RuBIN: A Russian Benchmark for Evaluating LLMs with Cultural Insights

Polina Lazukova[1], **Irina Piontkovskaya**[2]
[1]LiSGames*, [2]Lomonosov Research Institute
polina.lazukova2411@gmail.com, irapiont@yandex.ru
*work done when author was with LRI

Abstract

Understanding culture-specific knowledge is essential for developing language models that perform reliably across diverse social and linguistic settings. This work explores both methodological and practical aspects of evaluating culture-specific knowledge in large language models. Special attention is given to the multiple-choice question answering format as a tool for identifying and measuring such knowledge. An analysis of existing benchmarks reveals several limitations, including insufficient cultural sensitivity and the presence of uninformative distractor options. In response, the RuBIN benchmark is introduced – a dataset consisting of questions based on phrases that are widely known in Russian culture. The paper describes the process of selecting and filtering culturally relevant topics, generating plausible incorrect answers using LLMs, and annotating and testing the benchmark for cross-linguistic robustness. RuBIN helps identify current LLMs' weaknesses in transferring cultural knowledge and can serve as a tool for further adapting these models to diverse linguistic and cultural contexts.

Keywords: Evaluation Methodologies, Question Answering, Cultural Specifics, Benchmarking, LLM

1. Introduction

Evaluating LLMs on its cultural competence is considered to be a highly nuanced yet impactful task. As the number of languages spoken by LLMs increases, the need for its cultural appropriateness becomes more important.

In an attempt to capture the abstract nature of culture, researchers build various benchmarks, including such topics as customs and traditions (Myung et al., 2024), ethics (Rao et al., 2025), literature and history (Etxaniz et al., 2025), cuisine (Lavrouk et al., 2025), and many more. Nevertheless, it is still questionable whether the topic-based approach to benchmark development allows for a truly diverse evaluation of LLMs. An opposite approach, which is introduced by a number of researches (Chiu et al., 2025; Myung et al., 2024), relies on the experiences of native annotators who are instructed to report on any culture-related aspects. Although this method of data gathering overcomes the limitations of a first mentioned approach with strictly set topics, it introduces another issue – the lack of clear ground truth.

In order to address the shortcomings of existing topic-based and open-ended strategies for benchmark creation, this paper introduces an alternative framework for cultural benchmark development. Overall, our contributions are as follows:

- We construct RuBIN (**RU**ssian **B**enchmark of Cultural **IN**sights), a dataset to evaluate LLM acquisition of culturally specific knowledge in Russian. We propose a new framework that makes it possible to include a diverse set of topics while keeping ground-truth answers unambiguous and verifiable. Figure 1 presents a step-by-step construction process of RuBIN.

- We assess the benchmark using responses from native human annotators to ensure the relevance and resolvability of the developed tasks. After reviewing the range of existing benchmarking approaches, we conclude that human evaluation is essential to avoid including questions that fail to reflect cultural knowledge. To this end, we introduce RuBIN-Easy. In this version, we exclude questions that proved too difficult for Russian native speakers.

- To gain insight into the performance of models across cultural contexts, we performed an evaluation of the flagship LLMs.

2. Related Work

Benchmarking the multilingual abilities of LLMs has seen growing interest from the research community (Wu et al., 2025) with culture being one of the evaluated aspects. Despite that, the standards for a rigorous cultural dataset are not unified and the process of benchmark construction remains rather intuitive.

In this section we study the limitations and strengths of existing cultural benchmarks. Then we discuss the specifics of multiple-choice question answering tasks, taking into account the previous findings in order to develop a robust and reliable cultural benchmark.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4126–4140
11–16 May 2026.

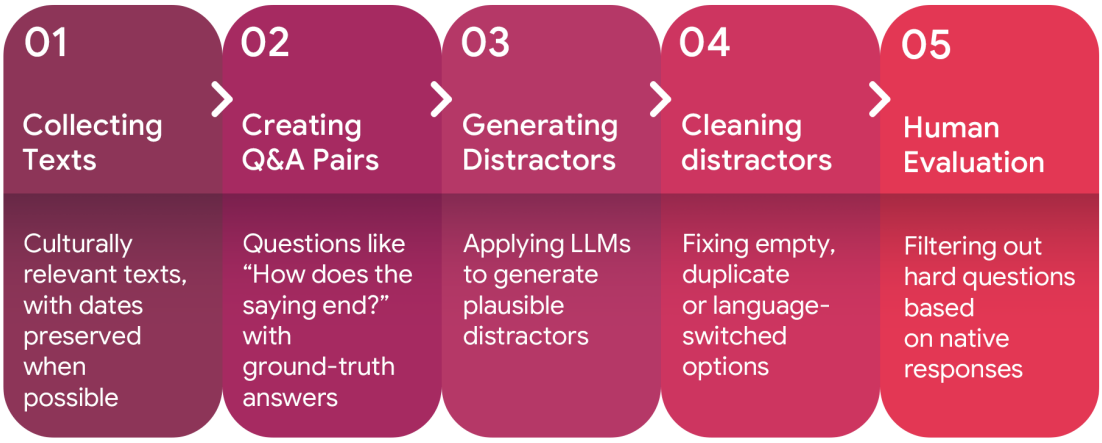

Figure 1: The RuBIN Framework.

2.1. Specifics of Cultural Benchmarking

Today, a variety of benchmarks exist, each offering different perspectives on culture evaluation. These benchmarks differ in sizes, domains, included languages, data collection strategies, and task types. We focus on benchmark size, data collection methods, and task types, as these dimensions most directly influence the design, scope, and interpretability of cultural evaluation tasks.

There is evidence that multilingual benchmarks are becoming significantly larger in size (Wu et al., 2025). As authors suggest, this is due to the emergence of foundation models that make large-scale evaluation possible. On the other hand, when dataset size is prioritized over other aspects, the quality of individual tasks often suffers due to a lack of cultural relevance. Extra-large benchmark sizes are achieved through automated data collection, reducing the feasibility of manual filtering. That results in the inclusion of flawed or misleading questions that significantly compromise the reliability of evaluation outcomes (Gema et al., 2025). As a result, researchers must strike a balance between dataset size and task quality; otherwise, the assessment credibility may be compromised.

As we discussed benchmark sizes, data collection strategies also merit attention. Upon reviewing several culture-related benchmarks, we identified two primary methods researchers use to obtain texts for task construction: automated collection of data in QA-format (Arora et al., 2025; Romanou et al., 2025; Rao et al., 2025; Hendrycks et al., 2021) and native speakers survey (Chiu et al., 2025; Myung et al., 2024). A large amount of data in the form of Question-Answer pairs is available online, including QA forums, quizzes, and exams. However, the majority of these tasks require reasoning skills for successful completion and do not evaluate cultural knowledge acquisition exclusively.

Another group of researchers constructs culture-related benchmarks by manually designing questions based on the lived experience of native speakers. This approach, exemplified by CulturalBench (Chiu et al., 2025) and BLEND (Myung et al., 2024), relies on native annotators to generate culturally relevant scenarios and reference answers. While this method improves cultural relevance by grounding tasks in authentic experience, it also introduces challenges related to subjectivity and representativeness. Individual perspectives may not adequately capture collective cultural knowledge, which limits the generalizability of the data.

Since the concept of "culture" encompasses a wide range of dimensions, researchers have focused on different domains when designing benchmarks to capture cultural significance. Existing benchmarks often target specific cultural aspects such as food (Lavrouk et al., 2025), holiday traditions (Myung et al., 2024), history and literature (Etxaniz et al., 2025), morality and ethics (Rao et al., 2025), or everyday habits (Chiu et al., 2025; Myung et al., 2024). This diversity reflects the multifaceted nature of culture and the varying approaches to evaluating culturally informed language understanding. However, this focus inevitably limits the range of topics covered and complicates the search for suitable resources.

Task types for LLM benchmarking vary widely, with examples such as multiple-choice question answering, text generation, reading comprehension, summarization, etc. While long-form tasks, such as those used in CALMQA (Arora et al., 2025), enable deeper insights into culturally grounded reasoning, they pose challenges for evaluation due to the lack of clear, comparable reference answers. In contrast, multiple-choice question answering offers a balance between scalability, interpretability, and evaluation reliability. By clearly defining answer op-

tions and correct responses, MCQA allows for large-scale, cross-model comparisons while still capturing essential aspects of cultural knowledge. For these reasons, we adopt the MCQA format in RuBIN. Importantly, recognition-based tasks are particularly well-suited for evaluating culturally shared knowledge, which often manifests as familiarity with conventionalized expressions rather than the ability to reconstruct them generatively. By constraining the space of possible answers, the MCQA format minimizes variability in surface realization and reduces evaluation noise. Moreover, the inclusion of distractors provides a structured foundation for future extensions of the benchmark, potentially enabling controlled manipulation of task difficulty and more fine-grained experimental analyses of cultural discrimination.

In the following subsection, we describe the specific design principles underlying our MCQA formulation, including the construction of answer options and the criteria used for selecting and validating distractors.

2.2. MCQA as a Benchmark Task

Any benchmark designed to evaluate a specific capability of a large language model is essentially a structured set of questions or tasks. The method of question formulation depends on the benchmark's goals and the nature of the underlying data. Recently, there has been growing interest in benchmarks using the multiple-choice question answering (MCQA) format. In MCQA, a model must choose the correct answer from several given options, based on a question and, optionally, additional context. This format is widely adopted due to its simplicity, ease of construction, and suitability for automatic evaluation, as model outputs can be directly compared to predefined ground-truth labels.

However, despite its practical advantages, MCQA has methodological limitations that can affect the reliability of evaluation. One key issue is the extraction of model predictions from generated outputs. As shown in recent work, different extraction methods – such as log-probability selection, regex-based parsing, or using pretrained classifiers like xFinder (Yu et al., 2025) – can lead to significant variation in performance metrics, particularly on non-technical questions where complex reasoning is not required (Molfese et al., 2025). In such cases, few-shot prompting combined with regex-based extraction is considered a reasonable compromise.

Another crucial factor is the quality of distractors – the incorrect but plausible answer choices. Studies have shown that poorly designed distractors can allow models to perform well by exploiting superficial patterns, without genuine understanding (Mozafari et al., 2025; Wang et al., 2025). For instance, models may succeed by eliminating obviously incorrect options rather than identifying the truly correct one. This reveals a risk of overestimating model capabilities when distractors are not carefully constructed. Therefore, when using MCQA for cultural evaluation, it is essential not only to validate correct answers, but also to ensure that distractors are culturally meaningful and non-trivial. In our benchmark, we address these known issues to ensure that RuBIN reliably measures cultural knowledge rather than test-taking shortcuts.

3. Benchmark Composition

3.1. Data Collection and Annotation

The development of the RuBIN benchmark began with collecting and preprocessing culturally specific texts relevant to the Russian-speaking context. Unlike benchmarks that create questions from scratch, RuBIN focuses on real texts deeply embedded in collective cultural memory. This includes widely recognized materials such as folklore, song lyrics, film quotes, memes and catchphrases – texts familiar to native speakers without additional context.

The data collection process for the RuBIN benchmark is grounded in a specific theoretical assumption about the nature of culture and its linguistic manifestation. We proceed from the premise that cultural knowledge is not limited to formal heritage artifacts, but is embedded in language through a multitude of conventionalized expressions that emerge and stabilize within a given community. Proverbs, idiomatic expressions, memes, song lyrics, film quotes, and other fixed or semi-fixed phrases serve as particularly salient carriers of such knowledge, as they reflect historically accumulated meanings that tend to circulate predominantly within one cultural space.

To identify these expressions, we adopted an iterative collection strategy. As new hypotheses emerged regarding domains in which culturally salient phrases might occur, we revisited and expanded the corpus. While some items were added manually when they fell outside previously defined categories, the majority of materials were gathered through semi-automated processing of encyclopedic websites, dictionaries, and curated music and film charts. Our underlying criterion was recognizability and collective salience: a phrase was considered culturally relevant if it appeared in organized compilations or popularity-based rankings, indicating that it is widely known rather than niche.

Importantly, RuBIN does not aim to exhaustively represent the entirety of a nation's cultural heritage. Instead, it is designed to assess the extent to which a language model possesses culturally embedded

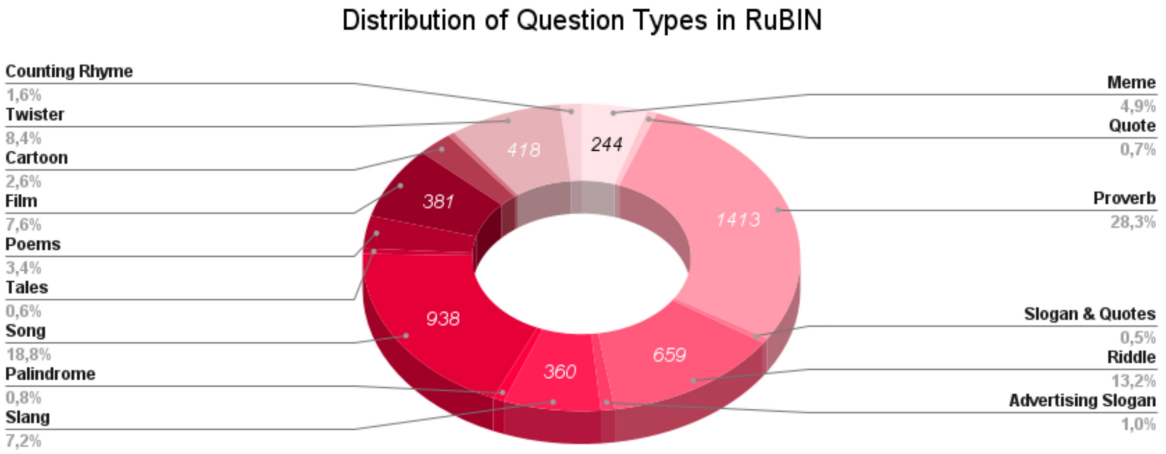

Figure 2: Cultural source types in the RuBIN benchmark.

knowledge that is intertwined with the Russian language and broadly shared within its linguistic community.

The compilation of texts was conducted directly by the authors themselves, employing field linguistics methodologies. In practice, this involved informal elicitation through conversations with native speakers, systematic observation of expressions used in spontaneous everyday speech, and the documentation of culturally salient phrases encountered in natural communicative contexts.

The majority of questions were designed in a pattern-based format like “Kak zakanchivayetsya fraza...?” – meaning “How does the phrase end ...?”, leveraging native speakers’ automatic recall and making it easier to test shared cultural knowledge. To ensure representativeness, sources were chosen for their wide public familiarity, cultural salience, and accessibility across age and social groups. For phrases that could not be naturally adapted to the “How does the phrase end?” format, questions were constructed manually, with particular attention to ensuring the presence of an unambiguous, factually verifiable ground truth answer.

The benchmark includes culturally significant materials from a wide range of sources. Where possible, items were annotated with the year of peak cultural relevance, such as release dates or award years. Traditional oral forms and slang, which lack precise dating, were included without temporal metadata. Detailed information on data sources is provided in Table 1.

A filtering procedure was applied to each topic, checking for cultural relevance, broad recognizability, appropriateness, and the feasibility of forming clear multiple-choice questions. Topics tied to recent or politically sensitive events were excluded. The goal was to ensure each question reflected culturally shared knowledge, was ethically sound, and had an unambiguous correct answer.

Memes Texts were selected from an online encyclopedia of internet memes, which records the year each meme appeared; this information was included in the dataset.
Folklore Texts were sourced from folklore dictionaries and collections. Due to the absence of reliable dating, no year was indicated.
Popular Songs Texts were selected from the lists of “Golden Gramophone” award winners and other music charts; the year of receiving the award was included in the dataset, as it most accurately reflects the moment of the work’s popularity.
Films and Cartoons Texts were chosen based on articles about the most popular and significant works; the year of release was specified.
Poems and Writers’ Quotes Texts were selected from works included in the school curriculum; no dates were provided.
Slang Texts included expressions from aggregator websites of slang terms; the year of origin of the expressions was not indicated.
Advertising Slogans The texts were collected from compilations of memorable television commercials; the approximate year of release was specified.

Table 1: Sources and Dating of Dataset Materials.

A collection of texts was initially gathered, which after a thorough filtering process resulted in a final dataset of 5,000 standardized questions. Each entry includes the question, four answer options, the correct label, topic, and year (when applicable). A breakdown of question domains is shown in Figure 2. We acknowledge the presence of class imbalance across domains, but given the availability of reliable sources, it could not be fully mitigated without compromising data authenticity.

Q.	How does the saying end: “Fingers and eggs - ...”?
A	don’t rub against the table
B	don’t fiddle in the grains
C	*don’t dip into the salt shaker*
D	don’t go into the nose

Table 2: Example of generated distractors. The correct answer is in italics.

3.2. Distractors creation

We adopted a multiple-choice question answering (MCQA) format because recognition-based tasks are particularly suitable for evaluating culturally shared knowledge, which typically operates at the level of familiarity and conventionalized recall rather than creative reconstruction. Unlike open-ended generation, the MCQA setup minimizes variability in surface formulations, thereby reducing evaluation noise. Additionally, the use of distractors provides a flexible framework for future extensions of the benchmark, making it possible to adjust task difficulty and design controlled experiments that probe different levels of cultural discrimination.

One of the key stages in the development of a multiple-choice question answering benchmark is the generation of distractors – incorrect but plausible answer options. The quality of these distractors directly impacts the reliability and validity of the evaluation: overly simplistic or obviously incorrect options may distort results by allowing models to achieve high scores through superficial heuristics rather than genuine comprehension.

In this study, distractors were generated using the QWEN2.5-14B language model, selected for its strong generative capabilities and flexibility in following instructions in Russian. The generation process involved providing the model with the question, the correct answer, and its assigned letter label (A, B, C, or D). The model was then prompted to generate three additional incorrect yet realistic options, producing a full set of four answers with a designated correct choice. This setup enabled direct integration of the output into the benchmark’s format. An example of the automatically generated distractors is provided in Table 2. The example is translated to English for demonstration purposes.

To prevent positional biases, the correct answer was randomly assigned to one of the four answer slots (A-D). This model-based generation approach allowed for automation of the distractor creation process while ensuring stylistic and thematic consistency among answer options and minimizing the risk of models identifying the correct answer based on surface cues alone.

All generated items underwent manual review to verify several quality criteria. Distractors were examined to ensure they were not obviously incorrect, nonsensical, or factually identical to the correct answer. Additional checks were conducted for stylistic consistency across options, including uniform grammatical agreement, case, and register. Items containing non-Russian distractors, duplicated or identical options, correct-answer duplicates, or empty answer slots were systematically removed.

3.3. Human evaluation

The development of a cultural knowledge benchmark loses its validity without the involvement of native speakers and cultural insiders. Since culture is inherently shared and collective, only verification by native speakers can confirm that the proposed items genuinely reflect cultural memory and recognizable context.

To ensure cultural validity, an initial version of the benchmark was evaluated by native Russian speakers via the Yandex Tasks platform. Eligibility criteria required participants to be over 20 years old and declare Russian as their native language.

To ensure response reliability, the time each annotator spent on the task was monitored; unusually short response times were flagged as potential indicators of low-quality or automated responses. Additionally, quality verification involved testing for a correlation between the question’s reference year and the respondent’s age group, as detailed in Section 3.4.

Each question was evaluated by an average of three annotators. Only questions with a correct response rate of at least 50% were retained, resulting in a filtered subset called RUBIN-EASY, consisting of 4,216 questions. The average human accuracy on this set was approximately 90%, supporting its cultural validity and indicating an appropriate level of difficulty for human respondents. RUBIN-EASY better represents widely shared cultural knowledge, making it a more reliable measure of what models should capture as common, rather than niche, information. This subset is recommended as the primary evaluation set for language models.

3.4. Verification of human evaluation

To evaluate the reliability of the collected annotations and confirm that the responses were provided by real individuals with diverse cultural and generational backgrounds, a chi-squared (χ^2) test of independence was performed between two categorical variables: the historical period referenced in a question and the age group of the respondent. The distribution of years and ages is presented in Figure 3.

The following hypotheses were tested:

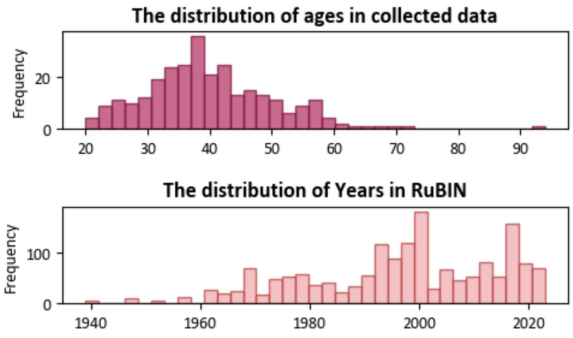

Figure 3: The distribution of ages in collected human data and questions' years in RuBIN.

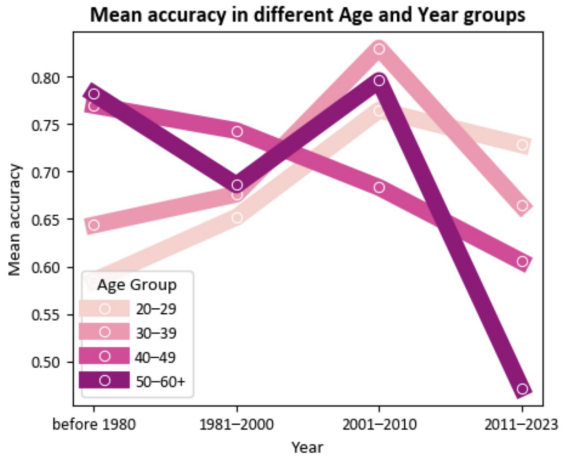

Figure 4: Mean accuracy in different Age and Year Groups. Notably, each temporal category shows a leading age group whose superior performance aligns logically with their generational experience.

H_0: The referenced year of a question and the respondent's age group are independent; the likelihood of a correct response is unrelated to the participant's age.

H_1: There is a statistically significant association between the referenced year of a question and the respondent's age group.

The resulting p-value (5.55×10^{-10}) is far below the significance threshold ($\alpha = 0.01$), leading to rejection of the null hypothesis. This indicates a statistically significant relationship between respondent age and question year. In other words, performance on the benchmark questions varies meaningfully with the age of the annotator, confirming that the annotations were provided by real individuals with different levels of cultural familiarity and historical knowledge.

The observed dependency is further illustrated in Figure 4, which presents the proportion of correct responses as a function of annotators' age and the time period to which the question content refers. These findings support the reliability of the collected responses and justify their use as a foundation for selecting benchmark questions and analyzing large language models' performance on the benchmark tasks.

4. Evaluation with LLMs

This section presents an experimental evaluation of various large language models on the RuBIN cultural knowledge benchmark. The primary objective of the evaluation is to identify differences in the models' abilities to handle tasks that require understanding culturally grounded expressions, idiomatic constructions, and other elements of collective memory specific to the Russian-speaking cultural context. In addition to assessing overall performance, the analysis examines variation across temporal and thematic dimensions, highlighting how each model responds to questions tied to different historical periods and content categories. Particular attention is given to YandexGPT5-Lite, which, despite its relatively modest size (8B parameters), demonstrates competitive results and exhibits strengths in processing Russian-language material, warranting closer inspection of its behavior and error patterns.

4.1. Conducting the Evaluation

The evaluation was conducted on the full version of the benchmark, which includes all question types. For each model, accuracy was calculated as the proportion of correctly answered questions, determined by comparing the model's predicted response with the reference answer. Predictions were extracted using regular expressions. For all models, the instruct versions were evaluated to ensure consistency in their response style and alignment with task instructions.

The evaluation involved 24 language models, both open-source and commercial, varying in size, architecture, and language focus.

The results shown in Table 3 reveal substantial variation in model performance on the RuBIN benchmark. Proprietary models such as GPT-4.1, DeepSeek-R1, and GEMINI achieved top scores, with Qwen2.5-MAX and CLAUDE 3 SONNET also performing strongly, suggesting that commercial API-accessible models tend to integrate Russian cultural knowledge more effectively. Among open-source models, LLAMA- 3.3-70B and Qwen2.5-72B demonstrated competitive results, highlighting the potential of large-scale open architectures with quality pretraining.

YandexGPT-8B achieved 59.94%, outperforming similarly sized multilingual models such as Qwen2.5-7B and Qwen3-8B, indicating a deeper adaptation to the Russian cultural context. Smaller

Model	Acc.	Acc. E	Δ	NA
Qwen2-0.5B	26.59	26.85	0.26	0.10
Qwen2-1.5B	31.04	31.95	0.91	0.22
Qwen2.5-0.5B	24.86	26.07	1.21	0.00
Qwen2.5-1.5B	31.65	32.47	0.82	0.00
Qwen2.5-7B	40.50	41.84	1.34	0.06
Qwen2.5-14B	54.03	56.45	2.42	0.00
Qwen2.5-72B*	57.39	-	-	0.00
Qwen2.5-Max	62.22	65.42	3.20	10.87
Qwen3-0.6B	26.02	26.49	0.47	3.41
Qwen3-1.7B	31.89	33.09	1.20	0.00
Qwen3-4B	38.29	39.97	1.68	4.63
Qwen3-8B	47.55	49.76	2.21	0.00
Llama-3-8B	35.81	36.95	1.14	0.06
Saiga-llama3-8B	35.38	36.98	1.60	0.00
Llama-3.2-1B	24.66	25.36	0.70	0.00
Llama-3.3-70B	57.55	59.75	2.20	0.00
Gemma-2-2b-it	30.55	32.02	1.47	0.33
YandexGPT-8B	59.94	63.83	3.89	0.06
Mistral-7B	31.71	32.92	1.21	0.35
GPT-4o-mini	54.56	57.52	2.96	0.22
GPT-4.1	78.33	80.91	2.58	0.00
DeepSeek-R1	80.50	83.35	2.85	0.00
Gemini-2.0-flash	74.51	77.40	2.89	0.02
Claude3 Sonnet	58.26	61.57	3.31	0.00

Table 3: The accuracy (%) of LLM evaluation on RuBIN (Acc.) and RuBIN-Easy (Acc. E), a subset of the full RuBIN benchmark that human participants solve with approximately 90% accuracy. Additionally, we report the percentage of questions that were not answered (NA), defined as cases where no answer tag could be extracted from the model output. Also we provide Δ - difference in accuracy between RuBIN and RuBIN-Easy (pp). It shows how much a model's performance drops when moving from easier to more challenging tasks, indicating its sensitivity to questions' complexity. We did not evaluate Qwen2.5-72B on the full version of the dataset due to budget constraints.

models (<2B) showed limited competence, underscoring their difficulty with culturally rich tasks.

The performance gap between the full RuBIN benchmark and RuBIN-Easy indirectly indicates whether a model's knowledge is likely derived from lexical resources or curated datasets rather than from genuine exposure to natural cultural usage. All models perform slightly better on the simplified version, which suggests that at least part of their knowledge is aligned with commonly encountered cultural information rather than niche or less frequent elements. However, the fact that their performance does not dramatically increase implies that a substantial portion of their cultural knowledge may stem from lexical resources, curated datasets, or synthetic corpora, rather than from genuine exposure to diverse, real-world language use.

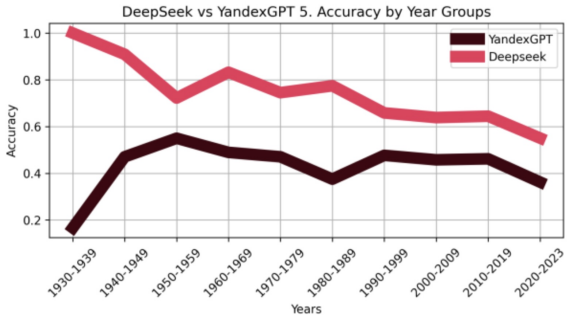

Figure 5: DeepSeek and YandexGPT-5 mean accuracy by time periods.

Most models left fewer than 0.5% of questions unanswered, except Qwen2.5-Max, Qwen3-0.6B, and Qwen3-4B. In the case of Qwen2.5-Max, this was primarily due to strict safety constraints: the model occasionally refrained from answering questions that could be indirectly associated with sensitive topics, such as politics, potentially offensive language, or criminal activity, even when the items themselves were culturally oriented and not intended to solicit harmful or inappropriate content; for others, formatting issues caused non-responses.

4.2. Insights from Model Performance

The following stage of the analysis considers the performance of the top-performing models, introducing a broader examination of how their outputs relate to the cultural characteristics of the questions.

To begin with, a key finding is that deep cultural adaptation during pre-training plays a crucial role: YandexGPT outperforms models of similar size, such as Saiga-Llama, which was further fine-tuned on Russian texts but still performs significantly worse. This highlights that comprehensive cultural alignment in the initial training phase is more important than later-stage fine-tuning for achieving strong performance on culture-specific benchmarks.

Another important observation is the tendency of large multilingual models to rely heavily on historical data. Figure 5 presents the comparative accuracy of the DeepSeek and YandexGPT models across questions categorized by decade, spanning 1930-1939 to 2020-2023. The accuracy dynamics of the two models differ notably across the temporal groups. DeepSeek shows a relatively stable performance, with only moderate fluctuations. In

Category	Accuracy	Category Size
Palindromes	0.238	42
Songs	0.437	938
Cartoons	0.442	129
Films	0.449	381
Advertising	0.451	51
Counting Rhymes	0.468	79
Lullabies	0.48	25
Twisters	0.55	418
Quotes	0.556	36
Memes	0.561	244
Proverbs	0.67	1413
Slogans & Quotes	0.692	26
Riddles	0.695	659
Poems	0.719	171
Slang	0.869	360
Tales	0.929	28

Table 4: YandexGPT5-Lite accuracy per category.

contrast, YandexGPT demonstrates a more pronounced variability, with a steep increase from the 1930s to its peak in the 1950-1969 range, followed by a consistent decline toward 2020-2023.

DeepSeek likely performs better on "older" topics because its training data may include a broader range of historical and archival sources, giving it greater exposure to older cultural references and language patterns. In contrast, YandexGPT may rely more heavily on contemporary datasets, which could limit its ability to recognize or interpret cultural elements from earlier decades, leading to lower accuracy on older material.

The next stage of the analysis examines the performance of YandexGPT across different content categories, as presented in Table 4. YandexGPT was selected for closer examination as it demonstrated noteworthy performance relative to its size category (8B parameters) and is specifically optimized for processing Russian-language content. The model's accuracy varies substantially, ranging from 0.238 for palindromes to 0.929 for tales. Categories characterized by complex linguistic structures or less common forms, such as palindromes and lullabies, show notably lower accuracy, suggesting limitations in the model's ability to process rare or unconventional textual patterns. In contrast, categories with more standardized or formulaic structures, such as slang and tales, demonstrate significantly higher performance. Intermediate accuracy levels are observed for widely represented categories like proverbs and riddles, where larger sample sizes may contribute to more stable results. These disparities indicate that the model's effectiveness is closely tied to both the linguistic properties of the category and the representativeness of such content in its training data.

The subsequent analysis addresses the distribution of incorrect answer selections produced by YandexGPT5, as summarized in Table 5. Among the erroneous responses, option A accounts for nearly half of all selections, while options B and D are chosen at relatively comparable rates. Option C is selected least frequently. This skewed distribution suggests a systematic bias in the model's response patterns, potentially reflecting tendencies in its decision-making process or imbalances in how it weighs particular answer positions. Such a pattern may indicate that the model either defaults to specific positions when uncertain or exhibits an implicit preference influenced by its training data.

Option Label	Answer Distribution, %
A	48.1
B	18.41
C	13.82
D	19.66

Table 5: Distribution of YandexGPT-5-Lite incorrect answer selections.

5. Conclusion

This paper introduces RuBIN, a novel benchmark designed to evaluate culturally grounded knowledge in large language models, with a particular focus on Russian linguistic and cultural specificity. Built using a dedicated framework that supports the structured generation of culturally rich, multiple-choice questions, RuBIN comprises 5,000 items spanning diverse domains. The underlying methodology used to construct RuBIN is language-agnostic and can be readily adapted to develop culturally grounded benchmarks for other languages and linguistic communities.

Using this benchmark, we conducted a large-scale evaluation across a wide range of contemporary models, allowing for a comprehensive comparison of their performance on culturally embedded tasks. Additionally, we performed an in-depth analysis of the top-performing models to better understand the factors contributing to their success.

We use language as a natural expression of culture, which grounds our benchmark in authentic cultural contexts. This approach ensures that the benchmark truly reflects culturally embedded knowledge, making it both meaningful and relevant. By focusing on language, we capture the nuances and subtleties that define cultural identity, enhancing the benchmark's validity.

Our underlying assumption is that the benchmark targets culturally conventionalized expressions that are stably reproduced within a linguistic community. Precisely because such expressions are conventionalized, they tend to occur online in relatively stable forms. However, as culturally embedded

units, they are disproportionately represented in Russian-language sources, including articles, comment sections, literature, and other locally produced materials. We therefore assume that strong performance on RuBIN reflects not merely incidental web exposure, but substantial contact with authentic Russian-language texts drawn from culturally grounded communicative contexts. In this sense, benchmark performance serves as an indirect indicator of the model's cultural adaptation to the Russian linguistic environment.

The analysis of model performance reveals several key insights. First, deep cultural adaptation during the pre-training phase is crucial: for instance, YandexGPT, which underwent thorough cultural alignment, outperforms Saiga-Llama – a model of the same size that was only fine-tuned on Russian texts but still lags behind. Second, large multilingual models tend to focus heavily on historical data, showing better performance on older questions, which implies their training relies mainly on established, archived content rather than recent cultural developments. Finally, although some large closed-source models demonstrate competitive results, none have yet reached human-level performance on these culture-specific benchmarks. These findings highlight the challenges and opportunities in developing language models with genuine cultural understanding. Future work should prioritize deeper cultural integration and more diverse, up-to-date training data to bridge the gap toward human-level performance.

Limitations

Although RuBIN introduces a scalable approach to building large and diverse cultural benchmarks, several limitations remain. First, while all questions underwent manual review to eliminate major flaws, we cannot guarantee that every automatically generated distractor is equally plausible. Some questions may still contain options that are less challenging, potentially influencing benchmark difficulty.

Second, despite the framework's scalability, adapting RuBIN to other languages still requires collaboration with native speakers. Cultural knowledge remains inherently language-specific, and steps such as formulating questions or refining distractors cannot be fully automated without risking inaccuracies or cultural inauthenticity.

Finally, not all RuBIN questions follow the "how does the saying end" format, as strictly adhering to this pattern would result in the loss of culturally relevant material. While this design choice preserves data diversity, it introduces variability in question structures, which could impact how models approach different question types.

Ethical Considerations

The RuBIN benchmark is released under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license, explicitly restricting its use to academic and non-commercial research purposes. Any derivative works based on RuBIN are subject to the same limitations and must comply with the original data access conditions.

The dataset contains no personally identifiable information or sensitive data, ensuring anonymity and privacy protection. All cultural and language-specific content has been carefully curated to avoid offensive or discriminatory material. Nonetheless, users should be aware that the benchmark focuses on culturally embedded linguistic phenomena, which may reflect certain biases or limitations inherent in the data collection process.

RuBIN is intended solely for research and evaluation of language models and should not be used for commercial applications. The authors acknowledge the potential risks associated with cultural generalizations and advise caution in interpreting results beyond the benchmark's defined scope.

Since we relied on crowd-sourcing to validate the benchmark tasks, we outline here the ethical considerations. No sensitive personal data was collected from participants during the study. All participants provided their informed consent before beginning the task, and they were informed that their responses would be used exclusively for research purposes. The recruitment was carried out through a crowd-sourcing platform, targeting people aged 20 or older who stated Russian as their native language, according to the demographic requirements of the study.

Participants were compensated at a fair rate, and the payment was calculated based on the estimated time needed to complete the task. No bonuses or penalties were applied.

6. Bibliographical References

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025.

CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2025. Bertaqa: how much do language models know about local culture? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Anton Lavrouk, Tarek Naous, Alan Ritter, and Wei Xu. 2025. What are foundation models cooking in the post-soviet world?

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18477–18494, Vienna, Austria. Association for Computational Linguistics.

Jamshid Mozafari, Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2025. Wrong answers can also be useful: Plausibleqa - a large-scale qa dataset with answer plausibility scores. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3832–3842, New York, NY, USA. Association for Computing Machinery.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya K Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan S Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia moakhar, Ayush Tarun, Azmine Toushik Wasi, Thenuka Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2025. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Representation Learning*, volume 2025, pages 83291–83322.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton

Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2025. LLMs may perform MCQA by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5852–5862, Abu Dhabi, UAE. Association for Computational Linguistics.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *CoRR*, abs/2504.15521.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024a. Qwen2 technical report.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu,

Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu li, Feiyu Xiong, Bo Tang, and Ding Chen. 2025. xfinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Representation Learning*, volume 2025, pages 59850–59892.

Appendix A. Examples of Benchmark Tasks

Palindromes	
Q.	How does the phrase end 'Pressed the boar…' ?
A	*On the eggplant*
B	Into the raspberry
C	Into the trap
D	Into the bottle
Songs	
Q.	How does the line from the song (performed by Lube) 'Taganskaya Station, the convict's fate…' end?
A	Here quietly and lonely flew the years of my youth
B	*White acacia blooms in the yard*
C	In winter and summer, with one music the pipes roar
D	The door slammed and I went to Taganskaya
Cartoons	
Q.	How does the quote from the cartoon end 'He flew away, but…' ?
A	Became a famous pilot
B	*Promised to return*
C	Left a gift for his friend
D	The last greeting was weak
Films	
Q.	How does the quote from the movie end 'For example, in London, a dog is a man's friend. – And with us…' ?
A	All animals are man's enemies
B	The police are man's friend
C	Mice are man's enemies
D	*The building manager is man's friend*
Advertising	
Q.	How does the phrase from the advertisement end 'Dimon, what a…' ?
A	Baboon
B	Raspberry
C	Sheep
D	*Ballet*
Counting Rhymes	
Q.	How does the counting rhyme end 'Beyond the seas, beyond the mountains, beyond the dense forests on the hill -…' ?
A	A den, a bear sits at home
B	*A little tower, a lock hangs on the door*
C	A hut on chicken legs
D	An onion garden, where grandmothers and grandfathers sit
Lullabies	
Q.	How does the line from the lullaby end 'Ay, luli-luli-luli, flew…' ?
A	Bullfinches
B	*Cranes*
C	Sparrows
D	Hares

Twisters	
Q.	How does the tongue twister 'We have a guest…' end?
A	*who took the cane*
B	who left the chambers
C	who left the spoon
D	who went to the tavern
Quotes	
Q.	How does the quote end 'Ill-cut, but…'?
A	Intended for centuries
B	Came out of the mold
C	Reliably assembled
D	*Tightly sewn*
Memes	
Q.	What ends the phrase from the video where a Shiba Inu dog falls into the river: 'This is a fiasco...' ?
A	buddy
B	doggo
C	Bobik
D	*bro*
Proverbs	
Q.	How does the proverb end 'A crow flew into the master's…'?
A	kingdom
B	yards
C	house
D	*chambers*
Slogans & Quotes	
Q.	How does the saying end 'Who comes to us with a sword…'?
A	With him we will fight
B	Will go back home
C	*That one will perish by the sword*
D	That one will not return to us
Riddles	
Q.	Guess the riddle: 'In winter and summer it is of the same color.'
A	Tulip
B	Sakura
C	*Spruce*
D	Raspberry
Poems	
Q.	How does the line from the poem 'A lonely sail is whitening…' end?
A	Beyond the islands of joyful dreams
B	In the heights, the shining moon sees us
C	*In the fog of the blue sea*
D	After the storm, the wave falls asleep

Slang	
Q.	Which slang word has this definition: 'A player who uses bugs or game world imbalances to gain personal advantage.' ?
A	Grinded
B	Screeper
C	Rusher
D	*Abuser*
Tales	
Q.	What is the name of the round character that the grandma and grandpa baked?
A	Fox Alisa
B	*Kolobok*
C	Teremok
D	Turnip

Table 6: Examples of multiple-choice questions with correct answers in italics.Translated into English for demonstration purposes.