TAPE: Assessing Few-shot Russian Language Understanding

Ekaterina Taktasheva1,2, Tatiana Shavrina1,3, Alena Fenogenova1∗, Denis Shevelev1, Nadezhda Katricheva1, Maria Tikhonova1,2, Albina Akhmetgareeva1, Oleg Zinkevich2, Anastasiia Bashmakova2, Svetlana Iordanskaia2, Alena Spiridonova2, Valentina Kurenshchikova2, Ekaterina Artemova4,5, Vladislav Mikhailov1
1SberDevices, 2HSE University, 3Artificial Intelligence Research Institute, 4Huawei Noah’s Ark lab, 5CIS 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 TAPE1 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 adaptation (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, 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...
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; Ethics₁/₂). 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 Toloka³ (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 (§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; Ethics₁/₂). 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 Toloka³ (Pavlichenko et al., 2021).
solved homonymy. In the resulting 2k+ sentences, homonymy is resolved automatically with UDPipe 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 dependent on the chosen candidate noun.

- **Context:** “Brosh’ iz Pompeii, 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 tochkii 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, grazhdanom 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 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, 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 oza Sevan v yego severo-zapadnyi chast’, 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 setup is similar to Jeopardy, where the player should answer
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**: “Puchehini” (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* (Ethics1). The second one is to evaluate the positive or negative implementation of these concepts with binary categories (Ethics2).

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 RusVectors 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 Ploeha syolsha, kotoryy pomog yey pereyti cherez ozhivlennoye shosse po peshekhodnomu perekhodu.” (100-year-old Greta Ploeh gave handmade cookies to a toddler who helped her cross a busy highway at a pedestrian crossing.)
- **Labels1**: ✓(Virtue) X(Law) X(Moral) ✓(Justice) ✓(Utilitarianism)
- **Labels2**: ✓(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 $\mathcal{D}_{\text{train}}$ with labeled examples and a test set $\mathcal{D}_{\text{test}}$. 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 $\mathcal{E}^i = (\mathcal{E}_{\text{train}}^i, \mathcal{D}_{\text{test}}^A)$ consists of $k$ demonstration examples $\mathcal{E}_{\text{train}}^i$ randomly sampled from $\mathcal{D}_{\text{train}}$ with replacement, and a single test $\mathcal{D}_{\text{test}}^A$ 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 $\mathcal{D}_{\text{test}}$ 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 $\mathcal{E}_{\text{train}}^i$ corresponds to $T + 1$ test variations, including the original $\mathcal{D}_{\text{test}}$ and $T$ adversarial test sets $\mathcal{D}_{\text{test}}^A$, acquired through the modification of $\mathcal{D}_{\text{test}}$. $T$ 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.** EMOMIFY 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) has proved to be efficient in fooling LMs on text classification tasks. We use two EDA configurations: swapping words (EDA\text{SWAP}) and deleting tokens (EDA\text{DELETE}).
- **Paraphrasing.** BACKTRANSLATION (Yaseen and Langer, 2021) allows to generate linguistic variations of the input without changing named entities. We use the OpusMT model\footnote{hf.co/Helsinki-NLP/opus-mt} 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\footnote{hf.co/THUMT/mGPT} 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.
<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Example</th>
<th>Adv. threshold</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spelling</td>
<td>BUTTERFINGERS</td>
<td>This is a sentence to test the code</td>
<td>0.15</td>
<td>82.72</td>
</tr>
<tr>
<td>Modality</td>
<td>EMOJIFY</td>
<td>This is a sentence to test the</td>
<td>0.4</td>
<td>84.27</td>
</tr>
<tr>
<td>Random</td>
<td>EDADELETE</td>
<td>This is a sentence to test the code</td>
<td>0.3</td>
<td>96.16</td>
</tr>
<tr>
<td></td>
<td>EDAWORD</td>
<td>code is a sentence to test the</td>
<td>0.3</td>
<td>93.99</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>BACKTRANSLATION</td>
<td>This sentence tests the code</td>
<td>0.5</td>
<td>95.38</td>
</tr>
<tr>
<td>Distraction</td>
<td>ADDSENT</td>
<td>This is a sentence to test the code; if you want to delete it</td>
<td>0.5</td>
<td>92.95</td>
</tr>
</tbody>
</table>

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 \(D_{train}\) and evaluated on \(D_{test}\).

**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 hyper-parameters.

**5.2 Neural Baselines**

We run zero-shot and few-shot evaluation of Russian GPT\(^ {12}\) LMs available under the HuggingFace library (Wolf et al., 2020). We consider three model versions: ruGPT\(^ {13}\) (125M), ruGPT\(^ {14}\) (350M), and ruGPT\(^ {15}\) (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.

\[
PPL(t) = \exp\left(-\frac{1}{|t|} \sum_{i=0}^{|t|} \log p_{\theta}(x_i | x_{<i})\right), \tag{1}
\]

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

| Dataset          | BF | EMJ | EDA | BT | AS | Subpopulations                                      | $|D|_{train}$ | $|D|_{test}$ | Metrics          | Domain         |
|------------------|----|-----|-----|----|----|----------------------------------------------------|-------------|-------------|-----------------|---------------|
| Winograd         | ✓  | ✓   | ✓   | ✓  | ✓  | • Gender, Number, • NumCandidates, • Class distribution, • Length | 60          | 986         | F1 / Acc.       | Misc.          |
| Ethics\textsubscript{1/2} | ✓  | ✓   | ✓   | ✓  | ✓  | • Domain, • Length, • Diversity, Readability     | 59/58       | 1k/1k       | F1 / Acc.       | News, fiction books |
| 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 |

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\textsuperscript{16} on a subset of the corresponding $D_{train}$. 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\textsuperscript{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 correct 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\textsuperscript{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\textsubscript{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

\textsuperscript{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.

\textsuperscript{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 $D_{train}$ and guaranteed to be unambiguous.

\textsuperscript{18}toloka.ai/docs/result-aggregation
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 $D_{test}$ 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 ruGPTs. 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 lower 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-
The lower, the better. The best ASR value is put in bold and the second best is underlined.

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.

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 $D_{\text{train}}$ and $D_{\text{test}}$.

**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


2481


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


Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalyne PaoloLada. 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).*


### A General Dataset Statistics

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

Table 1: General statistics for each dataset. NT refers to the total number of tokens. NU 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 Ethics1/2.

### B Winograd Queries

This appendix provides the list of queries that correspond to the RusCorpora query language 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
   2. “Istoriya o zhenschine, kotoraya...” (A story about a woman that...)
     - S & nom & sg & f
     - at the distance of 0 to 1 from
  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 obstanovi, 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 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 l 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 | ins | loc) & pl & m

- **Type 2**: Coincidence of nominative and accusative forms in the masculine gender.
  1. “Bal’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) & sg & m & inan

- **Type 3**: Coincidence of genitive and possessiveness.
  1. “Ikh deti razdrazhali” (Their kids were annoying / They were annoyed by kids)
     - they, (gen | gen2) & 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 4**: Two possible references to a pronoun.
  1. “Katya sprosila Mashu, delala li ona...” (Katya...
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.
Table 2: Prompt examples for the classification and QA tasks. The examples are translated into English for illustration purposes.

<table>
<thead>
<tr>
<th>Task</th>
<th>Template</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Winograd</td>
<td>V predlozhenii {CONTEXT} u Slovo {REFERENCE} otnositsya k slovu {CANDIDATE ANSWER}? [LABEL]. (In the sentence {CONTEXT} u Slovo {REFERENCE} refers to the word {CANDIDATE ANSWER} or not? [LABEL].)</td>
<td>Yes; No (✓)</td>
</tr>
<tr>
<td></td>
<td>V predlozhenii {CONTEXT} u Slovo &quot;{REFERENCE}&quot; otnositsya k slovu &quot;{CANDIDATE ANSWER}&quot;? [LABEL]. (In the sentence {CONTEXT} u Slovo &quot;{REFERENCE}&quot; refers to the word &quot;{CANDIDATE ANSWER}&quot; or not? [LABEL].)</td>
<td>Yes; No (✓)</td>
</tr>
<tr>
<td></td>
<td>V predlozhenii {CONTEXT} u Slovo &quot;{REFERENCE}&quot; otnositsya k slovu &quot;{CANDIDATE ANSWER}&quot; ili net? [LABEL]. (In the sentence {CONTEXT} u Slovo &quot;{REFERENCE}&quot; refers to the word &quot;{CANDIDATE ANSWER}&quot; or not? [LABEL].)</td>
<td>Yes; No (✓)</td>
</tr>
<tr>
<td></td>
<td>&quot;{CONTEXT}&quot; u Slovo &quot;{REFERENCE}&quot; otnositsya k slovu &quot;{CANDIDATE ANSWER}&quot; ili net? [LABEL]. (&quot;{CONTEXT}&quot; u Slovo &quot;{REFERENCE}&quot; refers to the word &quot;{CANDIDATE ANSWER}&quot; or not? [LABEL].)</td>
<td>Yes; No (✓)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RuOpenBookQA/RuWorldTree</th>
<th>{QUESTION} [CATEGtory] [QUESTION] A. B.</th>
<th>Generated answer</th>
</tr>
</thead>
</table>

Table 3: Prompt examples for Ethics2/3. We compare each target to possible output candidates: “Yes” (✓) and “No” (✗). The examples are translated into English for illustration purposes.
## D Constraints

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JEOPARDY</strong></td>
<td>Matching (1) noun phrases such as THIS FILM, THIS ACTOR, both UPPER- and lower-cased, (2) 'X', (3) named entities in parentheses</td>
<td>For the first time, <strong>THIS soda</strong> appeared in 1958 in Spain, the name of the drink is translated from the Esperanto language as “amazing”</td>
</tr>
<tr>
<td><strong>NAMED_ENTITIES</strong></td>
<td>Matching all the named entities in text</td>
<td>The singer from Turkey who impressed us all.</td>
</tr>
<tr>
<td><strong>REFERENTS</strong></td>
<td>Matching (1) the anaphoric pronoun, (2) all possible antecedents (3) all verbs referring to antecedents and anaphor</td>
<td>The singer from Turkey who impressed us all.</td>
</tr>
<tr>
<td><strong>MULTIHOP</strong></td>
<td>Constraint for multihop QA tasks. Matching all the bridge and main answers.</td>
<td><strong>Question:</strong> Where is the source of the river, the tributary of which is the Getar?</td>
</tr>
<tr>
<td></td>
<td><strong>Supporting Text:</strong> The <strong>Getar</strong> is a river in Armenia. It originates in the Kotayk region, flows through the central part of Yerevan and flows into the Hrazdan.</td>
<td><strong>Main Text:</strong> The <strong>Hrazdan</strong>, 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. <strong>Bridge answer:</strong> The Hrazdan <strong>Answer:</strong> Sevan</td>
</tr>
</tbody>
</table>
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**
  - 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**
  - 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) and justice (Ethics) 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.
<table>
<thead>
<tr>
<th>Task</th>
<th>IAA</th>
<th>Total</th>
<th>Pay rate</th>
<th>Verification</th>
<th>Overlap</th>
<th>NT</th>
<th>Npage</th>
<th>NC</th>
<th>ART</th>
</tr>
</thead>
<tbody>
<tr>
<td>RuOpenBookQA</td>
<td>96.55</td>
<td>$8.8</td>
<td>$1/hr</td>
<td>✗</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>48</td>
<td>75</td>
</tr>
<tr>
<td>RuWorldTree</td>
<td>97.45</td>
<td>$9.2</td>
<td>$0.8/hr</td>
<td>✗</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>55</td>
<td>89</td>
</tr>
<tr>
<td>Winograd</td>
<td>98.3</td>
<td>$84.9</td>
<td>$0.7/hr</td>
<td>✓</td>
<td>3-5</td>
<td>20</td>
<td>5</td>
<td>39</td>
<td>107</td>
</tr>
<tr>
<td>CheGeKa</td>
<td>129*</td>
<td>$24</td>
<td>$0.5/hr</td>
<td>✓</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>30</td>
<td>289</td>
</tr>
<tr>
<td>Ethics1</td>
<td>91.8</td>
<td>$136.5</td>
<td>$1.2/hr</td>
<td>✗</td>
<td>3-5</td>
<td>5</td>
<td>3</td>
<td>30</td>
<td>121</td>
</tr>
<tr>
<td>Ethics2</td>
<td>92.9</td>
<td>$130.9</td>
<td>$1.1/hr</td>
<td>✗</td>
<td>3-5</td>
<td>5</td>
<td>3</td>
<td>30</td>
<td>129</td>
</tr>
<tr>
<td>MultiQ</td>
<td>165*</td>
<td>$99.4</td>
<td>$1.2/hr</td>
<td>✓</td>
<td>3</td>
<td>14</td>
<td>3</td>
<td>40</td>
<td>146</td>
</tr>
</tbody>
</table>

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. **NT** is the number of training tasks. **Npage** denotes the number of examples per page. **NC** 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 Ethics1/2 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 Ethics1/2 tasks:

- **Ethics1**
  - Virtue: 93.39
  - Law: 95.89
  - Moral: 89.80
  - Justice: 93.54
  - Utilitarianism: 86.77

- **Ethics2**
  - 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
• 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 8: The instruction for the CheGeKa human evaluation project translated for illustration purposes.

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\textsubscript{1} 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.
• 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”

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”

Table 11: The instruction for the Ethics\textsubscript{2} human evaluation project translated for illustration purposes.
### F Diagnostic Analysis

#### Evaluation Report for RuGPT3-small, 0-shot

<table>
<thead>
<tr>
<th>Wordtree</th>
<th>Macro-F1</th>
<th>Accuracy</th>
<th>ASR</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddSent</td>
<td>0.13</td>
<td>0.04</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Back Trans</td>
<td>0.16</td>
<td>0.05</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Butter Fingers</td>
<td>0.18</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Del</td>
<td>0.18</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Ede</td>
<td>0.18</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Enronify</td>
<td>0.18</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Swap</td>
<td>0.18</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
</tbody>
</table>

#### Evaluation Report for RuGPT3-medium, 0-shot

<table>
<thead>
<tr>
<th>Wordtree</th>
<th>Macro-F1</th>
<th>Accuracy</th>
<th>ASR</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddSent</td>
<td>0.22</td>
<td>0.05</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Back Trans</td>
<td>0.25</td>
<td>0.05</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Butter Fingers</td>
<td>0.30</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Del</td>
<td>0.30</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Ede</td>
<td>0.30</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Enronify</td>
<td>0.30</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Swap</td>
<td>0.30</td>
<td>0.06</td>
<td>0.01</td>
<td>629</td>
</tr>
</tbody>
</table>

#### Evaluation Report for RuGPT3-large, 0-shot

<table>
<thead>
<tr>
<th>Wordtree</th>
<th>Macro-F1</th>
<th>Accuracy</th>
<th>ASR</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddSent</td>
<td>0.45</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Back Trans</td>
<td>0.48</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Butter Fingers</td>
<td>0.53</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Del</td>
<td>0.53</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Ede</td>
<td>0.53</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Enronify</td>
<td>0.53</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
<tr>
<td>Swap</td>
<td>0.53</td>
<td>0.08</td>
<td>0.01</td>
<td>629</td>
</tr>
</tbody>
</table>

**Figure 1**: Evaluation report for ruGPT models on the RuWorldTree task in the 0-shot setting.
Figure 2: Evaluation report for ruGPT models on the RuWorldTree task in the 1-shot setting.
Figure 3: Evaluation report for ruGPT models on the RuWorldTree task in the 4-shot setting.
Figure 4: Evaluation report for ruGPT models on the RuWorldTree task in the 8-shot setting.