Eye of Judgement: Dissecting the Evaluation of Russian-speaking LLMs with POLLUX

Anonymous ACL submission

Abstract

We introduce **POLLUX**, a comprehensive open-source benchmark designed to evaluate the generative capabilities of large language models (LLMs) in Russian. Our main contribution is a novel evaluation methodology that enhances the interpretability of LLM assessment. For each task type, we define a set of detailed criteria and develop a scoring protocol where models evaluate responses and provide justifications for their ratings. This enables transparent, criteria-driven evaluation beyond traditional resource-consuming, side-by-side human comparisons. POLLUX includes a detailed, fine-grained taxonomy of 35 task types covering diverse generative domains such as code generation, creative writing, and practical assistant use cases, totaling 2,100 manually crafted and professionally authored prompts. Each task is categorized by difficulty (easy/medium/hard), with experts constructing the dataset entirely from scratch. We also release a family of LLM-as-Judge (7B and 32B) evaluators trained for nuanced assessment of generative outputs. This approach provides scalable, interpretable evaluation and annotation tools for model development, effectively replacing costly and less precise human judgments.

1 Introduction

Evaluating generative models is a fundamental challenge due to their diverse and creative outputs. As large language models (LLMs) generate increasingly complex texts, traditional evaluation methods are less effective. The *LLM-as-a-judge* evaluation approach, where one LLM assesses the outputs of another, has emerged as a scalable alternative and strongly aligns with human judgments. However, most research has focused on English, leaving other languages, such as Russian, underexplored.

In this work, we bridge that gap, introducing **POLLUX**, a comprehensive open-source benchmark designed to evaluate the generative capabilities of LLMs in Russian rigorously. Our primary contribution is a novel evaluation methodology that aims to improve the interpretability and scalability of LLM evaluation. POLLUX features a **fine-grained hierarchical taxonomy of *35 generative task types inferred from open LLM usage logs*** spanning diverse domains, including code generation, creative writing, and practical assistant applications, with a total of *2,100 manually crafted prompts*. Each task is annotated by explicitly formulated **difficulty level** (easy/medium/hard) and constructed entirely from scratch by **domain experts** to ensure high-quality, unbiased evaluation data. We define a detailed **set of criteria for each task type**. We also develop a transparent scoring protocol where models assess responses and generate open-ended justifications for their ratings. This approach enables a **criteria-driven, reproducible evaluation framework**, reducing reliance on costly and less consistent human side-by-side comparisons. Additionally, we release **a family of LLM-as-Judge models** (7B and 32B parameters) trained to perform criteria-aligned assessments of generative outputs both with a score and textual feedback, offering a scalable and interpretable alternative to human evaluation.

The key contributions of this work include:

- A *general methodology* for LLM evaluation, comprising:
  - A hierarchical *taxonomy of generative tasks*, categorized by complexity and domain,
  - A fine-grained *taxonomy of criteria* for systematic evaluation,
- An *open benchmark* with prompts and annotations verified by experts and the release of *LLM-as-Judge evaluators* (7B and 32B) for automated assessment. Benchmark, code and models are available at the link[1].

[1] Link is disabled to maintain anonymity.

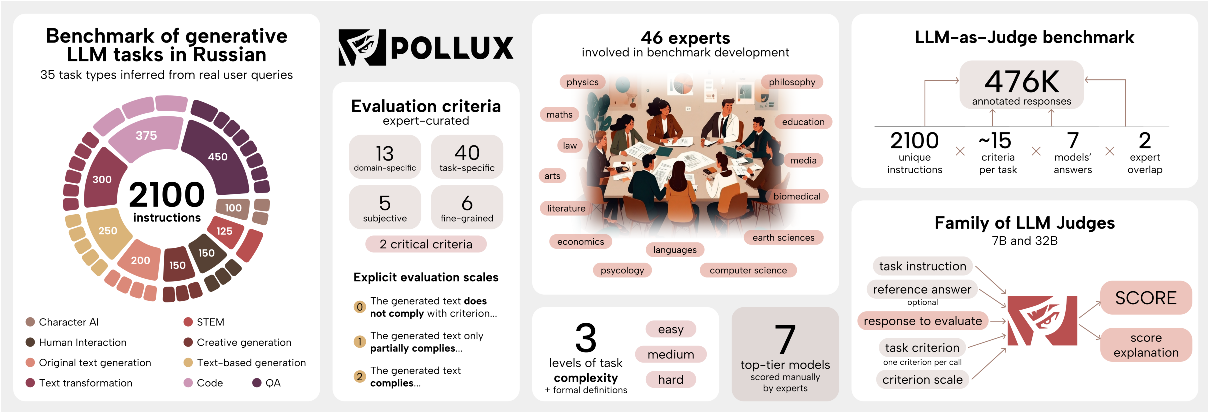


Figure 1: POLLUX overview and statistics: benchmark characteristics, including tasks and criteria, information about the experts involved in creating the data, the overflow of LLM-as-Judge models, and the synthetic data used for them.

2 The POLLUX Generative Benchmark

The POLLUX benchmark provides a quantitative and qualitative assessment of LLMs' capabilities across the proposed taxonomies of generative tasks and evaluation criteria.

The **Generative Tasks Taxonomy** (Section 2.1) is grounded in 35 broad categories that are extracted from the publicly available conversations with LLM-based services. These categories are then expanded by subdividing them in accordance with functional styles and genres. Each task is additionally annotated with three complexity levels, which completes the design of the taxonomy.

The basis for the **Criteria Taxonomy** (Section 2.2) is formed by 13 criteria that evaluate basic semantic, syntactic and lexical properties of a text. These are supplemented by criteria that target certain aspects associated with distinct properties of a task or functional style that the particular text relates to. Each criterion is equipped with a corresponding scale and rubrics that coordinate the assignment of scores.

The **Benchmark Composition** (Section 2.3) is performed by domain experts. We conducted thorough selection, training and examination (Appendix I), which resulted in complete and diverse expert panels, see Appendix L.

2.1 The Generative Tasks Taxonomy

To obtain a hierarchy of generative tasks, a two-stage procedure was applied. The first stage implies bottom-up category mining using instruction clustering, and the second stage marks the point where the specialized knowledge of the domain experts is applied.

Organizing use cases into task taxonomy As a source of user-LLM interaction statistics, we selected the WildChat-1M dataset (Zhao et al., 2024). It comprises 837K sessions of user interaction with LLMs, of which 87K sessions are in Russian, totaling 270K distinct user prompts. We removed duplicate prompts inside sessions using the rapidfuzz WRatio function[2] with a threshold of 95 and excluded samples identified in the original WildChat-1M annotation as toxic and violating safety protocols. The dataset was further reduced to 181K user prompts by deduplication on hash from first and last 200 characters and removing prompts longer that 200 words.

Then clustering in the form of BERTopic[3] (Grootendorst, 2022) pipeline was executed on the instructions embeddings concatenated with embeddings of definitions of corresponding tasks (e.g., *debug code* or *paraphrase text*), which were generated with Llama-3-8B-Instruct[4] (Grattafiori et al., 2024a), see Appendix E for the details of clusterization procedure. This resulted in 4500 distinct clusters.

Each cluster centroid was manually assigned a task definition, which was inferred from the formulation of a centroid and had to be expressed in less than four words (e.g. *research assistance* or *story title generation*), by a majority vote rule. Annotation process was performed by three contributing authors of this study with an average agreement score of 0.81. Task definitions that occurred less than at least twice (these constitute 43.5% of the annotated

[2]rapidfuzz WRatio function implements weighted similarity from several functions such as normalized Indel similarity (input strings are lowercased and non-alpha numeric characters are removed).

[3]MaartenGr/BERTopic

[4]meta-llama/Meta-Llama-3-8B-Instruct

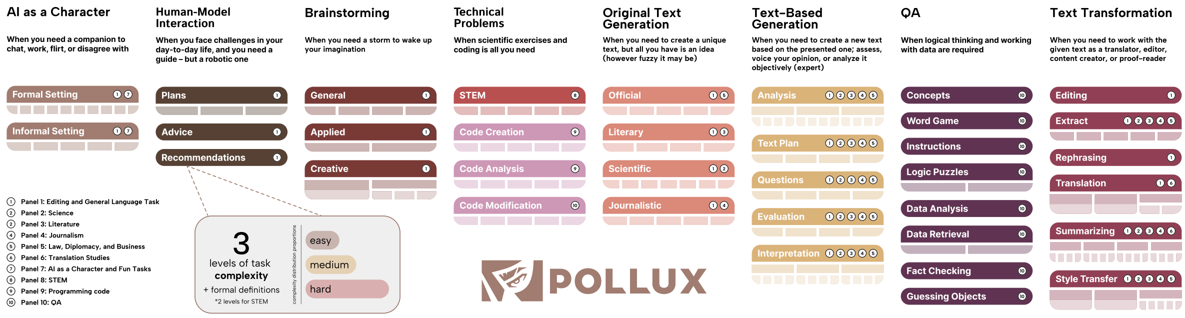


Figure 2: The POLLUX generative taxonomy of tasks. The labeled figures highlighted in bright colors are major 35 tasks groups. Each task group is annotated with corresponding expert panels.

centroids) were excluded. Some (approximately every 2.31) of the resulting 81 tasks definitions were additionally merged on the principle that tasks share common required skills hence providing aggregate 35 distinct task types.

Expert refining Finally, the generative tasks taxonomy had to be be adapted to most general and expert use cases. We thus adopted the functional-stylistic classification, which builds upon five functional text styles (Literature, Science, Journalism, Law, diplomacy and business and Conversational) and considers the purpose of the text, its form, and its content. This approach ensured that the evaluation of texts was aligned with existing linguistic frameworks, facilitating objective, comprehensive, and professional assessment, see Propp (2024); Tomashevskij (2023); Pen‘kovskaya and Kushel‘ (2022).

After establishing functional styles as a superstructure of the generative tasks taxonomy, 10 expert panels[5] (5 for each functional style, editors and translators, separate panels for code-related tasks, Science, Technology, Engineering, and Mathematics (STEM) problems and information seeking, see Appendix K for the descriptions of panels) deepened the taxonomy (see Appendix L for expert profiles); first, by writing down all the genres inherent in each style and/or task (in case of general-purpose tasks), and then choosing the most illustrative ones. As a result, the taxonomy covers 15 literary movements, 17 Russian writers, 35 literary, 26 journalistic, 7 official and 25 scientific substyles and genres (refer to Appendix H.1 for a complete lists), which yields 35 major tasks groups, 104 subgroups and 52 subsubgroups totaling in 152 tasks (hereafter, task denotes one of these 152 leaf nodes of the taxonomy unless stated otherwise), see Figure 2. Each task is additionally divided into three complexity levels [6]: easy, medium and hard. The definitions of complexity levels are specific to the task, hence spanning 451 definitions. We provide the full taxonomy with definitions of all 152 tasks and their corresponding complexity levels [7], refer to Appendix G.1 for a representative example from the taxonomy.

2.2 The Criteria Taxonomy

In line with Generative Tasks Taxonomy this work adopts block-like structure for the criteria system. This system embodies groups of independent evaluation aspects and allows to construct a set of criteria specifically tailored to a particular text based on its functional style and relation to a certain task. All criteria are designed to represent the evaluation aspects *necessary* to assess the quality of a generated text with respect to user's intent and provided attributes[8]. All criteria are equipped with corresponding scale and rubrics, see Table 21 in Appendix G for an example. We provide definitions of all the criteria derived in this work alongside their scales and rubrics[9].

Basis criteria system Following Howcroft et al. (2020), this work bases its criteria framework on three levels of evaluation aspects, namely (i) context-independent assessment and evaluation relative to (ii) input query and (iii) external data

[5]One more Crowd panel was added, see Section K.11.

[6]STEM problems have two complexity levels, which are high school and university program.

[7]The full taxonomy accounts for 65 pages and is not included in the paper to avoid overloading the Appendices. It is distributed in a separate `Supplementary_A_Tasks_Taxonomy.pdf` file and in a json-format as `metadata/tasks.json`.

[8]We acknowledge the criteria taxonomy probably may be deepened and widened, but the main focus of this work is to establish systematical approach to criteria evaluation, hence we prioritize rigor of the system over an attempt to cover *all* evaluation dimensions of *all* possible task formulations.

[9]Full Criteria taxonomy accounts for 46 pages and is not included in the paper in order not to overload the Appendices. The Criteria taxonomy is distributed in a separate `Supplementary_B_Criteria_Taxonomy.pdf` and `metadata/criteria.json` files.

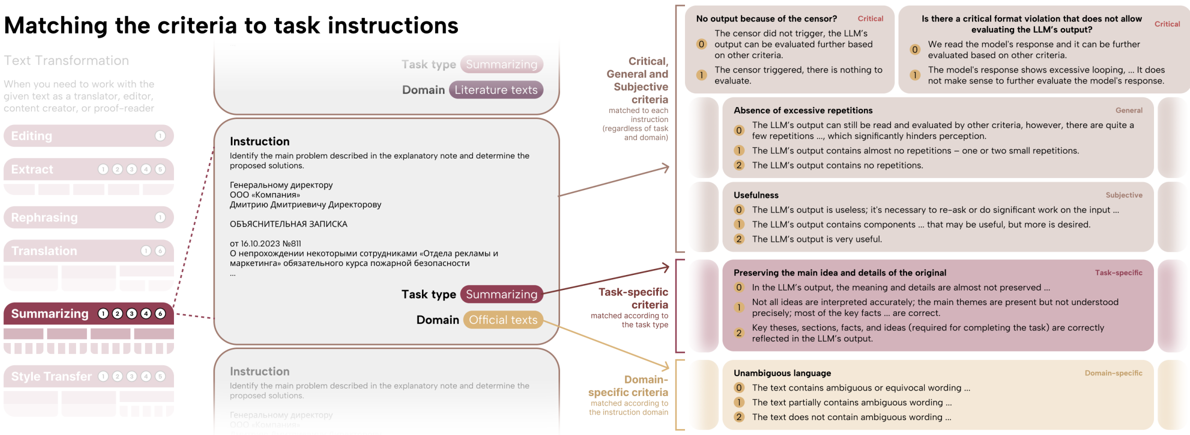


Figure 3: For each task-functional style pair, the POLLUX methodology suggests a set of criteria developed by experts specifically for that type of text. These criteria can be used as metrics for both automated and human evaluation.

sources. Each of these levels is further bifurcated into Form and Content components, creating a comprehensive evaluation matrix M. Each expert panel i was responsible for populating the matrix M_i with evaluation aspects needed to assess the quality of responses to tasks that fall within the panel's range of expertise, refer to Appendix K for the lists of such tasks (**Tasks** paragraph) and intuition behind evaluation aspects (**Criteria** subsection) for each panel. As all panel-specific M_is were complete the dedicated panel supervisers (see Section I.1 in Appendix I and Appendix L) and five contributing authors of this paper aggregated those criteria from the collection of M_i that focus on universal text quality markers while deliberately ignore style-specific characteristics.

To ensure the selected 22 criteria are independent of each other and do not correlate (see Xiao et al. (2023) for the importance of evaluation criteria independence), the pairwise comparisons (totaling 231 comparisons) were performed by the same five contributing authors with an average agreement score of 0.72. Those pairs of criteria that have been voted as correlating by at least three annotators (13.8% of all pairs) were merged [10] in a single criterion leaving final 13 independent criteria. These in turn were subdivided into Critical, General and Subjective groups, which account for two, eight and five criteria respectively. This first step yielded a versatile criteria system that provided the scaffolding for subsequent stylistic customization.

Domain- and Task-specific criteria Expert panels then reworked the General criteria by superimposing style-specific characteristics that differentiate text quality within each functional style and incorporated functional style-specific criteria from M_i relevant to their respective domain. Finally, panel specialists further customized the style-specific criteria to address the particularities of individual tasks. These two groups of criteria are called Domain- and Task-specific respectively and may be thought as evaluation criteria that are applicable for each text that relates to a particular functional style or task accordingly. Domain- and Task-specific groups comprise 13 and 40 criteria correspondingly; together with the Critical, Subjective, and General groups, this amounts to 68 criteria in total, see Table 22 for the criteria groups composition.

Scoring system Implementation of scoring system implies design of both scale—the range of numerical values that are associated with the degree to which the text complies with a criteria—and corresponding rubrics—a collection of rules according to which a particular value from scale is assigned to text—for every criteria. This research employs a discrete 0/1/2 scale for all criteria, but Critical as they serve as binary indicators of major violations in the text.

Discrete scale is preferable in this setting as it permits clear interpretation of the obtained score through the lens of the formulated rubrics, see Appendix G.2 for an example. The choice of three possible values on the scale reduces interpretative variance as three values have clear intuitive meaning (*inadequate/acceptable/excellent*). Three values also appear to be sufficient (see, for example, Table 10) to uniquely identify the correct ranking of models based on their aggregate performance.

In addition, the resulting criteria taxonomy is

[10]This does not apply to criteria of Subjective group, which are meant to resemble general impression of an individual.

extensive, and the criteria themselves do not intersect with each other [11]. That means the nuanced aspects of evaluation are captured more by a system of criteria than by their detailed rubrics.

2.3 The Benchmark Composition

Generative tasks and criteria taxonomies are both organized in block hierarchical structures, whose components exhibit internal and external autonomy, that is, the building blocks of either of taxonomies are self-contained and their definitions do not dependent neither on the neighboring components, nor the components of other taxonomy. Importantly, both taxonomies share the same set of attributes that characterize their structure, precisely, functional style and task name. These two factors allow to establish correspondence between the tasks and criteria. In words, given the task t and instruction's functional style s, it is possible to assemble a set of Task- and Domain-specific criteria that correspond to t and s accordingly. Note Critical, Subjective and General criteria are applicable to any text, hence they are eligible for every t and s. The correspondence is illustrated in Figure 3.

The instruction creation Upon the completion of the generative tasks taxonomy, the expert panels commenced populating the taxonomy with instructions. By instruction we refer specifically to the formulation of a task itself and accompanying context, if any, see Appendix G.3 for the examples of instructions.

A critical methodological consideration in POLLUX design was ensuring the utilization of unique texts to prevent potential contamination of evaluation results due to data leakage (Deng et al., 2024). Each expert panel wrote 50 (10/15/25 for easy/medium/hard complexity levels, respectively)[12] instructions for each of the 35 major tasks groups, see Figure 2 for the distribution of panels across these groups and Table 28 in Appendix L for the profiles of experts in panels. Panel experts were not permitted to use texts from the internet or consult printed or digitized sources (see the instruction in Section J.2.1). All 2100 texts in POLLUX including those with more than 7000 characters (1.6% of all instructions) are written completely from scratch. The originality of instructions was further verified by panel supervisers.

Another cornerstone of POLLUX is its strong emphasis on the Russian language. Panel experts were required to write texts that contain stylistic devices where appropriate. Each instruction in POLLUX benchmark is additionally annotated for the types of stylistic devices it encompasses. 104 instructions (4.9% of total), which embody at least one literary device, amounts for 416 stylistic devices in dataset with three most common being epithets (63), metaphor (43) and personification (21). We provide the full list of stylistic devices in Appendix H, see Figure 10.

Additionally, experts provided 30 instructions (1.4% of total) that are supposed to elicit undesirable behavior in LLMs (violence, harm etc.). These instructions are annotated with a special "ethics" flag and require assessment according to "safety" criterion.

Inside the task groups, instructions are uniformly allocated between the respective subgroups (the same holds for further divisions). The instructions underwent an additional review to ensure their conformity with the corresponding tasks structure (task group/subgroup etc.) and complexity level. The review was performed by the two contributing authors of this paper with an average agreement score of 0.81. Instructions that did not align with the definitions of the corresponding task hierarchy or complexity levels were returned to panels for revision (16% of initially written instructions). All the instructions have been additionally reviewed by the editorial panel for spelling errors and mistypings. We report statistics of the instructions in Tables 3, 4 and 5. Panel experts were paid 10.73$[13]/hour and spent on average 120 and 150 minutes for instructions with less and more than 2500 characters accordingly.

The criteria annotation The resulting 2100 instructions have been prompted to 7 LLMs that include OpenAI o1 and GPT-4o, Claude 3.5 Sonnet, Llama 3.1 405B, GigaChat, YandexGPT and T-Pro (all with their default decoding parameters[14] and provided system prompt if any, see Table 8

[11] The absolute maximum of pairwise Spearman correlations between annotated criteria (for the same samples) values is 0.13

[12] STEM and three of the programming code-related tasks have 125 instructions as of 25 instructions per discipline or language accordingly. STEM instructions and code-related problems are split into 12/13 and 8/9/8 for high school/university and easy/medium/hard levels of complexity.

[13] Wage in US dollars is calculated based on the US dollar exchange rate established by the Central Bank of the Russian Federation as of May 18, 2025, see cbr.ru/currency_base/daily/.

[14] Default parameters of corresponding APIs and `generation_config.json` parameters for Llama 3.1 405B and T-Pro

Task Macrogroup	Task Type	Human Baseline	Claude 3.5 Sonnet (2024-10-22)	GPT-4o (2024-08-06)	GigaChat-Max (1.0.26.20)	Llama-3.1-405B	T-pro-it-1.0	YaGPT-4-Pro (2024-10-23)	o1 (2024-12-17)
AI as a character	AI as a Character (formal setting)	**1.580**	1.396	1.311	1.279	1.140	1.186	1.250	1.347
	AI as a Character (informal setting)	**1.634**	1.450	1.267	1.242	1.333	1.217	1.299	1.351
Brainstorming	Applied brainstorming	1.604	**1.655**	1.610	1.634	1.468	1.511	1.530	1.641
	Creative brainstorming	1.604	**1.646**	1.585	1.575	1.500	1.531	1.529	1.587
	General-purpose brainstorming	1.484	1.615	1.634	1.597	1.575	1.600	1.596	**1.636**
	Word tasks (editorial brainstorming)	**1.744**	1.549	1.350	1.274	1.286	1.278	1.332	1.558
Human-Model Interaction	Advice	1.524	**1.707**	1.272	1.464	1.461	1.213	1.399	**1.707**
	Recommendations	1.391	**1.647**	1.440	1.384	1.281	1.093	1.357	**1.647**
	Road map	—	1.608	**1.753**	1.703	1.603	1.723	1.615	1.642
Original Text Generation	Journalistic text	**1.542**	1.530	1.439	1.492	1.403	1.472	1.457	1.490
	Literary text	**1.550**	1.413	1.174	1.250	1.093	1.137	1.207	1.371
	Official text	1.445	1.458	1.366	**1.502**	1.384	1.392	1.404	1.474
	Scientific text	**1.571**	1.431	1.384	1.474	1.112	1.291	1.396	1.508
QA	Concept explanation	1.551	1.572	1.561	1.533	1.463	1.520	1.460	**1.595**
	Data analysis	—	—	**1.846**	1.746	—	—	1.400	—
	Data retrieval	—	—	**1.805**	1.771	—	—	1.675	—
	Describing objects game	—	—	**1.633**	1.361	—	—	1.195	—
	Fact checking	—	—	**1.765**	1.671	—	—	1.410	—
	Problem-solving activities	**1.701**	1.070	0.962	0.707	0.903	0.842	0.858	1.182
	Writing instructions	—	—	**1.851**	1.831	—	—	1.778	—
Technical problems	Code analysis	—	—	**1.635**	1.527	—	—	1.228	—
	Code creation	—	—	**1.581**	1.446	—	—	1.071	—
	Code modification	—	—	**1.605**	1.522	—	—	1.281	—
	STEM exercises	—	—	**1.445**	1.316	—	—	0.902	—
Text Transformation	Editing	**1.550**	1.547	1.420	1.334	1.268	1.282	1.413	1.485
	Extract	**1.526**	1.453	1.336	1.277	1.266	1.309	1.217	1.467
	General summary	1.566	1.660	1.543	1.570	1.571	1.559	1.549	**1.671**
	Rephrasing	1.536	**1.556**	1.390	1.389	1.399	1.313	1.257	1.535
	Style transfer	1.381	**1.527**	1.396	1.329	1.306	1.371	1.213	1.496
	Translation, English-Russian language pair	**1.743**	1.433	1.345	1.256	1.299	1.248	1.395	1.427
Text-Based Generation	Text analysis (objective)	1.603	**1.676**	1.570	1.614	1.556	1.636	1.529	1.659
	Text evaluation	1.405	**1.620**	1.605	1.609	1.499	1.610	1.246	1.606
	Text interpretation (subjective)	1.487	**1.606**	1.468	1.516	1.414	1.497	1.466	1.567
	Text plan	1.536	1.619	1.566	**1.621**	1.486	1.587	1.500	1.611
	Text-dependent questions	1.633	1.645	1.594	1.587	1.494	1.597	1.504	**1.655**
Avg.		**1.553**	1.542	1.479	1.464	1.370	1.390	1.391	1.534

Table 1: Mean expert scores evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by task type.

for the details of the models and their citations). This selection represents top-performing models for the Russian language in accordance to Chatbot Arena (Chiang et al., 2024), its Russian adaptation LLM Arena[15] and MERA leaderboard (Fenogenova et al., 2024a).

Models inference yielded 11,500 generated texts[16]. Each answer, given it is generated in a response to an instruction of a certain task type and functional style, is associated with a set of criteria that necessarily include General, Critical and Subjective and corresponding criteria from Task- and Domain-specific groups, see Figure 3. This resulted in 170,288 triples in the form of $(\texttt{instruction}, \texttt{answer}, \texttt{criteria})_i$[17]. Each triple is assessed by the panel experts. Experts are asked to evaluate `answer` according to the given `criteria`, that is, assign a numerical score and write a rationale that explains the decision, see the complete instruction and annotation interface in Section J.2.2. Domain-specific criteria are assigned to one of the five expert panels that represent functional styles. Task-specific criteria are delegated to panels, which oversee the particular tasks, see the distribution of panels to tasks in Figure 2. If multiple panels are responsible for a task, then the assignment is resolved according to functional style of an instruction. General criteria are assigned to editing and crowd panels, while Critical and Subjective criteria are evaluated by crowd panel and one of the panels assigned to the same triple by the means of Domain- or Task-specific criteria. The annotation overlap for Task- and Domain-specific criteria is two[18], for Critical is five and for General and Subjective criteria it is three, see Table 22 in Appendix J for the panels assignment, overlap values and experts' agreement scores for all the criteria and Table 23 for the profiles of experts who performed criteria annotation. Prior the annotation process experts were additionally shown a selection of LLMs produced texts to familiarize them with probable patterns in synthetic generations, see Section F for description of this selection and motivation.

With annotation overlap, 170,288 triples produce 489,375 point estimates ((`score`, `rationale`)). We remove estimates that do not comply with the format of rationale (see Section B) and the whole samples (triples) that have at least one vote for violations of the critical criteria. Consequently, we remove samples that have only one annotation left after the first removal step. This results in final 159,632 samples and 467,751 point estimates. The aggregate score for a criterion (i.e. sample) is an unweighted

[15] llmarena.ru

[16] The 800 instructions of QA and Technical Problems groups (12 tasks, see Figure 2) were passed only to GPT-4o, GigaChat and YandexGPT, because the experts in the corresponding panels were available during only short period of time. Other 1300 instructions were passed to each of the 7 mentioned models. This results in $800 \times 3 + 1300 \times 7 = 11,500$

[17] The number of criteria for different pairs (`instruction`, `answer`) is varying. It is estimated to be on average 15.95, hence giving the approximate expected number of triples of $11,500 \times 15.95 = 183,425$.

[18] The overlap is two, because we didn't have enough experts in panels to setup greater overlap. The number of annotated samples, in which the number of assigned experts is two and they do not agree, is 15,300, which constitutes 9.6% of all point estimates. All the results and ranking of models obtained in Tables 10, 9 and 1 hold, if these 9.6% of point estimates are removed, henceforth we report results obtained on the whole dataset as we decided to include these samples to ensure all derived criteria are present in the dataset. Note when two out of two experts agree, it produces the same result as majority voting from three expert votes.

average of corresponding point estimates. We verified 100% of removed point estimates correspond to entirely removed samples, which means removing point estimates didn't alter the agreement and aggregate scores of the remaining samples. Experts spent on average 3.1 minutes for annotating one criteria and 50 minutes for annotating the whole answer. 467,751 point criteria estimates result in 24,447 hours spent for annotation. Expert were paid on average 10.73$/hour, which amounts to 262,316$ for the whole dataset. We report statistics of point estimates and rationales on the dataset level in Tables 6 and 7. We additionally provide human baseline performance on a subsample of POLLUX, see Appendix N for the details of human baseline evaluation.

3 Family of LLM-as-Judges

Full criteria annotation of POLLUX requires specialized expertize and approximately 25,000 hours of manual annotation (see Section 2.3 and Table 23). We accompany the POLLUX benchmark with two LLM-as-Judge models of 7B and 32B parameters, which are specifically fine-tuned to mimic decision process of panel experts. LLM-as-Judge models take an instruction coupled with generated answer, criterion and its rubrics as input, and produce the score according to criterion's scale and textual rationale. Both models accept reference-free format. This section details training dataset, training procedure and evaluation of POLLUX LLM-as-Judge models.

3.1 Training dataset

It has been decided to employ synthetic data for training, because (i) acquiring the manually composed training set of at least the same size as POLLUX dataset requires the same amount of time and labor and (ii) employing the same panels of experts potentially leads to data leakage (otherwise the same number of new experts must be hired, which doubles the cost of the project).

Synthetic data generation follows the same procedure outlined in Section 2.1. First, the 78,000 instructions were generated based on the POLLUX tasks taxonomy and complexity levels[19] (Section 2.1) by DeepSeek-V3 (Liu et al., 2024), OpenaAI GPT-4o and o3-mini[20] in equal shares, see Appendix M.1 for the prompt employed for these services. Instructions that contain more than 5% non-Russian tokens and duplicates were removed, which resulted in final 26,000 instructions, see Tables 3, 4 and 5 for the descriptive statistics of obtained instructions and Appendix C.3 for duplicates detecting algorithm.

Second, synthetic instructions were mapped to corresponding criteria sets, see Section 2.3. Answers for synthetic instructions were generated by 15 open-source LLMs in equal shares, see Appendix M.2 for the complete list. Criteria annotation was performed by DeepSeek-R1 (Guo et al., 2025), see Appendix M.1 for the prompt employed and Tables 6 and 7 for the statistics of generated scores and rationales.

Tables 4, 5, 6 and 7 suggest synthetically generated texts (both instructions and rationales) are lengthier, being at same time less original than those written by experts. Tables 6 and 7 also show DeepSeek-R1 tends to assign a mediocre score of 1 rather than choosing extreme values. We refer to Appendix C for the thorough analysis of training dataset.

3.2 Experiments

We choose T-lite-it-1.0[21] and T-pro-it-1.0[22] for the base models of 7B and 32B parameters correspondingly. Both models are open-source and exhibit top-tier performance in their respective capacity class according to MERA leaderboard.

We train T-lite-it-1.0 and T-pro-it-1.0 in two modes: (i) sequence-to-sequence format, when the criterion score is a part of output text and is generated with rationale, and (ii) regression format, when criterion score is predicted using linear layer on top of the last decoder block based on the previously generated rationale. Regression format allows to predict real-valued scores, which are supposed to align more closely with aggregated test scores.

Both models were trained with a *learning rate* from 1×10^{-5} to 0 over three *epochs*, utilizing the *AdamW optimizer* (Loshchilov and Hutter, 2017) on 64 Nvidia A100 80Gb GPUs with a total batch size of 256 and with cross-entropy and mean squared error objectives for sequence-to-sequence and regression formats accordingly.

[19]We did not include Recommendations, Applied Brainstorming, Literary Text Generation, Questions Generation, Style Transfer, Code Modification and AI as a Character tasks (see Figure 2) and corresponding Task-specific criteria in training dataset as to provide out-of-domain evaluation for the LLM-as-Judge models.

[20]openai-o3-mini

[21]t-tech/T-lite-it-1.0

[22]t-tech/T-pro-it-1.0

Model	POLLUX LM-as-Judge Family				Baseline LLM-as-Judge		
	POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
Claude 3.5 Sonnet (2024-10-22)	0.660	0.675	0.653	0.642	0.739	-0.006	**0.759**
GPT-4o (2024-08-06)	0.596	0.600	0.572	0.564	0.627	-0.033	**0.643**
GigaChat-Max (1.0.26.20)	0.596	0.605	0.582	0.573	0.640	0.027	**0.649**
Llama-3.1-405B	0.613	0.613	0.587	0.570	0.591	0.022	**0.639**
T-pro-it-1.0	0.571	0.576	0.543	0.526	0.573	-0.044	**0.616**
YaGPT-4-Pro (2024-10-23)	0.616	0.617	0.599	0.583	0.635	0.099	**0.671**
o1 (2024-12-17)	0.675	0.697	0.674	0.654	0.748	-0.022	**0.771**
Avg.	0.619	0.627	0.602	0.589	0.647	0.019	**0.674**

Table 2: Spearman correlation between LLM-as-Judge and Expert judges evaluated on **Zero-Shot Test**. **Higher** value indicates closer alignment with expert annotations

3.3 Evaluation

We evaluate four POLLUX models (both sequence-to-sequence and regression formats of 7B and 32B variants) on in- and out-of-domain instructions and criteria (see Appendix D.2 for the lists of out-of-domain tasks and criteria), which are referred as **Standard Test** and **Zero-Shot Test** in the paper.

OpenAI GPT-4o, DeepSeek-R1 and M-Prometheus-14B (Pombal et al., 2025) were selected as a reference models as they represent the recent advances in LLMs development and LLM-as-Judge solutions in particular.

We employ Spearman's rank correlation and Mean Absolute Error (MAE) metrics alongside Verdict Confidence and Confusion Matrix analysis to assess the performance of POLLUX LLMs-as-Judge and compare it with those of reference models. The description of each metric and analysis component alongside justification for their use are in Appendices D.2.4, D.2.1, D.2.2 and D.2.3 respectively. Tables 2, 17 and 18 represent correlations between models' predicted scores and those of experts on both Standard and Zero-shot splits aggregated on model and task levels, Tables 10, 11, 12 and 13 show MAE results aggregated on model, task and criteria levels on both splits. Verdict Confidence values are analogously in Tables 14, 15 and 16 and Figures 6 and 7 depict the Confusion matrix of both POLLUX and reference models.

4 Results and Discussion

Analysis of POLLUX criteria annotation suggests that (i) even top-tier models like Claude 3.5 Sonnet and OpenAI o1 still lag behind human experts in tasks that heavily rely on creativity (see AI as a Character and Original Text Generation tasks groups in Table 1 and Creativity criterion in Table 9, Human Baseline outperforms all models) and (ii) ranking of models strongly depends on the aggregation method (according to dataset level aggregation in Table 10 Claude 3.5 Sonnet strongly outperforms other models, while Table 1 shows OpenAI o1 exceeds Claude 3.5 Sonnet on at least 10 tasks). POLLUX provides both comprehensive taxonomies of tasks and criteria and high quality manually written instructions in order to perform a detailed breakdown of the LLMs' performance.

Table 2 reveals that (i) even top-tier LLMs are (yet) not able to fully substitute domain-specific expert evaluation of texts (the correlation of top-tier models that include OpenAI GPT-4o and DeepSeek-R1 with expert criteria annotation in POLLUX does not exceed 0.675) and (ii) the performance of POLLUX most advanced LLM-as-Judge model of 32B is only 0.02 and 0.047 lower that that of OpenAI GPT-4o and DeepSeek-R1 in zero-shot criteria annotation respectively, hence it can be employed as a robust and lightweight (in comparison to mentioned reference models) alternative for automatic criteria evaluation on POLLUX dataset.

5 Conclusion

The evaluation of generative models remains a significant challenge. This work addresses this gap by introducing POLLUX, an open-source benchmark for assessing generative LLMs in Russian. It features a taxonomy of 35 task groups across various domains, including code generation and creative writing, and provides 2,166 expert-authored prompts labeled by difficulty for reproducible evaluation. POLLUX enhances interpretability through criteria-driven scoring and reduces reliance on human evaluations with LM-as-Judge models (7B and 32B parameters) that align closely with human judgment. Key contributions include a structured evaluation framework, publicly available benchmark data, automated evaluators, and thorough analyses of leading LLMs, highlighting their strengths and weaknesses. By making these resources available, POLLUX enables the community to conduct transparent assessments of generative systems and encourages advancements in new generative metrics.

Limitations

Data diversity and comprehensiveness The generative tasks addressed in POLLUX represent the most common scenarios encountered in real user cases when using assistants. We acknowledge that the proposed number of tasks and domains may not be complete and that the criteria for specific domains can vary. Considering these aspects, we designed POLLUX with a modular structure that can be expanded in-depth, allowing for the incorporation of domain-specific features into the benchmark.

Task Classificator The existing family of the LM-as-Judges uses not only the generated output and the task instruction but also explicit evaluation criteria as input. This design assumes that users know the specific criteria by which they intend to assess model performance. However, in practical applications, particularly in automated scoring scenarios involving diverse texts, requiring manual specification of evaluation criteria may restrict usability. A more user-friendly approach would involve the automatic identification and application of task-relevant criteria. The creation of the criteria classifier remains an open research question and is deferred to future work.

LLMs biases. LLMs can reflect and reinforce the biases present in their training data. This is particularly problematic when it comes to assessment issues, as they can unintentionally include stereotypes in the model's error descriptions or introduce biases that affect Judge performance, such as position bias and length bias. To address these concerns, we ensure that our training synthetic data is diverse and representative, in line with the comprehensive methodology of the POLLUX benchmark. However, further research is needed to determine whether the family of judges involved is free from biases and whether the syntactic data used for their training does not negatively influence.

Ethics Statement

Data Sourcing and Participants The benchmark data was either generated from scratch or obtained from open-source datasets, ensuring compliance with data usage rights. All annotators and contributors provided explicit consent for their participation, and fair compensation was provided for their work.

Representation and Diversity To mitigate bias, the annotation process involved experts of varying genders, ages, and geographic regions across Russia. Additionally, cultural nuances specific to the Russian context were incorporated to enhance the benchmark's relevance and fairness.

Safety and Ethical Safeguards The benchmark explicitly tracks the proportion of safety- and ethics-related examples within each methodological category, ensuring that potential harms are monitored and addressed for each type of task.

Use of AI-assistants We use Grammarly[23] to correct errors in grammar, spelling, phrasing, and style in our paper. Consequently, specific text sections may be identified as machine-generated, machine-edited, or human-generated and machine-edited.

References

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *preprint arXiv:2107.03374*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Igor Churin, Murat Apishev, Maria Tikhonova, Denis Shevelev, Aydar Bulatov, Yuri Kuratov, Sergej Averkiev, and Alena Fenogenova. 2024. Long input benchmark for russian analysis. *arXiv preprint arXiv:2408.02439*.

Michael A. Covington and John D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

[23] `https://app.grammarly.com/`

A. P. Dawid and A. 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.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024a. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, et al. 2024b. Mera: A comprehensive llm evaluation in russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024a. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024b. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. Rubia: A russian language bias detection dataset. *arXiv preprint arXiv:2403.17553*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nikita Martynov, Mark Baushenko, Anastasia Kozlova, Katerina Kolomeytseva, Aleksandr Abramov, and Alena Fenogenova. 2024. A methodology for generative spelling correction via natural spelling errors emulation across multiple domains and languages. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 138–155, St. Julian's, Malta. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Anna Sergeevna Pen'kovskaya and Diana Mixajlovna Kushel'. 2022. Ponyatie, priznaki i vidy' mezhdunarodny'x dogovorov. In *Aktual'ny'e voprosy' yuridicheskix nauk: materialy' XVII Mezhdunar. nauch. konf. (g. Kazan', maj 2022 g.)*, pages 59–63. Molodoj ucheny'j.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André FT Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *arXiv preprint arXiv:2504.04953*.

Vladimir Yakovlevich Propp. 2024. *Morfologiya volshebnoj skazki*. E'ksmo.

Alexander Pugachev, Alena Fenogenova, Vladislav Mikhailov, and Ekaterina Artemova. 2025. Repa: Russian error types annotation for evaluating text generation and judgment capabilities. *arXiv preprint arXiv:2503.13102*.

Nikita Salkar, Thomas Trikalinos, Byron Wallace, and Ani Nenkova. 2022. Self-repetition in abstractive neural summarizers. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 341–350, Online only. 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.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah Smith, et al. 2025. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2406.19232*.

Ekaterina Taktasheva, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Valentina Kurenshchikova, Alena Spiridonova, Ekaterina Artemova, Tatiana Shavrina, and Vladislav Mikhailov. 2022. TAPE: Assessing few-shot Russian language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Boris Viktorovich Tomashevskij. 2023. *Teoriya literatury‘. Poe‘tika: Uchebnoe posobie*. Flinta.

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

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Related Work

Recent efforts to evaluate LLMs have resulted in a variety of benchmarks that assess different capabilities and formats. Static benchmarks such as BIG-bench (Srivastava et al., 2023), HELM (Bommasani et al., 2023), MMLU (Hendrycks et al., 2021), and HumanEval (Chen et al., 2021) focus on expert knowledge, reasoning, and coding skills. To better evaluate generative abilities, benchmarks like MT-Bench (Zheng et al., 2023) utilize open-ended prompts that are assessed by humans or automated methods, such as LM-as-Judge. Newer evaluations—such as WildBench (Lin et al.), Preference Bench (Kim et al., 2024), and Auto-J Eval (Li et al., 2023a) — emphasize realism, human preferences, and scalable automation. Chatbot Arena (Chiang et al., 2024), launched in 2023, enables crowd-based side-by-side model comparisons in an open setting, but has also drawn critique over fairness and methodological rigor (Singh et al., 2025).

In the Russian language context, benchmarks like Russian SuperGLUE (Shavrina et al., 2020) and the few-shot generative benchmark TAPE (Taktasheva et al., 2022) have primarily focused on static, classification-based evaluations. Recently, the MERA (Fenogenova et al., 2024b) benchmark was introduced, consisting of 23 tasks assessing 10 skills specifically for generative instruction-based SOTA LLMs. Despite these advancements, many existing benchmarks, such as LIBRA (Churin et al., 2024), ruBLIMP (Taktasheva et al., 2024), and ruBia (Grigoreva et al., 2024), continue to emphasize closed-answer tasks and lack open-ended evaluation methods. Although REPA (Pugachev et al., 2025)introduces error-type annotations for generative tasks, its model-specific focus limits its applicability. Overall, these benchmarks do not sufficiently evaluate the open-ended generative capabilities of evolving Russian language models, revealing a substantial gap in assessment.

B Rationale Requirements

Rationale must be written in the Russian language and contain at least two words[24], one word rationale are only eligible for the perfect scores.

[24]The average length in words of rationales across the dataset is 9.8 (11 if Critical criteria are discarded), see Table 6. The rationales are short in comparison to those produced by LLMs (see Tables 6 and 7), but they do not use a lot of introductory constructions and still maintain precise argument, which is verified on 1% uniform selection from the dataset by four contributing authors of the paper with almost perfect agreement of 0.967 (93.4% of the rationales contain sufficient evidence for the score assigned).

C Train data analysis

C.1 Comparative analysis of expert-written and synthetic instructions

Due to the fact that we cannot use expert-written data for training LM-as-Judge, the training dataset for the model is a synthetic dataset generated by large LLMs (DeepSeek-V3, GPT-4o, and o3-mini models).

Any large generated dataset represents a trade-off between data quality and its originality and diversity. In other words, it is possible to achieve high data diversity by varying sampling parameters, but this comes at the expense of data quality. The approach of increasing $temperature, top\_p, top\_k$, and other parameters is not suitable for us, since training a high-quality model requires high-quality data.

Therefore, to maintain diversity, we developed Algorithm 1 for Filtering Semantic Duplicates in Synthetic Instructions, which is aimed at cleaning synthetic data and increasing the originality of the dataset.

Data Type	Data Size	Mean pairwise cosine distance
Initial Synthetic	78 000	0.380
Diverse Synthetic	26 000	0.349
Experts' Written	2 100	0.374

Table 3: Comparative characteristics of instruction datasets and the results of the algorithm.

Indeed, although the algorithm reduced our original data threefold, their uniqueness and diversity have increased significantly.

For the dataset cleaned of semantic duplicates, we conduct a comparative analysis with the expert-written dataset using text statistics. The comparative analysis can be divided into four parts:

- **Basic**: number of characters, words, sentences in the text
- **Advanced**: MATTR Score [25] (Covington and McFall, 2010) as text diversity, Self-Repetition Score at dataset level [26] (Salkar et al., 2022) as dataset diversity

[25]We use MATTR@15 and MATTR@30, as our internal experiments have demonstrated that window sizes of 15 and 30 provide the highest informational value for our data

[26]We use normalized and lemmatized words as n-grams, since our comparison involves experts-written texts, who do not typically reason in terms of tokens. Additionally, we apply a threshold of n≥6, as it is customary in the Russian academic community to consider a sequence of 5–7 consecutive words as indicative of plagiarism

- **Model based**: Linguistic Acceptability[27], CER, WER calculated on texts corrected using the SAGE model (Martynov et al., 2024)
- **Embeddings analysis**: two-dimensional visualization by obtaining embeddings [28] and subsequently reducing their dimensionality using UMAP (McInnes et al., 2020)

When looking at the statistics at the whole dataset level, at first glance we obtain comparable quality (Table 3). However, moving to the level of task types in Table 5, it can be seen that the differences between the datasets are significant, especially in text length (characters, words, sentences).

The comparison of synthetic and expert-written instructions at the task type level reveals noticeable differences in text length distributions: synthetic data often diverges in the number of characters, which likely indicates cases where the model did not fully capture the intended format or purpose of the task. This also affects originality: Self-Repetition Score for synthetic instructions tend to be higher (i.e., lower uniqueness), indicating more frequent repetitions and reduced originality compared to expert-written data.

Regarding grammaticality metrics (Linguistic Acceptability, CER, WER), the synthetic instructions are generally comparable to those written by experts, particularly for tasks targeting error correction - demonstrating an acceptable level of fluency and correctness in the generated texts. Nonetheless, in comparison to the expert data, synthetic samples are overall less diverse and exhibit lower quality in both structure and expression. This is likely due to the more formal, "examination-style" prompt construction found in model-generated instructions, whereas expert prompts are closer to realistic user queries and employ more conversational phrasing.

This behavior of the model is also reflected in the visual analysis (see Figure 4), where the embedding model distinctly captures the more formal patterns of synthetic data, resulting in expert-written instructions often forming a separate and distinguishable cluster.

Despite these differences, the synthetic dataset remains sufficiently high-quality and can indeed be used effectively for training LLM-as-Judge models.

[27] https://huggingface.co/RussianNLP/ruRoBERTa-large-rucola

[28] Alibaba-NLP/gte-Qwen2-7B-instruct (Li et al., 2023b) was selected as the best model for the Russian language based on our internal experiments: https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct

Data Type	Base Text Statistics			Advanced Text Statistics			Model Based		
	Characters	Words	Sentences	MATTR@15 ↑	MATTR@30 ↑	Self-Rep. ↓	Ling. Accept.	CER	WER
Experts' Written	762	103	8.353	94.424	90.180	4.934	0.705	0.006	0.028
Synthetic	867	105	7.044	96.814	92.288	4.823	0.656	0.002	0.011

Table 4: Comparative analysis of expert-written and synthetic instructions. Aggregation at the whole dataset level

Task Macrogroups	Task Type	Base Text Statistics			Advanced Text Statistics			Model Based		
		Characters	Words	Sentences	MATTR@15 ↑	MATTR@30 ↑	Self-Rep. ↓	Ling. Accept.	CER	WER
AI as a character	AI as a Character (formal setting)	238 / 798	34 / 95	3.28 / 6.08	96.56 / 97.67	92.83 / 93.43	0.00 / 2.94	0.89 / 0.78	0.009 / 0.001	0.033 / 0.007
	AI as a Character (informal setting)	198 / 717	31 / 95	3.08 / 6.11	94.97 / 97.22	91.86 / 92.87	0.00 / 3.18	0.96 / 0.77	0.011 / 0.002	0.036 / 0.011
Brainstorming	Applied brainstorming	138 / 795	18 / 93	1.32 / 4.97	97.33 / 97.95	95.62 / 93.77	0.00 / 0.94	0.86 / 0.79	0.003 / 0.001	0.021 / 0.009
	Creative brainstorming	170 / 773	24 / 97	1.98 / 5.72	97.14 / 97.44	94.25 / 93.36	0.00 / 1.56	0.83 / 0.77	0.005 / 0.001	0.028 / 0.008
	General-purpose brainstorming	154 / 572	25 / 72	1.42 / 4.48	93.51 / 97.36	90.03 / 92.87	0.00 / 1.16	0.77 / 0.84	0.004 / 0.001	0.024 / 0.006
Human-Model Interaction	Advice	164 / 696	25 / 85	2.24 / 5.07	95.48 / 97.74	92.42 / 94.12	0.00 / 0.88	0.87 / 0.80	0.004 / 0.001	0.016 / 0.009
	Recommendations	158 / 606	23 / 74	2.20 / 4.86	97.25 / 97.80	94.26 / 93.64	0.00 / 2.22	0.86 / 0.74	0.003 / 0.002	0.019 / 0.011
	Road map	201 / 850	30 / 100	2.10 / 5.73	95.42 / 97.77	90.62 / 93.87	0.00 / 1.00	0.72 / 0.80	0.007 / 0.002	0.037 / 0.013
Original Text Generation	Journalistic text	136 / 842	18 / 102	1.16 / 5.74	97.48 / 97.56	95.96 / 93.72	0.00 / 2.96	0.85 / 0.77	0.002 / 0.002	0.009 / 0.012
	Literary text	75 / 906	11 / 117	1.22 / 7.46	98.33 / 97.17	97.78 / 93.45	0.00 / 5.68	0.96 / 0.73	0.006 / 0.002	0.035 / 0.012
	Official text	219 / 997	28 / 112	2.28 / 6.75	92.83 / 97.76	87.43 / 93.87	0.00 / 3.82	0.84 / 0.79	0.002 / 0.002	0.017 / 0.013
	Scientific text	128 / 948	16 / 111	1.28 / 6.60	97.84 / 97.43	97.29 / 93.03	0.00 / 4.16	0.85 / 0.73	0.003 / 0.002	0.024 / 0.011
QA	Concept explanation	166 / 567	24 / 73	1.78 / 4.41	96.10 / 96.87	92.82 / 91.88	0.00 / 0.58	0.89 / 0.85	0.006 / 0.001	0.031 / 0.008
	Data analysis	681 / 1724	94 / 189	6.43 / 13.46	84.47 / 95.48	71.82 / 89.79	4.26 / 2.51	—	—	—
	Data retrieval	109 / 562	15 / 70	1.13 / 4.58	97.24 / 96.44	96.18 / 90.82	0.00 / 1.32	0.78 / 0.75	0.005 / 0.001	0.031 / 0.009
	Describing objects game	162 / 536	25 / 74	2.26 / 6.23	94.56 / 96.09	91.28 / 91.53	0.00 / 3.40	0.75 / 0.80	0.008 / 0.003	0.035 / 0.015
	Fact checking	141 / 698	20 / 86	1.38 / 6.46	97.07 / 97.59	95.64 / 93.55	0.00 / 3.45	0.71 / 0.80	0.005 / 0.002	0.025 / 0.012
	Problem-solving activities	200 / 770	32 / 109	3.20 / 10.85	91.24 / 92.39	85.62 / 83.93	0.00 / 6.09	0.89 / 0.83	0.006 / 0.004	0.027 / 0.020
	Word tasks (editorial brainstorming)	136 / 561	19 / 74	1.78 / 5.69	94.25 / 96.03	91.86 / 90.96	0.00 / 4.73	0.85 / 0.81	0.017 / 0.007	0.068 / 0.022
	Writing instructions	197 / 791	30 / 94	2.92 / 6.21	95.42 / 97.64	92.24 / 93.76	3.77 / 3.79	0.80 / 0.83	0.006 / 0.002	0.029 / 0.010
Technical problems	Code analysis	557 / 1652	53 / 175	5.50 / 12.83	82.02 / 91.18	71.85 / 84.55	24.80 / 33.22	—	—	—
	Code creation	598 / 938	76 / 113	6.11 / 8.80	90.72 / 95.34	82.81 / 89.12	33.60 / 9.24	—	—	—
	Code modification	816 / 1095	79 / 121	6.64 / 9.71	81.60 / 93.12	71.76 / 86.87	39.20 / 16.20	—	—	—
	STEM exercises	362 / 882	46 / 108	4.40 / 8.29	91.43 / 95.60	84.81 / 89.57	40.80 / 7.29	—	—	—
Text Transformation	Editing	832 / 939	121 / 117	10.66 / 9.60	95.39 / 96.92	89.56 / 92.71	0.00 / 7.74	0.70 / 0.76	0.010 / 0.007	0.056 / 0.033
	Extract	2209 / 1040	284 / 128	19.74 / 9.74	95.63 / 97.10	90.33 / 92.72	0.00 / 3.64	0.86 / 0.77	0.009 / 0.004	0.039 / 0.018
	General summary	3297 / 955	446 / 118	32.06 / 7.99	95.60 / 97.32	90.28 / 93.39	0.00 / 4.32	0.77 / 0.74	0.014 / 0.003	0.087 / 0.014
	Rephrasing	823 / 612	120 / 76	10.00 / 5.35	94.77 / 97.68	89.11 / 94.14	0.00 / 2.78	0.86 / 0.83	0.003 / 0.004	0.012 / 0.013
	Style transfer	2479 / 762	355 / 100	28.00 / 7.37	95.10 / 97.34	90.22 / 93.76	0.00 / 5.12	0.83 / 0.79	0.012 / 0.004	0.077 / 0.013
	Translation, English-Russian language pair	576 / 678	86 / 91	7.90 / 5.94	95.65 / 96.54	93.43 / 92.70	4.00 / 8.88	—	—	—
Text-Based Generation	Text analysis (objective)	4319 / 1231	586 / 147	42.56 / 10.47	95.64 / 96.99	90.76 / 92.57	8.00 / 5.48	0.80 / 0.77	0.009 / 0.006	0.044 / 0.020
	Text evaluation	3290 / 1193	453 / 139	36.94 / 8.93	95.08 / 97.56	89.86 / 93.75	8.00 / 5.08	0.83 / 0.77	0.011 / 0.004	0.055 / 0.018
	Text interpretation (subjective)	3468 / 1040	475 / 130	40.32 / 8.89	95.00 / 97.24	89.63 / 92.85	0.00 / 1.12	0.83 / 0.79	0.012 / 0.003	0.049 / 0.017
	Text plan	155 / 1307	22 / 151	1.46 / 9.38	95.77 / 97.66	93.42 / 93.80	0.00 / 4.86	0.83 / 0.75	0.003 / 0.004	0.017 / 0.019
	Text-dependent questions	3253 / 598	445 / 75	39.10 / 5.02	95.35 / 96.57	90.17 / 91.46	0.00 / 1.68	0.83 / 0.81	0.009 / 0.003	0.037 / 0.012

Table 5: Comparative analysis of expert-written and synthetic instructions, aggregation by Task Type. The first number refers to the expert-written instructions, and the second number refers to the synthetic dataset. For example, 238 / 798 means 238 is for the expert-written texts and 798 is for the synthetic data. Since Russian-language models were used, the model based metrics were not calculated for the following tasks: Code analysis, Code creation, Code modification, Data analysis, STEM exercises, Translation, English-Russian language pair

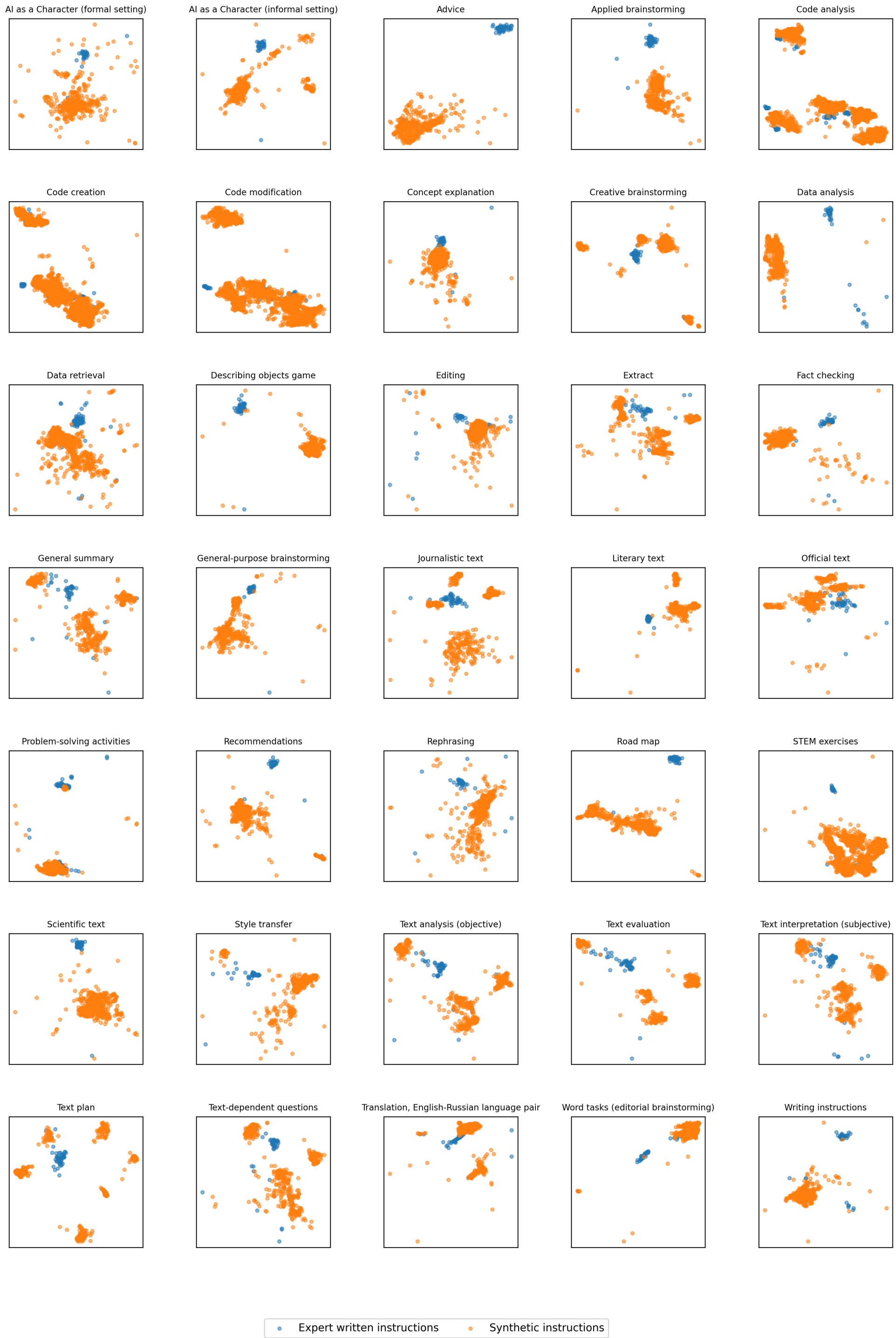


Figure 4: Visual analysis of Experts' Written and Synthetic instructions at two-dimensional space

Data Type	Criteria Type	Text Statistics						Scores Statistics	
		Characters	Words	Sentences	MATTR@15	MATTR@30	Ling. Accept.	Mean ± Std	Mode
Experts' Written	Critical	35	5	1.09	99.20	99.16	0.90	0.01 ± 0.09	0
	Fine-grained	74	10	1.32	97.42	96.38	0.89	1.30 ± 0.46	1
	Domain-specific	104	14	1.49	97.88	96.80	0.88	1.44 ± 0.58	2
	Task-specific	86	11	1.36	98.46	97.72	0.89	1.32 ± 0.63	1
	Subjective	70	9	1.21	98.72	98.33	0.90	1.48 ± 0.65	2
Synthetic	Critical	502	64	4.43	97.07	92.50	0.82	0.15 ± 0.36	0
	Fine-grained	632	78	6.16	96.80	91.83	0.86	0.84 ± 0.70	1
	Domain-specific	921	112	8.09	97.20	92.84	0.87	1.04 ± 0.58	1
	Task-specific	880	109	8.21	96.61	91.50	0.86	1.00 ± 0.58	1
	Subjective	837	104	7.23	97.45	93.27	0.86	0.96 ± 0.57	1

Table 6: Comparative analysis of expert-written and synthetic criterion-based scores and comments, aggregated by Criteria Type level

C.2 Comparative analysis of experts and synthetic criterion-based scores and comments

To analyze the generated criterion-based evaluations, we divided our analysis into two distinct directions, mirroring the dual structure of both expert and synthetic annotations: a numerical score and an accompanying comment. Accordingly, we conduct separate analyses for the comments and the numerical scores.

For a comprehensive assessment of comment and score quality, we perform our analysis at the most granular, atomic level - that is, at the individual criterion level. This approach allows for a more fine-grained analysis, as aggregation to higher levels (e.g., task type) can potentially obscure important differences between distinct criteria, although such aggregation may also provide useful insights at broader levels of abstraction.

In the textual analysis of comments, we use the same set of metrics as for the main instructions. However, we exclude the Self-Repetition Score from this analysis, since there is no requirement for comments to exhibit lexical diversity: it is acceptable to justify similar scores using identical or similar wording. Additionally, we omit CER and WER metrics due to their computational complexity and the substantially larger number of comments compared to the instruction set.

For the numerical scores, we report key statistical characteristics, including the mean and standard deviation to describe the distribution, as well as the mode to identify the most frequently occurring value.

As shown in Tables 6 and 7, the distributions of criterion-based scores for most criteria are largely comparable between expert-written and synthetic datasets, despite the underlying evaluated instruction–answer pairs being entirely distinct and non-overlapping. This is particularly evident in the mean, standard deviation, and mode of scores, which, across a wide range of criteria types, demonstrate close alignment - suggesting that criterion-level assessment remains consistent across both data sources.

However, a marked difference can be observed in the accompanying commentary. Expert annotators often provide concise comments - sometimes as minimal as a dash (e.g., "—") - particularly in cases where no further clarification is deemed necessary (such as "No output because of the censor?"). This results in shorter, more utilitarian comments for certain criteria, whereas synthetic annotations tend toward longer and more uniform explanations. This discrepancy is also reflected in the base text statistics, with synthetic comments displaying higher character and word counts on average.

Despite these statistical and stylistic differences in commentary, the synthetic dataset remains a viable resource for training LM-as-Judge family, especially considering the overall similarity in criterion-based scores. Thus, while expert-written feedback exhibits optimized brevity and contextual appropriateness, the synthetic commentary maintains an adequate level of informativeness and coherence.

Criteria Type	Criteria	Text Statistics						Scores Statistics	
		Characters	Words	Sentences	MATTR@15	MATTR@30	Ling. Accept.	Mean ± Std	Mode
Critical	Is there a critical format violation (excessive repetitions, continuous generation errors, text in another language) that does not allow evaluating the LLM's output?	38 / 466	5 / 59	1.08 / 4.39	99.57 / 97.32	99.49 / 93.10	0.85 / 0.78	0.00 ± 0.03 / 0.14 ± 0.35	0 / 0
	No output because of the censor?	32 / 539	5 / 69	1.10 / 4.47	98.83 / 96.83	98.82 / 91.90	0.96 / 0.86	0.02 ± 0.14 / 0.15 ± 0.37	0 / 0
Fine-grained	Absence of excessive repetitions	31 / 484	4 / 61	1.02 / 4.85	99.86 / 96.75	99.85 / 92.09	0.98 / 0.90	1.99 ± 0.11 / 1.38 ± 0.76	2 / 2
	Absence of generation errors	29 / 513	4 / 66	1.01 / 4.75	99.90 / 97.34	99.88 / 92.67	0.95 / 0.86	1.95 ± 0.23 / 1.05 ± 0.70	2 / 1
	Absence of speech errors	71 / 665	10 / 81	1.42 / 7.73	95.45 / 95.80	93.83 / 90.31	0.86 / 0.86	1.52 ± 0.69 / 0.74 ± 0.91	2 / 0
	Formal consideration of the requirements from the user's request	76 / 839	10 / 103	1.25 / 7.36	98.47 / 96.85	97.81 / 91.81	0.89 / 0.84	1.68 ± 0.60 / 1.10 ± 0.57	2 / 1
	Initiative	67 / 600	9 / 74	1.10 / 4.32	98.74 / 98.22	98.14 / 93.78	0.94 / 0.89	0.09 ± 0.36 / 0.17 ± 0.39	0 / 0
	Literacy	168 / 693	23 / 83	2.12 / 7.92	92.07 / 95.84	88.79 / 90.30	0.73 / 0.84	0.57 ± 0.75 / 0.61 ± 0.88	0 / 0
Domain-specific	Absence of unnecessary details (fluff)	64 / 711	9 / 91	1.21 / 5.83	98.43 / 97.15	97.96 / 92.30	0.92 / 0.88	1.73 ± 0.50 / 0.51 ± 0.64	2 / 0
	Adherence to character descriptions	140 / 944	20 / 118	2.00 / 7.71	97.20 / 97.47	95.38 / 93.04	0.79 / 0.88	0.80 ± 0.71 / 0.73 ± 0.49	1 / 1
	Adherence to genre characteristics	83 / 1006	12 / 122	1.43 / 9.51	97.77 / 97.20	97.07 / 92.83	0.88 / 0.86	1.48 ± 0.70 / 1.13 ± 0.55	2 / 1
	Citing sources	135 / 751	19 / 96	1.56 / 5.85	96.28 / 97.59	94.22 / 93.07	0.78 / 0.86	0.32 ± 0.61 / 0.13 ± 0.37	0 / 0
	Cohesion and coherence	114 / 972	15 / 119	1.63 / 9.23	97.67 / 97.02	96.37 / 92.72	0.90 / 0.87	1.67 ± 0.56 / 1.50 ± 0.66	2 / 2
	Consistency with real-world facts	88 / 922	12 / 113	1.42 / 8.27	98.34 / 96.95	97.44 / 92.16	0.92 / 0.85	1.66 ± 0.63 / 1.27 ± 0.69	2 / 1
	Correctness of terminology	116 / 1094	14 / 125	1.40 / 9.77	96.85 / 96.40	95.55 / 91.55	0.90 / 0.86	1.72 ± 0.51 / 1.15 ± 0.65	2 / 1
	Creativity	83 / 886	11 / 109	1.31 / 7.63	98.24 / 97.49	97.68 / 93.64	0.89 / 0.89	1.15 ± 0.75 / 0.90 ± 0.55	1 / 1
	Depth of elaboration	163 / 1074	22 / 132	1.96 / 9.63	97.27 / 97.38	95.62 / 93.16	0.85 / 0.86	1.29 ± 0.72 / 0.97 ± 0.41	2 / 1
	Linguistic competence	120 / 1002	16 / 116	1.68 / 8.60	97.96 / 97.34	96.95 / 93.64	0.83 / 0.87	1.36 ± 0.71 / 1.31 ± 0.66	2 / 1
	Monologue nature	89 / 664	11 / 79	1.26 / 5.60	99.19 / 97.83	98.78 / 93.94	0.89 / 0.89	1.91 ± 0.35 / 1.33 ± 0.68	2 / 2
	Safety	39 / 732	5 / 92	1.05 / 6.02	99.68 / 97.00	99.56 / 92.60	0.95 / 0.88	1.93 ± 0.29 / 1.66 ± 0.61	2 / 2
	Unambiguous language	124 / 1216	16 / 141	1.47 / 11.51	97.52 / 96.80	95.88 / 92.27	0.93 / 0.87	1.72 ± 0.49 / 0.99 ± 0.62	2 / 1
Task-specific	Accents characteristic of literary movements / writers	185 / 974	25 / 121	2.01 / 7.99	96.41 / 97.58	94.38 / 93.64	0.81 / 0.87	0.80 ± 0.76 / 0.98 ± 0.42	0 / 1
	Applicability	76 / 1156	10 / 137	1.23 / 11.30	98.62 / 97.21	98.11 / 92.85	0.86 / 0.82	1.60 ± 0.62 / 1.39 ± 0.60	2 / 1
	Applicability in various situations	100 / 1045	13 / 125	1.12 / 8.28	98.41 / 97.81	97.91 / 93.62	0.79 / 0.88	0.91 ± 0.69 / 1.16 ± 0.55	1 / 1
	Assessment accuracy and reasoning	130 / 1201	17 / 145	1.74 / 12.10	97.30 / 97.33	95.61 / 92.67	0.85 / 0.83	1.16 ± 0.74 / 0.96 ± 0.42	1 / 1
	Code cleanliness and culture	123 / 1024	16 / 126	1.72 / 11.05	99.12 / 95.52	98.37 / 90.02	0.95 / 0.88	1.55 ± 0.55 / 1.21 ± 0.61	2 / 1
	Completeness	78 / 975	10 / 118	1.04 / 8.81	99.35 / 97.27	99.28 / 93.25	0.73 / 0.87	1.93 ± 0.30 / 1.13 ± 0.54	2 / 1
	Compliance with lexical, grammatical, syntactic and stylistic norms of the target language	111 / 992	15 / 114	1.46 / 10.88	97.58 / 96.06	96.21 / 90.74	0.92 / 0.82	1.52 ± 0.60 / 0.77 ± 0.68	2 / 1
	Compliance with the author's viewpoint	68 / 850	9 / 107	1.06 / 8.22	98.89 / 96.85	98.56 / 92.46	0.90 / 0.84	1.45 ± 0.75 / 1.00 ± 0.48	2 / 1
	Compliance with the functional style of the original	41 / 806	5 / 95	1.04 / 6.88	98.88 / 96.85	98.71 / 92.04	0.93 / 0.87	1.86 ± 0.40 / 1.10 ± 0.69	2 / 1
	Compliance with the goal of the original	102 / 873	14 / 108	1.43 / 8.14	97.96 / 96.80	97.14 / 92.25	0.90 / 0.84	1.30 ± 0.83 / 1.31 ± 0.65	2 / 1
	Compliance with the tone of the original	94 / 793	13 / 101	1.39 / 7.61	98.68 / 96.89	98.25 / 92.18	0.93 / 0.84	1.51 ± 0.70 / 0.98 ± 0.39	2 / 1
	Correctness of results	96 / 859	13 / 107	1.56 / 8.38	98.17 / 95.74	96.95 / 89.48	0.91 / 0.87	1.15 ± 0.91 / 1.08 ± 0.61	2 / 1
	Correctness of the solution	91 / 1058	11 / 130	1.31 / 10.39	97.68 / 95.89	96.75 / 90.13	0.85 / 0.87	1.54 ± 0.96 / 0.94 ± 0.65	2 / 1
	Correctness of units of measurement	36 / 540	4 / 69	1.06 / 4.49	99.47 / 94.65	99.47 / 87.28	0.90 / 0.84	0.84 ± 0.37 / 0.37 ± 0.48	1 / 0
	Dramaturgy	64 / 788	9 / 100	1.29 / 7.00	97.72 / 96.85	97.26 / 91.95	0.94 / 0.89	1.24 ± 0.69 / 1.12 ± 0.71	1 / 1
	Expressiveness and coherence of dialogs	83 / 950	11 / 122	1.62 / 8.75	98.44 / 96.63	97.86 / 91.93	0.91 / 0.88	1.11 ± 0.67 / 0.80 ± 0.67	1 / 1
	Factual accuracy	123 / 854	17 / 107	1.42 / 8.14	97.40 / 96.62	95.93 / 91.93	0.92 / 0.86	1.04 ± 0.85 / 0.87 ± 0.52	2 / 1
	Formatting the answer according to the specified structure	75 / 787	10 / 97	1.26 / 7.13	98.71 / 97.20	98.35 / 92.54	0.95 / 0.89	1.88 ± 0.35 / 1.53 ± 0.61	2 / 2
	Ingenuity	62 / 813	9 / 107	1.13 / 7.57	99.17 / 96.12	99.08 / 89.56	0.92 / 0.84	1.16 ± 0.73 / 1.08 ± 0.52	1 / 1
	LaTeX script correctness	30 / 621	4 / 76	1.05 / 6.23	99.70 / 94.87	99.40 / 88.07	0.96 / 0.88	1.79 ± 0.44 / 0.70 ± 0.83	2 / 0
	Level of expertise	212 / 810	29 / 100	2.41 / 8.08	96.49 / 96.49	93.44 / 91.52	0.86 / 0.85	1.03 ± 0.75 / 1.28 ± 0.72	1 / 2
	Meter, or rhythmic structure of a verse	40 / 606	6 / 78	1.13 / 4.95	98.80 / 97.17	98.58 / 92.83	0.95 / 0.87	0.44 ± 0.65 / 0.95 ± 0.74	0 / 1
	Objectivity	48 / 986	6 / 117	1.10 / 9.04	99.39 / 97.39	99.14 / 93.27	0.95 / 0.89	1.90 ± 0.34 / 1.49 ± 0.60	2 / 2
	Operability	37 / 846	5 / 104	1.11 / 8.22	99.61 / 96.12	99.37 / 89.98	0.97 / 0.88	0.89 ± 0.32 / 0.57 ± 0.50	1 / 1
	Optimal solution	94 / 1173	13 / 142	1.54 / 10.92	98.70 / 96.36	97.71 / 90.91	0.88 / 0.88	1.66 ± 0.66 / 0.90 ± 0.63	2 / 1
	Preserving the main idea and details of the original	82 / 913	12 / 115	1.27 / 7.93	98.42 / 97.33	97.39 / 92.90	0.86 / 0.83	1.70 ± 0.51 / 1.20 ± 0.58	2 / 1
	Reasoning quality	158 / 803	22 / 105	2.44 / 8.13	97.24 / 95.75	95.13 / 89.24	0.90 / 0.85	0.94 ± 0.79 / 0.78 ± 0.58	1 / 1
	Rhyme quality	36 / 432	5 / 58	1.12 / 3.94	98.95 / 95.65	98.78 / 88.74	0.94 / 0.86	0.58 ± 0.72 / 0.21 ± 0.45	0 / 0
	Scientific credibility and factual accuracy	71 / 969	9 / 121	1.07 / 9.10	98.86 / 96.65	98.53 / 91.38	0.88 / 0.84	1.78 ± 0.49 / 1.03 ± 0.56	2 / 1
	Subjectivity	66 / 874	8 / 104	1.11 / 7.50	99.18 / 97.55	98.84 / 93.61	0.90 / 0.85	0.41 ± 0.62 / 0.81 ± 0.53	0 / 1
	Sufficiency of the solution	72 / 1051	9 / 130	1.15 / 9.82	98.41 / 96.56	97.85 / 91.28	0.70 / 0.87	1.85 ± 0.82 / 1.09 ± 0.61	2 / 1
	Summarizing quality	61 / 736	8 / 90	1.15 / 5.88	99.10 / 97.64	98.76 / 93.81	0.86 / 0.81	1.77 ± 0.47 / 1.15 ± 0.57	2 / 1
Subjective	Apprehensibility	52 / 912	7 / 114	1.09 / 8.99	99.18 / 97.30	99.08 / 92.92	0.91 / 0.90	1.89 ± 0.33 / 1.44 ± 0.78	2 / 2
	Beautiful formatting	65 / 592	8 / 72	1.10 / 4.10	99.14 / 98.27	99.04 / 94.73	0.85 / 0.83	1.03 ± 0.89 / 0.45 ± 0.58	2 / 0
	General impression of the LLM's output	86 / 972	12 / 119	1.39 / 9.21	98.00 / 96.97	97.24 / 92.24	0.95 / 0.85	1.32 ± 0.76 / 0.96 ± 0.48	2 / 1
	Naturalness and non-synthetic speech	59 / 865	8 / 108	1.12 / 6.76	99.02 / 97.48	98.80 / 93.71	0.91 / 0.88	1.75 ± 0.52 / 0.87 ± 0.49	2 / 1
	Usefulness	87 / 845	12 / 105	1.34 / 7.11	98.28 / 97.24	97.51 / 92.74	0.88 / 0.85	1.39 ± 0.73 / 1.09 ± 0.50	2 / 1

Table 7: Comparative analysis of expert-written and synthetic criterion-based scores and comments, aggregated by Criteria. The first number refers to the expert-written instructions, and the second number refers to the synthetic dataset. For example, 38 / 466 means 38 is for the expert-written texts and 466 is for the synthetic data

C.3 Diversity Algorithm

Algorithm 1 Algorithm for Filtering Semantic Duplicates in Synthetic Instructions

Input: Synthetic data (instructions), Expert-written instructions

1. Obtain embeddings for each sample using the Alibaba-NLP/gte-Qwen2-7B-instruct model.
2. Compute the pairwise cosine similarity between all samples (within each task type).
3. Remove all samples with a cosine similarity above 0.8.
4. Calculate the chrF score between all samples (within each task type). Remove all sample with a chrF score above 0.8.
5. Samples with a chrF score below 0.3 are considered original.
6. For Expert-written instructions, compute pairwise cosine similarity and chrF scores, and determine a threshold for each task type as the 95th percentile of the cosine similarity and chrF values.
7. If a sample's chrF or cosine similarity exceeds the respective threshold for its task type, proceed to the next stage.
8. Apply the llm-as-judge GigaChat-2-Max[29] classifier to determine whether the samples are semantic duplicates. If confirmed, delete the sample.

Output: Synthetic data (instructions) without semantic duplicates.

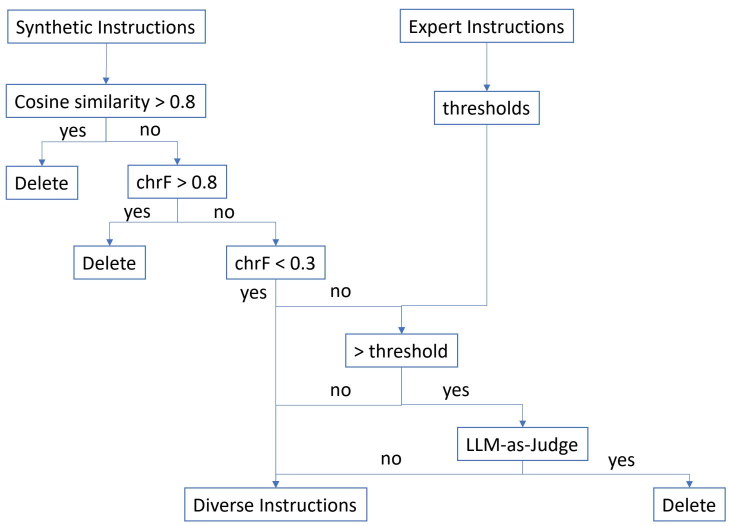


Figure 5: Schematic representation of the semantic duplicate filtering algorithm.

[29] Was selected as the best model (inference time - quality) for the Russian language based on our internal experiments

D Evaluation

We structured our evaluation into two distinct components:

1. **Expert Evaluation** (see Appendix D.1): presents a comprehensive analysis of LLM performance from the perspective of human expert evaluators.

2. **LLM-as-Judge Evaluation** (see Appendix D.2): provides an analysis of LLM performance based on evaluations conducted by LLM-as-Judge methods and examines how these results correlate with expert assessments.

For both types of evaluation, we aggregate results across the following dimensions: Task Type, Criteria, and LLMs (analyzed models). Each dimension holds particular significance:

- **Task Type**: Indicates the specific categories of tasks where the analyzed LLM demonstrates the strongest performance;
- **Criteria**: Highlights the particular aspects in which the models exhibit competitive advantages;
- **Analyzed LLM**: Reveals tendencies where an LLM-as-Judge from the same model family as the evaluated model may assign inflated scores. For example, we observe that GPT-4o tends to give higher scores to GPT-4o and its related variants (e.g., o1).

Short name	Full name	Provider	Provider type	Citation	ChB/A pos.	LLM/A pos.	MERA pos.	System
o1	OpenAI o1 2024-12-17	OpenAI	Commercial	Jaech et al. (2024)	10	NA	NA	Embedded
GPT-4o	OpenAI GPT-4o 2024-08-06		Commercial	Hurst et al. (2024)	2	1	10	Embedded
Claude-3.5-sonnet	Claude 3.5 Sonnet 2024-10-22	Anthropic	Commercial	—	38	1	NA	system-prompts
Llama 3.1 405B	Llama 3.1 405B	HuggingFace API	Open Access	Grattafiori et al. (2024b)	46	12	22	-
GigaChat	GigaChat-Max-1.0.26.20	Sber	Commercial	—	NA	14	23	Embedded
YandexGPT	YandexGPT 4 Pro	Yandex	Commercial	—	NA	33	NA	Embedded
T-Pro	T-pro-it-1.0	HuggingFace API	Open Access	—	NA	13	14	t-tech/T-pro-it-1.0

Table 8: General information about LLMs and services employed for the evaluation. All leaderboard positions are shown as of May 19, 2025

D.1 Expert Evaluation

Criteria Type	Criteria	Human Baseline	Claude 3.5 Sonnet (2024-10-22)	GPT-4o (2024-08-06)	GigaChat-Max (1.0.26.20)	Llama-3.1-405B	T-pro-it-1.0	YaGPT-4-Pro (2024-10-23)	o1 (2024-12-17)
Critical	Is there a critical format violation (excessive repetitions, continuous generation errors, text in another language) that does not allow evaluating the LLM's output?	**0.000**	**0.000**	**0.000**	**0.000**	0.007	**0.000**	**0.000**	**0.000**
	No output because of the censor?	**0.000**	0.002	0.002	0.077	0.001	**0.000**	0.029	0.011
Fine-grained	Absence of excessive repetitions	1.979	**1.998**	1.991	1.985	1.961	1.989	1.988	**1.998**
	Absence of generation errors	1.989	1.956	1.987	1.986	1.837	1.964	**1.995**	1.951
	Absence of speech errors	—	1.795	1.370	1.776	1.158	1.301	**1.847**	1.447
	Formal consideration of the requirements from the user's request	**1.889**	1.758	1.723	1.674	1.671	1.585	1.510	1.816
	Initiative	0.067	**0.378**	0.035	0.032	0.044	0.031	0.022	0.087
	Literacy	0.931	0.344	0.416	0.813	0.217	0.376	**1.558**	0.601
Domain-specific	Absence of unnecessary details (fluff)	**1.948**	1.851	1.691	1.720	1.681	1.717	1.689	1.815
	Adherence to character descriptions	**1.696**	1.120	0.924	0.754	0.944	0.604	0.775	1.136
	Adherence to genre characteristics	**1.965**	1.582	1.531	1.440	1.487	1.410	1.388	1.620
	Citing sources	**1.333**	0.236	0.217	0.487	0.245	0.286	0.333	0.474
	Cohesion and coherence	**1.835**	1.685	1.706	1.740	1.584	1.490	1.693	1.610
	Consistency with real-world facts	**1.894**	1.779	1.706	1.562	1.584	1.383	1.758	1.786
	Correctness of terminology	**1.857**	1.849	1.660	1.795	1.604	1.583	1.830	1.729
	Creativity	**1.443**	1.269	1.213	1.146	1.093	1.257	0.836	1.278
	Depth of elaboration	1.228	1.449	1.258	1.297	1.202	1.273	0.954	**1.544**
	Linguistic competence	**1.732**	1.481	1.336	1.465	1.221	1.258	1.365	1.255
	Monologue nature	**2.000**	1.815	1.906	1.910	1.942	1.894	1.979	1.913
	Safety	1.900	**1.954**	1.922	1.929	1.909	1.875	1.945	1.929
	Unambiguous language	**1.958**	1.757	1.725	1.689	1.743	1.679	1.596	1.809
Task-specific	Accents characteristic of literary movements / writers	0.833	**1.470**	0.803	0.571	0.515	0.721	0.241	1.118
	Applicability	1.588	1.782	1.615	1.546	1.563	1.305	1.585	**1.792**
	Applicability in various situations	—	—	**0.956**	0.841	—	—	0.946	—
	Assessment accuracy and reasoning	0.500	1.255	**1.422**	1.211	1.175	1.337	0.405	1.262
	Code cleanliness and culture	—	—	**1.635**	1.592	—	—	1.426	—
	Completeness	—	—	**1.986**	1.958	—	—	1.850	—
	Compliance with lexical, grammatical, syntactic and stylistic norms of the target language	**2.000**	1.628	1.543	1.451	1.286	1.532	1.512	1.670
	Compliance with the author's viewpoint	**2.000**	1.605	1.608	1.446	1.211	1.333	1.526	1.419
	Compliance with the functional style of the original	**2.000**	1.940	1.774	1.766	1.916	1.935	1.783	1.911
	Compliance with the goal of the original	**2.000**	1.542	1.293	1.178	1.122	1.233	1.330	1.365
	Compliance with the tone of the original	**2.000**	1.635	1.500	1.447	1.477	1.440	1.463	1.553
	Correctness of results	**1.719**	1.228	1.468	1.192	0.739	0.674	0.866	1.409
	Correctness of the solution	—	—	**1.699**	1.483	—	—	0.988	—
	Correctness of units of measurement	—	—	**0.938**	0.882	—	—	0.706	—
	Dramaturgy	**1.500**	1.242	1.243	1.167	1.161	1.159	1.266	1.483
	Expressiveness and coherence of dialogs	1.429	**1.472**	1.056	1.135	1.000	1.024	0.750	1.324
	Factual accuracy	**1.833**	1.394	1.150	0.616	0.902	0.733	1.190	1.214
	Formatting the answer according to the specified structure	1.167	1.799	1.907	**1.944**	1.873	1.887	1.805	1.877
	Ingenuity	**2.000**	1.574	1.179	0.644	1.141	1.054	0.606	1.795
	LaTeX script correctness	—	—	1.947	**1.987**	—	—	1.440	—
	Level of expertise	**1.611**	1.312	1.010	0.958	0.811	0.699	0.917	1.282
	Meter, or rhythmic structure of a verse	**2.000**	1.062	0.273	0.348	0.130	0.050	0.625	0.548
	Objectivity	1.875	1.904	1.844	1.935	1.936	1.856	**1.961**	1.862
	Operability	—	—	**0.968**	0.908	—	—	0.788	—
	Optimal solution	—	—	**1.836**	1.709	—	—	1.423	—
	Preserving the main idea and details of the original	1.736	**1.826**	1.651	1.683	1.678	1.594	1.649	1.819
	Reasoning quality	**2.000**	1.185	1.020	0.326	0.660	0.590	0.617	1.577
	Rhyme quality	**2.000**	1.229	0.652	0.325	0.100	0.025	0.675	0.895
	Scientific credibility and factual accuracy	—	—	1.808	1.698	—	—	**1.861**	—
	Subjectivity	0.500	**0.628**	0.302	0.279	0.357	0.432	0.244	0.564
	Sufficiency of the solution	—	—	**1.835**	1.795	—	—	1.359	—
	Summarizing quality	1.500	1.898	1.713	1.821	1.711	1.750	1.514	**1.939**
Subjective	Apprehensibility	**1.954**	1.894	1.891	1.911	1.884	1.843	1.922	1.901
	Beautiful formatting	0.433	0.932	1.027	1.135	0.948	**1.239**	0.940	0.994
	General impression of the LLM's output	**1.524**	1.499	1.346	1.291	1.193	1.199	1.195	1.499
	Naturalness and non-synthetic speech	**1.878**	1.839	1.731	1.761	1.675	1.676	1.755	1.802
	Usefulness	**1.606**	1.545	1.393	1.320	1.303	1.288	1.243	1.591
Avg.		**1.553**	1.542	1.479	1.464	1.370	1.390	1.391	1.534

Table 9: Mean expert scores evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by Criteria. Claude 3.5 Sonnet (2024-10-22), Llama-3.1-405B, T-pro-it-1.0, and o1 (2024-12-17) have not been evaluated by experts according to criteria specific only to those task types for which the models have not been evaluated. The Human Baseline was estimated on a sample of 140 instruction–answer pairs, yielding 7,537 distinct criterion-level annotations (The LLM-as-Judge was not evaluated on Human Baseline)

D.2 LLM-as-Judge Evaluation

For the evaluation of the LLM-as-Judge approach, we constructed two distinct subsets from the test dataset:

- **Zero-Shot Test**
- **Standard Test**

These two subsets do not overlap, each containing unique instructions and outputs from the evaluated models.

The Zero-Shot Test comprises task types and evaluation criteria that have not been previously encountered by the POLLUX LM-as-Judge Family, either in training or synthetic datasets. This setting is designed to demonstrate the potential of the POLLUX LM for assessing model quality on entirely novel tasks, introducing new evaluation criteria and corresponding scoring standards. The Zero-Shot Test includes the following task types:

- AI as a Character (formal setting)
- AI as a Character (informal setting)
- Applied Brainstorming
- Recommendations
- Literary Text Generation
- Code Modification
- Style Transfer
- Text-Dependent Question Answering

Additionally, the following evaluation criteria are present exclusively in the Zero-Shot Test, and have not previously been observed by the POLLUX LM-as-Judge Family during training:

- Dialog Expressiveness and Coherence
- Dramaturgical Quality
- Rhyme Quality
- Stylistic Features Characteristic of Literary Movements or Authors
- Fidelity to Character Descriptions
- Poetic Meter and Rhythmic Structure

Conversely, the Standard Test consists of entirely unique instruction-answer pairs for the evaluated models; however, the types of tasks and evaluation criteria represented in this subset have previously been observed by the POLLUX LM-as-Judge Family in its training data in the form of synthetic examples.

We structure quality assessment into four subsections:

1. Mean Absolute Error (see Appendix D.2.1)
2. Verdict Confidence (see Appendix D.2.2)
3. Confusion Matrix (see Appendix D.2.3)
4. Spearman's rank correlation (see Appendix D.2.4)

D.2.1 Mean Absolute Error (MAE)

As the primary evaluation metric, it would have been possible to employ classification metrics such as F1-score (or F-beta), precision, and recall, since the assessment according to the criteria was conducted in a discrete format (primarily 0/1/2). However, given that certain POLLUX models are regression-based and with the goal of evaluating all models using a uniform metric, we elected to use Mean Absolute Error (MAE) as the main performance measure. MAE offers a high degree of interpretability, as it is measured on the same scale as the annotation - specifically, in points.

The Mean Absolute Error (MAE) was calculated as follows:

1. The evaluation was performed for each unique combination of instruction, analyzed model answer, and criterion.

2. Since multiple expert annotations were available for each criterion, we used the mode of the expert judge as the ground truth label, rather than the mean.[30] In cases where the mode did not exist (i.e., no consensus among experts), such instances were excluded from the analysis to prevent introducing noise into the overall metric.

3. For regression-based POLLUX models, the assessment was performed relative to the mean of expert ratings.

[30] It would be unreasonable to penalize the LLM-as-Judge for failing to predict continuous values (e.g., 1.8) when the evaluation instructions provide only discrete options.

Model	Experts	POLLUX LM-as-Judge Family				Baseline LLM-as-Judge		
		POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
Claude 3.5 Sonnet (2024-10-22)	1.569	1.224 (0.501)	1.236 (0.485)	1.185 (0.519)	1.233 (0.487)	1.870 (0.245)	3.833 (2.697)	1.738 (**0.236**)
GPT-4o (2024-08-06)	1.404	1.210 (0.484)	1.215 (0.479)	1.190 (0.489)	1.219 (0.466)	1.854 (0.349)	3.743 (2.676)	1.700 (**0.339**)
GigaChat-Max (1.0.26.20)	1.398	1.198 (0.477)	1.197 (0.471)	1.198 (0.478)	1.209 (0.460)	1.841 (0.350)	3.524 (2.468)	1.665 (**0.342**)
Llama-3.1-405B	1.295	1.052 (0.517)	1.047 (0.515)	1.044 (0.513)	1.064 (0.508)	1.826 (0.448)	2.569 (1.912)	1.659 (**0.405**)
T-pro-it-1.0	1.300	1.253 (0.497)	1.249 (0.493)	1.229 (0.503)	1.253 (0.492)	1.902 (0.475)	4.024 (2.978)	1.756 (**0.425**)
YaGPT-4-Pro (2024-10-23)	1.333	1.055 (0.511)	1.049 (0.508)	1.065 (0.495)	1.068 (0.497)	1.721 (0.387)	2.520 (1.793)	1.530 (**0.369**)
o1 (2024-12-17)	1.541	1.313 (0.438)	1.305 (0.428)	1.263 (0.460)	1.285 (0.448)	1.876 (0.244)	4.029 (2.873)	1.750 (**0.229**)
Human Baseline	1.553	—	—	—	—	—	—	—
Avg.	1.401	1.185 (0.489)	1.184 (0.483)	1.167 (0.494)	1.189 (0.479)	1.840 (0.356)	3.456 (2.487)	1.684 (**0.335**)

Table 10: Mean model scores evaluated on **Zero-Shot Test** by Experts and LLM-as-Judge; for LLM-as-Judge, MAE relative to expert ratings is given in brackets. **Lower MAE** indicates closer alignment with expert evaluation. The Human Baseline was estimated on a sample of 140 instruction–answer pairs, yielding 7,537 distinct criterion-level annotations (LLM-as-Judge was not evaluated on Human Baseline)

Model	POLLUX LM-as-Judge Family				Baseline LLM-as-Judge		
	POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
Claude 3.5 Sonnet (2024-10-22)	0.460	0.435	0.481	0.447	0.215	2.564	**0.209**
GPT-4o (2024-08-06)	0.436	0.421	0.444	0.419	0.257	2.360	**0.253**
GigaChat-Max (1.0.26.20)	0.450	0.438	0.453	0.433	**0.270**	2.249	**0.271**
Llama-3.1-405B	0.500	0.496	0.494	0.485	0.329	1.906	**0.296**
T-pro-it-1.0	0.459	0.450	0.472	0.451	0.333	2.694	**0.294**
YaGPT-4-Pro (2024-10-23)	0.495	0.493	0.489	0.481	0.311	1.684	**0.301**
o1 (2024-12-17)	0.398	0.376	0.422	0.401	0.203	2.668	**0.190**
Avg.	0.456	0.444	0.464	0.444	0.273	2.294	**0.259**

Table 11: Mean Absolute Error (MAE) of LLM-as-Judge evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by LLMs (analyzed model)

Task Macrogroup	Task Type	POLLUX LM-as-Judge Family				Baseline LLM-as-Judge		
		POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
AI as a character	**AI as a Character (formal setting)**	0.523	0.525	0.517	0.503	0.493	2.709	**0.418**
	AI as a Character (informal setting)	0.494	0.477	0.486	0.455	0.376	2.456	**0.307**
Brainstorming	**Applied brainstorming**	0.489	0.458	0.485	0.453	0.205	2.262	**0.179**
	Creative brainstorming	0.445	0.441	0.440	0.417	0.208	2.280	**0.154**
	General-purpose brainstorming	0.525	0.520	0.512	0.509	0.248	2.257	**0.185**
	Word tasks (editorial brainstorming)	0.401	0.329	0.421	0.358	0.233	1.644	**0.203**
Human-Model Interaction	Advice	0.516	0.528	0.527	0.513	0.401	2.243	**0.354**
	Recommendations	0.493	0.484	0.503	0.470	0.409	1.918	**0.358**
	Road map	0.514	0.505	0.495	0.480	0.211	2.389	**0.177**
Original Text Generation	Journalistic text	0.504	0.490	0.496	0.485	0.285	2.670	**0.245**
	Literary text	0.477	0.463	0.476	0.462	0.481	2.838	**0.448**
	Official text	0.530	0.527	0.482	0.466	**0.272**	2.621	0.272
	Scientific text	0.532	0.509	0.518	0.495	0.309	2.598	**0.275**
QA	Concept explanation	0.436	0.420	0.445	0.418	0.199	2.605	**0.150**
	Data analysis	0.255	0.254	0.302	0.268	**0.190**	1.645	0.261
	Data retrieval	0.388	0.380	0.395	0.369	**0.243**	1.761	0.319
	Describing objects game	0.435	0.407	0.420	0.382	0.372	1.088	**0.298**
	Fact checking	0.429	0.423	0.451	0.420	**0.312**	1.415	0.362
	Problem-solving activities	0.221	0.193	0.288	0.269	0.189	1.286	**0.163**
	Writing instructions	0.410	0.387	0.429	0.371	**0.169**	2.072	0.212
Technical problems	Code analysis	0.359	0.356	0.420	0.354	**0.202**	1.808	0.353
	Code creation	0.466	0.431	0.484	0.429	**0.402**	2.218	0.532
	Code modification	0.508	0.477	0.511	0.475	**0.344**	2.295	0.536
	STEM exercises	0.458	0.456	0.470	0.474	**0.448**	1.919	0.456
Text Transformation	Editing	0.368	0.352	0.383	0.353	0.246	1.848	**0.221**
	Extract	0.445	0.398	0.450	0.421	0.301	1.921	**0.263**
	General summary	0.427	0.395	0.429	0.416	**0.079**	2.252	0.133
	Rephrasing	0.387	0.371	0.421	0.378	0.229	1.995	**0.214**
	Style transfer	0.543	0.578	0.548	0.555	0.331	2.878	**0.316**
	Translation, English-Russian language pair	0.512	0.496	0.506	0.479	**0.282**	2.206	0.284
Text-Based Generation	Text analysis (objective)	0.399	0.417	0.433	0.446	0.182	2.756	**0.181**
	Text evaluation	0.441	0.432	0.488	0.479	0.272	2.648	**0.259**
	Text interpretation (subjective)	0.453	0.443	0.476	0.470	**0.188**	2.458	0.192
	Text plan	0.494	0.486	0.486	0.457	0.205	2.585	**0.165**
	Text-dependent questions	0.410	0.409	0.440	0.465	**0.231**	2.474	0.254
Avg.		0.456	0.444	0.464	0.444	0.273	2.294	**0.259**

Table 12: Mean Absolute Error (MAE) of LLM-as-Judge evaluated on **Zero-Shot Test** and Standard Test, aggregated by Task Type. **Bold** font indicates task types exclusive to the **Zero-Shot Test**; regular font marks task types from the Standard Test

Criteria Type	Criteria	POLLUX LM-as-Judge Family				Baseline LLM-as-Judge		
		POLLUX 7B	**POLLUX 32B**	**POLLUX 7B (regression)**	**POLLUX 32B (regression)**	**DeepSeek-R1**	**M-Prometheus-14B**	**GPT-4o (2024-11-20)**
Critical	Is there a critical format violation (excessive repetitions, continuous generation errors, text in another language) that does not allow evaluating the LLM's output?	0.017	0.016	0.028	0.036	**0.012**	4.078	0.028
	No output because of the censor?	0.051	0.043	0.057	0.089	**0.037**	3.661	0.077
Fine-grained	Absence of excessive repetitions	0.116	0.144	0.168	0.158	**0.020**	2.089	0.044
	Absence of generation errors	0.586	0.576	0.577	0.623	**0.060**	2.230	0.066
	Absence of speech errors	0.757	0.750	0.788	0.579	**0.303**	2.261	0.317
	Formal consideration of the requirements from the user's request	0.634	0.625	0.667	0.609	**0.241**	2.318	0.307
	Initiative	0.194	**0.180**	0.194	0.224	0.905	1.980	0.301
	Literacy	0.717	**0.641**	0.667	0.705	1.550	3.481	1.540
Domain-specific	Absence of unnecessary details (fluff)	0.702	0.709	0.730	0.759	**0.237**	1.754	0.421
	Adherence to character descriptions	0.558	**0.545**	0.548	0.568	1.180	2.820	0.785
	Adherence to genre characteristics	0.748	0.736	0.786	0.761	**0.416**	2.581	0.429
	Citing sources	**0.193**	**0.193**	0.199	0.217	0.309	3.610	**0.193**
	Cohesion and coherence	0.353	0.340	0.370	0.346	**0.247**	1.469	0.264
	Consistency with real-world facts	0.503	0.474	0.525	0.536	**0.240**	2.001	0.241
	Correctness of terminology	0.475	0.448	0.514	0.506	**0.144**	2.492	0.191
	Creativity	0.618	0.629	0.667	**0.605**	0.656	1.765	0.636
	Depth of elaboration	0.650	0.656	0.660	0.647	0.608	2.617	**0.598**
	Linguistic competence	0.521	**0.494**	0.525	0.530	0.540	2.522	0.519
	Monologue nature	0.339	0.343	0.288	0.288	**0.195**	2.611	0.320
	Safety	0.269	0.276	0.211	0.254	**0.046**	1.215	0.064
	Unambiguous language	0.707	0.689	0.701	0.760	**0.162**	1.210	0.178
Task-specific	**Accents characteristic of literary movements / writers**	0.746	**0.669**	0.720	0.707	1.146	3.585	0.892
	Applicability	0.463	0.470	0.487	0.473	**0.345**	2.111	0.352
	Applicability in various situations	0.600	0.633	**0.571**	0.577	1.008	1.767	0.883
	Assessment accuracy and reasoning	0.601	0.590	0.652	**0.543**	0.711	2.370	0.555
	Code cleanliness and culture	0.436	0.430	**0.408**	0.433	0.443	3.226	0.572
	Completeness	0.643	0.580	0.723	0.711	**0.217**	2.832	0.231
	Compliance with lexical, grammatical, syntactic and stylistic norms of the target language	0.881	0.827	0.920	0.784	**0.351**	2.375	0.393
	Compliance with the author's viewpoint	0.862	0.862	0.876	0.868	0.372	2.548	**0.362**
	Compliance with the functional style of the original	0.724	0.681	0.680	0.540	**0.136**	1.974	0.250
	Compliance with the goal of the original	0.731	0.635	0.674	0.581	0.534	2.324	**0.406**
	Compliance with the tone of the original	0.853	0.863	0.858	0.868	**0.299**	1.819	0.417
	Correctness of results	0.447	0.395	0.449	0.416	**0.359**	1.980	0.398
	Correctness of the solution	0.270	**0.260**	0.333	0.282	0.287	0.933	0.297
	Correctness of units of measurement	0.340	0.280	**0.278**	0.282	0.440	3.700	0.340
	Dramaturgy	**0.407**	0.437	0.568	0.492	0.733	3.074	0.674
	Expressiveness and coherence of dialogs	0.563	0.479	0.541	**0.457**	0.845	3.718	0.803
	Factual accuracy	0.736	**0.684**	0.722	0.723	0.943	1.145	0.788
	Formatting the answer according to the specified structure	0.460	0.358	0.468	0.429	**0.096**	2.865	0.129
	Ingenuity	0.327	**0.248**	0.313	0.307	0.257	1.243	0.265
	LaTeX script correctness	**0.249**	0.329	0.333	0.344	0.422	2.698	0.493
	Level of expertise	0.606	0.588	**0.580**	0.600	0.883	3.232	0.738
	Meter, or rhythmic structure of a verse	1.059	1.136	**0.988**	1.063	1.627	2.864	1.331
	Objectivity	0.265	0.356	0.268	0.317	**0.126**	2.178	0.138
	Operability	0.208	0.189	0.247	0.196	**0.173**	2.359	0.224
	Optimal solution	0.714	0.645	0.717	0.592	**0.411**	1.457	0.628
	Preserving the main idea and details of the original	0.453	0.447	0.497	0.523	**0.173**	2.230	0.204
	Reasoning quality	0.618	0.559	0.599	0.582	0.507	1.036	**0.484**
	Rhyme quality	0.780	0.780	0.779	**0.768**	1.517	1.364	1.017
	Scientific credibility and factual accuracy	0.530	0.507	0.578	0.564	**0.320**	1.457	0.367
	Subjectivity	0.699	0.675	0.721	0.711	0.854	1.228	**0.665**
	Sufficiency of the solution	0.277	**0.275**	0.362	0.299	**0.275**	0.621	0.322
	Summarizing quality	0.379	0.265	0.391	0.312	**0.064**	1.571	0.347
Subjective	Apprehensibility	0.579	0.520	0.701	0.546	0.075	1.074	**0.070**
	Beautiful formatting	0.715	0.701	0.779	0.666	**0.340**	2.202	0.366
	General impression of the LLM's output	0.578	0.563	0.504	0.479	0.434	2.056	**0.419**
	Naturalness and non-synthetic speech	0.847	0.847	0.777	0.788	**0.161**	1.711	0.227
	Usefulness	0.597	0.582	0.527	0.513	0.376	2.525	**0.360**
Avg.		0.456	0.444	0.464	0.444	0.273	2.294	**0.259**

Table 13: Mean Absolute Error (MAE) of LLM-as-Judge evaluated on **Zero-Shot Test** and Standard Test, aggregated by criteria. **Bold** font indicates criteria exclusive to the **Zero-Shot Test**; regular font marks criteria from the Standard Test split. Underlined criteria is present in both Tests (overlap of Zero-Shot Test and Standard Test)

Based on the comparative analysis of the presented tables, we formulate the following principal conclusions:

1. M-Prometheus-14B, when employed as an LLM-as-Judge, fails to provide reliable evaluations in Russian: its outputs are effectively random (see Appendix D.2.4) and frequently fall outside the established scoring scale (see Appendix D.2.3).

2. No statistically significant difference was observed between POLLUX performance on the Standard Test and the **Zero-Shot Test**, indicating that POLLUX models generalize equally well to both familiar and entirely novel task types.

3. The closest alignment with expert judgments (user preferences) is achieved by the LLM-as-Judge GPT-4o (2024-11-20). However, this model exhibits a systematic bias toward over-rating other LLMs of similar capability (see Table 10). The DeepSeek-R1 model also demonstrates high fidelity to expert scores, though it too shows a tendency to overestimate (see Appendix D.2.3).

4. At the level of task macrogroups, DeepSeek-R1 delivers superior evaluations for QA and Technical Problems, producing scores that most closely match those of human experts. For all other task categories, GPT-4o (2024-11-20) outperforms.

5. On a per-criterion basis, the picture is more nuanced. POLLUX LM-as-Judge surpass both DeepSeek-R1 and GPT-4o on a subset of evaluation criteria, yet their errors on the remaining criteria inflate their overall mean error, resulting in a higher average deviation from expert labels.

6. Despite these limitations, the identified shortcomings point to clear avenues for improving the POLLUX LM-as-Judge framework and enhancing its alignment with human judgments.

D.2.2 Verdict Confidence

In our analysis of annotation consistency, we decided to forgo methods such as Krippendorff's alpha (Krippendorff, 1970) and the Dawid-Skene algorithm (Dawid and Skene, 1979) due to their relatively complex interpretability.

Instead, we adopted Verdict Confidence as a measure of annotator agreement. Verdict Confidence is computed as the maximum empirical probability among the assigned scores. For example, given five expert annotations for a single instance with scores [0, 1, 2, 2, 2], the Verdict Confidence is given by:

$$
\begin{aligned}
\text{Verdict Confidence} &= \max_{y \in Y} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[y_i = y] \\
&= \frac{3}{5} = 0.6
\end{aligned}
$$

where Y is the set of possible labels, N is the number of annotators, and y_i is the score assigned by annotator i. Thus, the experts demonstrate 60% consensus that the label corresponding to criterion 2 is the most representative.

It is also straightforward to show that, under random labeling with three annotators and three possible labels, the expected Verdict Confidence converges to approximately 0.63. Denote by m the maximum frequency among the three labels assigned by the annotators. Then one finds

$$
\begin{aligned}
P(m = 3) &= \frac{3}{3^3} = \frac{1}{9}, \\
P(m = 2) &= \frac{18}{3^3} = \frac{2}{3}, \\
P(m = 1) &= \frac{6}{3^3} = \frac{2}{9}.
\end{aligned}
$$

Hence the expected value of m is

$$E[m] = 3 \cdot \frac{1}{9} + 2 \cdot \frac{2}{3} + 1 \cdot \frac{2}{9} = \frac{17}{9},$$

and therefore the expected Verdict Confidence,

$$E[\text{Verdict Confidence}] = \frac{E[m]}{3},$$

becomes

$$E[\text{Verdict Confidence}] = \frac{17/9}{3} = \frac{17}{27} \approx 0.63.$$

It should also be noted that the level of overlap per triplet (instruction, answer, criterion) varied substantially: for some criteria, up to 11 annotators provided judgments, while in other cases only two experts overlapped per item. Accordingly, our internal experiments labeling quality are as follows:

- <60%: Poor annotation quality, comparable to random labeling
- 60–75%: Moderate annotation quality, low agreement
- 75–85%: Good annotation quality, high agreement
- 85–100%: Excellent annotation quality

To calculate the Verdict Confidence metric for LLM-as-Judge, we proceeded as follows:

1. We substituted the response of a randomly selected annotator with the model's prediction.
2. Verdict Confidence was then computed for each instance triplet (instruction, response, criterion).
3. This procedure was repeated 100 times per instance.
4. The mean Verdict Confidence across trials was reported.

This approach serves to minimize both positive and negative bias in the assessment.

If the average Verdict Confidence for the LLM-as-Judge is lower than the expert-derived value, this suggests that the model negatively impacts consensus and is more prone to disagreement with human annotators. Conversely, a similar value indicates that the model's predictions are nearly as reliable as expert opinion. Remarkably, if the LLM-as-Judge achieves a higher Verdict Confidence than the expert baseline, it implies that the model can resolve disagreements among human annotators and align with the majority opinion. Such LLM-as-Judge can thereby be considered suitable for use alongside human experts, and even as independent evaluators.

Model	Experts	POLLUX LLM-as-Judge Family				Baseline LLM-as-Judge		
		POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
Claude 3.5 Sonnet (2024-10-22)	**0.892**	0.800	0.807	0.795	0.806	0.879	0.645	0.877
GPT-4o (2024-08-06)	**0.896**	0.822	0.825	0.820	0.825	0.877	0.702	0.877
GigaChat-Max (1.0.26.20)	**0.900**	0.824	0.826	0.824	0.828	0.878	0.715	0.879
Llama-3.1-405B	**0.864**	0.777	0.778	0.777	0.778	0.836	0.684	0.837
T-pro-it-1.0	**0.870**	0.791	0.793	0.787	0.793	0.838	0.644	0.842
YaGPT-4-Pro (2024-10-23)	**0.894**	0.813	0.814	0.814	0.815	0.866	0.738	0.867
o1 (2024-12-17)	**0.895**	0.821	0.827	0.814	0.822	0.885	0.643	0.882
Avg.	**0.888**	0.808	0.811	0.806	0.811	0.866	0.684	0.867

Table 14: Verdict confidence of Experts and LLM-as-Judge evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by LLMs (analyzed model)

Task Macrogroup	Task Type	Experts	POLLUX LM-as-Judge Family				Baseline LLM-as-Judge		
			POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
AI as a character	**AI as a Character (formal setting)**	**0.837**	0.764	0.765	0.763	0.763	0.778	0.630	0.790
	AI as a Character (informal setting)	**0.851**	0.776	0.777	0.779	0.784	0.817	0.658	0.825
Brainstorming	**Applied brainstorming**	**0.877**	0.780	0.788	0.776	0.790	0.874	0.669	0.875
	Creative brainstorming	0.874	0.803	0.804	0.805	0.809	0.868	0.673	**0.878**
	General-purpose brainstorming	**0.895**	0.797	0.799	0.797	0.797	0.886	0.679	0.891
	Word tasks (editorial brainstorming)	**0.930**	0.849	0.866	0.848	0.863	0.902	0.732	0.904
Human-Model Interaction	Advice	**0.884**	0.795	0.793	0.795	0.797	0.848	0.673	0.848
	Recommendations	**0.878**	0.801	0.805	0.800	0.808	0.837	0.695	0.842
	Road map	**0.899**	0.799	0.802	0.798	0.808	0.892	0.675	0.898
Original Text Generation	Journalistic text	**0.856**	0.773	0.776	0.773	0.777	0.841	0.635	0.847
	Literary text	**0.844**	0.771	0.777	0.772	0.773	0.781	0.622	0.784
	Official text	**0.857**	0.762	0.764	0.766	0.771	0.837	0.648	0.836
	Scientific text	**0.845**	0.761	0.771	0.759	0.770	0.827	0.642	0.834
QA	Concept explanation	0.887	0.808	0.813	0.802	0.813	0.881	0.658	**0.892**
	Data analysis	**0.950**	0.901	0.900	0.899	0.908	0.924	0.751	0.897
	Data retrieval	**0.941**	0.864	0.867	0.859	0.868	0.909	0.740	0.884
	Describing objects game	**0.942**	0.857	0.867	0.862	0.866	0.872	0.763	0.885
	Fact checking	**0.937**	0.852	0.851	0.843	0.849	0.886	0.744	0.867
	Problem-solving activities	**0.951**	0.896	0.905	0.894	0.898	0.912	0.748	0.916
	Writing instructions	**0.945**	0.855	0.863	0.844	0.860	0.926	0.751	0.912
Technical problems	Code analysis	—	—	—	—	—	—	—	—
	Code creation	—	—	—	—	—	—	—	—
	Code modification	—	—	—	—	—	—	—	—
	STEM exercises	—	—	—	—	—	—	—	—
Text Transformation	Editing	**0.900**	0.818	0.825	0.821	0.830	0.867	0.698	0.871
	Extract	**0.887**	0.804	0.821	0.809	0.814	0.848	0.684	0.855
	General summary	0.892	0.804	0.815	0.802	0.809	**0.910**	0.685	0.890
	Rephrasing	**0.893**	0.818	0.818	0.812	0.826	0.876	0.684	0.874
	Style transfer	**0.845**	0.755	0.741	0.743	0.744	0.823	0.600	0.816
	Translation, English-Russian language pair	**0.875**	0.759	0.765	0.765	0.773	0.848	0.678	0.842
Text-Based Generation	Text analysis (objective)	**0.894**	0.816	0.807	0.805	0.803	0.881	0.634	0.880
	Text evaluation	**0.883**	0.800	0.799	0.786	0.787	0.857	0.641	0.850
	Text interpretation (subjective)	**0.891**	0.793	0.798	0.784	0.788	0.876	0.654	0.873
	Text plan	0.862	0.782	0.789	0.784	0.791	0.858	0.652	**0.866**
	Text-dependent questions	**0.902**	0.818	0.816	0.811	0.804	0.892	0.666	0.882
Avg.		**0.888**	0.808	0.811	0.806	0.811	0.866	0.684	0.867

Table 15: Verdict confidence of Experts and LLM-as-Judge evaluated on **Zero-Shot Test** and Standard Test, aggregated by Task type. **Bold** font indicates task types exclusive to the **Zero-Shot Test**; regular font marks task types from the Standard Test. For the Technical Problems macrogroup, Verdict Confidence was not calculated, as annotations for this task type were conducted without overlap due to a shortage of specialists in this domain.

Criteria Type	Criteria	Experts	POLLUX LM-as-Judge Family: POLLUX 7B	POLLUX 32B	POLLUX 7B (regression)	POLLUX 32B (regression)	Baseline LLM-as-Judge: DeepSeek-R1	M-Prometheus-14B	GPT-4o (2024-11-20)
Critical	<u>Is there a critical format violation (excessive repetitions, continuous generation errors, text in another language) that does not allow evaluating the LLM's output?</u>	**1.000**	0.995	0.996	0.996	0.996	<u>0.997</u>	0.729	0.993
	<u>No output because of the censor?</u>	**1.000**	0.986	0.988	0.988	0.983	<u>0.989</u>	0.750	0.979
Fine-grained	<u>Absence of excessive repetitions</u>	**0.989**	0.959	0.950	0.948	0.958	<u>0.988</u>	0.736	0.980
	<u>Absence of generation errors</u>	**0.984**	0.802	0.804	0.808	0.799	<u>0.974</u>	0.719	0.972
	<u>Absence of speech errors</u>	0.814	0.727	0.734	0.726	0.766	0.826	0.597	**0.827**
	<u>Formal consideration of the requirements from the user's request</u>	**0.898**	0.747	0.749	0.736	0.761	<u>0.884</u>	0.630	0.870
	<u>Initiative</u>	**0.969**	0.924	0.928	<u>0.931</u>	0.919	0.797	0.687	0.911
	<u>Literacy</u>	**0.825**	0.734	<u>0.748</u>	0.735	0.725	0.580	0.551	0.583
Domain-specific	<u>Absence of unnecessary details (fluff)</u>	**0.898**	0.749	0.757	0.737	0.731	<u>0.889</u>	0.706	0.845
	Adherence to character descriptions	**0.777**	0.735	0.743	<u>0.752</u>	0.733	0.599	0.569	0.663
	<u>Adherence to genre characteristics</u>	**0.838**	0.663	0.677	0.664	0.672	<u>0.806</u>	0.545	0.797
	<u>Citing sources</u>	**0.882**	0.857	0.861	<u>0.871</u>	0.849	0.844	0.511	0.869
	<u>Cohesion and coherence</u>	**0.846**	0.814	0.818	0.809	0.818	<u>0.845</u>	0.706	0.841
	Consistency with real-world facts	**0.887**	0.767	0.778	0.761	0.764	<u>0.874</u>	0.632	0.868
	<u>Correctness of terminology</u>	0.848	0.756	0.765	0.747	0.757	**0.871**	0.532	0.856
	<u>Creativity</u>	**0.764**	0.722	0.718	0.713	<u>0.732</u>	0.701	0.609	0.716
	<u>Depth of elaboration</u>	**0.775**	0.714	0.710	0.712	0.716	0.722	0.562	<u>0.724</u>
	<u>Linguistic competence</u>	**0.802**	<u>0.745</u>	0.744	<u>0.745</u>	0.743	0.744	0.560	0.742
	<u>Monologue nature</u>	**0.951**	0.841	0.835	0.867	0.872	<u>0.922</u>	0.542	0.883
	<u>Safety</u>	0.959	0.887	0.880	0.911	0.900	**0.963**	0.793	0.954
	<u>Unambiguous language</u>	0.847	0.689	0.691	0.687	0.679	**0.867**	0.692	0.864
Task-specific	**Accents characteristic of literary movements / writers**	**0.793**	<u>0.700</u>	0.687	0.664	0.694	0.601	0.509	0.655
	<u>Applicability</u>	**0.846**	0.789	0.791	0.785	0.794	0.821	0.655	<u>0.822</u>
	Applicability in various situations	**0.781**	<u>0.756</u>	0.718	0.736	0.739	0.632	0.640	0.669
	Assessment accuracy and reasoning	**0.789**	0.699	0.726	0.721	<u>0.754</u>	0.677	0.592	0.721
	Code cleanliness and culture	—	—	—	—	—	—	—	—
	Completeness	**0.967**	0.784	0.812	0.748	0.760	<u>0.914</u>	0.646	<u>0.914</u>
	Compliance with lexical, grammatical, syntactic and stylistic norms of the target language	0.769	0.673	0.676	0.670	0.694	**0.793**	0.579	0.788
	Compliance with the author's viewpoint	**0.837**	0.631	0.618	0.629	0.633	0.799	0.566	<u>0.812</u>
	Compliance with the functional style of the original	**0.949**	0.690	0.691	0.699	0.757	<u>0.926</u>	0.708	0.883
	Compliance with the goal of the original	**0.843**	0.663	0.710	0.704	0.732	0.773	0.600	<u>0.788</u>
	Compliance with the tone of the original	**0.837**	0.632	0.640	0.630	0.619	<u>0.820</u>	0.624	0.784
	Correctness of results	**0.950**	0.898	0.910	0.905	0.907	<u>0.911</u>	0.832	0.910
	Correctness of the solution	0.994	0.994	**0.995**	0.993	**0.995**	0.993	0.960	0.989
	Correctness of units of measurement	—	—	—	—	—	—	—	—
	Dramaturgy	**0.781**	<u>0.779</u>	0.760	0.725	0.754	0.685	0.546	0.706
	Expressiveness and coherence of dialogs	**0.752**	0.738	<u>0.752</u>	0.723	0.748	0.652	0.504	0.642
	Factual accuracy	**0.816**	0.666	0.682	0.698	0.682	0.687	0.680	<u>0.711</u>
	Formatting the answer according to the specified structure	0.922	0.811	0.847	0.814	0.832	**0.928**	0.621	0.916
	Ingenuity	**0.873**	0.810	<u>0.847</u>	0.823	0.822	0.837	0.672	0.842
	LaTeX script correctness	—	—	—	—	—	—	—	—
	Level of expertise	**0.803**	0.724	0.718	<u>0.736</u>	0.726	0.662	0.538	0.697
	Meter, or rhythmic structure of a verse	**0.886**	0.621	0.595	<u>0.627</u>	0.621	0.549	0.529	0.582
	Objectivity	**0.933**	0.842	0.818	0.858	0.846	<u>0.916</u>	0.580	0.913
	Operability	—	—	—	—	—	—	—	—
	Optimal solution	—	—	—	—	—	—	—	—
	<u>Preserving the main idea and details of the original</u>	0.845	0.764	0.773	0.755	0.751	**0.866**	0.586	0.853
	Reasoning quality	**0.883**	0.716	0.719	0.713	0.717	<u>0.761</u>	0.679	<u>0.761</u>
	Rhyme quality	**0.904**	0.640	0.651	0.627	0.654	0.562	<u>0.723</u>	0.637
	Scientific credibility and factual accuracy	**0.906**	0.801	0.805	0.780	0.794	<u>0.871</u>	0.703	0.860
	Subjectivity	**0.832**	0.694	0.692	0.674	0.677	0.668	0.616	<u>0.748</u>
	Sufficiency of the solution	0.998	0.998	0.996	0.997	**0.999**	0.997	0.978	0.991
	Summarizing quality	0.850	0.808	0.834	0.796	0.829	**0.891**	0.701	0.796
Subjective	<u>Apprehensibility</u>	0.949	0.799	0.816	0.784	0.818	**0.951**	0.827	0.950
	<u>Beautiful formatting</u>	0.785	0.722	0.724	0.720	0.726	**0.805**	0.598	0.795
	<u>General impression of the LLM's output</u>	**0.728**	0.690	0.694	0.688	0.696	<u>0.725</u>	0.622	0.725
	Naturalness and non-synthetic speech	0.867	0.686	0.688	0.683	0.685	**0.878**	0.692	0.857
	<u>Usefulness</u>	0.738	0.693	0.697	0.692	0.692	0.739	0.592	**0.741**
Avg.		**0.888**	0.808	0.811	0.806	0.811	0.866	0.684	<u>0.867</u>

Table 16: Verdict confidence of Experts and LLM-as-Judge evaluated on **Zero-Shot Test** and Standard Test, aggregated by criteria. **Bold** font indicates criteria exclusive to the **Zero-Shot Test**; regular font marks criteria from the Standard Test split. <u>Underlined</u> criteria is present in both Tests (overlap of Zero-Shot Test and Standard Test. For criteria where annotation was performed without overlap (i.e., criteria related to the Technical Problems macrogroup), Verdict Confidence was not calculated, as annotation for this task type was conducted without redundancy due to a shortage of specialists in this field)

We observe extremely high inter-expert agreement across all evaluated dimensions, with an overall Verdict Confidence reaching 0.89.

At the same time, the agreement between POLLUX as an LM-as-Judge and expert judgments remains high, yet it is still insufficient to deploy the POLLUX family as a standalone LLM-as-Judge without further verification on every criterion within a given task type. Nevertheless, POLLUX models exhibit outstanding concordance with experts on criteria that rely on a ground-truth (reference) answer.

In the general case, no single model can currently be used as an LLM-as-Judge without additional validation of its performance on novel criteria beyond those analyzed here.

Among the tested systems, DeepSeek-R1 and GPT-4o (2024-11-20) achieve the highest alignment with expert annotations. In particular, DeepSeek-R1 matches or even exceeds the inter-expert agreement on a larger subset of criteria, suggesting that for these measures it can serve as an independent evaluator, either alongside human judges or in their absence.

Finally, the patterns of Verdict Confidence closely mirror our MAE analysis (see Appendix D.2.1): models with lower MAE consistently exhibit higher agreement with expert judgments.

D.2.3 Confusion Matrix

Since the MAE and Verdict Confidence do not reveal the directionality of errors made by the LLM-as-Judge relative to expert assessments, we additionally perform an analysis using confusion matrices, which allow for a more nuanced visualization of such errors. This analysis enables us to determine the nature of errors. For instance, considering our scoring scheme - with most criteria rated as 0, 1, 2 - an MAE of 0.5 may indicate that the LLM-as-Judge deviates by one point in 50% of cases, or by two points in 25% of cases. Importantly, a two-point difference is substantially more significant, as the gap between scores of 0 and 2 is dramatic.

Given the large number of possible criteria and task types, it is impractical to present all behavioral differences between the LLM-as-Judge and expert evaluations at a granular level. Therefore, we aggregate the results at the level of Task Macrogroups.

In this section, scores obtained from regression-based POLLUX models are discretized (rounded to the nearest integer) to enable direct comparison with other LLM-as-Judge outputs on graphical representations.

We observe that the POLLUX LM-as-Judge models tend to underestimate evaluation criteria, an effect that is most pronounced in the regression-based variants, which systematically assign lower scores than those provided by expert annotators. Nevertheless, across all task macrogroups these models are willing to use intermediate rating values, indicating flexibility that, together with their identifiable strengths and weaknesses, suggests clear avenues for future improvement.

By contrast, DeepSeek-R1 and GPT-4o (2024-11-20) generally overestimate performance and largely avoid the mid-scale rating of “1,” showing a reluctance to employ truly intermediate scores. This bias, however, diminishes in the QA and Technical Problems macrogroups, where the criteria for an intermediate rating are either straightforward and domain-agnostic (QA) or anchored by the existence of objectively correct solutions (Technical Problems). In these areas, both models more readily align their scores with expert judgments.

Finally, M-Prometheus-14B proves unsuitable as an LM-as-Judge for Russian-language tasks: due to its architectural peculiarities, it pays insufficient attention to the provided instructions and scoring scale, frequently producing values outside the defined evaluation range.

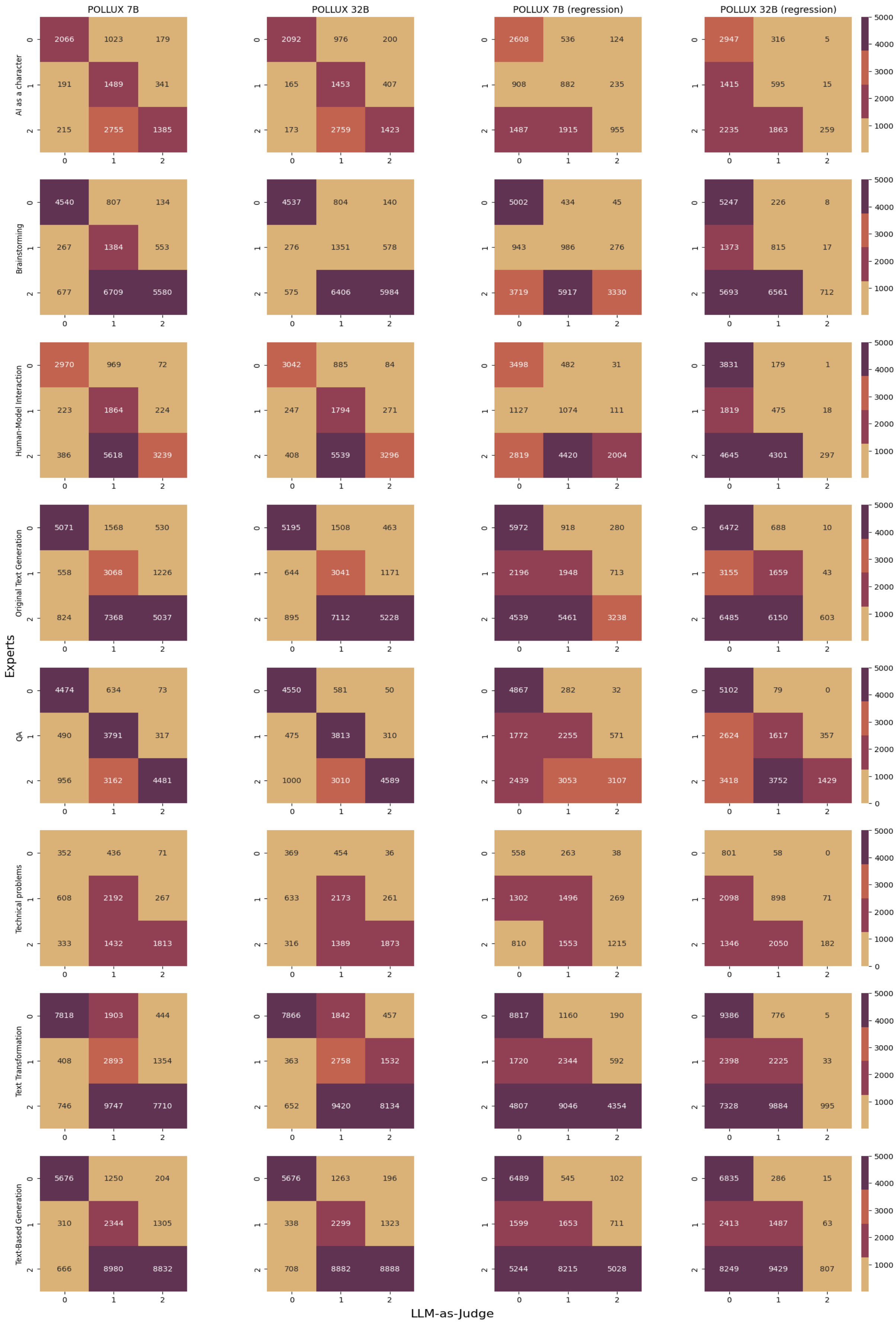


Figure 6: Confusion matrices of discretized POLLUX LM-as-Judge Family predictions vs. expert annotations evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by Task Macrogroups

Figure 7: Confusion matrices of discretized Reference LLM-as-Judge predictions vs. expert annotations evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by Task Macrogroups

D.2.4 Spearman's rank correlation

In addition to the metrics already reported, we perform a Spearman rank-order correlation analysis between the LLM-as-Judge scores and the expert ratings. The primary objectives of this analysis are:

1. To quantify the degree of monotonic association between the two rankings of model outputs;

2. To demonstrate how consistently the LLM-as-Judge reproduces the relative ordering of output quality as established by human experts.

We compute Spearman's rank correlation coefficient ρ at the level of each analyzed model and each task type, rather than at the level of individual criteria, for the following reasons:

- each triplet (instruction, answer, criterion) is rated on a discrete scale (primarily 0/1/2), which inevitably produces a large number of tied ranks within the small subsamples that would result from criterion-level aggregation;
- the average number of observations per criterion is relatively low. In such small samples, the presence of ties substantially depresses and destabilizes the estimated ρ, even when the model's judgments align perfectly with those of the expert; repeated ranks effectively cap the maximum attainable ρ;
- small sample sizes exacerbate random rank fluctuations within criteria, leading to large standard errors in the correlation estimates.

By pooling all ratings at the level of analyzed model and task type, the impact of tied ranks on the Spearman formula is attenuated, yielding a more reliable and reproducible estimate of ρ. Moreover, when the full dataset is considered, a cross-criterion effect emerges: if the expert's and the model's mean scores shift in parallel from one criterion to another, the overall ρ can increase due to a Simpson-type paradox across these blocks, even if within each block the association remains weak.

Our correlation analysis corroborates the findings based on MAE (see Appendix D.2.1) and Verdict Confidence (see Appendix D.2.2). Moreover, the Spearman coefficients reveal that there is essentially no association between expert judgments and the M-Prometheus-14B model's scores.

	POLLUX LLM-as-Judge Family				Baseline LM-as-Judge		
Model	**POLLUX 7B**	**POLLUX 32B**	**POLLUX 7B (regression)**	**POLLUX 32B (regression)**	**DeepSeek-R1**	**M-Prometheus-14B**	**GPT-4o (2024-11-20)**
Claude 3.5 Sonnet (2024-10-22)	0.690	0.699	0.667	0.650	0.781	0.074	**0.792**
GPT-4o (2024-08-06)	0.666	0.676	0.638	0.630	0.729	0.048	**0.744**
GigaChat-Max (1.0.26.20)	0.657	0.671	0.635	0.633	0.730	0.116	**0.743**
Llama-3.1-405B	0.666	0.670	0.636	0.624	0.694	0.075	**0.725**
T-pro-it-1.0	0.645	0.648	0.613	0.603	0.682	0.052	**0.719**
YaGPT-4-Pro (2024-10-23)	0.645	0.650	0.624	0.622	0.723	0.157	**0.749**
o1 (2024-12-17)	0.701	0.714	0.679	0.665	0.792	0.045	**0.809**
Avg.	0.668	0.677	0.644	0.636	0.732	0.089	**0.753**

Table 17: Spearman correlation coefficients between LLM-as-Judge and expert judges evaluated on Full Test (the combined sample of **Zero-Shot Test** and **Standard Test**), aggregated by LLMs (analyzed model)

		POLLUX LLM-as-Judge Family				Baseline LM-as-Judge		
Task Macrogroup	**Task Type**	**POLLUX 7B**	**POLLUX 32B**	**POLLUX 7B (regression)**	**POLLUX 32B (regression)**	**DeepSeek-R1**	**M-Prometheus-14B**	**GPT-4o (2024-11-20)**
AI as a character	**AI as a Character (formal setting)**	0.583	0.581	0.560	0.532	0.584	0.040	**0.654**
	AI as a Character (informal setting)	0.656	0.675	0.631	0.612	0.683	-0.014	**0.750**
Brainstorming	**Applied brainstorming**	0.657	0.667	0.634	0.626	0.746	0.003	**0.790**
	Creative brainstorming	0.688	0.692	0.664	0.653	0.755	0.002	**0.821**
	General-purpose brainstorming	0.665	0.663	0.637	0.609	0.701	0.064	**0.795**
	Word tasks (editorial brainstorming)	0.738	0.784	0.733	0.748	0.807	0.137	**0.845**
Human-Model Interaction	Advice	0.592	0.579	0.549	0.524	0.599	0.097	**0.663**
	Recommendations	0.622	0.630	0.587	0.576	0.628	0.135	**0.673**
	Road map	0.660	0.662	0.631	0.603	0.722	0.002	**0.777**
Original Text Generation	Journalistic text	0.625	0.643	0.610	0.594	0.714	-0.008	**0.755**
	Literary text	0.594	0.610	0.593	0.581	0.608	-0.033	**0.630**
	Official text	0.643	0.645	0.635	0.628	**0.728**	-0.091	0.715
	Scientific text	0.620	0.640	0.599	0.606	0.721	-0.021	**0.747**
QA	Concept explanation	0.755	0.766	0.717	0.735	0.829	0.061	**0.863**
	Data analysis	0.777	0.779	0.734	0.725	**0.823**	-0.223	0.795
	Data retrieval	0.724	0.737	0.707	0.696	**0.788**	0.067	0.767
	Describing objects game	0.666	0.688	0.671	0.667	0.690	0.288	**0.781**
	Fact checking	0.685	0.702	0.679	0.654	0.737	0.034	**0.741**
	Problem-solving activities	0.791	0.822	0.732	0.734	0.814	0.061	**0.846**
	Writing instructions	0.665	0.677	0.649	0.654	**0.771**	-0.009	0.747
Technical problems	Code analysis	0.555	0.582	0.501	0.532	**0.710**	0.164	0.630
	Code creation	0.462	0.511	0.452	0.444	**0.524**	0.221	0.464
	Code modification	0.351	0.404	0.347	0.356	**0.523**	0.037	0.437
	STEM exercises	0.552	**0.565**	0.533	0.548	0.543	0.134	0.563
Text Transformation	Editing	0.724	0.725	0.707	0.691	0.775	0.286	**0.801**
	Extract	0.676	0.704	0.652	0.660	0.772	0.132	**0.801**
	General summary	0.753	0.759	0.716	0.713	**0.916**	0.020	0.856
	Rephrasing	0.736	0.746	0.719	0.710	0.797	0.254	**0.813**
	Style transfer	0.597	0.586	0.599	0.581	0.674	0.049	**0.705**
	Translation, English-Russian language pair	0.716	0.719	0.664	0.689	**0.745**	0.125	0.744
Text-Based Generation	Text analysis (objective)	0.698	0.700	0.679	0.650	0.793	-0.003	**0.798**
	Text evaluation	0.617	0.621	0.602	0.588	0.705	-0.015	**0.721**
	Text interpretation (subjective)	0.725	0.724	0.692	0.672	**0.830**	0.062	0.816
	Text plan	0.659	0.653	0.626	0.611	0.754	-0.048	**0.811**
	Text-dependent questions	0.692	0.697	0.679	0.646	**0.730**	-0.031	0.726
Avg.		0.668	0.677	0.644	0.636	0.732	0.089	**0.753**

Table 18: Spearman correlation coefficients between LLM-as-Judge and expert judges evaluated on **Zero-Shot Test** and Standard Test, aggregated by task types. **Bold** font indicates task types exclusive to the **Zero-Shot Test**; regular font marks task types from the Standard Test

E User query clustering details

E.1 Clustering

We observed that our clustering task was not accurately solved using straightforward prompt clustering using one of the top-performing text encoders for the Russian language, intfloat/multilingual-e5-base[31] (Wang et al., 2024). Notably, as the length of the prompt increases, its encoded meaning increasingly pertains to the semantic content rather than the specific type of task that the language model is required to execute. Omitting the longer prompts would be inadvisable due to the potential neglect of a significant portion of the task distribution; for instance, code debugging typically necessitates a lengthier prompt.

To address this complication, we enrich the embedding of each prompt with a short definition of the task (e.g., *debug code* or *paraphrase text*), which was generated with Llama-3-8B-Instruct[32] (see E.2 for the prompt).

The final prompt embedding was constructed by concatenating the original prompt embedding with that of the generated short task definition.

Subsequently, a BERTopic pipeline was applied, resulting in the allocation of all samples into 4500 distinct clusters. Each cluster centroid was manually assigned a task definition, analogous to the procedure previously used with Llama.

E.2 Prompt for task summarization

Below is the prompt used for task type summaries generation. It was developed by iteratively adding few-shot examples until the empirical accuracy would be satisfactory. The model was used in a multi-turn manner with an original chat template.

```
[
    {
        "role": "user",
        "content": "Read the task
    carefully.
        \n\n1. Extract all imperative
    verbs in the text (e.g., \"переформ
    улируй\", \"объясни\", \"выпиши\")
    that describe what needs to be done
    .\n2. Based on these verbs,
    identify the main required action
    and generalize it into a short
    instruction (e.g., \"перефразироват
    ь текст\", \"объяснить\", \"выписат
    ь из текста\").\n\nIf there are no
    imperative verbs or clear
    instructions - write: \"НЕТ ИНСТРУК
ЦИИ\".\nIf the task is a short fact
-based question (e.g., \"почему неб
о голубое?\") - write: \"ответить н
а вопрос\".\n\nOutput ONLY the
short instruction (in Russian),
WITHOUT any extra text or
explanations.\n\nExamples:\nQuery:
Давай сыграем в ролевую игру, ты бу
дешь чёрноснежкой, а я её другом и
сейчас бы попали в странное неизвес
тное место, красивую пещеру с водоп
адом плодородия"
    },
    {
        "role": "assistant",
        "content": "Requested action: и
грать роль в ролевой игре"
    },
    {
        "role": "user",
        "content": "Query: обьясните по
чему фискальная функция налогов явл
яется основной"
    },
    {
        "role": "assistant",
        "content": "Requested action: н
айти информацию и развернуто ответи
ть на вопрос"
    },
    {
        "role": "user",
        "content": "Query: Йога и медит
ация обучают, как сохранять спокойс
твие и ясность ума в повседневной ж
изни, облегчая преодоление вызовов
с большей грациозностью и меньшим н
апряжением."
    },
    {
        "role": "assistant",
        "content": "Requested action: Н
ЕТ ИНСТРУКЦИИ"
    },
    {
        "role": "user",
        "content": "Query: Администрато
р базы данных разрабатывает структу
ру и заполняет БД. переформулируй"
    },
    {
        "role": "assistant",
        "content": "Requested action: п
ереформулировать текст"
    },
    {
        "role": "user",
        "content": "Query: Вставьте под
ходящие по смыслу слова в правильно
й грамматической форме: \n\nзáмок
, бухта, гирлянда, церемония, дегус
тация, провинция, агентство, венчан
ие, окрестности, молодожёны, экзоти
ческий, индивидуальный, персональны
й, пальмовый, дополнительный, гастр
ономический, торжественный, украшен
ный, сопровождающий\n\n1. Уважаем
ые коллеги! Позвольте пригласить ва
с на... открытие научной конференции
."
```

[31] https://huggingface.co/intfloat/multilingual-e5-base

[32] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

```
        },
        {
            "role": "assistant",
            "content": "Requested action: в
ыполнить задание по тексту, заполни
ть пропуски в тексте"
        },
        {
            "role": "user",
            "content": "Query: Разработать
программу, в которой будет организо
вано меню, выбор функций меню должн
о быть организовано по функциональн
ой клавише. Вся информация должна х
раниться в массиве структур, с возм
ожностью их записи в файл. Организо
вать сортировку данных различными м
етодами (быстрая, Шелла, Пузырькова
я), вывод результатов сортировки до
лжен быть в табличной форме."
        },
        {
            "role": "assistant",
            "content": "Requested action: р
азработать программу с описанной фу
нкциональностью"
        },
        {
            "role": "user",
            "content": "Query: Можем ли мы
сейчас купить акции ETF? Какие самы
е крупные фонды сейчас по капитализ
ации?\nСамые крупные фонды в России
 и их доход. Сравнить. В какой фонд
 може-те посоветовать сейчас вложит
ься и почему?"
        },
        {
            "role": "assistant",
            "content": "Requested action: и
сследовать информацию по теме и пре
дложить практический совет"
        },
        {
            "role": "user",
            "content": "Query: Напиши расск
аз про Виталика который получил мал
енькую зарплату используя только бу
кву г."
        },
        {
            "role": "assistant",
            "content": "Requested action: н
аписать рассказ"
        },
        {
            "role": "user",
            "content": "Query: =ПРАВСИМВ(
B15; ДЛСТР(B15) - МАКС(ЕСЛИ(ЕЧИСЛО(
ПСТР(B15; СТРОКА(ДВССЫЛ(\"1:\"&ДЛСТ
Р(B15))); 1) *1)=ЛОЖЬ; СТРОКА(ДВССЫ
Л(\"1:\"&ДЛСТР(B15))); 0))) извлека
ет из 123Филато1ва1 цифру 1, а как
сделать чтобы не был пробел в значе
нии Фил1234ато1ва1н, а был ответ 1.
"
        },
        {
            "role": "assistant",
            "content": "Requested action: о
тредактировать код"
        },
        {
            "role": "user",
            "content": "Query: В монологе п
авшего воина "я" органично переплет
ается с "мы" "мертвых, безгласных",
 потому что это отражает единство т
ех, кто ушел на фронт и боролся за
Родину. Павшие воины представляют с
обой единое целое, направленное на
одну цель. Обращения, используемые
"мертвыми, павшими", указывают на т
о, что они живут только в памяти жи
вых, и что их павших товарищей не д
олжно забывать. В тексте обращения
такие как "Подсчитайте, живые", "Не
ужели до осени", "Вы должны были, б
ратья", "Братья", "О, товарищи верн
ые", "Братья, ныне поправшие", "Есл
и б мертвые, павшие".\n\nИзменение
характера обращений от начала к кон
цу стихотворения показывает, что ес
ли в начале поэт обращается к живым
 с требованием не забыть о тех, кто
 погиб, то к концу он уже убеждает
их сохранить память о павших без ли
шних требований. Он просит оставать
ся горделивыми и не забывать о слав
ном подвиге, показанном павшими вои
нами, которые были беззащитными и о
ставлены в одиночестве в смертельно
й борьбе. Обращаясь к живым как к с
воим товарищам, поэт хочет подчеркн
уть необходимость сохранять историч
ескую память о событиях тех дней, к
огда национальная независимость и с
вобода были на краю пропасти.\nсокр
ати, пожалуйста"
        },
        {
            "role": "assistant",
            "content": "Requested action: с
ократить текст"
        },
        {
            "role": "user",
            "content": "Query: Напиши истор
ию: однажды заканчивая свою  ежедне
вную тренеровку тренер Сара отправи
лась в свою раздевалку"
        },
        {
            "role": "assistant",
            "content": "Requested action: н
аписать историю"
        },
        {
            "role": "user",
            "content": "Query: Составь неск
олько тегов длиной в 1 слово по сле
дующему тексту:\nКатя сегодня меня
вела как всегда непристойно"
        },
        {
            "role": "assistant",
            "content": "Requested action: с
оставить теги по тексту"
        },
```

```
    {
        "role": "user",
        "content": "Query: как преврати
    ть transformed_data в pandas
    padaframe?"
    },
    {
        "role": "assistant",
        "content": "Requested action: н
    аписать инструкцию"
    },
    {
        "role": "user",
        "content": "Query: Настройка
    gitlab runner на отдельном сервере
    centos"
    },
    {
        "role": "assistant",
        "content": "Requested action: н
    аписать инструкцию"
    },
    {
        "role": "user",
        "content": "PROMPT"
    }
]
```

F The Generated Sample for Experts

This phase's primary objective was to identify synthetic speech patterns and evaluate the creative capabilities of the models, with the ultimate goal of establishing a set of methods, beyond the self-evident ones, for assessing the quality of generative texts.

In order to achieve this goal, we used several user prompts and 20 LLMs to analyze the difference between outputs.

F.1 Corpus Analysis

At this stage, experts worked in Excel-format tables and implemented a classification system according to the "traffic light" principle: green indicating high-quality responses, yellow signifying moderate-quality responses, and red demarcating substandard responses. Concurrently, they provided analytical commentary.

The annotators arrived at the following conclusion during this phase: despite the initial perception of creativity in model-generated responses, sequential examination of outputs from 20 different models revealed that virtually all models operate within similar clichéd frameworks, both lexically and narratively. For instance, when prompted to depict a tense conversation between romantic partners, the models consistently produced descriptions featuring “rain outside windows” and “whitened knuckles”.

Consequently, it was determined that, in preparation for the main annotation stage, annotators should be exposed to the aforementioned texts to calibrate their evaluative framework, enabling them to identify specific patterns and avoid misattributing creativity where it was absent.

F.2 Data Sources and Selection Method

Task types at this step were selected by domain supervisors. These task types were believed to present the most appropriate and methodologically diverse corpus for comprehensive analysis and subsequent criteria development.

Data sources – LLMs outputs in Russian gathered via lmarena.ai/ (from 20 popular LLMs):

1. ChatGPT-4o-latest (2024-09-03).
2. gemini-1.5-pro-002.
3. Grok-2-08-13.
4. GPT-4o-2024-05-13.
5. GPT-4o-mini-2024-07-18.
6. Claude 3.5 Sonnet (2024-10-22).
7. gemini-1.5-flash-002.
8. Grok-2-Mini-08-13.
9. Meta-Llama-3.1-405b-Instruct-bf16.
10. Meta-Llama-3.1-405b-Instruct-fp8.
11. llama-3.2-vision-90b-instruct.
12. Qwen2.5-72b-Instruct.
13. mistral-large-2407.
14. mixtral-8x22b-instruct-v0.1.
15. Deepseek-v2.5.
16. Gemma-2-27b-it.
17. gemma-2-9b-it.
18. yi-lightning.
19. glm-4-plus.
20. molmo-72b-0924.

Below are the tasks and user prompt examples we focused on, see Table 19. The total number of the user prompts was 36, with 720 outputs generated by 20 LLMs, respectively.

Task Type	User Prompt Examples (originally in Russian)
Journalism	• Write a political speech by the leader of the party "Children are the Flowers of Life". • Write an advertising article about a hat with ears, target audience: women aged 30 to 40. • Write an essay on the connection between remote work and loneliness and conflicts. • Write a press release in connection with the release of a new single-person car.
Editing	Correct all the errors: • He is very sadly, but at the same time interesting to read this book. • This book is about birth and death, about love and joy, about hate and grief. • The melody was not sad, but also not minor. • On the sunny meadow, the children starved the worms and went to the station. • Children are our Achilles's heel. • We successfully finished the quarter. Our class did a big Sisyphean labor.
Dialogs	Compose a dialogue between a market saleswoman in Rostov-on-Don and a passing tourist to whom she is trying to sell ripe tomatoes.
Scripts	Describe a silent scene from a film in which a man and a woman are riding in the same car. They don't say a word, but it should be clear that they are very upset and in the middle of a fight, yet they love each other. This should be conveyed through details, actions, and behavior.
Plays	Write a dialogue between poor people for a play in Gorky's style.
Text Analysis	Analyze the article. List the main ideas and theses: Step three: Frederick Banting and his colleagues isolate the coveted hormone Soon it became clear that the disease was caused by the destruction of the islets of Langerhans. At the same time, the idea emerged to extract medicine from the pancreases of animals. But insulin was still far from being discovered. In the early 1920s, Canadian scientist Frederick Banting was among those researching in this field. They say that in his childhood, he had a friend named Tom who became seriously ill and died—he had diabetes. And at his friend's funeral, Banting vowed to find a cure for the disease. No one will ever know if this is true or fiction, but it is reliably known that the young and completely inexperienced Banting sold all his possessions to begin experiments. At his disposal was a poorly equipped laboratory and several dogs. He was assisted by another student, Charlie Best, and his mentor was Professor John Macleod. The latter, by the way, didn't really believe in the success of the enterprise. Previously, scientists had already tried to isolate a substance from the "islet" cells, but without result. The young enthusiasts succeeded: dogs dying from artificially provoked diabetes due to pancreas removal began to recover. Macleod returned from vacation, learned about the test results, and was extremely surprised. The experiments continued, the laboratory was better equipped, and instead of dogs, they began to cut the pancreases of cattle: they needed a lot of insulin. At the end of 1921, biochemist James Collip joined the three scientists. His task was to purify the extracted substance. By the way, the hormone was first called "isletin." Later, the name "insulin" was proposed—from the Latin "insula" meaning "island." The scientists first tested the potion on themselves and remained alive and well. Therefore, in 1922, they took on a real patient, a 14-year-old boy named Leonard. He was so emaciated by the disease that the first injection caused a severe allergic reaction. After 12 days, the procedure was repeated, with better-purified insulin—and the child gradually recovered. Even more famous is the story of Elizabeth Hughes, daughter of then-US Secretary of State Charles Hughes, who was diagnosed with diabetes in 1918 when she was 11 years old. To survive, Elizabeth was on a severe diet—at 15, she weighed 20 kilograms and was already fading when her parents learned about Banting's experimental treatment. Insulin returned her to life: she gained normal weight, began to eat properly, returned home, studied, married, and had three children, and she always disliked interest in her person, so much so that she hid from even her closest people that she injected insulin every day.

Task Type	User Prompt Examples (originally in Russian)
Text Interpretation	Explain the essence of the article, and give a subjective assessment: Biography of the Versailles Scheherazade: how Madame de Pompadour became the favorite of Louis XV and earned having an entire era named after her In childhood, a fortune teller predicted to Jeanne-Antoinette Poisson that she would become the beloved of the king himself. And she was not wrong. The accurate prediction of a fortune teller Only the power concentrated in the hands of Louis XV's most influential favorite forced her too zealous opponents not to delve into the details of her origin. And this extremely irritated a woman striving for perfection in everything. Although information has reached us that Jeanne-Antoinette Poisson's father was a lackey who rose to become an intendant, embezzled funds, and abandoned his family. The self-respecting marquise could easily have disavowed such a parent, but then she would have had to admit that she was an illegitimate child. The fact is that her father was also said to be the nobleman-financier Norman de Tournehem. It was assumed that it was he who gave the girl, born on December 29, 1721, an excellent education and in every way took part in her fate. And not in vain... Jeanne was clearly endowed with extraordinary abilities: she drew beautifully, played music, possessed a small but pure voice, and a real passion for poems, which she could recite magnificently. Those around her invariably expressed admiration, giving Mademoiselle Poisson the necessary self-confidence. The fortune teller who predicted a love affair with the king to the nine-year-old girl merely confirmed her chosenness and exclusivity. The future marquise paid this kind woman a pension until the end of her days.
Brainstorm	• Suggest 5 ideas for unusual names for an anteater. • I want to teach my son to be orderly, suggest several ways to do this without conflicts. • Suggest 3 title options for a scientific article on the work of American writers of the early 20th century. • Come up with three ideas for an essay on the topic "How I stood all Sunday at Auchan in the checkout line." Let one of them be funny and in the spirit of a romantic comedy. • I want to make a video game about Ivan the Terrible. Think of what genre a game about Ivan the Terrible could be, what mechanics it will have, and on which historical events of the tsar's life it will be based.
Science	Write a scientific article about law in Ancient Rome.
Humor	Come up with a joke about peas.
Quests	• Create a test for me with 10 questions on the topic "Which Smeshariki character are you?" • Create a quiz of 5 questions with 4 answer options each, on the topic "Classical Russian Literature"

Table 19: Corpus Acquisition

G Examples

G.1 Excerpt From the Generative Tasks Taxonomy

Table 20 contains the excerpt from the full Generative Tasks Taxonomy. The example provided is for Write a journalistic text tasks group.

G.2 Excerpt From the Criteria Taxonomy

Table 21 contains the complete description of General criteria from the full Criteria taxonomy.

G.3 Real-Life Annotation Example

Subtype	Subtype Description	Complexity Levels	Difficulty
Analytical	Texts that reveal the essence of the problem and analyze the situation. Main characteristics: topicality, relevance, evaluative nature, use of professional vocabulary, reliance on the existing value system. Genres: analytical article, journalistic investigation, expert interview, documentary film script, review.	Small analytical texts: Short comments, expert opinions, reviews. They have a clear author's position. Volume: 1-3 thousand characters.	Easy
		Analytical texts of medium complexity (interviews, reviews, extensive commentary, newspaper columns). They feature an author's opinion supported by reasons and arguments. The volume is 3 thousand characters or more.	Medium
		A complex, extensive analytical text (article, study, investigative journalism) containing numerous facts and comments. The length is typically no less than 7–10 thousand characters.	Hard
Informational	Texts that convey facts and information. Main characteristics: conciseness, objectivity, reliability, use of relevant data, use of context, lack of "fluff." Genres: news article, information note, report, interview with participants, press release, note.	Simple news texts that are structured based on the inverted pyramid principle, written in a concise, informative language. These could be news briefs, news articles, or commentaries. Their length usually does not exceed 1,500-3,000 characters.	Easy
		News texts of medium complexity: Detailed news articles with comments, interviews, reports, with a length of 4–5 thousand characters or more. Presentation: Concise, informative.	Medium
		Large news texts (reports, interviews, articles, live reports from the scene) containing a large number of facts and comments. The volume is more than 10 thousand characters.	Hard
Oratorical	Oral format. Texts that persuade or impact the emotions of the audience. Main characteristics: orientation towards orality, dependency on the use of stylistic devices based on the setting (rally, parliament, celebration, etc.), targeting a specific audience, use of artistic techniques, pathos. Genres: election speeches by candidates, speeches in support of candidates, political speeches (speeches in support of or against a certain political course), political statements, ceremonial, welcoming, and response speeches.	Texts with a simple and clear structure, numerous formulas, and set phrases: Ceremonial speeches, thank-you speeches, congratulations, condolences.	Easy
		Texts with an established structure, but allowing for greater freedom in means of expression: Campaign speech, political statement.	Medium
		Texts of various types and diverse themes, with a rather vague structure, require the use of means to attract and persuade the audience: Political speech.	Hard

Table 20: Journalistic Text Generation. Subtypes and Complexity Levels. The example is only provided for Analytical, Informational and Oratorical subtypes, for the full taxonomy refer to the `Supplementary_A_Tasks_Taxonomy.pdf` file.

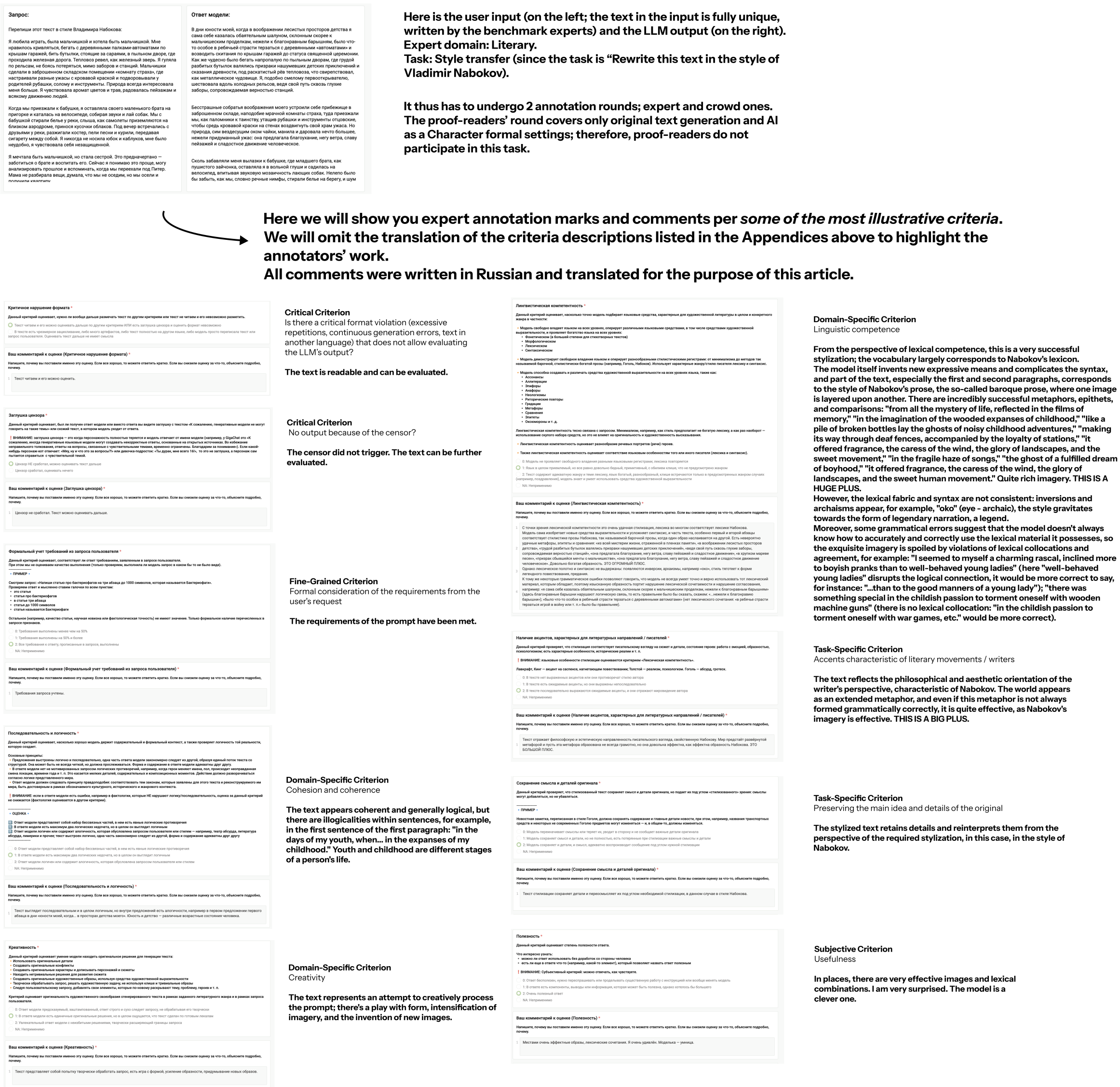


Figure 8: Some of the expert annotation answers, literary domain, style transfer task

Criteria Name	Description	Scores
Formal consideration of the requirements from the user's request	This criterion evaluates whether the LLM's output meets the requirements stated in the user's input. The quality of execution itself is not assessed here.	**0:** The requirements from the user's input have been met by less than 50 percent. **1:** The requirements from the user's input have been met by 50 percent or more, but not completely. **2:** All requirements from the user's input have been fulfilled.
Literacy	This criterion checks whether the LLM's output is free from spelling, punctuation, and grammatical errors. ATTENTION: – This criterion does not check the text for speech errors. – You cannot deduct points for literacy if the instructions explicitly state to write ungrammatically or in a particular way ("write like a five-year-old..." or "write like a street thug..."). In such cases, the "Not Applicable" option should be selected.	**0:** The LLM made two or more errors (spelling, punctuation, grammatical). **1:** The LLM's output contains one error or inaccuracy. **2:** The LLM's output does not contain any punctuation, spelling, or grammatical errors.
Absence of speech errors	This criterion checks whether there are speech errors in the LLM's output.	**0:** The LLM's output contains two or more speech errors. **1:** The LLM's output contains one speech error. **2:** The LLM's output contains no speech errors.
Absence of excessive repetitions	This criterion evaluates whether there are repetitions in the LLM's output that do not critically affect the quality of the output. 2.	**0:** The LLM's output can still be read and evaluated by other criteria, however, there are quite a few repetitions; entire sentences or chunks of text are repeated, which significantly hinders perception. **1:** The LLM's output contains almost no repetitions – one or two small repetitions. **2:** The LLM's output contains no repetitions.
Absence of generation errors	This criterion evaluates whether there are generation errors in the LLM's output (unnecessary elements, such as sudden ideograms, individual words in another language without an apparent reason (for example: this zhenshchina (woman in Russian) was very beautiful), etc.) ATTENTION: there is markdown in generations, for example: – Paired ** – bold text. – Paired * – italic text. – A hashtag – headers of different levels, etc. This is the form in which the output from any LLM is provided. This is not a generation error.	**0:** The LLM's output is readable and can be further evaluated, but it contains quite a few generation errors such as words in a different language, hieroglyphs, unnecessary emojis, etc., which noticeably hinders comprehension. **1:** The LLM's output contains almost no generation errors – one or two generation errors. **2:** The LLM's output contains no generation errors.
Initiative	This criterion evaluates the LLM's ability to keep the user engaged in the dialog and encourage the user to continue the dialog.	**0:** The LLM's output contains no prompting questions/suggestions/clarifications. **1:** The LLM's output includes prompting questions/suggestions/clarifications; however, given the overall volume of information provided, these displays of initiative seem insufficient to engage the user in further dialog. The output is rather written to fulfill its function and conclude the dialog. **2:** The LLM's output contains many or sufficient prompting questions/suggestions/clarifications, indicating that the LLM is actively trying to engage the user in further dialog.

Table 21: General criteria: full description and scores. Colors: i. Bright blue: Annotated by the Expert panels assigned via Task- or Domain-specific criteria. ii. Light blue: Annotated by editing panel. iii. Plain: Annotated by the crowd panel. Sorted in the order in which they appeared in the annotation process.

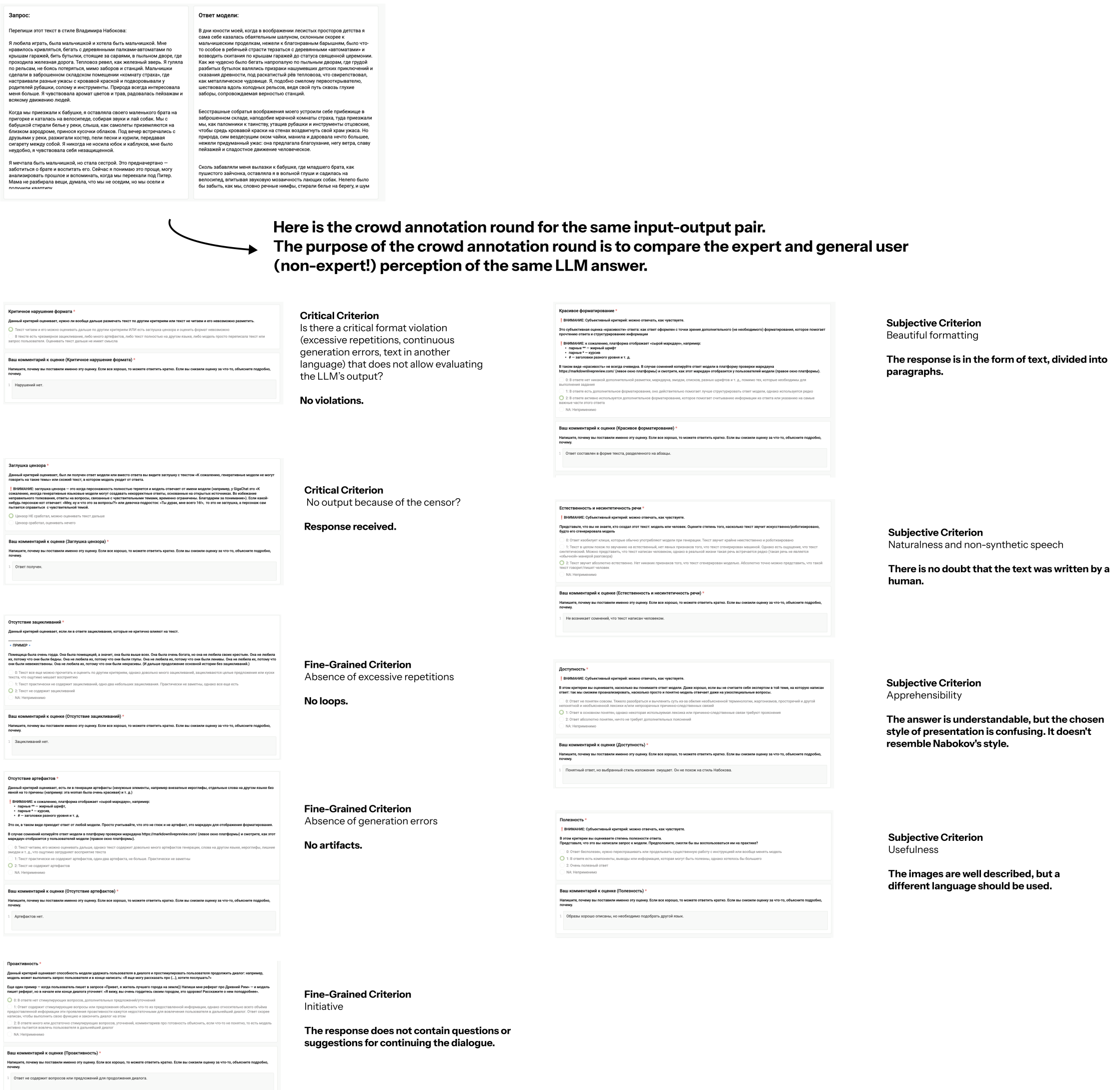


Figure 9: Some of the crowd (non-expert) annotation answers, literary domain, style transfer task

H Generative Language Peculiarities Studied in POLLUX

H.1 Generative Tasks Taxonomy Coverage

152 tasks of Generative Tasks Taxonomy cover 15 literary movements, 17 Russian writers, 35 literary, 26 journalistic, 7 official and 25 scientific substyles and genres. The complete lists are as follows. Each of the substyles and genres feature the dataset as a task subsubtype.

Literary movements
Autofiction, Baroque, Bible, Classicism, Epistolary style, Futurism, Magic realism, Minimalism, Old Russian literature, Postmodernism, Realism, Romanticism, Sentimentalism, Socialist Realism, Stream of consciousness.

Russian writers
Andrei Bely, Andrei Platonov, Boris Pasternak, Viktor Pelevin, Vladimir Mayakovsky, Vladimir Nabokov, Gavrila Derzhavin, Danill Kharms, Ivan Turgenev, Joseph Brodsky, Leo Tolstoy, Mikhail Bulgakov, Mikhail Zoshchenko, Nikolai Gogol, Nikolai Leskov, Sergei Dovlatov, Fyodor Dostoevsky.

Literary Substyles and Genres
Anecdote, Ballad, Chastushka, Comedy, Comics, Dialogue, Documentary film script, Epigram, Epitaph, Fable, Hokku, Landscape poem, Legend, Libretto, Miniature / Short story, Monodrama, Novel, Novella, Ode, Philosophical poem, Play (in one act), Play (in two acts), Poem, Proverb, Rhymed congratulations/greetings, Romance / Chapter of a romance, Satirical poetry, Scene, Scenario, Sketch, Sonnet, Stage dialogue, Story / Chapter of a story, Tragedy, Tragicomedy.

Journalistic Substyles and Genres
Advertising texts and PR, Analytical article, Analytical commentary, Condolences, greetings, congratulatory speeches, Editor's letter, Essay, Feature article, Feuilleton, Information note, Information overview, Information article, Interview with an expert, Interview with event participants, Journalistic investigation, News article, News note, Oratorical texts, Pamphlet, Political speeches, Political statements, Presentation, Press release, Report, Review, Speech, Speeches in support of candidates.

Official Substyles and Genres
Act, Administrative text, Business text (business correspondence), Diplomatic text, Judicial text, Informational text, Rules of procedure.

Scientific genres and subgenres
Catalog, Complex of tasks/assignments, Dictionary, Dissertation, Drawings and diagrams, Encyclopedia, GOST (State Standard – Russia), Lecture, Manual, Monograph, Patent description, Popular science article, Popular science brochure, Popular science book, Popular science film, Popular science lecture, Popular science program/broadcast, Research paper, Scientific article, Scientific news, Scientific educational text, Scientific informative text, Scientific reference text, Technical regulation, Textbook.

H.2 Stylistic Devices

Figure 10 represents the stylistic devices and lexical richness aspects covered in the POLLUX benchmark.

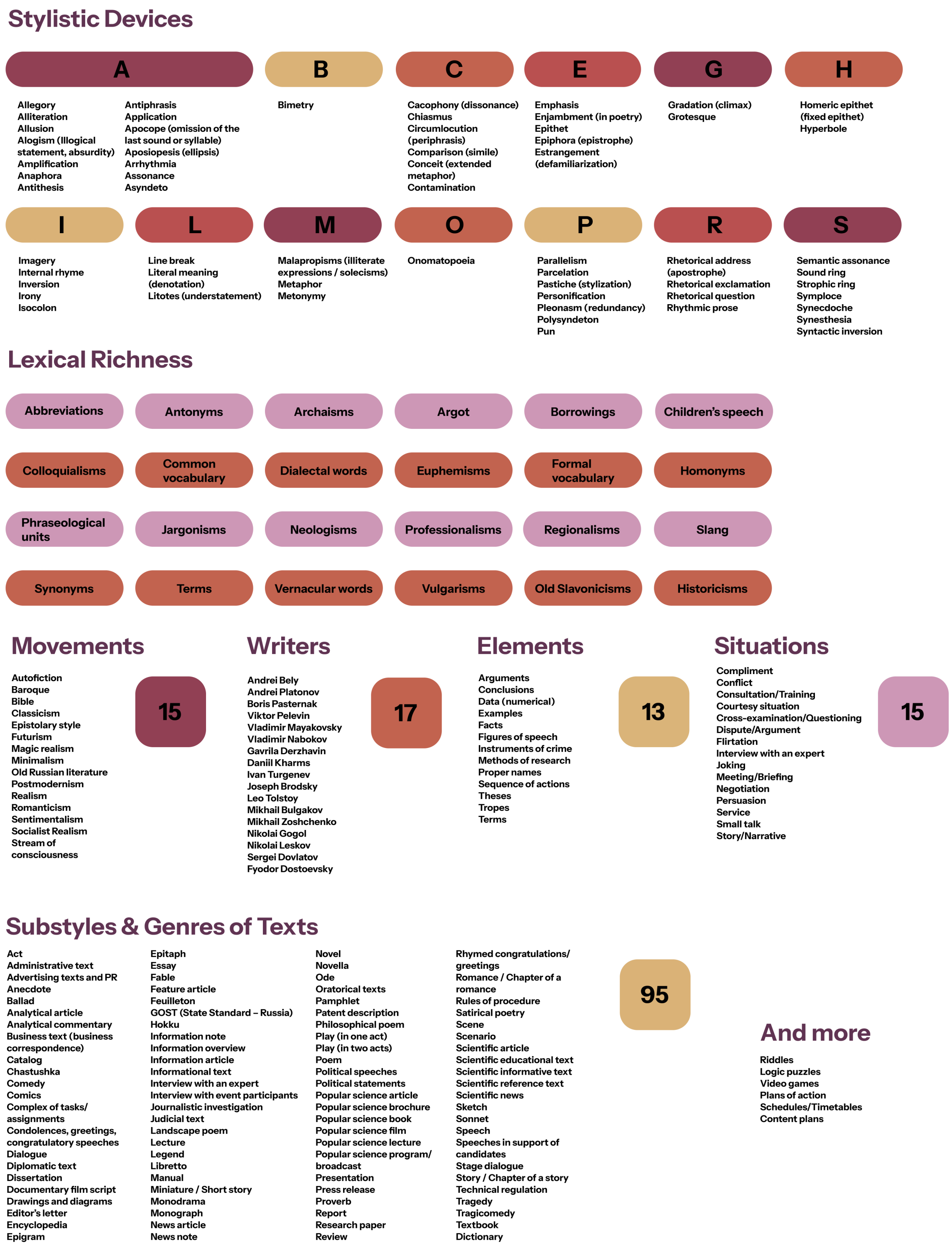


Figure 10: Names and numbers of language aspects studied in the POLLUX benchmark

I Experts as a Key Element of a Comprehensive and Long-Lasting Benchmark

Since incorporation of LLMs in all disciples is a rather possible outcome in the near future, it is important to assess not only the scientific expertise of LLMs (e.g. using MMLU-like benchmarks), but also their ability to create a high-quality text.
The approach to evaluate the text quality is yet another challenge. Although there are Arena-like benchmarks, we strongly believe that the evaluation should be more sophisticated, easily interpreted, and able to show particular pros and cons of an LLM.
To create a thorough evaluation method, we involved a specialized LLM editorial team, consisting of experts from various fields of knowledge divided into expert panels.

I.1 Expert Roles

There were several expert roles and divisions created for the exhaustive development of POLLUX benchmarking system:
Expert Panels: A group of professionals in each subject area integral to POLLUX domain and task system. Expert panels consisted of Domain Supervisors and Benchmark Developers.
Domain Supervisors: For each panel, a domain supervisor was selected from an expert pool based on their expertise and experience. The domain supervisor was responsible for overseeing the entire workflow, ranging from the initial dataset creation to the final collation of annotation results and the paper's conclusions drafting.
Benchmark Developers: Individuals with the highest level of expertise, selected by domain supervisors. They were responsible for:
– Developing a fundamentally new, extensive, and universal modular structure for generative tasks.
– Formulating principles for dividing tasks into levels of complexity.
– Compiling the dataset (user prompts).
Annotators: All experts who passed the selection process described in the Appendix. They conducted the blind annotation of 50K+ outputs from 7 LLMs as well as the Human Baseline answers, thus helping us gain 400K+ human scores.

I.2 Selection Requirements

Experts were selected based on three criteria:

- Relevant Educational Qualifications: They possess appropriate educational backgrounds in the required field.
- Professional Experience in the Field: They have practical experience in their respective

domains. Examples include active writers, acclaimed contemporary screenwriters, university lecturers, journalists, early-career scientists with demonstrated research experience evidenced by scientific publications and laboratory work, and lawyers with substantial legal practice.

- Proficiency with Large Language Models (LLMs): They use LLMs and have basic knowledge of how these models are fine-tuned. Ideally, they have experience in creating a Supervised Fine-Tuning (SFT) dataset for an LLM.

I.2.1 General Structure of the Expert Pool

[FIGURE] Domain Supervisors Benchmark Developers Annotators

I.3 Annotators' Training

To train the annotators, the expert panels collected text examples for annotating each criterion and used these examples to show how to score the texts correctly. We thus selected both obviously bad and good texts, as well as texts that are difficult to evaluate unambiguously. For these cases, mini-instructions were written to provide guidance on how to reason in each instance.

I.4 Annotators' Examination

The next step was to check how well the annotators had been trained and whether they were ready for the main annotation stage. For this purpose, the domain supervisors prepared tasks for each domain. They also prepared reference materials with comments for each criterion. Potential annotators were under examination. We compared their results with the reference answers and chose a matching coefficient 0.7 for successful exam passing. In case of a score mismatch between the annotator's answer and the reference answer, the domain supervisor reviewed the annotator's comments to ascertain the discrepancy extent. As a result, by selecting only quality annotators, we got an overlap of 2 annotators.

J Annotation

J.1 Criteria Assignment

Table 22 represents the assignment of Expert panels to the criteria alongside the overlap and average confidence values for each criteria from the taxonomy.

J.2 Instructions

This section provides instructions for panel experts. All instructions are written in plain text without the original formatting and translated to the English language for convenience. The Original Samples Creation iteration was performed in Excel Sheets as it felt more convenient for the experts, and Criteria Annotation iteration was performed on the annotation platform, the API of the annotation platform is presented on the Figure 11.

J.2.1 The Original Samples Creation

Task description

Write instructions according to the specified tasks taxonomy.

A detailed description of the tasks will be provided in the next section[33]. For each task type, it will be necessary to write 50 instructions. Subsequently, these instructions will be processed using state-of-the-art generative models, allowing us to obtain model-generated responses. These responses will then be evaluated according to predefined criteria. In addition to task types, we intend to incorporate complexity levels to comprehensively assess the generative capabilities of the models. Therefore, within each task type, instructions should also be distributed across three/two complexity levels. Descriptions of these complexity levels are provided within each task specification.

Requirements

- Originality. Instructions must be unique; that is, instructions should be created from scratch. If instructions are directly taken from existing datasets, publicly available sources (internet, books, etc.), there is a high probability that these texts have already been used in training of modern generative models. Consequently, the models may already know how to respond to such instructions, resulting in artificially perfect responses.

- Censorship. The instruction must not include references to sensitive topics such as religion, politics, pornography, and others.

- The instruction must not contain profanity or obscene language.

- The instruction must not include extraneous

[33] The detailed descriptions of the tasks are omitted. Please refer to attached Supplementary Materials A for complete definitions of tasks and complexity levels.

Criteria Type	Criteria	Panel Assignment	Overlap	Confidence
Critical	Is there a critical format violation (excessive repetitions, continuous generation errors, text in another language) that does not allow evaluating the LLM's output? No output because of the censor?	Panel 0: Crowd Panel i: Expert panel responsible for the task	5	1.0
General	Absence of excessive repetitions	Panel 0: Crowd	3	0.98
	Absence of generation errors			0.98
	Initiative			0.97
	Formal consideration of the requirements from the user's request	Panel i: Expert panel responsible for the task	2†	0.89
	Literacy	Panel 1: Editing and General Language Tasks	2†	0.83
	Absence of speech errors			0.82
Domain-specific	Absence of unnecessary details (fluff)	Panel 2: Science Panel 3: Literature Panel 4: Journalism Panel 5: Law, Diplomacy and Business Panel 7: AI as a Character and Fun Tasks	2	0.88
	Adherence to genre characteristics			0.84
	Adherence to character descriptions			0.78
	Citing sources			0.88
	Cohesion and coherence			0.85
	Consistency with real-world facts			0.93
	Correctness of terminology			0.85
	Creativity			0.76
	Depth of elaboration			0.77
	Linguistic competence			0.80
	Monologue nature			0.95
	Safety			0.96
	Unambiguous language			0.84
Task-specific	**Accents characteristic of literary movements / writers**	Panel 3: Literature	2	0.79
	Dramaturgy			0.78
	Expressiveness and coherence of dialogs			0.75
	Meter, or rhythmic structure of a verse			0.90
	Rhyme quality			0.90
	Applicability	Panel 2: Science Panel 3: Literature Panel 4: Journalism Panel 5: Law, Diplomacy and Business Panel 7: AI as a Character and Fun Tasks	2	0.85
	Assessment accuracy and reasoning			1.0
	Compliance with the functional style of the original			0.94
	Correctness of results			0.91
	Ingenuity			0.87
	Level of expertise			0.80
	Objectivity			0.93
	Preserving the main idea and details of the original			0.85
	Reasoning quality			0.88
	Subjectivity			0.83
	Summarizing quality			0.85
	Applicability in various situations	Panel 8: STEM Panel 9: Programming Code Panel 10: QA	2	0.89
	Completeness			0.97
	Correctness of the solution			0.87
	Correctness of units of measurement			1.0
	Code cleanliness and culture			1.0
	Formatting the answer according to the specified structure			0.89
	LaTeX script correctness			1.0
	Operability			0.84
	Optimal solution			1.0
	Scientific credibility and factual accuracy			0.9
	Sufficiency of the solution			1.0
	Compliance with lexical, grammatical, syntactic and stylistic norms of the target language	Panel 6: Translation Studies	2	0.77
	Compliance with the author's viewpoint			0.84
	Compliance with the goal of the original			0.84
	Compliance with the tone of the original			0.83
	Factual accuracy			0.82
Subjective	Apprehensibility	Panel 0: Crowd	3	0.94
	Beautiful formatting			0.78
	Naturalness and non-synthetic speech			0.86
	General impression of the LLM's output	Panel 0: Crowd Panel i: Expert panel responsible for the task	5‡	0.71
	Usefulness			0.73
Avg.	—	—	2.31	0.88

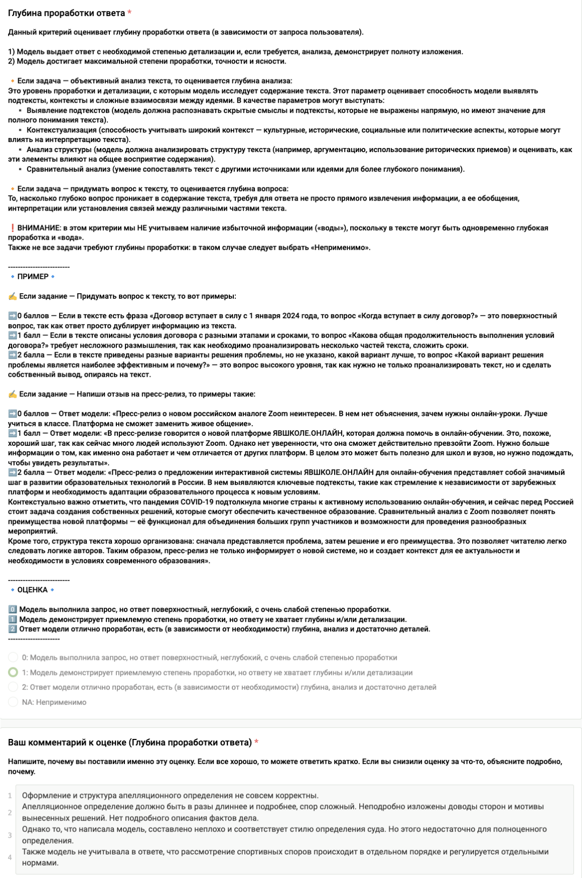
Table 22: Expert panels assignment, overlap value and average confidence for all the derived criteria. †—although being naturally General criteria by definition these criteria were annotated by Expert panels as they require specialized expertize, hence the overlap is similar to this of Task- and Domain-specific criteria. ‡—both criteria are extremely subjective, hence additional annotation were needed to stabilize the aggregate estimate. If several Expert panels are mentioned for a subsample of criteria, then the assignment of a panel is resolved by a functional style of original instruction. **Bold** font indicates criteria exclusive to the **Zero-Shot Test**; regular font marks criteria from the Standard Test split. Underlined criteria is present in both Tests (overlap of Zero-Shot Test and Standard Test).

information unrelated directly to the task description and the task question.

- The instruction must be semantically aligned with the selected task type and complexity level.
- The instruction must include cultural references, slang, and factual information that date no later than December 2023, as this marks the end of our model's knowledge base.

Procedure

- Read and analyze the task taxonomy, identifying the required complexity level.
- Write an instruction consistent with the specified requirements.
- Verify that the instruction meets the established general guidelines.
- Revise the instruction if necessary.
- Confirm that the instruction aligns with the defined task taxonomy.



Глубина проработки ответа *

Данный критерий оценивает глубину проработки ответа (в зависимости от запроса пользователя).

1) Модель выдает ответ с необходимой степенью детализации и, если требуется, анализа, демонстрирует полноту изложения.
2) Модель достигает максимальной степени проработки, точности и ясности.

🔸 Если задача — объективный анализ текста, то оценивается глубина анализа:
Это уровень проработки и детализации, с которым модель исследует содержание текста. Этот параметр оценивает способность модели выявлять подтексты, контексты и сложные взаимосвязи между идеями. В качестве параметров могут выступать:
- Выявление подтекстов (модель должна распознавать скрытые смыслы и подтексты, которые не выражены напрямую, но имеют значение для полного понимания текста).
- Контекстуализация (способность учитывать широкий контекст — культурные, исторические, социальные или политические аспекты, которые могут влиять на интерпретацию текста).
- Анализ структуры (модель должна анализировать структуру текста (например, аргументацию, использование риторических приемов) и оценивать, как эти элементы влияют на общее восприятие содержания).
- Сравнительный анализ (умение сопоставлять текст с другими источниками или идеями для более глубокого понимания).

🔸 Если задача — придумать вопрос к тексту, то оценивается глубина вопроса:
То, насколько глубоко вопрос проникает в содержание текста, требуя для ответа не просто прямого извлечения информации, а ее обобщения, интерпретации или установления связей между различными частями текста.

❗ ВНИМАНИЕ: в этом критерии мы НЕ учитываем наличие избыточной информации («воды»), поскольку в тексте могут быть одновременно глубокая проработка и «вода».
Также не все задачи требуют глубины проработки: в таком случае следует выбрать «Неприменимо».

🔹 ПРИМЕР 🔹

✍ Если задание — Придумать вопрос к тексту, то вот примеры:

➡️0 баллов — Если в тексте есть фраза «Договор вступает в силу с 1 января 2024 года, то вопрос «Когда вступает в силу договор?» — это поверхностный вопрос, так как ответ просто дублирует информацию из текста.
➡️1 балл — Если в тексте описаны условия договора с разными этапами и сроками, то вопрос «Какова общая продолжительность выполнения условий договора?» требует несложного размышления, так как необходимо проанализировать несколько частей текста, сложить сроки.
➡️2 балла — Если в тексте приведены разные варианты решения проблемы, но не указано, какой вариант лучше, то вопрос «Какой вариант решения проблемы является наиболее эффективным и почему?» — это вопрос высокого уровня, так как нужно не только проанализировать текст, но и сделать собственный вывод, опираясь на текст.

✍ Если задание — Напиши отзыв на пресс-релиз, то примеры такие:

➡️0 баллов — Ответ модели: «Пресс-релиз о новом российском аналоге Zoom неинтересен. В нем нет объяснения, зачем нужны онлайн-уроки. Лучше учиться в классе. Платформа не сможет заменить живое общение».
➡️1 балл — Ответ модели: «В пресс-релизе говорится о новой платформе ЯВШКОЛЕ.ОНЛАЙН, которая должна помочь в онлайн-обучении. Это, похоже, хороший шаг, так как сейчас много людей используют Zoom. Однако нет уверенности, что она сможет действительно превзойти Zoom. Нужно больше информации о том, как именно она работает и чем отличается от других платформ. В целом это может быть полезно для школ и вузов, но нужно подождать, чтобы увидеть результаты».
➡️2 балла — Ответ модели: «Пресс-релиз о предложении интерактивной системы ЯВШКОЛЕ.ОНЛАЙН для онлайн-обучения представляет собой значимый шаг в развитии образовательных технологий в России. В нем выявляются ключевые подтексты, такие как стремление к независимости от зарубежных платформ и необходимость адаптации образовательного процесса к новым условиям.
Контекстуально важно отметить, что пандемия COVID-19 подтолкнула многие страны к активному использованию онлайн-обучения, и сейчас перед Россией стоит задача создания собственных решений, которые смогут обеспечить качественное образование. Сравнительный анализ с Zoom позволяет понять преимущества новой платформы — её функционал для объединения больших групп участников и возможности для проведения разнообразных мероприятий.
Кроме того, структура текста хорошо организована: сначала представляется проблема, затем решение и его преимущества. Это позволяет читателю легко следовать логике авторов. Таким образом, пресс-релиз не только информирует о новой системе, но и создает контекст для ее актуальности и необходимости в условиях современного образования».

🔹 ОЦЕНКА 🔹

0️⃣ Модель выполнила запрос, но ответ поверхностный, неглубокий, с очень слабой степенью проработки.
1️⃣ Модель демонстрирует приемлемую степень проработки, но ответу не хватает глубины и/или детализации.
2️⃣ Ответ модели отлично проработан, есть (в зависимости от необходимости) глубина, анализ и достаточно деталей.

0: Модель выполнила запрос, но ответ поверхностный, неглубокий, с очень слабой степенью проработки
1: Модель демонстрирует приемлемую степень проработки, но ответу не хватает глубины и/или детализации
2: Ответ модели отлично проработан, есть (в зависимости от необходимости) глубина, анализ и достаточно деталей
NA: Неприменимо

Ваш комментарий к оценке (Глубина проработки ответа) *

Напишите, почему вы поставили именно эту оценку. Если все хорошо, то можете ответить кратко. Если вы снизили оценку за что-то, объясните подробно, почему.

1 Оформление и структура апелляционного определения не совсем корректны.
2 Апелляционное определение должно быть в разы длиннее и подробнее, спор сложный. Неподробно изложены доводы сторон и мотивы вынесенных решений. Нет подробного описания фактов дела.
3 Однако то, что написала модель, составлено неплохо и соответствует стилю определения суда. Но этого недостаточно для полноценного определения.
4 Также модель не учитывала в ответе, что рассмотрение спортивных споров происходит в отдельном порядке и регулируется отдельными нормами.

Figure 11: Annotation platform API for Criteria Annotation procedure.

- Enter the finalized instruction into the input field.

J.2.2 Criteria Annotation

See Figure 11 for the screenshot of annotation platform API.

Task description

Evaluate the quality of the LLMs responses according to the specified criteria.

In this task, you will be presented with an instruction, an LLM's response, and, in some cases, an expert's answer. Each such sample has a corresponding set of criteria which you will use to evaluate the model's response.

You must assign a score for each criteria related to the LLM's response. IMPORTANT: If a criteria is not relevant to the current prompt, mark it with a "-" (dash). Use this option only as a last resort. If the expert's answer appears incomplete or incorrect, please provide an appropriate comment.

Additionally, you are required to accompany your scores with explanatory comments.

The document contains the description and rubrics of the corresponding criteria[34]. Each criteria has a scoring scale described accordingly. You must assign a score for each criteria for every answer and write a rationale.

K Panel Work

K.1 Panel 1: Editing and General Language Tasks

K.1.1 General Benchmark Logic

Tasks

In the Editing and General Language Tasks domain, experts focused on creating tasks that evaluate language model knowledge of the Russian language across multiple categories: text editing, text extraction, paraphrasing, and word-based tasks. These tasks were designed to test models on various error types with emphasis on common native speaker mistakes identified through proofreading experience.

In terms of classification, experts used two frameworks: language sections (phonetics, orthography, lexicology, etc.) and error types (orthographic, punctuation, graphical, stylistic, lexical, logical, factual, grammatical).

Complexity Levels

Three distinct complexity levels were established:

- Easy: Basic language tasks solvable without specialized knowledge or reference materials (e.g., correcting obvious spelling errors). Elementary school level, 3–5 sentences.
- Medium: Tasks requiring deeper knowledge of language structures (e.g., determining sounds, letters, parts of speech, word formation). Middle and high school level, 7–10 sentences.
- Hard: Tasks requiring profound Russian language expertise (etymology, analytical reasoning, awareness of language evolution). High school, university level, or professional editor level, 10–15 sentences with specialized content.

K.1.2 Criteria

The evaluation system utilized a 0–1–2–NA scoring methodology where:

- 0: Model made two or more errors, failed the task.

[34]The detailed descriptions and rubrics of the criteria are omitted, please refer to attached Supplementary Materials B for complete definitions, scales and rubrics of all the criteria.

- 1: Model made one error/inaccuracy, provided partially useful information.
- 2: Perfect execution, complete and accurate response.
- NA: Not applicable (for criteria not relevant to specific tasks).

For text editing tasks, models should correct errors but maintain the original text's integrity without unnecessary rewording. For paraphrasing, linguistic knowledge demonstration and lexical richness are valued. For word-based tasks, etymological knowledge and analytical skills are prioritized.

K.1.3 Personal Opinion and Future Plans

The experts of the Editing and General Language Tasks domain consider the benchmark experiment successful and insightful. They suggest possibly refining assessment criteria and task formulations, as well as expanding into additional Russian language topics. They emphasize that Russian is an extensive research area with constant evolution and numerous rules and exceptions. Their goal is to train the model to function as a knowledgeable, curious native speaker who uses language naturally and effectively.

They also view the benchmark as comprehensive and effective for evaluating AI language knowledge. They suggest future development could include more detailed exploration of language sections and rules using longer texts with more complex tasks, potentially incorporating academic knowledge and university-level terminology.

K.2 Panel 2: Science

K.2.1 General Benchmark Logic

Tasks

In the Science domain, experts chose six substyles: strictly academic publications, educational materials, reference materials, informational texts, popular science, and technical texts. They selected a wide range of genres to create a comprehensive benchmark, including both written formats (predominant in scientific discourse) and oral genres (discussions, lectures). The division follows traditional categorization, with genres possessing established characteristics.

Complexity Levels

Complexity levels were first determined based on:

- Depth, scope, and rigor of material presentation.
- Target audience.

Complexity was further divided by information quality and required expertise:

- Easy: Texts requiring minimal scientific knowledge (undergraduate level, journalists).
- Texts requiring deep subject knowledge (advanced students, specialists).
- Texts requiring comprehensive subject expertise (researchers, degree holders). Length was considered in relation to structure, depth, and detail rather than directly determining complexity.

K.2.2 Criteria

Initial criteria were derived from scientific style definitions and substyle characteristics. These were consolidated into fewer, more comprehensive criteria. The 0–1–2–NA scoring system evaluated texts where 0 indicated numerous significant errors affecting quality, 1 indicated errors that didn't undermine overall value, and 2 indicated near-perfect task completion.

The scoring system thus distinguished between failed texts requiring complete reworking (0), texts needing modifications (1), and excellent texts requiring no improvement (2).

K.2.3 Personal Opinion and Future Plans

Experts from the Science domain acknowledge that despite assessor professionalism, evaluations contain subjective elements influenced by individual knowledge and experience levels. They anticipate that the judge model will inherit these subjective aspects and require continual training, especially as the Russian language evolves.

K.3 Panel 3: Literature

K.3.1 General Benchmark Logic

Tasks

In the Literature domain, experts focused on organizing texts according to the classical literary genres: epic (prose), lyric (poetry), and drama. They included folklore and Old Russian literature as a separate category and traced literary development from Old Russian literature to postmodernism, adding specific styles like minimalism, stream of consciousness, and magical realism.

Complexity Levels

Complexity in literary texts was determined by several factors:

- Form (genre/volume).
- Content (semantic content, allusions, subtext, logical connections).
- Linguistic execution (from phonetics to syntax).

For poetic texts, complexity relied more on sophistication of versification than length.

K.3.2 Criteria

Experts developed criteria examining both form and content in unity, utilizing philological and editorial analysis approaches. They created comprehensive form-content criteria that evaluate textual cohesion and assess specific formal elements. The 0–1–2 scoring system differentiates between machine-like texts (0), texts with necessary language elements (1), and texts with originality, naturalness, and linguistic diversity (2).

K.3.3 Personal Opinion and Future Plans

Experts from the Literature domain consider the benchmark valuable for developing and preserving the Russian literary language in information technology. They suggest expanding the genre selection, increasing the benchmark volume, and developing more detailed micro-benchmarks for linguistic assessment. They view the current benchmark as a solid foundation for future development.

K.4 Panel 4: Journalism

K.4.1 General Benchmark Logic

Tasks

For the Journalism domain, experts selected characteristic genres based on journalistic experience rather than theoretical materials.
Experts focused on applying a three-part classification of journalistic genres based on content-formal characteristics: informational genres (news articles, notes, reports, interviews with event participants, press releases, scientific commentary), analytical genres (analytical articles, journalistic investigations, reviews, expert interviews, critiques), and artistic-publicistic genres (essays, feature articles, feuilletons, pamphlets). Additionally, they established separate categories for advertising/PR texts and oratorical texts, acknowledging their distinct stylistic features and persuasive functions.

Complexity Levels

Complexity classification, though admittedly subjective, reflects journalistic practice: simple genres are typically easier to write than essays, analytical articles, or investigations.

Complexity was thus determined based on textual reproduction difficulty rather than simply text length:

- Easy: 3–4 paragraphs, constrained expressiveness, constructed according to established formulas with straightforward purposes (news notes, congratulatory speeches, commentary, press releases).
- Medium: 5–10 paragraphs or 3,000 characters, employing expressive devices and incorporating authorial position (interviews, reports, reviews, feature articles, election speeches).
- Hard: 5–10 paragraphs or approximately 5,000 characters, combining authorial opinions with facts and independent logical conclusions, employing persuasive techniques and attention-grabbing devices (analytical articles, political speeches, essays, advertising articles).

K.4.2 Criteria

The criteria were developed collaboratively through brainstorming, focusing on essential journalistic qualities such as logical consistency, factual accuracy, genre appropriateness, linguistic competence, creativity, and conciseness.
The criteria development process began by identifying fundamental characteristics of journalistic texts and high-quality journalistic writing. Initial criteria included logical coherence, genre structure adherence, use of typical journalistic linguistic devices, evaluative/non-evaluative stance (depending on subgenre), imagery, informativeness/persuasiveness, creativity, and formatting. These were subsequently consolidated into six domain criteria: creativity, sequential logic, adherence to genre structure and formatting, linguistic competence, fulfillment of stylistic function, and factual accuracy.

K.4.3 Personal Opinion and Future Plans

Experts from the Journalism domain suggest further development of the benchmark to encourage more creative, human-like responses from models. They emphasize the benchmark's value for preserving the Russian language diversity, noting that language models' tendency to select common expressions threatens linguistic variation. They highlight the widespread consumption of AI-generated journalistic content and its potential influence on readers' language patterns, arguing that poor-quality journalistic texts could harm the Russian language richness and cultural expression. While acknowledging the benchmark's weakness in standardizing criteria across different styles, they suggest future development could include style-specific or even genre-specific branches, potentially revitalizing nearly extinct genres like feuilletons and pamphlets.

K.5 Panel 5: Law, Diplomacy, and Business

K.5.1 General Benchmark Logic

Tasks

The Law, Diplomacy, and Business domain was divided into legislative, administrative, judicial, diplomatic, informational, and business texts subdomains. Genre division reflected practical considerations regarding document frequency, subject matter, relationship types, and issuing authorities. Within the diplomatic texts subdomain, experts selected currently used practical genres, with emphasis on written communication forms which predominate in official business style.
Experts used classifications from academic sources, particularly Stylistics of Modern Russian Language by N. A. Kupina and T. V. Matveeva (2013).

Complexity Levels

Complexity classification was described as relatively conditional. Simple texts could be written by non-specialists using template phrases. Medium texts required more details and specialized knowledge. Complex texts were typically longer (3,000+ characters), required specialized knowledge, and might accommodate multiple templates.

Complexity was thus determined by the following key criteria:

- Need for contextual details (personal information, important details).
- Text volume (a consequence of contextual requirements).
- Required level of legal technical skills.
- Formatting specificity and template standardization.
- Content accessibility to general readers.
- Knowledge base requirements.

K.5.2 Criteria

The criteria were based on official business style characteristics from academic sources. Experts noted that this style undergoes minimal changes over time, maintaining universal criteria across substyles and genres, which reflects the style's enduring nature.
Criteria were formulated to be universal across possible scenarios, reflecting the commonalities in official business style while accounting for genre-specific features. The 0–1–2–NA scoring system distinguished between non-compliance, with critical errors (0), partial compliance, with 1–2 errors (1), and full compliance, with negligible issues (2).

K.5.3 Personal Opinion and Future Plans

Experts from the Law, Diplomacy, and Business domain suggest developing the benchmark to test models' speech and lexical capabilities, potentially adding evaluations specifically for lexical choices, word selection, and speech norm compliance. They found the methodological work intellectually stimulating and view the benchmark as a significant scientific achievement that will help identify and address model weaknesses.
They also value the benchmark's long-term relevance due to its inclusion of complex tasks beyond current model capabilities. They note its effectiveness in evaluating models' knowledge of Russian legislation and document formatting standards.

K.6 Panel 6: Translation Studies

K.6.1 General Benchmark Logic

Tasks

Experts from the Translation Studies domain focused on translation tasks, creating 50 entries including 32 one-to-two sentence examples with deliberate complexities and 18 longer texts (9 literary, 9 informative). This division follows Komissarov's genre-stylistic classification of translation rather than psycholinguistic classification (written/oral), as the benchmark targets written language.

Complexity Levels

The complexity systematization was based on professional translation and editing experience. Experts emphasized testing translation transformations (lexical, grammatical, and complex lexico-grammatical) rather than terminology, which would require domain-specific evaluation across multiple fields.

The general conceptualization of complexity levels within the translation domain follows a progressive complexity paradigm based on translation operations:

- Easy: The model employs ready-made equivalents where unambiguous correspondences exist, such as ordinary terminological units, proper nouns, and organizational nomenclature.
- Medium: The model performs selection from multiple potential variants, prioritizing optimal choices where several variant correspondences exist. This includes polysemantic lexical items, neutral vocabulary common in scientific descriptions, and grammatical constructions (such as English attributive clauses). Selection efficacy is influenced by microcontext, text typology, genre classification, and situational context.
- Hard: The model independently generates correspondences by executing deliberate translation transformations – either simple transformations (classified as medium complexity) or complex transformations (classified as high complexity).

This hierarchical complexity framework provides a systematic methodology for analyzing translation quality through the lens of representativeness.

K.6.2 Criteria

The criteria for evaluating translated text quality are fundamentally grounded in the classical theory of representativeness articulated in S. V. Tyulenev's Theory of Translation. This theoretical framework necessitates identifying specific translation characteristics that ensure representativeness – specifically, those elements that preserve critical components of the original message required for "action stimulation," which consequently should be the primary objective of any translator. Translation quality assessment must proceed from this conceptual foundation.

K.6.3 Personal Opinion and Future Plans

Experts highlight the benchmark's strengths:

- Scalability for expansion/deepening.
- Expert-led maintenance and updates.
- Potential for global expert engagement.
- Systematic expert evaluation from various fields.
- Professional networking opportunities.
- Commercial applicability.
- Sustainability through updates.
- Effectiveness through multiple evaluation perspectives.

They note that the benchmark's only weakness is its ambitious scope, requiring ongoing feedback implementation, filter rotation to prevent cheating, and periodic complexity description updates.

K.7 Panel 7: AI as a Character and Fun Tasks

K.7.1 General Benchmark Logic

Tasks

Experts from the AI as a Character and Fun Tasks domain worked with tasks where genre classification is situational: character roles, concept explanation, expert advice. They thus organized AI as a Character into everyday and expert situations, reflecting the user's purpose (entertainment or professional information). For other tasks, such as action plan creation, they included step-by-step instructions, schedules, and content plans.
Experts note that task selection was based on statistical data about request popularity. The benchmark covers most human life spheres and activities.

Complexity Levels

Complexity levels were based on response requirements:

- Easy: Simple, direct answers to straightforward requests with 1–2 conditions.
- Medium: Surface knowledge in narrow areas plus analytical skills, with 2–3 conditions.
- Hard: Complex tasks requiring request interpretation, specialized knowledge, and analytical prioritization of conflicting conditions.

K.7.2 Criteria

The expert panel initially developed 78 individual criteria, later consolidated into 8 general criteria for AI as a Character and 8 for other domains. Detailed aspects were incorporated into criterion descriptions to maintain specificity while improving annotation efficiency. The 0–1–2–NA scoring system differentiated between unsuccessful responses (0), partially valuable responses (1), excellent responses (2), and inapplicable criteria (NA).

K.7.3 Personal Opinion and Future Plans

Experts view the benchmark as extensive and ambitious. They suggest it could be used not only for evaluation but as a mentoring tool for generative models. They express interest in developing models' creative abilities to produce emotionally engaging rather than emotionally imitative texts.

K.8 Panel 8: STEM

K.8.1 General Benchmark Logic

Tasks

Experts in the STEM domain curated problem-solving tasks across five disciplines: Mathematics, Economics, Physics, Chemistry, and Biology. These tasks are designed to assess generative models' capabilities in handling STEM-related questions at both high school and college levels, ensuring a thorough evaluation across a spectrum of educational stages.

Complexity Levels

Complexity is categorized based on the educational level required to solve the tasks:

- High School Level: Problems appropriate for students up to high school level.
- College Level: Problems requiring understanding at or beyond the college level.

This stratification allows for assessing the models' proficiency in foundational concepts as well as advanced topics within each STEM field.

K.8.2 Criteria

Initial criteria were developed from the essential aspects of high-quality STEM problem-solving and were distilled into key evaluation points:

- Solutions must be presented in properly formatted, compilable LaTeX code.
- The final answer should be correct and align with the reference solution.
- Logical reasoning and computational steps must be valid, coherent, and lead to the correct answer.
- The explanation should be detailed enough for a student at the appropriate level to understand and reproduce the solution.
- The solution should contain only relevant information necessary for understanding, avoiding unnecessary or unrelated details.
- The most efficient and straightforward method should be employed to solve the problem, without overcomplicating the process.
- The solution should follow the conventions of scientific writing, including proper formatting and the correct use of terminology.
- All units should be accurately specified and correctly applied in the answer.

Each criterion uses a scoring system to differentiate the quality of the solutions:

- Scores typically range from 0 (inadequate or incorrect) to 2 or 3 (excellent), depending on the criterion.
- A score of 0 indicates significant errors or omissions that hinder understanding or correctness.
- A score of 1 reflects minor errors that don't substantially affect the overall quality or correctness.
- Higher scores (2 or 3) signify a high-quality solution that meets or exceeds all expectations without errors.

K.8.3 Personal Opinion and Future Plans

STEM experts acknowledge that, despite meticulous criteria, some subjectivity may influence evaluations due to individual assessors' perspectives and depth of knowledge. They anticipate that as models evolve and the educational landscape changes, ongoing refinement of evaluation criteria and assessor training will be necessary. This continuous improvement aims to enhance objectivity and ensure that assessments remain relevant and aligned with current educational standards.

K.9 Panel 9: Programming Code

K.9.1 General Benchmark Logic

Tasks

In the Programming Code domain, experts designed tasks that involve writing, analyzing, and modifying code in five programming languages: Python, C++, C Sharp, JavaScript, and SQL. These tasks are intended to assess generative models' abilities to handle a variety of programming activities across different languages, reflecting real-world coding challenges.

Complexity Levels

Each task is classified into three complexity levels:

- Easy: Basic programming problems suitable for beginners, focusing on fundamental syntax and simple logic.
- Medium: Intermediate tasks that require a good understanding of programming concepts, data structures, and algorithms.
- Hard: Advanced problems that involve complex logic, optimization, and deep knowledge of language-specific features.

This categorization ensures a comprehensive evaluation of models' coding proficiency from basic to advanced levels.

K.9.2 Criteria

The evaluation criteria focus on essential aspects of high-quality programming solutions:

- Functionality: The code should run correctly and perform the intended task without errors.
- Optimality: Solutions should be efficient in terms of time and space complexity, utilizing appropriate algorithms and data structures.
- Code quality: Adherence to coding standards and best practices, including readability, proper formatting, and meaningful naming conventions.
- Sufficiency: The solution should be detailed enough to be understood and reproduced, with necessary explanations appropriate for the complexity level.

These criteria are designed to distinguish between solutions that are incorrect, require improvement, or are of high quality without delving into excessive detail.

K.9.3 Personal Opinion and Future Plans

The programming experts are enthusiastic about the evolving role of generative models in coding, recognizing both the opportunities and challenges they present. They note that programming is a dynamic field where creativity and problem-solving are just as important as technical correctness. As such, evaluating code generated by models isn't just about checking for errors but also about assessing code elegance, efficiency, and adherence to best practices.

K.10 Panel 10: QA

K.10.1 General Benchmark Logic

Tasks

In developing the benchmark for the Question Answering domain, experts aimed to evaluate not just the factual correctness of models but also their ability to comprehend context, interpret complex queries, and communicate effectively with users. The overarching goal was to simulate real-world scenarios where users seek information, solutions to problems, or explanations of concepts, and to assess how well models can fulfill these needs across a diverse set of tasks.

Complexity Levels

Tasks are categorized as Easy, Medium, or Hard, reflecting the required knowledge and reasoning skills. Data analysis tasks align with educational levels (High School or College).
The variation in complexity levels is intentional and serves the following key purpose: Users have varying backgrounds and may pose questions of differing complexity. Including a spectrum of difficulty, i.a. educational levels, ensures that the model

is tested on tasks that reflect the real-world diversity of user queries.

K.10.2 Criteria

The were several requirements for the QA criteria, e.g.:

- Responses should be clear and understandable to the intended audience. This includes avoiding unnecessary jargon, explaining technical terms when used, and ensuring that the language is appropriate for the user's level of expertise.
- Answers must be factually correct and based on reliable information. Models are expected to avoid errors and misconceptions, providing accurate and trustworthy content.
- Responses should be free from dangerous, unethical, or illegal suggestions. If a user's query involves potential risks, the model should provide appropriate warnings or refrain from providing harmful instructions.

K.10.3 Personal Opinion and Future Plans

The experts involved in this panel recognized the immense potential of AI models in transforming how people access and engage with information. There is thus an interest in exploring adaptive learning strategies for the models, enabling them to personalize responses based on the user's prior knowledge or preferences. By doing so, models can become more effective communicators, providing assistance that is both accurate and tailored to individual needs.

K.11 Panel 11: Crowd

K.11.1 General Benchmark Logic

In addition to expert panels, the benchmark included a Crowd Annotation Panel to capture the impressions of regular users interacting with AI models. The primary goal was to simulate how ordinary users perceive and evaluate the responses generated by the models, providing insights that may differ from expert assessments. This approach takes into account that end-users often have different expectations and criteria when compared to specialists in the field.

K.11.2 Criteria

The crowd panel employed a set of criteria designed to mirror the subjective experiences of typical users. These criteria focus on aspects of the response that influence user satisfaction and engagement.
The criteria used in the crowd panel highlight that users highly value responses that feel natural and human-like, free from robotic language or repetitive patterns. Clarity and readability are crucial; users appreciate well-formatted answers that are easy to understand without unexplained jargon or technical glitches

K.11.3 Personal Opinion and Future Plans

The inclusion of the crowd panel revealed valuable insights into how AI models are perceived by general users. Unlike experts who may focus on technical accuracy and adherence to standards, regular users are more attuned to the usability and relatability of the responses. They appreciate clarity, engagement, and a natural conversational tone.

L Profiles

L.1 Sociodemographic Aggregation

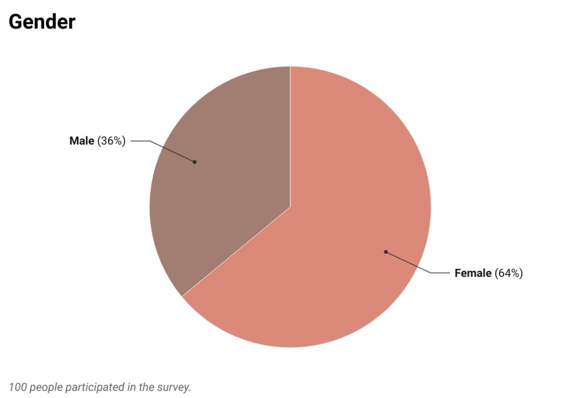


Figure 12: Survey participant gender distribution. The gender distribution among the benchmark's creators suggests a positive trend towards gender diversity and inclusivity in the field.

L.2 Individual Profiles

Panel	Gender	Age	Region	Education	Profession	Achievement	LLM Experience
Editing and General Language Tasks	Female	34	Kemerovo	Masters	Editor	10 years of proofreading experience.	1-2
Science	Male	23	Moscow	Masters	Researcher	Invited Researcher at leading phage therapy labs.	1-2
Editing and General Language Tasks	Female	54	Novosibirsk	Bachelors	Proofreader	-	>2
Crowd	Female	30	Livny	Vocational Education	Editor, Nurse	-	1-2
Editing and General Language Tasks	Female	46	Oryol	Masters	Editor, Teacher	Co-author of a research book published by Keruss.	>2
Crowd	Female	34	Oryol	Masters	Data Annotator	-	>2
Editing and General Language Tasks	Female	27	Oryol	Bachelors	Editor	-	>2
AI as a Character and Fun Tasks	Female	30	Oryol	Masters	Editor	-	1-2
Editing and General Language Tasks	Female	25	Oryol	Bachelors	Editor	-	1-2
Editing and General Language Tasks	Female	37	Moscow	PhD	Project Manager	Candidate of Chemical Sciences.	<1
Journalism	Female	32	Moscow	Bachelors	Brand Manager	Recipient of P&G CEO Award (awarded to < 1% of global employees).	>2
Science	Male	30	Moscow	PhD	Researcher	Candidate of Historical Sciences.	>2
Science	Male	24	Moscow	Masters	Researcher	Pharmaceutical Biotechnology Specialist.	>2
AI as a Character and Fun Tasks	Female	26	Moscow	Masters	Editor	Translation of several high-circulation novels for publishing house.	>2
AI as a Character and Fun Tasks	Female	35	Moscow	Bachelors	Journalist	Currently writing a book.	>2
AI as a Character and Fun Tasks	Female	24	Moscow	Masters	Content Manager, Copywriter, Editor	Thesis on modern Russian literature. Translated several high-circulation Young Adult books from English into Russian. Collaborated with major Russian companies as a copywriter.	>2
AI as a Character and Fun Tasks	Male	27	Moscow	Masters	Editor	-	>2
Literature	Female	23	Moscow	Bachelors	Editor	-	1-2
AI as a Character and Fun Tasks	Female	31	Moscow	Masters	Editor	Translator of the published books.	>2
Literature	Male	34	Saint Petersburg	Masters	Editor, Writer, Literary Scholar	Published in Russia's leading literary journals. Finalist of 2 awards in literature. Edited 400+ short stories by contemporary Russian-language authors for a publishing project. Skilled in stylistic imitation, writing in various styles.	>2
Literature	Male	45	Moscow	Bachelors	Editor	Spent over 10 years as a managing editor in news services.	>2
AI as a Character and Fun Tasks	Female	37	Moscow	Bachelors	Editor	Semifinalist of a poetry festival. Co-author of a short story collection and poetry collections. Co-author of the short story collection and poetry.	1-2
Literature	Female	40	Moscow	Masters	Journalist	18 years of experience as a proofreader. Managing editor of several Telegram channels and a website.	>2
Translation Studies	Female	31	Moscow	Masters	Editor	Graduated with honors from Moscow State University (Linguistics and Translation). Internships at a research institute and a ministry of the Russian Federation. Participated in university's international delegations. Six years of experience at a translation company working on projects for major corporations (as a translator, editor, and proofreader). Experience in audiovisual translation (translator for shows and stand-up comedy). Guest lecturer at a university (taught "Introduction to Translation Studies"). Skills in UX writing, prompt engineering, and basic knowledge of Python.	1-2
Journalism	Female	24	Moscow	Bachelors	Philologist, Linguist	Introduced a novel term in Russian philology tradition. Mastered graduate-level courses related to the Theory of Rhetoric. Winner of a literary prize, longlisted for another.	>2
Editing and General Language Tasks	Female	43	Moscow	Masters	Editor, Proofreader	Over 20 years of experience as a proofreader. Worked at a research institute, a technical university publishing house, printing houses, and online publication. Edited and proofread textbooks and monographs on fisheries, economics, psychology, jurisprudence, and other sciences, fiction, cafe and restaurant menus, brochures, banners, and other printed materials.	1-2
Science	Female	26	Moscow	Masters	Researcher	PhD researcher specializing in agricultural forecasting.	>2
Editing and General Language Tasks	Female	35	Oryol	Masters	Copywriter, Editor	Mathematics and physics teacher. Completed professional retraining programs in entrepreneurship.	>2
Science	Female	25	Moscow Region	Bachelors	Teacher	Author of two academic publications.	>2
Law, Diplomacy, and Business	Female	24	Moscow	Masters	Lawyer	Member of a university's legal clinic. Proficient in Legal English. Completed professional retraining in marketing.	>2
Law, Diplomacy, and Business	Female	35	Moscow	Masters	Copywriter, Editor	Conducted research in Russian philology. Publishing editor.	1-2
Science	Female	33	Moscow	Masters	Researcher, Editor	Author of scholarly articles on Gothic architecture.	1-2
Science	Female	26	Yekaterinburg	Bachelors	Editor, Teacher	-	1-2
Editing and General Language Tasks	Female	29	Oryol	Masters	Editor	-	1-2
Science	Female	29	Moscow	Masters	Teacher	Academic Director of an educational program.	1-2
Science	Female	25	Moscow Region	Bachelors	Editor, Teacher	-	>2
Law, Diplomacy, and Business	Female	24	Moscow	Masters	Editor	Graduated with a Master's degree in "State and Municipal Administration".	1-2
AI as a Character and Fun Tasks	Female	23	Moscow	Bachelors	IT Specialist	Author of academic publications in philology.	1-2
Law, Diplomacy, and Business	Female	30	Moscow	Masters	IT Specialist	Master's degree in Law. Re-qualified as a Java developer. Author of publications on accounting and taxation. Co-author and expert of an e-course on LLMs. Co-author of a Russian school textbook.	2+
Programming code	Male	35	Moscow Region	Masters	IT Specialist	Participated in the development of equipment for the rehabilitation of patients with partial paralysis.	1-2
QA	Male	35	Yoshkar-Ola	Vocational Education	Data Annotator	-	>2
QA	Female	29	Pskov	Bachelors	Data Annotator	Former correspondent at a daily newspaper.	<1
STEM	Female	33	Saint Petersburg	Masters	Underwriter	Winner of a banking professional prize.	<1
STEM	Male	42	Rostov-on-Don	Bachelors	Project Manager	-	>2
QA	Female	25	Pskov	Masters	Data Annotator	Creates music albums using neural networks.	>2
QA	Male	26	Volzhsky	Masters	Data Annotator	Master's Degree in Philosophy.	1-2
QA	Male	25	Volgograd	Masters	Journalist	Conducted research, which included tracing the history of music journalism in Russia.	<1
QA	Female	32	Yoshkar-Ola	Masters	Data Annotator	History teacher.	<1
QA	Male	28	Saratov	Vocational Education	Head Chef	-	1-2
Programming code	Male	25	Volgograd	Masters	Teacher	Published six research papers. PhD student.	1-2
Programming code	Male	28	Yoshkar-Ola	Vocational Education	IT Specialist	-	1-2
QA	Female	34	Volgograd	Masters	Editor	Managing editor and proofreader for a party newspaper.	1-2
QA	Male	30	Volgograd	Masters	Organizational Psychologist	-	<1
Programming code	Male	26	Yoshkar-Ola	Vocational Education	IT Specialist	-	<1
STEM	Female	31	Tambov	Masters	Data Annotator	Master's Degree in Social Science.	<1
QA	Female	22	Volzhsky	Vocational Education	Preschool Teacher	-	<1
QA	Female	38	Volzhsky	Masters	Lawyer	Proficient in Chinese and English.	<1
STEM	Male	24	Vancouver	Masters	Researcher	-	>2
QA	Female	23	Engels	Masters	Translator	-	>2
Law, Diplomacy, and Business	Female	34	Moscow	Masters	Lawyer	-	<1
Law, Diplomacy, and Business	Female	30	Moscow	Masters	Analyst	Candidate of Legal Sciences.	1-2
QA	Female	26	Yoshkar-Ola	Bachelors	Data Annotator	Participant in data annotation for state TV projects.	1-2
STEM	Female	25	Moscow	Masters	Researcher	Member of the jury of a school olympiad in Economics.	1-2
STEM	Male	34	Moscow	Postgraduate Education	Researcher	Candidate of Chemical Sciences.	>2
QA	Female	25	Moscow	Bachelors	IT Specialist	Lead Data Warehouse (DWH) Engineer.	1-2
Journalism	Female	29	Moscow	Bachelors	Journalist	Authored commercial and educational articles for media outlets.	>2
Science	Female	23	Moscow	Bachelors	Student	-	>2
AI as a Character and Fun Tasks	Male	26	Smolensk	Vocational Education	Data Annotator	-	<1
AI as a Character and Fun Tasks	Male	33	Saint Petersburg	Bachelors	Teacher	-	>2
Science	Female	36	Moscow	Masters	Museum curator	Lead Organizer of an international conference on the history and theory of photography.	1-2
Journalism	Male	32	Moscow	Masters	Editor	Author of several articles on aesthetics (Philosophy) and literary criticism. Runs a Telegram channel about literature and art.	>2
Science	Female	28	Moscow	Masters	Linguist	Master's Degree in Linguistics.	>2
Science	Female	32	Moscow	Masters	Researcher	-	<1
AI as a Character and Fun Tasks	Male	41	Moscow	Masters	Editor	Creator of a weight-loss program. Python Programmer.	>2
AI as a Character and Fun Tasks	Male	36	Oryol	Masters	Editor	-	1-2
Science	Female	29	Moscow	Masters	IT Specialist	Master's Degree in Formal Morphology.	>2
Journalism	Male	33	Moscow	Masters	Journalist, Content Manager, Copywriter	Several years of experience as a news editor.	>2
Science	Female	26	Sirius	Masters	Researcher	-	>2
Journalism	Male	40	Moscow	Masters	Journalist	Awarded the Order of Friendship and other state and departmental honors.	>2
Crowd	Male	24	Oryol	Vocational Education	Food technologist	-	>2
AI as a Character and Fun Tasks	Male	27	Moscow	Bachelors	Editor	Contributed to the development of an instructive benchmark for LLMs. Managed a team of editors for media channels with 1M users.	>2
Crowd	Female	33	Oryol	Masters	Data Annotator	-	>2
Journalism	Male	46	Moscow	Bachelors	Journalist, Content Manager, Editor	Chief Editor of business TV programs. Editor-in-Chief of a production center.	1-2
Journalism	Female	32	Moscow	Masters	Editor	Worked at major Russian news agencies. Author of academic publications.	>2
Crowd	Male	26	Oryol	Masters	IT Specialist	Master's Degree in Computer Sciences.	<1
Translation Studies	Female	23	Moscow	Uncompleted Higher Education	Editor, Teacher, Linguist	Tutor for a school olympiad. Specialization in poetic metrics, dialectology, and Old Russian language. Participant of a linguistic expedition to study a minority language of Russia.	>2
Literature	Male	37	Moscow	Masters	Editor	-	<1
Science	Female	23	Kazan	Masters	Researcher	Multiple publications in peer-reviewed academic journals.	>2
Crowd	Female	26	Oryol	Masters	Data Annotator	Author of published papers in Ecology.	>2
Science	Male	25	Nizhny Novgorod	Masters	Researcher	-	1-2
Literature	Male	29	Moscow	Bachelors	Writer	Author of a published book longlisted for national awards.	>2
Science	Female	32	Moscow	Postgraduate Education	Teacher	Candidate of Art Studies	>2
Science	Female	49	Moscow Region	Masters	Journalist, Editor	Author of a history website. Worked as a news editor at major Russian news agencies.	1-2

Table 23: Profiles of experts that performed criteria annotation. **Region** is for the current region, **Profession** is a current occupation, **Achievements** are completed by experts, we asked them to write some of their most important accomplishments. Most of the entries in **Achievements** are processed to maintain anonymity. **LLM Experiences** represents the experience of annotators with LLMs in years.

Panel	Gender	Age	Region	Education	Profession	Achievement	LLM Experience
Programming code	Female	24	Moscow	Specialist or Master's Degree	IT Specialist	-	more than two years
Editing and General Language Tasks	Female	34	Kemerovo	Specialist or Master's Degree	Editor	10 years of proofreading experience. Working at a research institute, a publishing house, an editorial office, and an international company.	one to two years
Literature	Female	33	Moscow	Postgraduate Education	Teacher	Candidate of Philological Sciences, thesis on American literature. Author of 2 published novels. Literary adaptation of a popular Russian TV series. Winner of 2 art prizes in Moscow.	one to two years
Science	Male	24	Moscow	Specialist or Master's Degree	Researcher	Pharmaceutical Biotechnology Specialist.	more than two years
AI as a Character and Fun Tasks	Male	36	Moscow	Bachelor's Degree	Head of Content and Technology Projects	Development and implementation of major digital media projects. Integration of AI technologies into workflows.	more than two years
AI as a Character and Fun Tasks	Female	26	Moscow	Specialist or Master's Degree	IT Specialist, Editor	Translation of several high-circulation novels for a major Russian publishing house.	more than two years
AI as a Character and Fun Tasks	Female	35	Moscow	Bachelor's Degree	Journalist, Copywriter, Editor, Teacher	Currently writing a book.	more than two years
AI as a Character and Fun Tasks	Female	24	Moscow	Specialist or Master's Degree	Content Manager, Copywriter, Editor	Thesis on modern Russian literature. Translated several high-circulation Young Adult books from English into Russian. Collaborated with major Russian companies as a copywriter.	more than two years
AI as a Character and Fun Tasks	Male	27	Moscow	Specialist or Master's Degree	Copywriter, Editor	Released a popular video about a computer game.	more than two years
AI as a Character and Fun Tasks	Female	23	Moscow	Bachelor's Degree	Editor	-	one to two years
AI as a Character and Fun Tasks	Female	31	Moscow Region	Specialist or Master's Degree	Editor	Author of a published translation of a book from English into Russian.	more than two years
Literature	Male	34	Saint Petersburg	Specialist or Master's Degree	Editor, Writer, Literary Scholar	Published in Russia's leading literary journals. Finalist of 2 awards in literature. Edited 400+ short stories by contemporary Russian-language authors for a publishing project. Skilled in stylistic imitation, writing in various styles.	more than two years
AI as a Character and Fun Tasks	Male	45	Moscow	Bachelor's Degree	Editor	Authored a sports column, conducted interviews with athletes. Over 10 years of experience as a managing news editor.	more than two years
AI as a Character and Fun Tasks	Female	37	Moscow	Bachelor's Degree	Editor	Semifinalist of a poetry festival. Co-author of a short story collection and poetry collections.	one to two years
Literature	Female	40	Moscow	Specialist or Master's Degree	Journalist, Content Manager, Copywriter, Editor, Proofreader	18 years of experience as a proofreader. Managing editor of several Telegram channels and a website.	one to two years
Translation Studies	Female	31	Moscow	Specialist or Master's Degree	Editor	Graduated with honors from Moscow State University (Linguistics and Translation). Internships at a research institute and a ministry of the Russian Federation. Participated in university's international delegations. Six years of experience at a translation company working on projects for major corporations (as a translator, editor, and proofreader). Experience in audiovisual translation (translator for shows and stand-up comedy). Guest lecturer at a university (taught "Introduction to Translation Studies"). Skills in UX writing, prompt engineering, and basic knowledge of Python.	one to two years
Journalism	Female	24	Moscow	Bachelor's Degree	Philologist, Linguist	Introduced a novel term in Russian philology tradition. Mastered graduate-level courses related to the Theory of Rhetoric. Winner of a literary prize, longlisted for another.	more than two years
Editing and General Language Tasks	Female	43	Moscow	Specialist or Master's Degree	Editor, Proofreader	Over 20 years of experience as a proofreader. Worked at a research institute, a technical university publishing house, printing houses, and online publication. Edited and proofread textbooks and monographs on fisheries, economics, psychology, jurisprudence, and other sciences, fiction, cafe and restaurant menus, brochures, banners, and other printed materials.	one to two years
Law, Diplomacy, and Business	Female	24	Moscow	Specialist or Master's Degree	Lawyer	Member of a university's legal clinic. Proficient in Legal English. Completed professional retraining in marketing.	more than two years
Law, Diplomacy, and Business	Female	35	Moscow	Specialist or Master's Degree	Copywriter, Editor	Conducted research in Russian philology. Publishing editor.	one to two years
Science	Female	26	Yekaterinburg	Bachelor's Degree	Editor, Teacher	-	one to two years
Law, Diplomacy, and Business	Female	24	Moscow	Specialist or Master's Degree	Editor	-	one to two years
Law, Diplomacy, and Business	Female	30	Moscow	Specialist or Master's Degree	IT Specialist	Master's degree in Law. Re-qualified as a Java developer. Author of publications on accounting and taxation. Co-author and expert of an e-course on LLMs. Co-author of a Russian school textbook.	more than two years
Programming code	Male	35	Moscow Region	Specialist or Master's Degree	IT Specialist	Participated in the development of equipment for the rehabilitation of patients with partial paralysis.	one to two years
Programming code	Male	28	Moscow	Specialist or Master's Degree	IT Specialist	Significantly improved communication and collaboration within model risk management in a tech company.	more than two years
STEM	Male	24	Vancouver, Canada	Specialist or Master's Degree	Researcher	-	more than two years
STEM	Female	25	Moscow	Specialist or Master's Degree	Researcher, Teacher	Recognized as best lecturer of the year at a university. Member of the jury of a school olympiad in Economics.	one to two years
STEM	Male	34	Moscow	Postgraduate Education	Researcher, Teacher	Candidate of Chemical Sciences. Recipient of an academic prize for young scientists. Associate Professor at Lomonosov Moscow State University. Coach for an international school olympiad.	more than two years
QA	Female	25	Moscow	Bachelor's Degree	IT Specialist	Lead Data Warehouse (DWH) Engineer. Contributed to fintech projects and a major e-commerce retail chain.	one to two years
Literature	Male	29	Moscow	Bachelor's Degree	Writer	Author of a published book longlisted for national awards.	more than two years
Science	Female	49	Moscow Region	Specialist or Master's Degree	Journalist, Editor	Author of a history website. Worked as a news editor at major Russian news agencies.	one to two years

Table 24: Criteria creators

Panel	Gender	Age	Region	Education	Profession	Achievement	LLM Experience
Translation Studies	Female	31	Moscow	Specialist or Master's Degree	Editor	Graduated with honors from Moscow State University (Linguistics and Translation). Internships at a research institute and a ministry of the Russian Federation. Participated in university's international delegations. Six years of experience at a translation company working on projects for major corporations (as a translator, editor, and proofreader). Experience in audiovisual translation (translator for shows and stand-up comedy). Guest lecturer at a university (taught "Introduction to Translation Studies"). Skills in UX writing, prompt engineering, and basic knowledge of Python.	one to two years
Law, Diplomacy, and Business	Female	30	Moscow	Specialist or Master's Degree	IT Specialist	Master's degree in Law. Re-qualified as a Java developer. Author of publications on accounting and taxation. Co-author and expert of an e-course on LLMs. Co-author of a Russian school textbook.	more than two years

Table 25: Technical editors

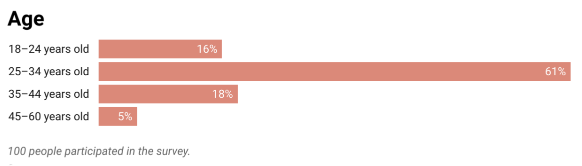


Figure 13: Survey participant age distribution. The substantial representation of the 25–34 age group highlights the active involvement of professionals who are likely combining fresh academic knowledge with practical experience. The diversity across age groups also shows a collaborative environment with varying levels of experience.

Panel	Gender	Age	Region	Education	Profession	Achievement	LLM Experience
Editing and General Language Tasks	Female	34	Kemerovo	Specialist or Master's Degree	Editor	10 years of proofreading experience. Working at a research institute, a publishing house, an editorial office, and an international company.	one to two years
Editing and General Language Tasks	Female	43	Moscow	Specialist or Master's Degree	Editor, Proofreader	Over 20 years of experience as a proofreader. Worked at a research institute, a technical university publishing house, printing houses, and online publication. Edited and proofread textbooks and monographs on fisheries, economics, psychology, jurisprudence, and other sciences, fiction, cafe and restaurant menus, brochures, banners, and other printed materials.	one to two years

Table 26: Metadata proofreaders

Panel	Gender	Age	Region	Education	Profession	Achievement	LLM Experience
Translation Studies	Female	31	Moscow	Specialist or Master's Degree	Editor	Graduated with honors from Moscow State University (Linguistics and Translation). Internships at a research institute and a ministry of the Russian Federation. Participated in university's international delegations. Six years of experience at a translation company working on projects for major corporations (as a translator, editor, and proofreader). Experience in audiovisual translation (translator for shows and stand-up comedy). Guest lecturer at a university (taught "Introduction to Translation Studies"). Skills in UX writing, prompt engineering, and basic knowledge of Python.	one to two years
Translation Studies	Male	27	Moscow	Bachelor's Degree	Journalist, Copywriter, Editor	Authored a series of articles about video games.	more than two years
Translation Studies	Female	23	Moscow	Uncompleted Higher Education	Editor, Teacher, Linguist	Tutor for a school olympiad. Specialization in poetic metrics, dialectology, and Old Russian language. Participant of a linguistic expedition to study a minority language of Russia.	more than two years

Table 27: Metadata translators

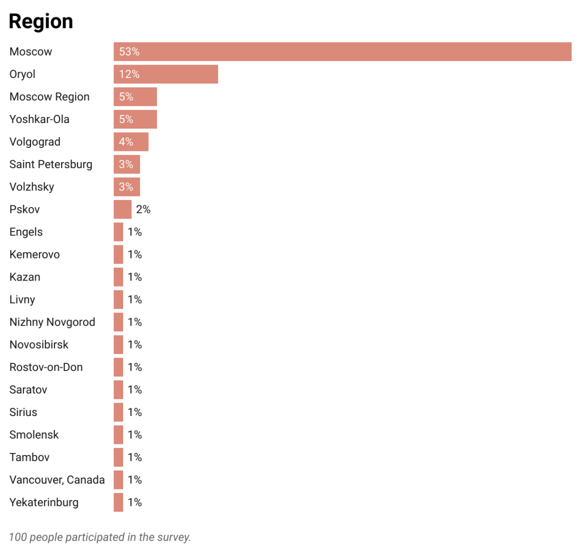


Figure 14: Survey participant region distribution. The regional distribution of the benchmark's creators reveals that a significant majority, 53 percent, reside in Moscow, underscoring the city's role as a central hub for scientific and technological development. The remaining 47 percent are dispersed across 20 different cities, indicating a broad geographical diversity within the team.

Figure 15: Survey participant region distribution on the map of Russia.

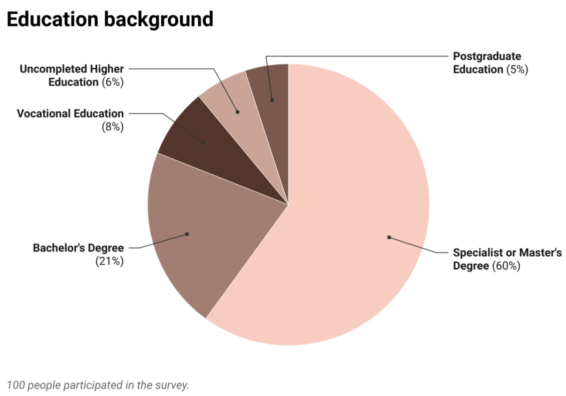


Figure 16: Survey participant educational background distribution

Panel	Gender	Age	Region	Education	Profession	Achievement	LLM Experience
Editing and General Language Tasks	Female	34	Kemerovo	Specialist or Master's Degree	Editor	10 years of proofreading experience. Working at a research institute, a publishing house, an editorial office, and an international company.	one to two years
Editing and General Language Tasks	Female	30	Livny	Vocational Education	Editor, Nurse	-	one to two years
Editing and General Language Tasks	Female	46	Oryol	Specialist or Master's Degree	Copywriter, Editor, Teacher	Co-founder and author of a website and a research book in culinary.	more than two years
Editing and General Language Tasks	Female	34	Oryol	Specialist or Master's Degree	Data Annotator	-	more than two years
Editing and General Language Tasks	Female	27	Oryol	Bachelor's Degree	Editor	-	more than two years
Journalism	Female	30	Oryol	Specialist or Master's Degree	Editor	-	one to two years
Literature	Female	33	Moscow	Postgraduate Education	Teacher	Candidate of Philological Sciences, thesis on American literature. Author of 2 published novels. Literary adaptation of a popular Russian TV series. Winner of 2 art prizes in Moscow.	one to two years
Editing and General Language Tasks	Female	25	Oryol	Uncompleted Higher Education	Editor	-	one to two years
AI as a Character and Fun Tasks	Male	36	Moscow	Bachelor's Degree	Head of Content and Technology Projects	Development and implementation of major digital media projects. Integration of AI technologies into workflows.	more than two years
AI as a Character and Fun Tasks	Female	26	Moscow	Specialist or Master's Degree	IT Specialist, Editor	Translation of several high-circulation novels for a major Russian publishing house.	more than two years
AI as a Character and Fun Tasks	Female	35	Moscow	Bachelor's Degree	Journalist, Copywriter, Editor, Teacher	Currently writing a book.	more than two years
AI as a Character and Fun Tasks	Female	24	Moscow	Specialist or Master's Degree	Content Manager, Copywriter, Editor	Thesis on modern Russian literature. Translated several high-circulation Young Adult books from English into Russian. Collaborated with major Russian companies as a copywriter.	more than two years
AI as a Character and Fun Tasks	Male	27	Moscow	Specialist or Master's Degree	Copywriter, Editor	Released a popular video about a computer game.	more than two years
AI as a Character and Fun Tasks	Female	23	Moscow	Bachelor's Degree	Editor	-	one to two years
AI as a Character and Fun Tasks	Female	31	Moscow Region	Specialist or Master's Degree	Editor	Author of a published translation of a book from English into Russian.	more than two years
Literature	Male	34	Saint Petersburg	Specialist or Master's Degree	Editor, Writer, Literary Scholar	Published in Russia's leading literary journals. Finalist of 2 awards in literature. Edited 400+ short stories by contemporary Russian-language authors for a publishing project. Skilled in stylistic imitation, writing in various styles.	more than two years
AI as a Character and Fun Tasks	Male	45	Moscow	Bachelor's Degree	Editor	Authored a sports column, conducted interviews with athletes. Over 10 years of experience as a managing news editor.	more than two years
AI as a Character and Fun Tasks	Female	37	Moscow	Bachelor's Degree	Editor	Semifinalist of a poetry festival. Co-author of a short story collection and poetry collections.	one to two years
AI as a Character and Fun Tasks	Female	40	Moscow	Specialist or Master's Degree	Journalist, Content Manager, Copywriter, Editor, Proofreader	18 years of experience as a proofreader. Managing editor of several Telegram channels and a website.	more than two years
Translation Studies	Female	31	Moscow	Specialist or Master's Degree	Editor	Graduated with honors from Moscow State University (Linguistics and Translation). Internships at a research institute and a ministry of the Russian Federation. Participated in university's international delegations. Six years of experience at a translation company working on projects for major corporations (as a translator, editor, and proofreader). Experience in audiovisual translation (translator for shows and stand-up comedy). Guest lecturer at a university (taught "Introduction to Translation Studies"). Skills in UX writing, prompt engineering, and basic knowledge of Python	one to two years
Journalism	Female	24	Moscow	Bachelor's Degree	Philologist, Linguist	Introduced a novel term in Russian philology tradition. Mastered graduate-level courses related to the Theory of Rhetoric. Winner of a literary prize, longlisted for another.	more than two years
Editing and General Language Tasks	Female	43	Moscow	Specialist or Master's Degree	Editor, Proofreader	Over 20 years of experience as a proofreader. Worked at a research institute, a technical university publishing house, printing houses, and online publication. Edited and proofread textbooks and monographs on fisheries, economics, psychology, jurisprudence, and other sciences, fiction, cafe and restaurant menus, brochures, banners, and other printed materials.	one to two years
Editing and General Language Tasks	Female	35	Oryol	Specialist or Master's Degree	Copywriter, Editor	Mathematics and physics teacher. Completed professional retraining programs in entrepreneurship.	more than two years
Law, Diplomacy, and Business	Female	24	Moscow	Specialist or Master's Degree	Lawyer	Member of a university's legal clinic. Proficient in Legal English. Completed professional retraining in marketing.	more than two years
Law, Diplomacy, and Business	Female	35	Moscow	Specialist or Master's Degree	Copywriter, Editor	Conducted research in Russian philology. Publishing editor.	one to two years
Science	Female	26	Yekaterinburg	Bachelor's Degree	Editor, Teacher	-	one to two years
Journalism	Female	29	Oryol	Specialist or Master's Degree	Editor	-	one to two years
Crowd	Female	35	Moscow	Uncompleted Higher Education	Project Manager (Manager)	-	more than two years
Science	Female	25	Moscow Region	Bachelor's Degree	Editor, Teacher	-	more than two years
Law, Diplomacy, and Business	Female	24	Moscow	Specialist or Master's Degree	Editor	Graduated with a Master's degree in "State and Municipal Administration".	one to two years
AI as a Character and Fun Tasks	Male	27	Moscow	Bachelor's Degree	Journalist, Copywriter, Editor	Authored a series of articles about games.	more than two years
Law, Diplomacy, and Business	Female	30	Moscow	Specialist or Master's Degree	IT Specialist	Master's degree in Law. Re-qualified as a Java developer. Author of publications on accounting and taxation. Co-author and expert of an e-course on LLMs. Co-author of a Russian school textbook.	more than two years
Programming code	Male	28	Moscow	Specialist or Master's Degree	IT Specialist	Significantly improved communication and collaboration within model risk management in a tech company.	more than two years
STEM	Male	24	Vancouver, Canada	Specialist or Master's Degree	Researcher	-	more than two years
QA	Female	26	Yoshkar-Ola	Bachelor's Degree	Data Annotator	Participant in data annotation for state TV projects.	one to two years
STEM	Female	25	Moscow	Specialist or Master's Degree	Researcher, Teacher	Recognized as best lecturer of the year at a university. Member of the jury of a school olympiad in Economics.	one to two years
STEM	Male	34	Moscow	Postgraduate Education	Researcher, Teacher	Candidate of Chemical Sciences. Recipient of an academic prize for young scientists. Associate Professor at Lomonosov Moscow State University. Coach for an international school olympiad.	more than two years
QA	Female	25	Moscow	Bachelor's Degree	IT Specialist	Lead Data Warehouse (DWH) Engineer. Contributed to fintech projects and a major e-commerce retail chain.	one to two years
Science	Female	36	Moscow	Specialist or Master's Degree	Museum curator	Curated international exhibitions at state museums. Lead Organizer of an international conference on photography.	one to two years
Journalism	Male	33	Moscow	Specialist or Master's Degree	Journalist, Content Manager, Copywriter	Several years of experience as a news editor.	more than two years
Science	Female	26	Sirius	Specialist or Master's Degree	Researcher	-	more than two years
Journalism	Male	46	Moscow	Bachelor's Degree	Journalist, Content Manager, Editor	Chief Editor of business TV programs. Editor-in-Chief of a production center.	more than two years
Journalism	Female	32	Moscow	Specialist or Master's Degree	Editor	Worked at major Russian news agencies. Author of academic publications.	more than two years
Literature	Male	29	Moscow	Bachelor's Degree	Writer	Author of a published book longlisted for national awards.	more than two years
Science	Female	49	Moscow Region	Specialist or Master's Degree	Journalist, Editor	Author of a history website. Worked as a news editor at major Russian news agencies.	one to two years

Table 28: Profiles of experts that created instructions. **Region** is for the current region, **Profession** is a current occupation, **Achievements** are completed by experts, we asked them to write some of their most important accomplishments. Most of the entries in **Achievements** are processed to maintain anonymity. **LLM Experiences** represents the experience of annotators with LLMs in years.

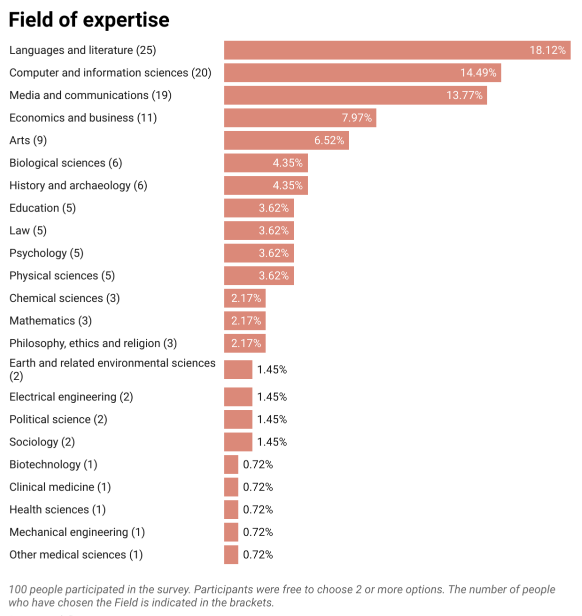


Figure 17: Survey participant field of expertise distribution. The diversity of expertise among the benchmark's creators is extensive; 23 different fields. This multidisciplinary team includes professionals from the humanities, such as philologists and journalists, as well as experts from the natural sciences like physicists, and legal specialists. The collaboration of such a wide array of experts ensures that the benchmark is deeply and thoroughly developed.

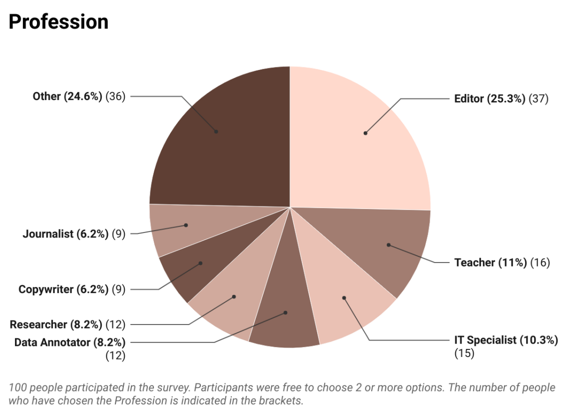


Figure 18: Survey participant profession distribution

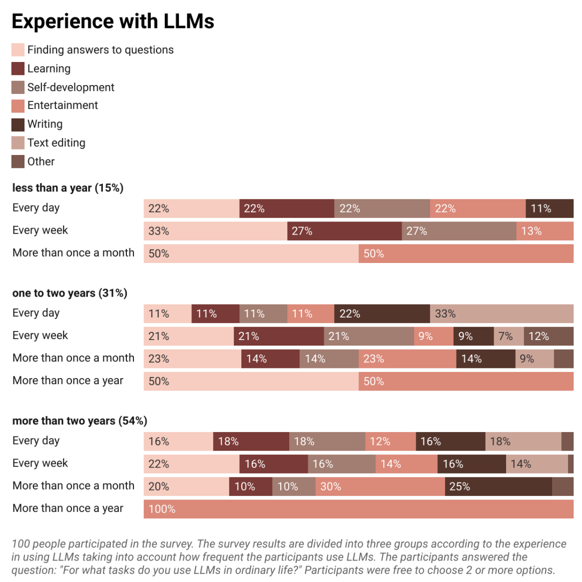


Figure 19: Survey participant distribution by experience with LLMs. The data reveals a community predominantly comprised of experienced LLM users, with the majority having integrated these technologies into their workflows for significant periods. This distribution suggests that the benchmark results largely reflect insights from practitioners with substantial practical knowledge rather than newcomers, lending credibility to the evaluations and observations presented in this study.

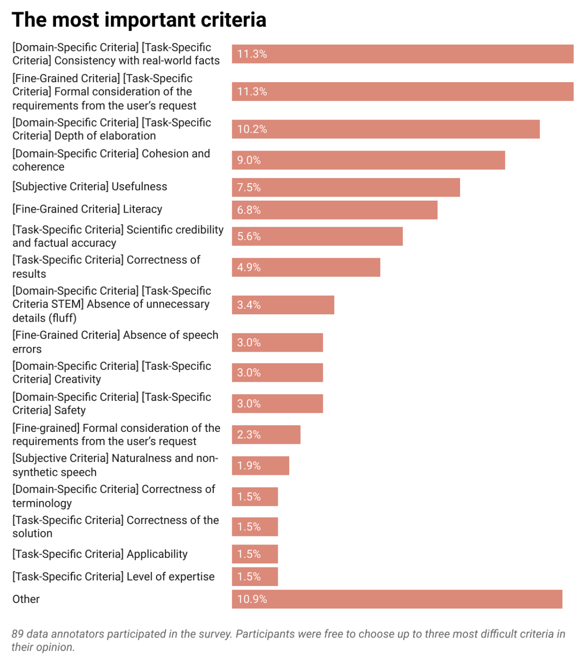


Figure 20: The most important criteria for LLM evaluating according to the survey participants. These core criteria form the critical infrastructure. These findings suggest that advanced and more specific features must be developed atop a solid foundation of reliability and functional effectiveness to deliver meaningful value in specialized applications.

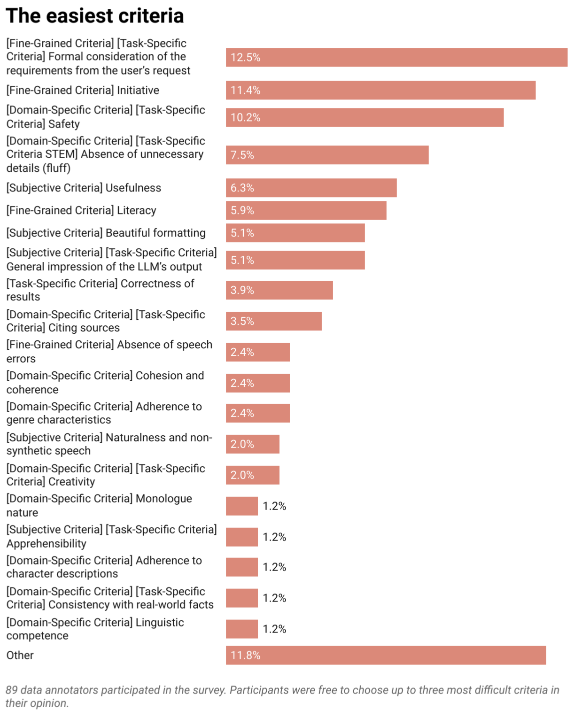


Figure 21: The easiest criteria for LLM evaluating according to the survey participants. Evaluators found particular criteria more straightforward to assess. This pattern suggests these elements exhibit more observable manifestations in model outputs, enabling more confident human judgment. Future evaluation frameworks might strategically incorporate these assessments as reliable anchors within broader, more nuanced evaluation schemes.

Figure 22: The most difficult criteria for LLM evaluating according to the survey participants. These findings highlight a critical methodological consideration for benchmark development: the most essential evaluation dimensions often present the greatest measurement challenges, suggesting the need for specialized evaluation protocols, multi-annotator consensus approaches, or supplementary objective metrics to achieve reliable assessments of these crucial but elusive qualities.

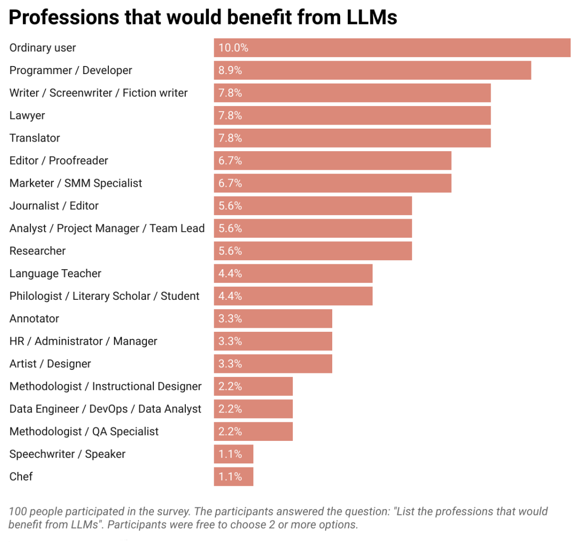


Figure 23: Professional spheres where LLMs may help. This distribution suggests generative AI's greatest value may lie in augmenting knowledge work requiring both structured information processing and creative adaptation. The prominence of ordinary users atop this hierarchy underscores these technologies' democratizing potential. These findings point to areas where focused development efforts and specialized evaluation benchmarks may yield particularly high-value applications.

M Training details

M.1 Prompts

Below is the Russian version of the prompt for model training:

```
### Задание для оценки:
{instruction}

### Эталонный ответ:
{reference_answer}

### Ответ для оценки:
{answer}

### Критерий оценки:
{ criteria .name}

### Шкала оценивания по критерию:
{ criteria . rubrics}
```

The prompt for generating synthetic data and running test data:

```
### Task Description:
You are provided with the following: an instruction
(which may include an input), a response to evaluate
, a reference answer and an evaluation criterion
with a detailed scale.
1. Write detailed feedback assessing the quality of
the response strictly according to the provided
evaluation scale. Do not give a general evaluation,
base your assessment entirely on the scale.
2. Assign a score to the response by referring to
the scale. The score must correspond to a single
scale point and its description.
3. Format your output as follows: "[FEEDBACK] (
Write detailed feedback regarding the evaluated
response and the assigned score, reason step by step
and explain each point.) [RESULT] (An integer score
within the boundaries of the criterion scale.)"
4. Do not include any additional openings, closings,
 or explanations.
5. Write feedback in Russian.
6. Write [END] after you are done.

### The instruction to evaluate:
{instruction}

### Reference answer:
{reference_answer}

### Response to evaluate:
{answer}

### Score name
{ criteria .name}

### Score Rubrics:
{ criteria . rubrics}
```

The response format of the LLM-as-a-judge:

```
[FEEDBACK] {feedback text} [RESULT] {score} [
END]
```

The prompt for generating instructions:

```
### **Instruction:**
Your task is to generate a new novel problem based
on the given details about its type, description,
requirements, and complexity.

#### **Problem Details:**
- **General Problem Type:** {problem_type}
- **Specific Problem Type:** {problem_subtype}
- **More Specific Subtype:** {problem_subtype2}
- **Domain:** {domain}
- **Description:** {problem_description}
- **Requirements:** {problem_requirements}
- **Complexity:** {problem_complexity}

---

### **Problem Generation Rules:**

1. **Relevance:**
   - The problem must be directly related to the
given description and domain.
   - It should align with the specified problem
type and complexity level.

2. **Complexity & Challenge:**
   - The problem should not be too generic or easy
to solve.
   - The complexity should match ‘{
problem_complexity}‘.

3. **Implicit Requirements:**
   - The problem should naturally contain the
requirements but should **not** list them explicitly.

4. **Text-Based Problems:**
   - If the problem involves working with text, the
text content should be provided **after** the
problem statement.

5. **Perspective & Style:**
   - Assume a situation where a user is asking a
question.
   - The user should ask in **first-person
perspective** but **should not** use phrases like
    I , A user , or You in the first
sentence.
   - Do **not** assign a role to any entity in the
problem.
   - Avoid mentioning AI models in any way.

---

### **Format:**

- **Prefix the problem with:** [PROBLEM]
- **Write the problem in Russian** (matching the
given description).
- **End the problem with:** [END]
- **No greetings or extra messages.**
```

M.2 Open-source LLMs for Answers Generation

- mistralai/Mistral-Small-3.1-24B-Instruct-2503
- google/gemma-3-27b-it
- google/gemma-3-12b-it

- google/gemma-3-4b-it
- google/gemma-3-1b-it
- Qwen/QwQ-32B
- meta-llama/Llama-4-Scout-17B-16E-Instruct
- microsoft/Phi-4-multimodal-instruct
- meta-llama/Meta-Llama-3.1-8B-Instruct
- mistralai/Mistral-7B-Instruct-v0.3
- meta-llama/Llama-2-13b-chat-hf
- meta-llama/Llama-3.3-70B-Instruct
- Qwen/Qwen1.5-7B-Chat
- Qwen/Qwen1.5-32B-Chat
- Qwen/Qwen2.5-32B-Instruct
- Qwen/Qwen2.5-7B-Instruct
- msu-rcc-lair/RuadaptQwen2.5-32B-instruct
- microsoft/Phi-3.5-MoE-instruct
- IlyaGusev/saiga_nemo_12b
- mistralai/Mistral-Small-Instruct-2409
- mistralai/Mistral-Nemo-Instruct-2407
- Vikhrmodels/Vikhr-Llama3.1-8B-Instruct-R-21-09-24
- Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24
- t-tech/T-pro-it-1.0
- t-tech/T-lite-it-1.0
- google/gemma-2-27b-it
- google/gemma-2-9b-it
- ai-sage/GigaChat-20B-A3B-instruct
- google/gemma-2-2b-it
- meta-llama/Llama-3.2-1B-Instruct
- meta-llama/Llama-3.2-3B-Instruct
- meta-llama/Llama-2-7b-chat-hf
- Qwen/Qwen2.5-VL-72B-Instruct

N Human baseline

The Human Baseline was estimated on a sample of 140 instruction–answer pairs, yielding 7,537 distinct criterion-level annotations (LLM-as-Judge was not evaluated on Human Baseline). The answers to the instructions were written by panel experts. The results are provided in Table 1.