acts annotated as positive and the number of acts with various disadvantages in terms of the ethical norms. In practice, this leads to moderate inter-annotator agreement and approximate human and model performance estimates.

Risks related to ethics. We acknowledge that approaches to evaluating LMs' ability to perform ethical judgments about text situations have been criticized (Talat et al., 2022). While we use a similar set of ethical concepts (Hendrycks et al., 2021), we collect annotations according to the five criteria that describe the aspects of the annotators' attitude towards the deed. The attitude can be determined by various individual and social aspects. Here, we have analyzed metadata of our Ethics$_1$ annotators available via the Toloka interface. There are 481 Russian speakers across 16 different countries, who can be grouped by their age as follows: $18 - 30$ (163 annotators); $30 - 50$ (265 annotators), and $50 - 78$ (53 annotators). Thus, we will further take into account specific risks arising within the annotation process:

Social properties: the diffusion of norms in the Russian-speaking communities has been the object of rapid changes (Casier, 2022). This can be expressed in a shift in attitude towards actions that have different interpretations from the point of view of regional cultural norms, cultures of small peoples, religious norms, and normative behavior for classes of society.

Legal properties: as the "legality" of a deed in a text can change over time, we are sure to see a growing annotation inconsistency in individual examples that reflect societal changes after some years. The risks are partially mitigated by the prior training of the annotators and annotator's performance control. Running the annotation experiments from year to year is reasonable to understand possible norm shifts, measuring the variation in annotators' opinions about aspects of the described actions. Furthermore, other data-dependent risks can be indicated, such as genre bias and author's bias in specific publicly available text sources.

Societal impact. The TAPE's design allows us to alleviate the problems of a large carbon footprint (Bender et al., 2021) and keep computational costs accessible to academic and industrial fields (Couldry and Mejias, 2020). In particular, our evaluation approach does not consider LMs' fine-tuning and relies on a limited amount of episodes, while the number of attacks and perturbations can be adjusted based on the user's needs.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021.

Deep Reinforcement Learning at the Edge of the Statistical Precipice. *Advances in Neural Information Processing Systems*, 34.

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, et al. 2021. RAFT: A Real-World Few-Shot Text Classification Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2020a. Learning to few-shot learn across diverse natural language classification tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5108–5123, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020b. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2019. Few-shot Text Classification with Distributional Signatures. In *International Conference on Learning Representations*.

Adriano Barbosa-Silva, Simon Ott, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *arXiv preprint arXiv:2203.04592*.

Iz Beltagy, Arman Cohan, Robert Logan IV, Sewon Min, and Sameer Singh. 2022. Zero- and few-shot NLP with pretrained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 32–37, Dublin, Ireland. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. FLEX: Unifying Evaluation for Few-Shot NLP. *Advances in Neural Information Processing Systems*, 34:15787–15800.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Casier. 2022. Russia and the diffusion of political norms: the perfect rival? *Democratization*, 29(3):433–450.

Kenneth Ward Church and Valia Kordoni. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering*, 28(2):249–269.

Nick Couldry and Ulises A Mejias. 2020. The Costs of Connection: How Data Are Colonizing Human Life and Appropriating It for Capitalism.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – russian reading comprehension dataset: Description and analysis. In *Lecture Notes in Computer Science*, pages 3–15. Springer International Publishing.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Alena Fenogenova. 2021. Russian paraphrasers: Paraphrase with transformers. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 11–19, Kiyv, Ukraine. Association for Computational Linguistics.

Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. 2020. Read and reason with MuSeRC and RuCoS: Datasets for machine reading comprehension for Russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6481–6497, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Taisia Glushkova, Alexey Machnev, Alena Fenogenova, Tatiana Shavrina, Ekaterina Artemova, and Dmitry I Ignatov. 2020. DaNetQA: a yes/no Question Answering Dataset for the Russian Language. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 57–68. Springer.

Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the NLP evaluation landscape. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch Critch, Jerry Li Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. In *International Conference on Learning Representations*.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT Really Robust? A Strong Baseline For Natural Language Attack On Text Classification And Entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Andrey Kutuzov and Elizaveta Kuzmenko. 2017. *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*, pages 155–161. Springer International Publishing, Cham.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4208–4215. International Joint Conferences on Artificial Intelligence Organization.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv preprint arXiv:2107.13586*.

James Martineau. 2006. *Types of ethical theory*. Cosimo, Inc.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Elena Mikhalkova. 2021. A russian jeopardy! data set for question-answering systems.

Swaroop Mishra and Anjana Arunkumar. 2021. How Robust are Model Rankings: A Leaderboard Customization Approach for Equitable Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13561–13569.

Subhabrata Mukherjee, Xiaodong Liu, Guoqing Zheng, Saghar Hosseini, Hao Cheng, Greg Yang, Christopher Meek, Ahmed Hassan Awadallah, and Jianfeng Gao. 2021. Few-Shot Learning Evaluation in Natural Language Understanding. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel,

Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *the Journal of machine Learning research*, 12:2825–2830.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models. *Advances in Neural Information Processing Systems*, 34:11054–11070.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ivan Rybin, Vladislav Korablinov, Pavel Efimov, and Pavel Braslavski. 2021. Rubq 2.0: an innovated russian question answering dataset. In *European Semantic Web Conference*, pages 532–547. Springer.

Gerard Salton and C. S. Yang. 1973. On the Specification of Term Values in Automatic Indexing. *Journal of Documentation*, 29:351–372.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:«taiga» syntax tree corpus and parser. *Proceedings of the "Corpora*, pages 78–84.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779, Seattle, United States. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to Grow a Mind: Statistics, Structure, and Abstraction. *science*, 331(6022):1279–1285.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021a. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2021b. Measure and Improve Robustness in NLP Models: A Survey. *arXiv preprint arXiv:2112.08313*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks. *arXiv preprint arXiv:2204.07705*.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2021. Disembodied Machine Learning: On the Illusion of Objectivity in NLP. *arXiv preprint arXiv:2101.11974*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual*

Representation Learning, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Liang Xu, Xiaojing Lu, Chenyang Yuan, Xuanwei Zhang, Huilin Xu, Hu Yuan, Guoao Wei, Xiang Pan, Xin Tian, Libo Qin, et al. 2021. FewCLUE: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

Usama Yaseen and Stefan Langer. 2021. Data Augmentation for Low-Resource Named Entity Recognition Using Backtranslation. *CoRR*, abs/2108.11703.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Li Jian, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2022. FewNLU: Benchmarking state-of-the-art methods for few-shot natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 501–516, Dublin, Ireland. Association for Computational Linguistics.

A General Dataset Statistics

	Size	N_T	N_U	Label Distribution
Winograd				
Train	804	14547	6417	66.3/33.7
Test	976	20933	9559	58.0/42.0
Ethics$_1$				
Train	254	6978	63769	31.9/39.0/44.9/5.9/38.2
Test	1000	195247	4091	31.0/34.8/36.8/15.3/39.0
Ethics$_2$				
Train	259	32845	14368	69.1/65.3/78.4/40.9/23.9
Test	1000	194097	63091	64.7/63.5/78.9/53.0/27.9
RuWorldTree				
Train	118	3659	1881	26.3/22.9/28.8/22.0
Test	629	21931	7406	22.1/27.5/25.6/24.8
RuOpenBookQA				
Train	2339	53553	15556	31.4/23.6/21.8/23.2
Test	500	1078	660	25.2/27.6/22.0/25.2
MultiQ				
Train	1056	166570	29917	✗
Test	1000	209509	30083	✗
CheGeKa				
Train	29376	419744	112606	✗
Test	520	13893	7620	✗

Table 1: General statistics for each dataset. N_T refers to the total number of tokens. N_U denotes the number of unique tokens. Label distribution by target class is presented in %. We report the distribution of the positive class for each category in Ethics$_{1/2}$.

B Winograd Queries

This appendix provides the list of queries that correspond to the RusCorpora query language[19] and examples in natural language.

- Type 1: Noun phrase & subordinate clause with "that" in the same gender and number.
1. *"Bulochka iz pekarni, kotoraya..."* (A bun from a bakery that...)
    - S & nom & sg & f
    - at the distance of 0 to 1 from
    - at the distance of 0 to 1 from S & (gen | gen2) & sg & f
    - at the distance of 0 to 1 from which, sg & f
2. *"Istoriya o zhenshchine, kotoraya.."* (A story about a woman that...)
    - S & nom & sg & f
    - at the distance of 0 to 1 from about
    - at the distance of 0 to 1 from loc & sg & f

3. *"Kameya bez opravy, kotoraya..."* (Rimless Cameo that...)
    - S & nom & sg & f
    - at the distance of 0 to 1 from without
    - at the distance of 0 to 1 from gen & sg & f
    - at the distance of 0 to 1 from that
    - (nom | dat | acc | ins | loc) & sg & f
4. *"Smena obstanovki, kotoraya..."* (A change of scenery that...)
    - S & nom & sg & f
    - at the distance of 0 to 1 from gen & sg & f
    - at the distance of 0 to 1 from gen & sg & f
    - at the distance of 0 to 1 from that
    - (nom | dat | acc | ins | loc) & sg & f
5. *"Gruppy studentov, kotoryye..."* (Groups of students that...)
    - S & nom & pl
    - at the distance of 0 to 1 from
    - from | of| without at the distance of 1 from S & gen & pl
    - at the distance of 0 to 1 from that
    - at the distance of 1 from that
    - (nom | dat | acc | acc2 | ins | loc) & pl
- Type 2: Coincidence of nominative and accusative forms in the masculine gender.
1. *"Bul'var ukrashayet gorod"* (The boulevard decorates the city / The boulevard is decorated by the city)
    - S & (nom| acc) & sg & m & inan
    - at the distance of 1 from V & indic & (praes | praet) & tran & sg & m
    - at the distance of 1 from S & (nom | acc) & sg & m & inan
2. *"Flagi ukrashayut goroda"* (Flags adorn the cities / Flags are adorned by the cities)
    - S & (nom | acc) & pl & m
    - at the distance of 1 from V & indic & (praes | praet) & tran & pl
    - at the distance of 1 from S & (nom | acc) & pl & m
- Type 3: Coincidence of genitive and possessiveness.
1. *"Ikh deti razdrazhali"* (Their kids were annoying / They were annoyed by kids)
    - they, (gen | gen2) & pl
    - at the distance of -1 to 1 from S & nom & pl
    - at the distance of 0 to 1 from V & indic & (praes | praet) & tran & pl
- Type 4: Two possible references to a pronoun.
1. *"Katya sprosila Mashu, delala li ona..."* (Katya

asked Masha if she...)
- – S & nom & f & anim
- – at the distance of 1 to 4 from S & (gen | dat | acc | ins | loc) & sg & f & anim
- – at the distance of 1 from she, (nom | gen | dat | acc | ins | loc) & sg & f

2. *"Ivan sprosil Petra, delal li on..."* (Ivan asked Peter if he...)
   - – S & nom & m & anim
   - – at the distance of 1 to 4 from S & (gen | dat | acc | ins | loc) & sg & m & anim
   - – at the distance of 1 from he, (nom | gen | dat | acc | ins | loc) & sg & m

3. *"Uchitelya sprashivayut uchenikov, delali li oni..."* (Teachers ask students if they...)
   - – S & nom & pl & anim
   - – at the distance of 1 to 4 from S & (gen | dat | acc | ins | loc) & pl & anim
   - – at the distance of 1 to 2 from them, (nom | gen | dat | acc | ins | loc) & pl

C Prompt formats

We design the prompt templates based on the task specifics and format (see Table 2, Table 3). The choice of the prompts is based on the preliminary experiments on the corresponding training set and manual analysis of the results.

- **Winograd**: we use "yes" and "no" label encoding.
- **RuOpenBookQA** and **RuWorldTree**: we unite the question or the sentence prefix with each of the possible choices.
- **Ethics$_{1/2}$**: we regard each category as a separate binary target, which we encode as "yes" or "no" and, therefore, use different prompts for each category. We manually crafted a large pool of templates and selected between 1 and 3 best prompts for each target, which yields the best F1-score on a subset of the training set.
- **MultiQ** and **CheGeKa**: we use generative baselines and format the prompts so that the LMs better capture the task.

Task	Template	Output
Winograd	V predlozhenii {CONTEXT}\n Slovo {REFERENCE} otnositsya k slovu {CANDIDATE ANSWER}? {LABEL}. (In the sentence {CONTEXT}\n The word {REFERENCE} refers to the word {CANDIDATE ANSWER}? {LABEL}.) V predlozhenii "{CONTEXT}"\n Slovo "{REFERENCE}" otnositsya k slovu "{CANDIDATE ANSWER}" {LABEL}. (In the sentence "{CONTEXT}"\n The word "{REFERENCE}" refers to the word "{CANDIDATE ANSWER}"? {LABEL}.) V predlozhenii "{CONTEXT}"\n Slovo "{REFERENCE}" otnositsya k slovu "{CANDIDATE ANSWER}" ili net? {LABEL}. (In the sentence "{CONTEXT}"\n The word "{REFERENCE}" refers to the word "{CANDIDATE ANSWER}" or not? {LABEL}.) "{CONTEXT}"\n Slovo "{REFERENCE}" otnositsya k slovu "{CANDIDATE ANSWER}" ili net? {LABEL}. ("{CONTEXT}"\n The word "{REFERENCE}" refers to the word "{CANDIDATE ANSWER}" or not? {LABEL}.) "{CONTEXT}"\n Slovo "{REFERENCE}" otnositsya k slovu "{CANDIDATE ANSWER}"? {LABEL}. ("{CONTEXT}"\n The word "{REFERENCE}" refers to the word "{CANDIDATE ANSWER}"? {LABEL}.)	Yes; No
RuOpenBookQA/ RuWorldTree	{QUESTION} {CANDIDATE ANSWER}	A; B; C; D
MultiQ	Tekst: {MAIN TEXT} Vopros: {QUESTION} Otvet: (Text: {MAIN TEXT} Question: {QUESTION} Answer:)	Generated answer
CheGeKa	ChGK. Tema: {CATEGORY} Vopros: {QUESTION} Otvet: ChGK. Topic: {CATEGORY} Question: {QUESTION} Answer:	Generated answer

Table 2: Prompt examples for the classification and QA tasks. The examples are translated into English for illustration purposes.

Target	Template
Ethics₁	
Virtue	Tekst: {TEXT}\n Vopros: upominayetsya li v etom tekste dobro ili zlo? Otvet: {LABEL}. (Text: {TEXT}\n Question: does this text mention good or evil? Answer: {LABEL}.) {TEXT}\n Kasayetsya li etot tekst i proiskhodyashcheye v nem ch'kh-to dobrykh/zlykh namereniy? Otvet: {LABEL}. ({TEXT}\n Does this text and what is happening in it refer to someone's good/evil intentions? Answer: {LABEL}.)
Law	Tekst: {TEXT}\n Kasayetsya li etot tekst zakona, rasporyadka ili tseremoniala? Otvet: {LABEL}. (Text: {TEXT}\n Does this text deal with law, regulation, or ceremonial? Answer: {LABEL}.)
Moral	{TEXT}\n Upominayutsya li v etom tekste voprosy morali? Otvet: {LABEL}. (Text: {TEXT}\n Does this text mention moral issues? Answer: {LABEL}.)
Justice	Tekst: {TEXT}\n Vopros: upominayetsya li v etom tekste karma ili torzhestvo spravedlivosti? Otvet: {LABEL}. (Text: {TEXT}\n Question: does this text mention karma or the triumph of justice? Answer: {LABEL}.) Tekst: {TEXT}\n Vopros: kasayetsya li etot tekst karmy ili torzhestva spravedlivosti? Otvet: {LABEL}. (Text: {TEXT}\n Question: does this text refer to karma or the triumph of justice? Answer: {LABEL}.)
Utilitarianism	Tekst: {TEXT}\n Vopros: kasayetsya li etot tekst i proiskhodyashcheye v nem priobriteniy ili poter' (kak material'nykh, tak i emotsional'nykh)? Otvet: {LABEL}. (Text: {TEXT}\n Question: does this text and what happens in it deal with gains or losses (both material and emotional)? Answer: {LABEL}.) {TEXT}\n Kasayetsya li etot tekst priobriteniy ili poter'? Otvet: {LABEL}. ({TEXT}\n Does this text refer to gains or losses? Answer: {LABEL}.)
Ethics₂	
Virtue	Tekst: {TEXT}\n Vopros: geroi etogo teksta proyavlyayut svoi luchshiye kachestva? Otvet: {LABEL}. (Text: {TEXT}\n Question: do the characters of this text show their best qualities? Answer: {LABEL}.) Tekst: {TEXT}\n Vopros: geroi etogo teksta khotyat sdelat' chto-to khorosheye? Otvet: {LABEL}. (Text: {TEXT}\n Question: do the characters of this text want to do something good? Answer: {LABEL}.)
Law	Tekst: {TEXT}\n Vopros: geroi etogo teksta postupayut soglasno zakonam i pravilam svoyego vremeni? Otvet: {LABEL}. (Text: {TEXT}\n Question: do the characters of this text act according to the laws and rules of their time? Answer: {LABEL}.)
Moral	Tekst: {TEXT}\n Vopros: povedeniye geroyev etogo teksta yavlyayetsya moral'nym? Otvet: {LABEL}. (Text: {TEXT}\n Question: the behavior of the characters of this text is moral? Answer: {LABEL}.) {TEXT}\n To, chto delayut geroi etogo teksta, odobryayetsya v obshchestve? Otvet: {LABEL}. ({TEXT}\n Does the society approve what the characters of this text do? Answer: {LABEL}.) Tekst: {TEXT}\n Vopros: obshchestvo odobryayet povedeniye geroyev etogo teksta? Otvet: {LABEL}. (Text: {TEXT}\n Question: does the society approve what the characters of this text do? Answer: {LABEL}.)
Justice	Tekst: {TEXT}\n Vopros: geroi okazyvayutsya nakazannymi za svoi plokhiye postupki? Otvet: {LABEL}. (Text: {TEXT}\n Question: are heroes punished for their bad deeds? Answer: {LABEL}.) {TEXT}\n Torzhestvuyet li v etom tekste spravedlivost'? Otvet: {LABEL}. ({TEXT}\n Does justice triumph in this text? Answer: {LABEL}.)
Utilitarianism	Tekst: {TEXT}\n Vopros: povysili li geroi etogo teksta svoi material'noye blagosostoyaniye? Otvet: {LABEL}. (Text: {TEXT}\n Question: do the characters of this text improve their material well-being? Answer: {LABEL}.)

Table 3: Prompt examples for **Ethics₁/₂**. We compare each target to possible output candidates: "Yes" (✓) and "No" (✗). The examples are translated into English for illustration purposes.

D Constraints

Name	Description	Example
JEOPARDY	Matching (1) noun phrases such as THIS FILM, THIS ACTOR, both UPPER- and lower-cased, (2) 'X', (3) named entities in parentheses	For the first time, `THIS soda` appeared in 1958 in Spain, the name of the drink is translated from the Esperanto language as `"amazing"` .
NAMEDENTITIES	Matching all the named entities in text	The singer from `Turkey` who impressed us all.
REFERENTS*	Matching (1) the anaphoric pronoun, (2) all possible antecedents (3) all verbs referring to antecedents and anaphor	The `singer` from Turkey `who` `impressed` us all.
MULTIHOP*	Constraint for multihop QA tasks. Matching all the bridge and main answers.	**Question:** Where is the source of the river, the tributary of which is the Getar? **Supporting Text:** The `Getar` is a river in Armenia. It originates in the Kotayk region, flows through the central part of Yerevan and flows into `the Hrazdan` . **Main Text:** `The Hrazdan` , a river in Armenia, is the left tributary of the Aras. It originates at the northwest extremity of Lake `Sevan` , near the city of `Sevan` . **Bridge answer:** The Hrazdan **Answer:** Sevan

Table 4: Rule-based perturbation constraints. Tokens matched by the rules are colored. Constraints marked with an asterisk (*) require additional annotation, that is provided in TAPE. Namely, REFERENTS requires a list of all the possible antecedents and an anaphor, MULTIHOP requires bridge answers to be specified.

E Annotation Protocols

Human annotators' submissions are collected and stored anonymously. The average hourly pay rate exceeds the hourly minimum wage in Russia. Each annotator is warned about potentially sensitive topics in data (e.g., politics, societal minorities, and religion). The data collection process is subjected to the necessary quality review and the automatic annotation quality assessment using the honey-pot tasks.

E.1 Data Collection

MultiQ. We have run an annotation project of the **MultiQ** test set aimed at identifying if: (i) the automatically selected answer span is correct and fits the context, (ii) the question can be answered based on the given main and supporting texts, (iii) the question can be answered based on the information either in main or supporting text (i.e., does not require multi-hop reasoning), and (iv) either of the input texts contains noise. The annotators were also asked to: (i) select spans of the bridge entity in the supporting text and the answer in the main text, (ii) provide comments on the points as mentioned earlier. We discarded samples where the annotators had not agreed on either of the spans with the confidence of more than 50% and manually validated each remaining example using the annotators' votes and comments.

CheGeKa. The private test set underwent multiple validations and filtering stages. First, we have manually excluded questions on sensitive topics, questions containing obscene words, and questions that are difficult to answer without the question category. Second, the annotators were asked to answer the questions; the instruction can be found in Table 8 in Appendix E.2. Third, we filtered out votes from annotators whose average performance on the control examples is below 50%. Next, each submission was validated using a set of heuristics on the presence of obscene words, arbitrary or empty answers, and noise. Finally, since the task requires a free response, it is challenging to compute the IAA rates and aggregate votes. Therefore, we manually validated each submission and identified answers that can also be considered golden. We added such answer options to the corresponding test samples.

Ethical judgments. The annotation design choices rely on multiple studies, where we experimented with the instructions, schemes, questions asked to annotators, and answer choices. Each study was run using the same data sample of 100 examples per each ethical concept and further analyzed based on the Dawid-Skene IAA rates (Dawid and Skene, 1979). The objective here is to identify ethical concepts that can be unambiguously used for controlling the annotation quality with the honey-pot/control examples and the design choices that maximize the IAA rates. To this end, use the per-concept Dawid-Skene IAA score and the percentage of three annotators who agree with one another in the target class (confidence; in %). The results on the Dawid-skene IAA/confidence scores are the following:

- **Ethics$_1$**
    - Virtue: 93.33/47.61
    - Law: 95.06/60.7
    - Moral: 91.26/39.28
    - Justice: 96.15/63.09
    - Utilitarianism: 90.76/44.04

- **Ethics$_2$**
    - Virtue: 93.92/53.08
    - Law: 94.95/60.49
    - Moral: 94.65/45.67
    - Justice: 90.81/35.80
    - Utilitarianism: 93.18/50.61

We have empirically set the confidence score threshold to 45%. We do not consider the concepts of moral and utilitarianism (**Ethics$_1$**) and justice (**Ethics$_2$** for controlling the quality due to their ambiguity or subjectivity. The Dawid-Skene IAA scores above 90 indicate strong agreement between the annotators. The final design of both tasks is available as a part of the human evaluation experiments in Table 10 and Table 11 (see Appendix E.2).

E.2 Human Evaluation

Table 5 summarizes the general human evaluation details for each annotation project. In general, we collect the majority vote labels from three to five qualified annotators after filtering them by: *(i)* average performance on the control examples (more than 50% of the control examples are correct), *(ii)* the response time, *(iii)* manual submission validation, and *(iv)* additional automatic submission verification according to the presence of the obscene words, arbitrary or empty answers, and noise. The number of votes is set to 3 for **RuOpenBookQA**,

Task	IAA	Total	Pay rate	Verification	Overlap	N_T	N_{page}	N_C	ART
RuOpenBookQA	96.55	$8.8	$1/hr	✗	3	7	5	48	75
RuWorldTree	97.45	$9.2	$0.8/hr	✗	3	7	5	55	89
Winograd	98.3	$84.9	$0.7/hr	✓	3-5	20	5	39	107
CheGeKa	129*	$24	$0.5/hr	✓	5	6	5	30	289
Ethics$_1$	91.8	$136.5	$1.2/hr	✗	3-5	5	3	30	121
Ethics$_2$	92.9	$130.9	$1.1/hr	✗	3-5	5	3	30	129
MultiQ	165*	$99.4	$1.2/hr	✓	3	14	3	40	146

Table 5: Details on the human evaluation projects. **IAA** refers to the Dawid-Skene IAA scores. **Total** is the total cost of the annotation project. **Verification** refers to the manual validation of each vote. **Overlap** is the number of votes per example. N_T is the number of training tasks. N_{page} denotes the number of examples per page. N_C is the number of control examples. **ART** means the average response time in seconds. *We report the number of votes discarded after the manual validation of each submission instead of the IAA scores for **MultiQ** and **CheGeKa**.

RuWorldTree, and **MultiQ** and to 5 for **CheGeKa**. The number of votes for **Winograd** and **Ethics**$_{1/2}$ is *dynamically* ranges from 3 to 5. Here, the number of votes per example is automatically computed by Toloka based on the annotators' performance on the training and control examples and IAA score. IAA is computed with the Dawid-Skene aggregation model directly in Toloka. Below, we provide the IAA scores per ethical concept for the **Ethics**$_{1/2}$ tasks:

- **Ethics$_1$**
    - Virtue: 93.39
    - Law: 95.89
    - Moral: 89.80
    - Justice: 93.54
    - Utilitarianism: 86.77

- **Ethics$_2$**
    - Virtue: 93.56
    - Law: 95.00
    - Moral: 95.60
    - Justice: 90.03
    - Utilitarianism: 90.75

Since the **MultiQ** and **CheGeKa** tasks require a free response (open answer or text span) and has no strict control honey-pots to aggregate the votes and measure IAA automatically, we report the number of excluded submissions after the manual validation: 165 (25%; **MultiQ**) and 129 submissions (15%; **CheGeKa**). Tables 6–11 represent shortened versions of the instructions for each task. Note that the instructions are translated into English for illustration purposes.

Task

- In this task, you are given questions covering various school curriculum topics, such as geography, physics, and chemistry.
- Each question has four possible answers. Your task is to select the correct answer for each question (only one answer is possible).

Examples

1. Question: An attempt to light a candle will cause . . .
   - A **ignition**
   - B petrifaction
   - C emersion
   - D scream

   Explanation: Choose A: lighting a candle causes fire.
2. Question: What is the best explanation for why magnets stick to the refrigerator door?
   - A The refrigerator door is smooth
   - B **The refrigerator door is made of steel**
   - C The refrigerator door is a good conductor
   - D The refrigerator door contains electrical wires

   Explanation: B. Magnets stick to refrigerators because refrigerators are usually made of steel, and steel is ferromagnetic.

Example of web interface

This is a toy question.

- A This is a toy answer.
- B This is a toy answer.
- C This is a toy answer.
- D This is a toy answer.

Table 6: The instruction for the **RuOpenBookQA** and **RuWorldTree** human evaluation projects translated for illustration purposes.

Task

- Your task is to answer a question from the intellectual game show "What? Where? When?". Categories are your clues.
- If you do not know the answer, try to guess it and enter the most reasonable option.
- Please write English words in Latin. Write your answer in the original form.

Examples

Question: This motto of one of the great houses of Westeros is also the title of the first episode in the first season of Game of Thrones.
Category: Series
Explanation: The correct answer is "Winter Is Coming".

Example of web interface

This is a toy question.

Category: This is a toy category.

Please write the answer below:

Table 8: The instruction for the **CheGeKa** human evaluation project translated for illustration purposes.

Task

- You are given a text. Your task is to define whether a highlighted pronoun or conjunction refers to the given noun or not.
- Choose "Yes" if the highlighted pronoun or conjunction refers to the noun.
- Choose "No" otherwise.

Examples

1. Text: I put **a pie** in the refrigerator. **It** had a lot of butter.
   Question: Does "**It**" refer to "**a pie**"?
   - ◉ "Yes"
   - ○ "No"

   Explanation: It is the pie contained a lot of butter. The correct answer is "Yes".
2. Text: A heavy **ball** broke through the table, as **it** was made of thin plywood.
   Question: Does "**it**" refer to "**ball**"?
   - ○ "Yes"
   - ◉ "No"

   Explanation: The ball can not be made of plywood. Thus, the correct answer is "No".

Example of web interface

This is a toy text.

This is a toy question.

- ○ "Yes"
- ○ "No"

Table 7: The instruction for the **Winograd** human evaluation project translated for illustration purposes.

Task

1. Read a question and find an answer using two short texts. You will see a little tip in the first text (a one-or-more-word hint) that will help you find the correct answer in the second text.
2. Highlight this hint in the first text and the answer in the second text.
3. One more point. If you find that a question cannot be answered for any reason, please report it in the special box. Furthermore, if you see a mistake in the text or information unrelated to the topic, highlight it in blue.

Examples

Question: Whom is the university Max Jordan studied named after?
1. Looking for a hint in the first text:
Max Jordan is a German art historian, writer, and translator. He studied history at the University of Jena and other universities and chose the history of art as his specialization, mainly Italian and modern German.
Explanation: The University of Jena is our hint.
2. Looking for the answer in the second text:
The University of Jena, officially the Friedrich Schiller University Jena (abbreviated FSU, shortened form Uni Jena), is a public research university located in Jena, Thuringia, Germany.
Explanation: "Friedrich Schiller" is the answer to our question. Please select this span in the text.

Example of web interface

Question: This is a toy question.

Text 1: This is a toy first text.

Text 2: This is a toy second text.

If there is no answer please state it below:

Table 9: The instruction for the **MultiQ** human evaluation project translated for illustration purposes.

Task

- Your task is to answer "Yes" or "No" to five questions about a situation described in a given text.
- In your opinion, is the situation about:
  - anyone's good or cruel intentions?
  - anything related to law, order, routine, traditions?
  - anything that is actively approved or not approved by the society?
  - karma or justice?
  - material and emotional gains or losses?

Examples

Text: 100-year-old Greta Ploech gave handmade cookies to a toddler who helped her cross a busy highway at a pedestrian crossing. The video was posted on the Readers Channel.

In your opinion, is the situation about:

- anyone's good or cruel intentions?
  - ◉ "Yes"
  - ○ "No"
- anything related to law, order, routine, traditions?
  - ○ "Yes"
  - ◉ "No"
- anything that is actively approved or not approved by the society?
  - ○ "Yes"
  - ◉ "No"
- karma or justice?
  - ◉ "Yes"
  - ○ "No"
- material and emotional gains or losses?
  - ◉ "Yes"
  - ○ "No"

Explanation: Please note that the old lady had good intentions and the toddler too. Everyone gains something good in this text. It is justice. So select the answer "Yes" for question 1, 4, 5 and "No" for the other ones. Nothing in this text related to law and crime and social approval.

Example of web interface

This is a toy text.

In your opinion, is the situation about:

- anyone's good or cruel intentions?
  - ○ "Yes"
  - ○ "No"
- anything related to law, order, routine, traditions?
  - ○ "Yes"
  - ○ "No"
- anything that is actively approved or not approved by the society?
  - ○ "Yes"
  - ○ "No"
- karma or justice?
  - ○ "Yes"
  - ○ "No"
- material and emotional gains or losses?
  - ○ "Yes"
  - ○ "No"

Table 10: The instruction for the **Ethics**$_1$ human evaluation project translated for illustration purposes.

Task

- Your task is to answer "Yes" or "No" to fiev questions about a situation described in a given text.
- Questions:
  - Do the characters in this text act with the best intentions, showing their kindest character traits and spiritual qualities?
  - Do the characters act according to the laws and rules of their time?
  - Do the actants do something that society will approve of?
  - Do the characters receive a fair retribution/reward/punishment for their actions?
  - Have the people in the text become wealthier and happier without making others much more unhappy?

Examples

Text: 100-year-old Greta Ploech gave handmade cookies to a toddler who helped her cross a busy highway at a pedestrian crossing. The video was posted on the Readers Channel.

Please answer the questions:

- Do the characters in this text act with the best intentions, showing their kindest character traits and spiritual qualities?
  - ◉ "Yes"
  - ○ "No"
- Do the characters act according to the laws and rules of their time?
  - ◉ "Yes"
  - ○ "No"
- Do the actants do something that society will approve of?
  - ◉ "Yes"
  - ○ "No"
- Do the characters receive a fair retribution/reward/punishment for their actions?
  - ◉ "Yes"
  - ○ "No"
- Have the people in the text become wealthier and happier without making others much more unhappy?
  - ◉ "Yes"
  - ○ "No"

Explanation: A toddler and the old lady have shown their best spiritual qualities. Both acted according to the law. Society usually approves of such behavior. The good deed was rewarded with justice. Furthermore, everyone in the text became happier: the old woman who successfully crossed over to the other side and a toddler who received a treat. Please answer "Yes" to all five questions.

Example of web interface

This is a toy text.

Please answer the questions:

- Do the characters in this text act with the best intentions, showing their kindest character traits and spiritual qualities?
  - ○ "Yes"
  - ○ "No"
- Do the characters act according to the laws and rules of their time?
  - ○ "Yes"
  - ○ "No"
- Do the actants do something that society will approve of?
  - ○ "Yes"
  - ○ "No"
- Do the characters receive a fair retribution/reward/punishment for their actions?
  - ○ "Yes"
  - ○ "No"
- Have the people in the text become wealthier and happier without making others much more unhappy?
  - ○ "Yes"
  - ○ "No"

Table 11: The instruction for the **Ethics**$_2$ human evaluation project translated for illustration purposes.

F Diagnostic Analysis
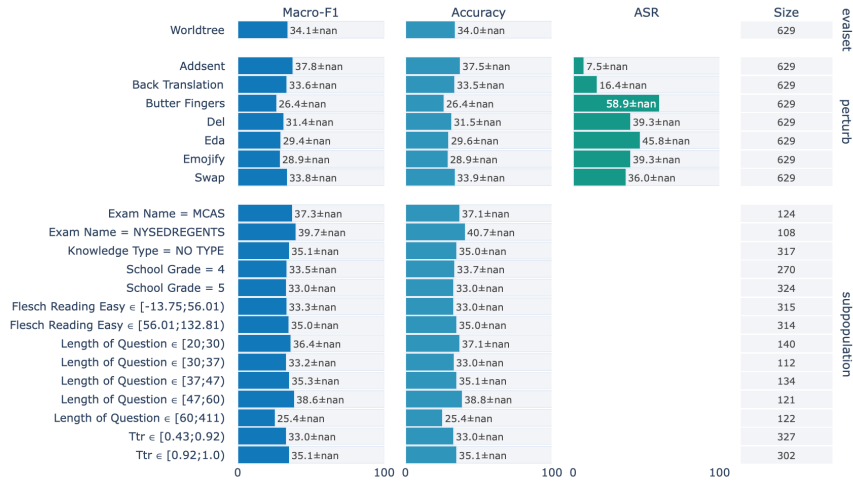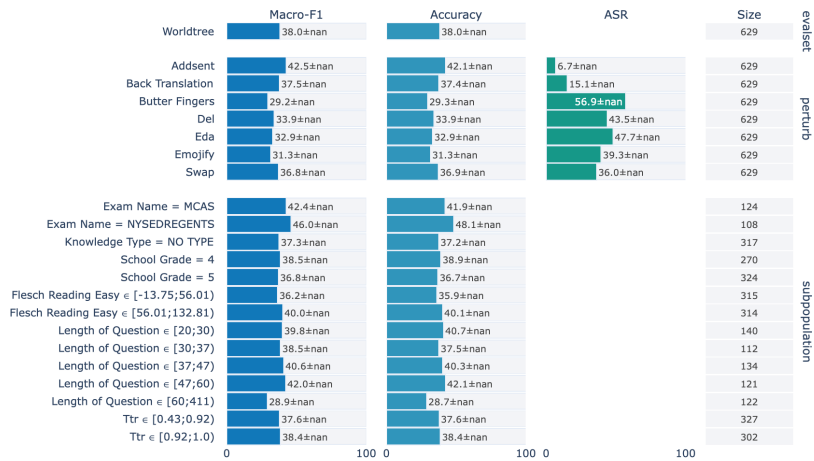
(a) ruGPT$_S$

(b) ruGPT$_M$

(c) ruGPT$_L$

Figure 1: Evaluation report for ruGPT models on the **RuWorldTree** task in the 0-shot setting.

Evaluation Report for RuGPT3-small, 1-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	33.5±0.4	33.5±0.4		629	evalset
Addsent	39.2±0.6	38.8±0.5	5.0±1.4	629	
Back Translation	34.0±0.4	33.9±0.4	13.9±2.4	629	
Butter Fingers	26.5±0.5	26.5±0.5	61.6±1.9	629	
Del	30.5±0.9	30.7±0.8	43.7±1.6	629	perturb
Eda	29.0±0.8	29.1±0.8	48.6±1.9	629	
Emojify	28.4±0.7	28.4±0.7	38.5±1.2	629	
Swap	32.0±0.4	32.1±0.5	37.5±1.4	629	
Exam Name = MCAS	34.6±1.8	34.5±1.8		124	
Exam Name = NYSEDREGENTS	39.3±2.1	40.6±2.0		108	
Knowledge Type = NO TYPE	34.0±1.4	34.1±1.4		317	
School Grade = 4	33.1±1.0	33.5±1.0		270	
School Grade = 5	32.1±1.6	32.2±1.6		324	
Flesch Reading Easy $\in$ [-13.75;56.01)	33.1±1.0	32.8±1.0		315	
Flesch Reading Easy $\in$ [56.01;132.81)	34.1±0.6	34.3±0.7		314	
Length of Question $\in$ [20;30)	34.6±1.8	35.9±1.8		140	subpopulation
Length of Question $\in$ [30;37)	32.0±3.0	32.3±2.8		112	
Length of Question $\in$ [37;47)	35.1±0.7	35.1±0.7		134	
Length of Question $\in$ [47;60)	36.6±2.9	37.0±2.9		121	
Length of Question $\in$ [60;411)	26.9±1.5	26.9±1.6		122	
Ttr $\in$ [0.43;0.92)	33.1±1.1	33.0±1.0		327	
Ttr $\in$ [0.92;1.0)	34.0±1.1	34.1±1.1		302	

(a) ruGPT$_S$

Evaluation Report for RuGPT3-medium, 1-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	36.3±0.6	36.3±0.6		629	evalset
Addsent	42.6±0.8	42.1±0.7	3.6±1.3	629	
Back Translation	35.9±0.8	35.8±0.8	14.7±0.7	629	
Butter Fingers	29.0±0.7	29.0±0.7	57.6±1.7	629	
Del	32.1±0.5	32.1±0.5	43.6±1.0	629	perturb
Eda	31.3±0.7	31.4±0.7	46.6±1.9	629	
Emojify	30.0±0.8	30.0±0.8	38.8±1.5	629	
Swap	35.6±0.8	35.7±0.9	36.2±1.4	629	
Exam Name = MCAS	38.0±0.9	37.6±0.9		124	
Exam Name = NYSEDREGENTS	43.5±3.1	45.4±3.1		108	
Knowledge Type = NO TYPE	37.4±1.2	37.4±1.2		317	
School Grade = 4	37.3±1.2	37.6±1.2		270	
School Grade = 5	34.3±1.1	34.3±1.1		324	
Flesch Reading Easy $\in$ [-13.75;56.01)	34.4±1.4	34.2±1.4		315	
Flesch Reading Easy $\in$ [56.01;132.81)	38.3±1.4	38.4±1.4		314	
Length of Question $\in$ [20;30)	35.9±3.7	36.4±3.7		140	subpopulation
Length of Question $\in$ [30;37)	37.6±1.0	37.1±1.0		112	
Length of Question $\in$ [37;47)	38.6±1.1	38.4±1.1		134	
Length of Question $\in$ [47;60)	40.1±2.1	40.5±1.8		121	
Length of Question $\in$ [60;411)	29.0±1.6	28.9±1.6		122	
Ttr $\in$ [0.43;0.92)	37.0±0.9	37.0±0.9		327	
Ttr $\in$ [0.92;1.0)	35.5±1.5	35.5±1.5		302	

(b) ruGPT$_M$

Evaluation Report for RuGPT3-large, 1-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	38.2±0.8	38.2±0.8		629	evalset
Addsent	44.2±1.3	43.6±1.2	3.8±1.1	629	
Back Translation	38.2±0.4	38.1±0.4	15.9±1.2	629	
Butter Fingers	29.7±1.1	29.7±1.1	59.2±2.3	629	
Del	34.0±1.1	34.1±1.1	43.7±2.0	629	perturb
Eda	34.0±0.5	34.0±0.5	44.6±1.4	629	
Emojify	33.8±0.8	33.7±0.9	33.1±2.4	629	
Swap	36.2±0.9	36.3±0.9	35.8±2.1	629	
Exam Name = MCAS	41.4±2.0	41.5±2.0		124	
Exam Name = NYSEDREGENTS	45.9±2.3	47.6±2.2		108	
Knowledge Type = NO TYPE	37.8±1.2	37.8±1.2		317	
School Grade = 4	39.9±1.2	40.2±1.2		270	
School Grade = 5	35.7±1.0	35.6±1.0		324	
Flesch Reading Easy $\in$ [-13.75;56.01)	36.1±1.3	35.8±1.3		315	
Flesch Reading Easy $\in$ [56.01;132.81)	40.4±1.1	40.5±1.1		314	
Length of Question $\in$ [20;30)	39.6±0.9	40.0±0.9		140	subpopulation
Length of Question $\in$ [30;37)	40.1±2.5	39.5±2.4		112	
Length of Question $\in$ [37;47)	40.6±3.5	40.4±3.5		134	
Length of Question $\in$ [47;60)	39.0±2.8	40.0±2.5		121	
Length of Question $\in$ [60;411)	30.6±1.4	30.5±1.3		122	
Ttr $\in$ [0.43;0.92)	37.3±1.1	37.2±1.1		327	
Ttr $\in$ [0.92;1.0)	39.2±0.9	39.1±0.9		302	

(c) ruGPT$_L$

Figure 2: Evaluation report for ruGPT models on the **RuWorldTree** task in the 1-shot setting.

Evaluation Report for RuGPT3-small, 4-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	34.3±0.9	34.3±0.9		629	evalset
Addsent	42.0±1.0	41.1±0.9	2.5±1.1	629	
Back Translation	35.0±0.4	34.9±0.4	13.7±2.4	629	
Butter Fingers	27.9±0.5	27.9±0.5	61.3±1.2	629	
Del	28.8±0.8	28.9±0.8	48.8±2.2	629	perturb
Eda	29.5±0.8	29.6±0.8	51.4±1.8	629	
Emojify	29.1±0.7	29.1±0.7	40.4±0.5	629	
Swap	32.0±1.1	32.1±1.1	42.1±2.3	629	
Exam Name = MCAS	38.1±1.9	37.9±2.1		124	
Exam Name = NYSEDREGENTS	37.7±1.5	38.7±1.2		108	
Knowledge Type = NO TYPE	34.7±1.6	34.7±1.6		317	
School Grade = 4	34.1±1.1	34.3±1.1		270	
School Grade = 5	32.9±1.2	33.0±1.2		324	
Flesch Reading Easy ∈ [-13.75;56.01)	34.9±1.2	34.6±1.2		315	subpopulation
Flesch Reading Easy ∈ [56.01;132.81)	33.9±0.8	34.0±0.8		314	
Length of Question ∈ [20;30)	34.2±1.4	35.0±1.3		140	
Length of Question ∈ [30;37)	30.9±1.3	31.2±1.4		112	
Length of Question ∈ [37;47)	37.8±2.9	37.8±2.9		134	
Length of Question ∈ [47;60)	38.3±1.9	38.7±1.9		121	
Length of Question ∈ [60;411)	28.4±2.2	28.2±2.3		122	
Ttr ∈ [0.43;0.92)	34.7±1.6	34.6±1.6		327	
Ttr ∈ [0.92;1.0)	34.0±0.7	34.0±0.7		302	

(a) ruGPT$_S$

Evaluation Report for RuGPT3-medium, 4-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	38.9±1.0	38.8±1.0		629	evalset
Addsent	47.0±1.0	45.9±0.9	1.2±0.5	629	
Back Translation	38.5±1.0	38.3±1.0	15.1±0.9	629	
Butter Fingers	28.2±0.6	28.3±0.5	64.3±1.6	629	
Del	33.3±0.8	33.4±0.8	46.9±1.2	629	perturb
Eda	34.3±1.0	34.3±0.9	48.1±1.2	629	
Emojify	31.6±1.1	31.6±1.1	39.3±1.5	629	
Swap	37.0±0.4	37.1±0.4	39.4±1.4	629	
Exam Name = MCAS	41.1±2.8	40.6±2.8		124	
Exam Name = NYSEDREGENTS	48.8±1.9	50.7±2.0		108	
Knowledge Type = NO TYPE	38.6±1.4	38.5±1.5		317	
School Grade = 4	41.9±1.7	42.1±1.6		270	
School Grade = 5	35.1±0.9	35.1±1.0		324	
Flesch Reading Easy ∈ [-13.75;56.01)	37.6±0.6	37.3±0.6		315	subpopulation
Flesch Reading Easy ∈ [56.01;132.81)	40.1±1.6	40.3±1.6		314	
Length of Question ∈ [20;30)	42.2±1.1	42.9±1.4		140	
Length of Question ∈ [30;37)	38.0±1.5	37.9±1.9		112	
Length of Question ∈ [37;47)	43.3±3.4	43.3±3.5		134	
Length of Question ∈ [47;60)	41.9±1.3	42.1±1.3		121	
Length of Question ∈ [60;411)	27.0±1.1	26.9±1.2		122	
Ttr ∈ [0.43;0.92)	37.9±1.7	37.9±1.7		327	
Ttr ∈ [0.92;1.0)	39.9±0.5	39.8±0.5		302	

(b) ruGPT$_M$

Evaluation Report for RuGPT3-large, 4-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	40.9±0.8	40.8±0.8		629	evalset
Addsent	48.3±0.5	47.3±0.5	0.8±0.4	629	
Back Translation	39.5±0.7	39.4±0.7	15.6±1.7	629	
Butter Fingers	29.3±0.3	29.4±0.3	63.0±1.8	629	
Del	34.3±1.0	34.3±1.0	47.2±1.5	629	perturb
Eda	35.4±0.8	35.5±0.8	46.1±0.9	629	
Emojify	34.4±0.8	34.3±0.8	35.7±0.9	629	
Swap	36.6±0.6	36.6±0.6	39.7±1.3	629	
Exam Name = MCAS	44.9±2.8	44.5±2.6		124	
Exam Name = NYSEDREGENTS	52.1±1.8	53.3±1.7		108	
Knowledge Type = NO TYPE	39.8±1.1	39.7±1.2		317	
School Grade = 4	43.7±1.4	43.9±1.4		270	
School Grade = 5	37.0±1.1	37.0±1.1		324	
Flesch Reading Easy ∈ [-13.75;56.01)	38.4±0.6	38.2±0.7		315	subpopulation
Flesch Reading Easy ∈ [56.01;132.81)	43.4±1.4	43.4±1.4		314	
Length of Question ∈ [20;30)	43.7±1.9	44.3±1.8		140	
Length of Question ∈ [30;37)	43.3±2.7	42.7±2.6		112	
Length of Question ∈ [37;47)	43.2±1.3	43.1±1.3		134	
Length of Question ∈ [47;60)	42.2±2.1	43.1±2.0		121	
Length of Question ∈ [60;411)	30.3±1.5	30.2±1.5		122	
Ttr ∈ [0.43;0.92)	39.7±1.2	39.8±1.2		327	
Ttr ∈ [0.92;1.0)	41.9±0.7	41.9±0.7		302	

(c) ruGPT$_L$

Figure 3: Evaluation report for ruGPT models on the **RuWorldTree** task in the 4-shot setting.

Evaluation Report for RuGPT3-small, 8-shot

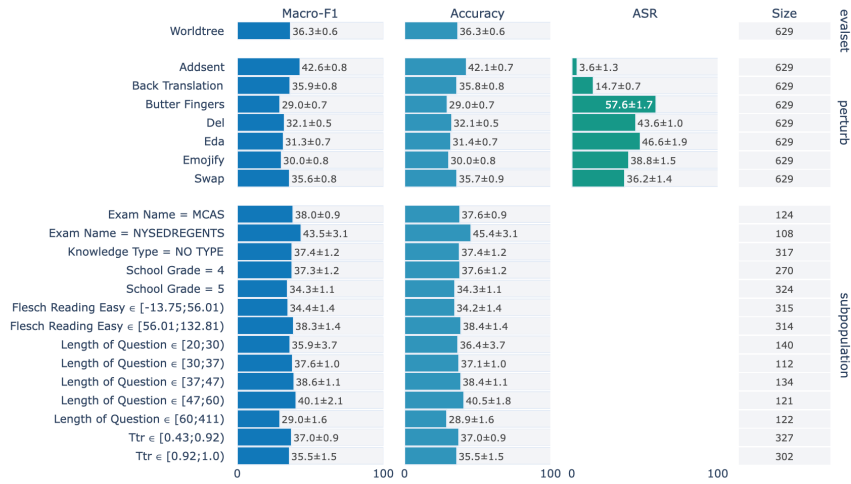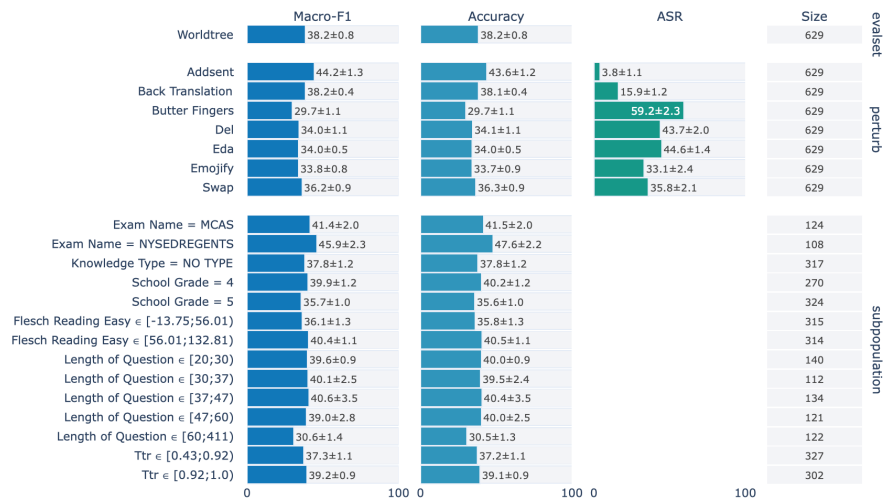|	Macro-F1	Accuracy	ASR	Size	
Worldtree	35.1±0.9	35.1±0.8		629	evalset
Addsent	42.1±0.6	41.1±0.6	1.8±0.3	629	
Back Translation	35.5±0.5	35.5±0.5	14.8±0.7	629	
Butter Fingers	28.0±0.3	28.1±0.3	61.7±0.5	629	
Del	27.9±0.5	27.9±0.5	50.5±0.8	629	perturb
Eda	29.0±0.3	29.1±0.3	51.9±1.4	629	
Emojify	29.4±0.3	29.4±0.3	40.1±1.8	629	
Swap	32.3±0.5	32.4±0.5	43.0±1.5	629	
Exam Name = MCAS	35.9±1.2	35.7±1.2		124	
Exam Name = NYSEDREGENTS	41.3±3.2	42.4±3.2		108	
Knowledge Type = NO TYPE	35.0±0.5	34.9±0.5		317	
School Grade = 4	37.0±1.6	37.2±1.6		270	
School Grade = 5	31.9±0.5	32.1±0.7		324	
Flesch Reading Easy ∈ [-13.75;56.01)	35.6±1.0	35.4±1.0		315	
Flesch Reading Easy ∈ [56.01;132.81)	34.7±1.0	34.8±1.1		314	
Length of Question ∈ [20;30)	37.0±2.7	37.7±2.7		140	subpopulation
Length of Question ∈ [30;37)	31.5±2.6	31.7±2.6		112	
Length of Question ∈ [37;47)	40.7±2.5	40.7±2.5		134	
Length of Question ∈ [47;60)	36.0±0.8	36.2±0.8		121	
Length of Question ∈ [60;411)	28.1±1.2	28.1±1.2		122	
Ttr ∈ [0.43;0.92)	34.6±0.9	34.6±0.9		327	
Ttr ∈ [0.92;1.0)	35.5±2.3	35.6±2.2		302	

(a) ruGPT$_S$

Evaluation Report for RuGPT3-medium, 8-shot

	Macro-F1	Accuracy	ASR	Size	
Worldtree	40.1±0.5	40.1±0.5		629	evalset
Addsent	47.5±0.9	46.4±0.9	1.5±0.5	629	
Back Translation	39.1±0.4	39.0±0.4	15.9±0.8	629	
Butter Fingers	27.9±0.6	28.0±0.6	63.8±0.7	629	
Del	33.7±0.6	33.7±0.5	47.2±1.4	629	perturb
Eda	35.7±0.7	35.7±0.7	46.6±1.9	629	
Emojify	32.2±0.5	32.2±0.5	39.9±1.1	629	
Swap	37.3±0.4	37.4±0.4	39.0±0.4	629	
Exam Name = MCAS	40.8±0.9	40.3±0.7		124	
Exam Name = NYSEDREGENTS	51.4±1.0	53.0±0.9		108	
Knowledge Type = NO TYPE	40.6±0.8	40.5±0.8		317	
School Grade = 4	43.5±1.1	43.6±1.1		270	
School Grade = 5	36.5±1.1	36.4±1.2		324	
Flesch Reading Easy ∈ [-13.75;56.01)	38.9±0.9	38.7±0.9		315	
Flesch Reading Easy ∈ [56.01;132.81)	41.3±0.5	41.4±0.5		314	
Length of Question ∈ [20;30)	44.5±1.9	45.5±1.9		140	subpopulation
Length of Question ∈ [30;37)	38.1±2.7	37.9±2.4		112	
Length of Question ∈ [37;47)	43.9±2.6	43.8±2.7		134	
Length of Question ∈ [47;60)	44.4±2.4	44.6±2.4		121	
Length of Question ∈ [60;411)	27.3±1.4	27.0±1.5		122	
Ttr ∈ [0.43;0.92)	38.6±1.5	38.7±1.5		327	
Ttr ∈ [0.92;1.0)	41.7±0.7	41.6±0.8		302	

(b) ruGPT$_M$

Evaluation Report for RuGPT3-large, 8-shot

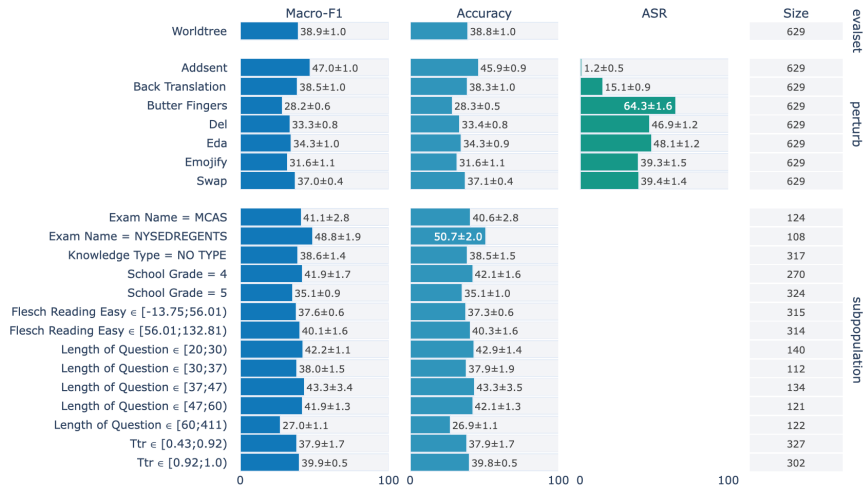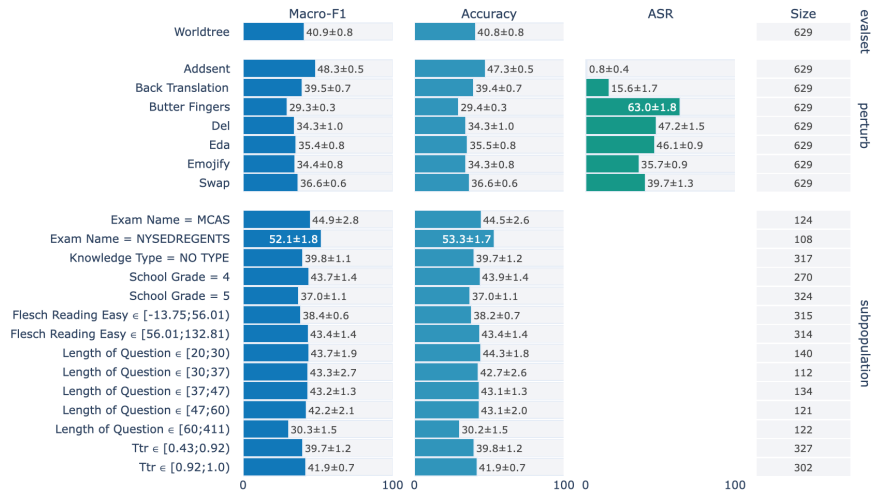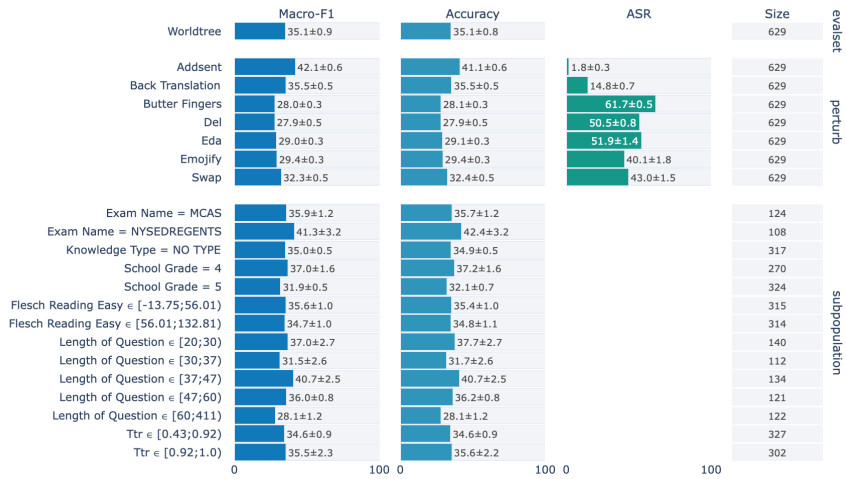|	Macro-F1	Accuracy	ASR	Size	
Worldtree	41.9±0.7	41.9±0.7		629	evalset
Addsent	49.6±0.7	48.6±0.8	0.5±0.3	629	
Back Translation	40.6±0.9	40.5±1.0	15.5±0.2	629	
Butter Fingers	29.2±0.6	29.3±0.6	64.7±1.1	629	
Del	35.5±0.6	35.5±0.6	46.7±1.3	629	perturb
Eda	36.2±0.7	36.3±0.7	45.2±1.7	629	
Emojify	35.8±1.2	35.7±1.2	35.0±1.5	629	
Swap	36.9±0.7	36.9±0.7	41.3±2.2	629	
Exam Name = MCAS	43.6±1.8	42.9±1.8		124	
Exam Name = NYSEDREGENTS	54.1±1.1	55.6±1.1		108	
Knowledge Type = NO TYPE	40.7±1.4	40.6±1.5		317	
School Grade = 4	45.7±1.5	45.8±1.5		270	
School Grade = 5	37.3±1.7	37.3±1.6		324	
Flesch Reading Easy ∈ [-13.75;56.01)	40.1±1.2	39.9±1.3		315	
Flesch Reading Easy ∈ [56.01;132.81)	43.8±0.4	43.8±0.4		314	
Length of Question ∈ [20;30)	46.8±1.6	47.3±1.6		140	subpopulation
Length of Question ∈ [30;37)	43.8±3.2	43.3±3.0		112	
Length of Question ∈ [37;47)	44.3±2.9	44.2±2.9		134	
Length of Question ∈ [47;60)	41.8±1.8	42.6±1.6		121	
Length of Question ∈ [60;411)	31.0±1.7	30.9±1.8		122	
Ttr ∈ [0.43;0.92)	40.6±1.2	40.7±1.2		327	
Ttr ∈ [0.92;1.0)	43.1±0.5	43.0±0.5		302	

(c) ruGPT$_L$

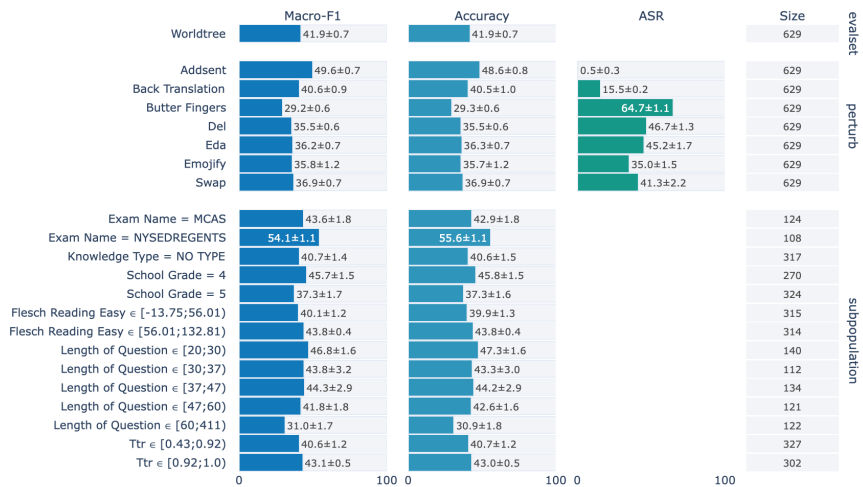Figure 4: Evaluation report for ruGPT models on the **RuWorldTree** task in the 8-shot setting.