DOI: 10.14529/jsfi250307

Do Open Large Language Models Know What, Where, and When? A Case Study with Quiz-Style Questions

Anna V. Kuznetsova[1], ***Viktor A. Byzov***[2], ***Ilias V. Aslanov***[1], ***Evgeny V. Kotelnikov***[1]

Large language models (LLMs) are increasingly tested on reasoning-intensive benchmarks, yet their performance on complex quiz-style tasks remains underexplored. In this paper we evaluate modern open-source LLMs on the Russian intellectual game *What? Where? When?*, a challenging format requiring fact recall, associative reasoning, and interpretation of hidden clues. We introduce a new dataset of 2600 questions (2018–2025), enriched with empirical human team success rates and annotated with structural and thematic clusters. We benchmark 14 recent open models accessible via API using both automatic metrics (Exact Match, BLEU, ROUGE) and an LLM-as-a-Judge framework. The best system, `Qwen3-235B-A22B-Thinking`, achieved 32.4% accuracy, but still lagging behind the average human team success rate (45.8%). Large-scale reasoning-enabled models consistently outperformed non-reasoning or smaller counterparts, particularly in domains such as technology, ancient world, psychology, and nature. However, omission, wordplay, and proper-name questions remained difficult across all systems. Comparison with *CheGeKa* (MERA leaderboard) shows that our dataset is substantially harder: while leading proprietary and open models reach EM of 0.534–0.645 and 0.442 on *CheGeKa*, respectively, the strongest model in our benchmark achieves only 0.255 EM. Correlation analysis indicates that human and model perceptions of difficulty only weakly align, suggesting different problem-solving strategies. Qualitative case studies further show that models excel more in fact recall than in reconstructing hidden logic. Our findings highlight both the progress of open LLMs and their current limitations in quiz-style reasoning. The new dataset offers a complementary and more challenging benchmark for Russian-language evaluation.

Keywords: Large Language Models (LLMs), question answering, reasoning, evaluation metrics, quiz datasets, LLM-as-a-judge, human-AI comparison.

Introduction

Recently, there has been a growing interest in evaluating large language models (LLMs) on tasks that require: (1) extensive and well-organized memory; (2) ability to hold multiple information units in working memory while tracking chains of thought; (3) logical-associative reasoning; and (4) fluid intelligence – the capacity to tackle novel problems beyond rote patterns [3, 11, 14, 17, 21]. These competencies are particularly well-tested in intellectual games such as *What? Where? When?* [8]. *What? Where? When?* (Russian: *Chto? Gde? Kogda?*) is a long-running Russian intellectual quiz game similar to *Jeopardy*, where teams of players are given one minute to answer complex, riddle-like questions that often involve hidden clues, wordplay, and multi-step reasoning [2]. Such games offer a challenging testbed to assess LLM reasoning ability.

While many widely used benchmarks (such as *BIG-Bench* [20], *MMLU* [10], and *GSM-8K* [7]) offer multi-step reasoning challenges, they are often too general or already included in model training data. Quiz-style datasets like *TriviaQA* [12], *QANTA (Quizbowl)* [19], and *HotpotQA* [23] provide more direct parallels to trivia competitions, yet still focus mainly on fact recall or retrieval. More recent resources, including *modeLing* (Linguistics Olympiad puzzles) [6], *TurnBench-MS* (multi-turn logic games) [24], and *QUENCH* (open-domain quizzing

[1]European University at St. Petersburg, St. Petersburg, Russian Federation
[2]Vyatka State University, Kirov, Russian Federation

with masked rationales) [13], better mirror the cognitive demands of *What? Where? When?* and allow for more robust assessment of reasoning and inference beyond standard static tasks.

However, existing studies on LLM performance in intellectual quiz-style tasks are limited. Prior work has often relied on outdated datasets, such as those likely already included in model pre-training corpora, and on models that no longer represent the state of the art. Moreover, there remains a lack of comparative benchmarks against human teams, and insufficient analysis of how performance varies with question themes and structural features.

In this paper, we address these gaps by presenting a refreshed and rigorous evaluation of modern, open LLMs in the context of *What? Where? When?* game. First, we construct a novel dataset of 2600 questions and answers from *What? Where? When?*, annotated with human team answer-rates. We then perform structural and thematic clustering of the dataset, enabling fine-grained analysis. Employing an "LLM-as-a-Judge" methodology, we evaluate model responses across question clusters for 14 open models accessible via API, identifying both strengths and systematic errors, and complement this analysis with automatic metrics such as Exact Match, BLEU, and ROUGE. Through this analysis, we demonstrate how LLM success correlates with question type and topic, and highlight key reasoning limitations. We also compare model performance to human teams using the recorded success rates and analyze how question difficulty aligns across humans and models. Finally, we present qualitative case studies – successes and failures – that illustrate typical reasoning patterns and types of errors.

Our contributions are as follows:

- We created a new dataset of 2600 *What? Where? When?* questions with human team success rates.
- We applied structural and thematic clustering to enable fine-grained analysis of reasoning requirements.
- We evaluated 14 recent open-access LLMs using LLM-as-a-Judge and automatic metrics, and compared their performance to human teams.
- We analyzed results across clusters, identifying strengths, recurring reasoning failures, and systematic challenges for LLMs.
- We complemented the quantitative results with qualitative case studies that illustrate characteristic successes and errors.

The remainder of this paper is organized as follows. Section 1 reviews previous work on quiz-style question answering and reasoning evaluation of large language models. Section 2 describes the construction and annotation of our What? Where? When? dataset. Section 3 details the methodology of our experiments, including model selection, evaluation metrics, and analysis procedures. Section 4 presents and discusses the results, highlighting model performance across structures, topics, and comparisons with human teams. Finally, the Conclusion summarizes the key findings and outlines directions for future research.

1. Previous Work

Research on LLMs in quiz-style question answering spans two main directions. On the one hand, traditional trivia corpora such as *TriviaQA* [12], *QANTA (Quizbowl)* [19], and *HotpotQA* [23] provide large-scale collections of factoid or multi-hop questions. These datasets are valuable for measuring knowledge coverage and retrieval ability, but they only partially reflect the associative reasoning and hidden-clue structure characteristic of intellectual games. On the other hand, a new generation of reasoning-oriented benchmarks (including *modeLing* [6],

TurnBench-MS [24], and *QUENCH* [13]) explicitly targets logical inference, multi-turn reasoning, and puzzle-like inference. These resources move closer to the spirit of *What? Where? When?*, though they remain English-centric and do not capture its cultural specificity.

In the Russian-language context, the most significant contribution to date is the *CheGeKa* dataset [17, 21], which contains nearly 29 375 Jeopardy-style questions annotated by topic and difficulty. *CheGeKa* distinguishes between factoid and reasoning tasks and introduced a scoring system adapted to gameplay. It has become the reference benchmark for Russian quiz QA, with a public leaderboard [1] where human teams still lead (token-wise F1=0.719, EM=0.645). Among proprietary models, `Gemini 1.5 Pro` (F1=0.630, EM=0.534) and `Claude 3.7 Sonnet` (F1=0.630, EM=0.526) perform best, while the strongest open model, `DeepSeek-V3-0324`, trails behind (F1=0.531, EM=0.442).

Other studies have explored model behavior on quiz questions in smaller settings. Lifar et al. [14] tested `LLaMA3-405B` on a 416-question *CheGeKa* sample and showed that multi-agent prompting strategies such as self-consistency and suggesterdiscriminator improved Exact Match by about 8 percentage points over single-agent baselines. Aßenmacher et al. [3] introduced *wwm-german-18k*, a German multiple-choice dataset, and found that accuracy remained high on easy questions but dropped to near-random on the hardest levels. Hu et al. [11] proposed a dynamic benchmark of interactive games (*Akinator*, *Taboo*, *Bluffing*) for testing deductive, abductive, and inductive reasoning, showing that different frontier models excel in different reasoning modes.

Our work extends this literature by introducing a new Russian dataset of 2600 *What? Where? When?* questions (2018–2025), enriched with empirical human success rates – an element absent in prior corpora. Unlike earlier studies, we combine structural and thematic clustering, evaluation of 14 recent open models, and comparison to human teams, complemented by qualitative case studies of successes and failures.

2. Dataset

The *What? Where? When?* quiz questions were collected from the IQ Game website[3]. The initial dataset contained 3526 entries, which were then preprocessed: blitz questions, multi-part questions with limited answering time, questions with accompanying materials, and rarely used questions were removed. Specifically, we excluded questions that had been played fewer than 100 times on the platform.

After filtering, the final dataset[4] included 2600 unique questions spanning 2018–2025, with an average length of 29.1 words. An example question is shown in Fig. 1.

Using regular expressions, we identified five structural clusters of questions (Tab. 1). Each question was assigned only to the first matching cluster in a priority hierarchy, where more specific patterns had higher precedence.

We employed a two-step procedure for thematic clustering:

1. generation of a list of topics using `BERTopic` [9], `HDBSCAN` algorithm [4], and embeddings from `FRIDA`[5] model;
2. assignment of questions to the identified topics using `Qwen3-235B-A22B` model.

To avoid bias from structural patterns, these were stripped from the questions before thematic clustering. Embedding dimensionality was reduced from 1536 to 50 using `UMAP` [16]. Op-

[3] `https://iqga.me`

[4] `https://github.com/kotelnikov-ev/quiz-dataset`

[5] `https://huggingface.co/ai-forever/FRIDA`; we used "`categorize_topic: `" prefix.

Example Question

id: 3453

Question: According to the construction plan, the scene depicting *THIS* was supposed to be located high above. To better capture the perspective, Antonio asked to hoist a donkey. Answer in two words: what is *THIS*?

Answer: Nativity of Christ

Accepted answers: Birth of Jesus; Christmas Nativity

Commentary: During the construction of the Sagrada Familia, Gaud asked to hoist a donkey by a winch to the height where the Nativity scene was planned to be placed.

Answer rate: 82/188 (44%)

Season: 2024–2025

Figure 1. Illustrative example of a quiz question

Table 1. Structural clusters of questions

Cluster	Number	Share, %	Examples of questions (answer)
Word substitution (HE, SHE, X, SUCH, DOING THIS, ...)	1363	52.4	In a story by John Coetzee, a savage brought HIM to life with his breath. Name HIM (fire).
Answer format (answer in N words, consecutive letters, etc.)	477	18.3	A riddle of Turkic nomads: "I sit on a hill, stepping on copper bowls." Name these bowls in one word (stirrups).
Omission (missing word/letter, abbreviation)	128	4.9	A jewelry studio is called Room. Write the two Latin letters that we omitted in the name of this studio (au[room]).
Name (proper name required)	85	3.3	Name the person who, according to Michel Pastoureau, was often depicted with black lips (Judas).
Other	547	21.0	What did Tsvetan Angelov call the spears of the snow army? (icicles).
Total	**2600**	**100**	

timal `UMAP` and `HDBSCAN` parameters were selected via the Tree-structured Parzen Estimator implemented in `optuna`[6], targeting silhouette maximization and noise minimization.

This process yielded 30 preliminary topic clusters (silhouette score: 0.331). To improve interpretability, semantically similar clusters were identified and merged with the assistance of the `Claude Sonnet 4` model, which compared the most frequent terms and representative questions for each cluster. The LLM was provided with 50 most frequent terms from each

[6] `https://optuna.org`

cluster along with 15 randomly selected questions. As a result, we obtained 16 topic clusters with automatically generated names. Subsequently, we assigned all questions to the identified topics using `Qwen3-235B-A22B` model (Tab. 2).

Although it is impossible to completely exclude the possibility of training data overlap, the overall performance of the evaluated models on our dataset (see Section 4.1) suggests that large-scale memorization of the questions is unlikely. If the dataset had been substantially included in pretraining corpora, accuracy levels would plausibly be much higher.

3. Methodology

Our study included the following main stages:

1. Selection of a judge model from several proprietary models.
2. Obtaining answers to questions from several open models accessible via API.
3. Analysis of the answers.

3.1. Judge Model Selection

To evaluate the quality of the answers, we employed both automatic evaluation metrics (such as Exact Match, BLEU [18], ROUGE-1 and ROUGE-L [15]) and the LLM-as-a-Judge approach [25]. BLEU was computed as a precision-oriented metric based on n-gram overlap, while ROUGE-1 and ROUGE-L were calculated as F1-scores combining precision and recall. Recent studies indicate that metrics based solely on surface overlap (n-grams, exact spans) are limited in their applicability to more complex QA tasks – for example, Chen et al. (2019) show that F1 and similar metrics may not capture answer quality beyond extraction or simple generation tasks [5]. Similarly, Xian et al. (2025) demonstrate that in long-form question answering the style, length and category of answers can heavily bias traditional automatic metrics, and LLM-based evaluators exhibit significantly higher consistency with human judgments [22]. We therefore adopt the LLM-as-a-Judge approach as a complementary method, while acknowledging its own limitations and the need for further review of semantic-based and embedding-based evaluation metrics.

For the LLM-as-a-Judge, a random sample of 10% of the dataset (260 questions) was selected to compare evaluations from candidate LLM judges against human annotators. For these questions, answers were obtained from five open models: `Gemma-3-27b-it`, `QwQ-32B`, `Phi-4-multimodal`, `Llama-4-Scout-17B-16E`, and `Qwen3-32B` (52 answers per model). Their correctness was independently evaluated by two human annotators as well as by several proprietary LLMs (available via API) considered as candidates for the role of judge. The human annotators first labeled the answers of the open models independently. The initial inter-annotator agreement was high, with only a few discrepancies that were discussed and reconciled to obtain a consensus gold-standard set of labels. We then measured which candidate judge model aligned best with this reconciled human annotation set, using Cohens kappa coefficient (Tab. 3). The complete prompt used to instruct the judge model during automatic evaluation is shown below.

Table 2. Topic clusters of questions

Cluster	Number	Share, %	Examples of questions (answer)
Literature	443	17.0%	In the "Aeneid" it is said that Styx forms THEM. Name THEM in two words (nine circles)
History	337	13.0%	Interestingly, potatoes first appeared in China during the rule of... Which dynasty? (Ming)
Art	258	9.9%	Who demanded to rename his painting to "Love in the Bin"? (Banksy)
Nature	226	8.7%	In the illustration to the first chapter of Dr. Komarovskys parenting guide, a plant is shown. Which one? (cabbage)
Science	222	8.5%	"First the sails, then the ships hull." Aristotle used this observation to prove... What? (sphericity of Earth)
Etymology	189	7.3%	In Norway THIS is called krøllalfa–curly alpha. Name THIS (@)
Cinema	160	6.2%	Maria Scholl writes that the Swiss remain true to tradition and still announce IT in local cinemas. Name IT (intermission)
Technology	152	5.8%	In the late 1970s a famous company hired young people to stroll around Tokyo. What were they advertising? (Sony Walkman)
Ancient World	100	3.8%	Whom did the Aztecs equate with warriors fallen in battle? (women who died in childbirth)
Games	94	3.6%	At an event, a chess player from a TV series is asked to say "queen." Instead of which word? (cheese)
Sports	88	3.4%	MMA fighter Diana Avsaragova threw in the towel already during IT. Name IT (weigh-in)
Geography	69	2.7%	What name was given to the land where Indians wore moccasins of roughly tanned hides? (Patagonia)
Psychology	50	1.9%	A wealthy patient of Sigmund Freud had a phobia because of which he literally... did what? (laundered money)
Design	48	1.8%	One ATM in Vienna is stylized as HER. Name HER with a hyphenated word (piggy-bank)
Numismatics	11	0.4%	The first German radio listener paid 350 marks for the right to use a receiver. In this question we omitted nine of THEM. Name THEM (zeros)
Other	153	5.9%	A Russian tattoo salon is called "Yes and No." Which two rhyming words did we replace? (wants, hurts)
Total	**2600**	**100**	

```
You are an expert evaluating answers in an intellectual quiz game.
Your task is to assess a list of answers to questions. Evaluate each
answer independently of the others. For each question, you are given:
- "id" - question identifier,
- "question" - the question text,
- "answer" - the answer to be evaluated,
- "correct_answer" - the correct answer,
- "variations" - acceptable alternative answers that should also be
considered correct.
Return only JSON, without any additional comments:  a list of evaluations,
where each item is a dictionary with the keys:
- "id" - question identifier,
- "is_correct" - a boolean value indicating whether the answer is correct.
```

Table 3. Agreement of proprietary models with annotators (Cohen's kappa).
Best value is in **bold**, second-best is underlined

Judge model	Cohen's κ
GPT-4.1	**0.9370**
Gemini-2.5-flash	0.9348
Claude-sonnet-4	0.8984
Gemini-2.0-flash-001	0.7444
GPT-4.1-mini	0.6907
GPT-4o-mini	0.6429
Claude-3.5-haiku	0.5448

The best results were demonstrated by GPT-4.1 and Gemini-2.5-flash. Since at the time of the study the API cost of GPT-4.1 was several times higher than that of Gemini-2.5-flash, the latter was chosen as the judge model, as the quality difference was negligible.

3.2. Answer Generation

We evaluated 14 open-source models, focusing on recently released models accessible via API:

- **DeepSeek family:**
  - DeepSeek-R1-0528: Mixture-of-Experts (MoE) architecture; reasoning-first RL model from the V3 family.
  - DeepSeek-V3-0324: MoE, 671B total / 37B active.
  - DeepSeek-V3.1: MoE, hybrid thinking / non-thinking variant.
- **Qwen family:**
  - Qwen3-235B-A22B-Thinking: MoE, 235B total / 22B active; reasoning-oriented.
  - Qwen3-235B-A22B: MoE, 235B total / 22B active; hybrid (instruction + reasoning).
  - Qwen3-30B-A3B: MoE, 30B total / 3B active; hybrid (instruction + reasoning).
  - Qwen3-32B: dense, 32B; hybrid (instruction + reasoning).
  - QwQ-32B: dense, 32B; reasoning-oriented.
- **Llama family:**

  - `Llama-4-Maverick-17B-128E`: MoE; 128 experts, long-context optimization.
  - `Llama-4-Scout-17B-16E`: MoE; 16 experts, efficiency-focused.
- **Kimi family:**
  - `Kimi-K2-Instruct`: MoE; ∼1T total / 32B active; long-context model.
- **GPT-OSS family:**
  - `GPT-OSS-120B`: MoE; ∼117B total / 5.1B active; reasoning-tuned.
- **GLM family:**
  - `GLM-4.5-Air`: MoE; ∼106B total / 12B active; hybrid reasoning.
- **Gemma family:**
  - `Gemma-3-27B-it`: dense; 27B instruction-tuned model for dialogue and QA.

We focused on open-source models because they can, in principle, be reproduced or fine-tuned by the community, unlike proprietary counterparts. At the same time, many of the strongest open models require substantial computational resources to run locally. Since our access to hardware was limited, we relied on those open models that are available via the DeepInfra API[7]. This setup enabled systematic evaluation across families and scales at a fraction of the cost of operating large models in-house.

The same prompt was used for all models:

```
You are participating in an intellectual quiz game.
Please briefly reason about the following question and provide an answer.
Question: {question}.
Output your reasoning and answer in JSON format:
{ "reasoning": "your reasoning here",
"answer": "your answer here" }
```

All models were used in their default inference configuration as provided in the official model documentation. The temperature was set to zero to improve determinism. Each model was allowed up to five attempts per question, not to introduce variability, but to handle cases where a model produced no valid response due to output looping or incorrect formatting. If no answer was generated, the maximum output length was increased by 1000 tokens at each retry (starting from 2000 tokens). Despite these multiple attempts, in some cases models still failed to produce any valid answer; such cases were recorded as unanswered and treated as incorrect in subsequent evaluation.

3.3. Answer Analysis

Answer evaluation was carried out using two approaches: (1) automatic metrics (Exact Match, BLEU, ROUGE-1, ROUGE-L), and (2) evaluation by the judge model (`Gemini-2.5-flash`).

Before applying automatic metrics, answers were lemmatized, lowercased, and stripped of punctuation. Both the exact reference answers and acceptable variants provided in the dataset were considered correct. For the LLM-as-a-Judge evaluation, the judge model classified each response as correct or incorrect relative to the reference answers, yielding a binary decision. From these judgments we computed *Accuracy*, defined as the proportion of correctly classified responses.

[7] `https://deepinfra.com`

After preprocessing and evaluation with both automatic metrics and the judge model, we analyzed the results along several dimensions: overall model performance, variation across structural and thematic clusters, alignment with human team success rates, and representative qualitative examples.

4. Results and Discussion

4.1. Overall Model Performance

Table 4 summarizes the performance of 14 open-source LLMs on our benchmark, evaluated with both automatic metrics and an LLM-as-a-Judge approach. The results show a considerable variation across models, reflecting differences in reasoning capability, training paradigms, and model scale.

Table 4. Performance of open models

Model	Accuracy	Reasoning	EM	BLEU	R-1	R-L
`DeepSeek-R1-0528`	30.00	✓	0.223	0.255	0.290	0.289
`DeepSeek-V3-0324`	29.00		0.222	0.250	0.287	0.286
`DeepSeek-V3.1`	29.65		0.227	0.258	0.292	0.291
`Qwen3-235B-A22B-Thinking`	**32.42**	✓	**0.255**	**0.290**	**0.320**	**0.319**
`Qwen3-235B-A22B`	20.31		0.156	0.181	0.208	0.207
`Qwen3-30B-A3B`	6.12		0.047	0.052	0.065	0.065
`Qwen3-32B`	8.58		0.062	0.072	0.087	0.086
`QwQ-32B`	12.62	✓	0.084	0.100	0.124	0.123
`Llama-4-Maverick-17B-128E`	21.77		0.172	0.199	0.227	0.226
`Llama-4-Scout-17B-16E`	13.81		0.100	0.123	0.148	0.147
`Kimi-K2-Instruct`	19.77		0.136	0.158	0.189	0.188
`GPT-OSS-120b`	13.65	✓	0.095	0.107	0.128	0.127
`GLM-4.5-Air`	12.42	✓	0.091	0.108	0.129	0.129
`Gemma-3-27b-it`	12.23		0.085	0.103	0.126	0.125

Among all evaluated systems, `Qwen3-235B-A22B-Thinking` achieved the best overall performance, with the highest accuracy (32.42%) and superior results on all automatic metrics. Importantly, this model explicitly incorporates a reasoning mode, which appears to contribute significantly to its advantage over the non-reasoning counterpart `Qwen3-235B-A22B`, which reached only 20.31% accuracy. This contrast highlights the effectiveness of explicit reasoning strategies for complex question answering tasks, where solutions often require multi-step inference rather than surface-level retrieval.

The `DeepSeek` family demonstrated relatively strong performance, with accuracies around 2930%. The reasoning-enabled `DeepSeek-R1-0528` slightly outperformed the non-reasoning `DeepSeek-V3` and `DeepSeek-V3.1` variants in terms of *Accuracy*, again underscoring the importance of reasoning traces. However, the performance gap between the reasoning and non-reasoning `DeepSeek` models was narrower than that observed in the `Qwen3` family, suggesting that other architectural or training factors may also play a role.

By contrast, smaller-scale models such as `Qwen3-30B-A3B`, `Qwen3-32B`, and `QwQ-32B` achieved substantially lower accuracies (613%), with weak scores on EM, BLEU, and ROUGE. `QwQ-32B`,

explicitly positioned as a reasoning-oriented variant, outperformed its standard dense counterpart `Qwen3-32B` (12.62% vs. 8.58%), showing that reasoning specialization can bring relative gains even at moderate scale. However, the absolute performance of both models remained low, far behind the large reasoning-enabled `Qwen3-235B-A22B-Thinking`. This suggests that while reasoning traces improve results, they cannot compensate for limited model size and knowledge coverage.

Other evaluated families, including `Llama-4`, `Kimi-K2`, `GPT-OSS`, `GLM-4.5-Air`, and `Gemma-3`, demonstrated moderate to low performance (1222% accuracy). While some of these models occasionally produced plausible answers, their overall metrics remained below those of the strongest `DeepSeek` and `Qwen` variants. Notably, both `GPT-OSS-120B` (∼117B total / 5.1B active) and `GLM-4.5-Air` (∼106B total / 12B active) are reasoning-enabled Mixture-of-Experts architectures, yet their accuracy (1213%) was far below that of the much larger `Qwen3-235B-A22B-Thinking` (235B total / 22B active, 32.42% accuracy). This contrast highlights that scale and effective integration of reasoning capabilities are critical: smaller MoE models with reasoning signals did not achieve competitive performance.

To illustrate typical failure patterns of lower-performing models, Tab. 5 shows two representative examples where all `DeepSeek` models and `Qwen3-235B-A22B-Thinking` produced correct answers, while some weaker models, for example `Kimi-K2-Instruct` and `Qwen3-32B`, failed. Both questions were among the easiest for human participants (answered correctly by over 96% of them), which highlights the qualitative gap between the higher- and lower-performing models.

Table 5. Examples of failure cases of lower-performing models

Question: The symbiotic relationship attributed to certain birds is merely a legend. In fact, these birds catch flies that appear in meat leftovers rather than DO THIS. What does DO THIS mean? **Correct answer:** clean crocodiles teeth **Kimi-K2-Instruct / Qwen3-32B:** remove parasites from an animal
Question: When composer John Tesh came up with a good melody, he was in a hotel and could not write the music down. To preserve the melody, he called his home number of... whom? **Correct answer:** himself **Kimi-K2-Instruct / Qwen3-32B:** his wife

4.2. Performance by Question Structure

Figure 2 presents model accuracies broken down by the main structural categories of questions: word substitution, answer format, omission, name, and other (see Tab. 1). The results reveal systematic differences in difficulty across categories, as well as clear trends in how reasoning-enabled models perform relative to their non-reasoning counterparts.

Word substitution questions (e.g., replacing pronouns or phrases) are the most frequent type and generally yielded the highest accuracies across models (except for the *Other* category). The best results were achieved by `Qwen3-235B-A22B-Thinking` (32.1%) and `DeepSeek-R1-0528` (31.2%), with other `DeepSeek` variants following closely. Even medium models such as `Llama-4-Maverick` (22.2%) and `Kimi-K2-Instruct` (19.1%) achieved moderate success here,

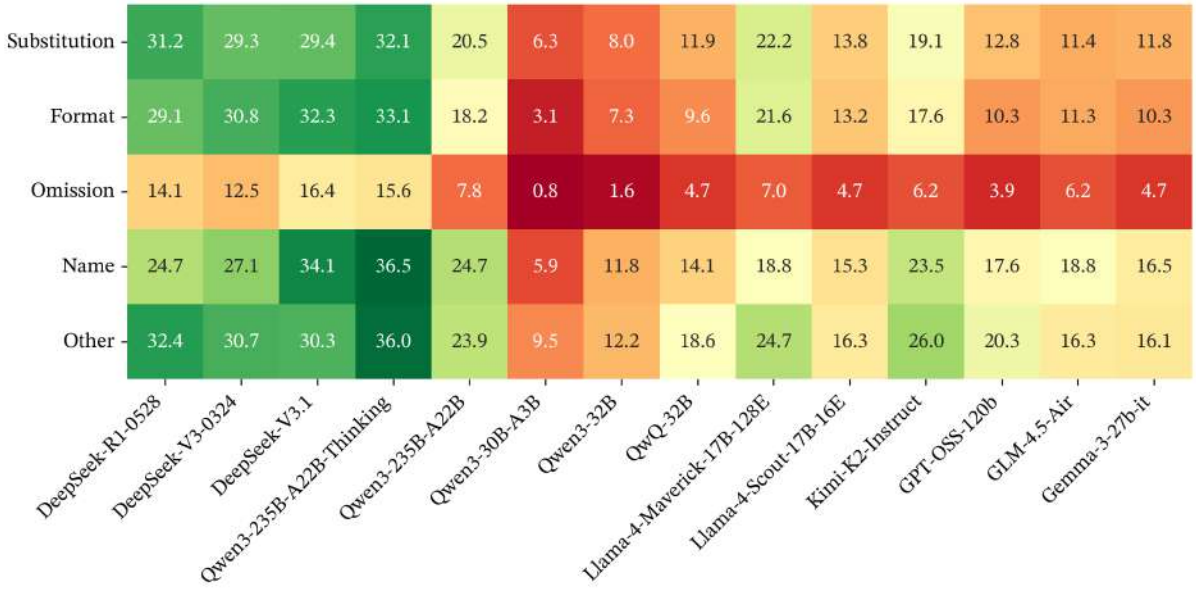

Figure 2. Accuracy of LLMs across question structures

indicating that substitution tasks benefit from lexical flexibility and do not always require deep multi-step reasoning.

Answer format questions (e.g., constrained by number of words or letter sequences) proved more challenging, and the strongest results were achieved by `Qwen3-235B-A22B-Thinking` (33.1%) and `DeepSeek-V3.1` (32.3%). The gap between reasoning and non-reasoning models is visible here: `Qwen3-235B-A22B` scored only 18.2%, suggesting that explicit reasoning helps models interpret and enforce output constraints.

Omission questions (requiring restoration of missing words, letters, or abbreviations) were the most difficult across all the models. Even the strongest models did not exceed 17% accuracy. This category appears particularly challenging because it requires precise contextual recall or cultural knowledge rather than general reasoning ability.

Name questions (requiring specific proper names) posed a considerable challenge. The reasoning-enabled `Qwen3-235B-A22B-Thinking` achieved the best result (36.5%), clearly outperforming both its non-reasoning counterpart `Qwen3-235B-A22B` (24.7%) and large non-reasoning models such as `DeepSeek-V3.1` (34.1%). This contrast highlights that while scale alone can bring solid performance, explicit reasoning traces provide an additional advantage for tasks demanding precise entity recall. Mid- and small-scale models struggled markedly (e.g., `Qwen3-30B-A3B` at 5.9%, `Qwen3-32B` at 11.8%), confirming that both scale and reasoning capabilities are crucial for reliably handling proper-name questions.

Finally, the Other category, which aggregates more heterogeneous question types, confirmed the general advantage of reasoning-enabled large models. `Qwen3-235B-A22B-Thinking` reached 36.0%, followed by `DeepSeek-R1-0528` at 32.4%, whereas most mid-scale models remained below 26%.

4.3. Performance by Question Topic

Figure 3 reports accuracies by thematic category. The analysis highlights both the relative difficulty of different knowledge areas and the advantage of reasoning-enabled models across most topics.

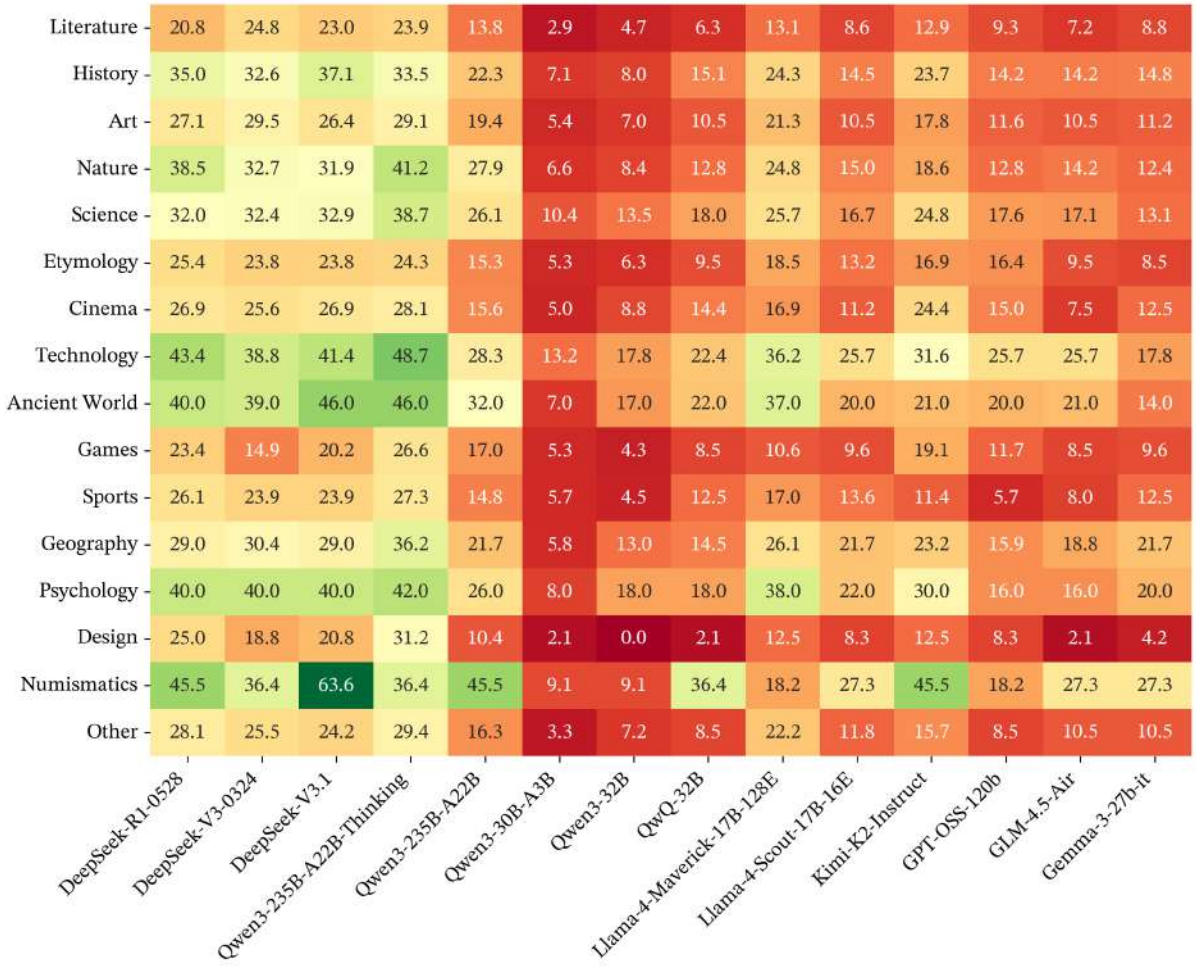

Figure 3. Accuracy of LLMs across question topics

Questions from literature and art proved moderately challenging, with accuracies not exceeding 30%. The best results in these categories were achieved by `DeepSeek-V3-0324` (24.8% in literature; 29.5% in art), followed by `Qwen3-235B-A22B-Thinking` (23.9% / 29.1%), while smaller dense models often remained below 15%. In history and the ancient world, reasoning-capable and large-scale models performed much better: `DeepSeek-V3.1` and `Qwen3-235B-A22B-Thinking` reached 33–46%, and even mid-size models such as `Llama-4-Maverick` achieved moderate accuracy (2437%), suggesting that historical knowledge is relatively well represented in training corpora.

Nature and science questions achieved relatively high accuracies. `Qwen3-235B-A22B-Thinking` scored 41.2% in nature and 38.7% in science, while `DeepSeek-R1` also performed strongly (38.5% and 32.0%). Smaller dense models again fell below 20%, indicating that large MoE architectures with reasoning support are especially effective for factual and explanatory domains. In technology, results were even stronger: `Qwen3-235B-A22B-Thinking` reached 48.7%, the best score across all categories, with `DeepSeek-R1` and `DeepSeek-V3.1`

exceeding 40%. This likely reflects both rich representation of contemporary technological concepts in pretraining data and the reasoning-friendly structure of such questions.

Performance in specialized domains varied widely. In numismatics, `DeepSeek-V3.1` achieved 63.6% accuracy (the single highest category-level result), but this figure is based on only 11 questions, so it should be interpreted with caution. Design questions proved difficult for all models, with a maximum of 31.2% by `Qwen3-235B-A22B-Thinking`. Etymology also challenged most systems, with top results below 26%.

In more culturally grounded or popular categories such as cinema, games, and sports, even the strongest models rarely exceeded 27%. Here, reasoning-enabled models (`Qwen3-235B-A22B-Thinking` and `DeepSeek-R1`) maintained a relative edge but still lagged behind their performance in science and technology. Geography and psychology showed stronger outcomes: `Qwen3-235B-A22B-Thinking` reached 36.2% and 42.0% respectively, while smaller dense models rarely surpassed 20%.

Finally, in the heterogeneous "Other" category, large reasoning-enabled models again outperformed their non-reasoning counterparts (`Qwen3-235B-A22B-Thinking` at 29.4% vs. 16.3% for `Qwen3-235B-A22B`), while mid-scale models typically stayed around 1522%.

Overall, the thematic breakdown confirms that reasoning-enabled large-scale MoE models consistently lead across domains, but their relative advantage varies depending on the knowledge area, with particularly strong gains in technology, ancient world, psychology, and nature.

4.4. Comparison with Human Teams

Since the dataset includes empirical measures of human team performance for each question, we first computed the *average team success rate*, defined as the proportion of correct answers across all teams and all questions. This quantity reflects the expected probability that a randomly selected team would answer a randomly selected question correctly.

A direct comparison of this measure with model accuracy, however, is not fully appropriate. The *average human success rate* aggregates performance across a population of teams and captures the distribution of abilities in the sample, whereas *model accuracy* describes the performance of a single agent answering each question once. Thus, while both metrics are probabilities of success on a random question, they represent different types of averages: one collective, the other individual. For this reason, numerical values cannot be interpreted as strictly equivalent.

Nevertheless, with these limitations in mind, the overall level of performance of the models can be interpreted against the human benchmark. In our case, the mean success rate of human teams was **45.8%**, whereas the best-performing model reached **32.4%** accuracy. This indicates that the model underperforms relative to the average human team, although the comparison should be interpreted with caution.

To assess whether models and humans perceive question difficulty in a comparable way, we examined the relationship between human success rates per question and model outcomes. We computed *Pearsons correlation* (sensitive to linear relationships) and *Spearmans rank correlation* (robust to monotonic but non-linear dependencies). Both coefficients converged to the same result: a weak but statistically significant positive correlation ($r \approx 0.19$, $p < 10^{-22}$). This indicates that questions which are easier for humans tend to be somewhat easier for the model as well, though the strength of the relationship is limited.

We further stratified questions into ten groups according to human success rate (from 0–10% up to 90–100%) and plotted model accuracy in each bin (Fig. 4). The barplot shows a clear

upward trend: model accuracy rises from about 13% on the hardest questions (0–10% human success) to about 62% on the easiest questions (90–100% human success). However, the model consistently lags behind human teams, most strikingly on easy questions, where human success approaches 100% while the model remains far below this ceiling.

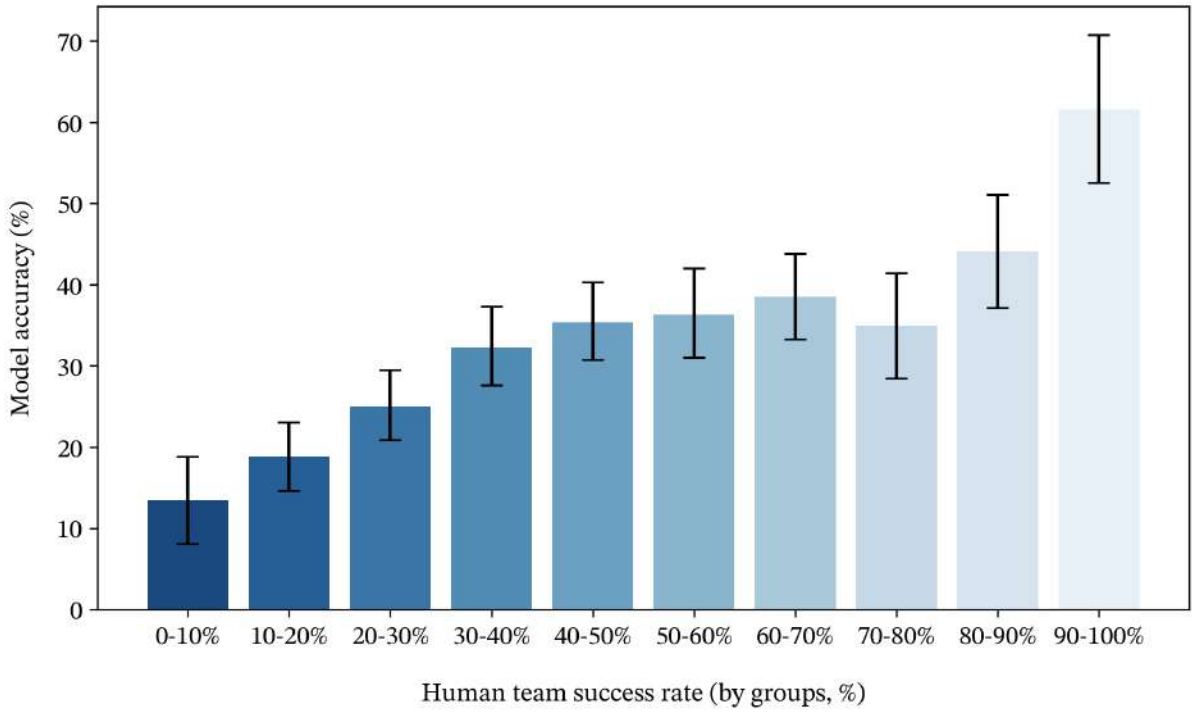

Figure 4. Accuracy of the model on questions grouped by human team success rate (%). Darker colors indicate more difficult questions, lighter colors indicate easier ones. Error bars show 95% bootstrap confidence intervals

In summary, models exhibit systematically lower average performance compared to human teams and only partially align with human judgments of question difficulty. While there is evidence of a shared gradient of difficulty (harder questions for humans also tend to be harder for models), the relatively low correlations and persistent performance gap indicate substantial differences in underlying problem-solving strategies.

4.5. Illustrative Case Studies

To complement the aggregate statistics, we present qualitative examples that illustrate where human teams and models converge or diverge. We organize the cases into four categories (Tab. 6):

- difficult for humans, answered correctly by the model;
- easy for humans, answered incorrectly by the model;
- difficult for both humans and the model;
- easy for both.

These examples suggest that models tend to perform better on questions that hinge on recalling well-known facts or stable cultural associations, while they are less reliable when success depends on reconstructing hidden logic, metaphors, or wordplay.

Table 6. Examples of questions illustrating differences between human and model performance. Answers produced by `Qwen3-235B-A22B-Thinking` model are shown. Correct model answers are highlighted in green, incorrect ones in red

Question	Human success	Model answer	Correct answer
Difficult for humans, answered correctly by the model			
"The achievement a player receives for the first queen checkmate on chess.com has the same name as a 1974 work. Name this work."	0.0%	"Killer Queen"	Killer Queen
"The slogan Let your fingers do the walking belonged to THEM. In a note about THEM, Dennis Yu mentioned several tens of millions of trees. Give THEIR name or say what THEY are."	2.5%	"Yellow Pages"	Yellow Pages
Easy for humans, answered incorrectly by the model			
"Recently, a prisoner serving a life sentence demanded release after HER. Name HER in two words."	96.8%	"Constitutional Court"	clinical death
"On the tram line between Dsseldorf and Duisburg, at certain times of the day, THEY used to attach HIM to the trains. Name HIM."	95.1%	"trailer car"	dining car
Difficult for both humans and the model			
"Above the Goshin grove rise several bare tree trunks, as if struck by lightning, though that could not have happened. Goshin is one of the most famous examples of group of THESE. Name THIS with an indeclinable word."	0.0%	¡No answer¿	bonsai
"Folk etymology traces THIS WORD to a numeral. Researchers, however, believe it goes back to a Proto-Indo-European root meaning to cause pain. Name THIS WORD."	0.5%	"trizna" (Old Russian funeral feast)	duel
Easy for both humans and the model			
"The national motto of Switzerland emerged in the second half of the 19th century. The motto emphasizes that, despite the division of the country into many cantons, if something happens, each canton will support the others. Name this motto."	100.0%	"Unus pro omnibus, omnes pro uno"	One for all and all for one
"The airtight armor made for Henry VIII in the 16th century allowed him to move quite freely. Dan Snow writes that several centuries later, Henry VIIIs armor attracted the interest of specialists working on a commission for which organization?"	99.0%	"NASA"	NASA

Conclusion

In this study, we introduced a new dataset of 2600 *What? Where? When?* questions collected from 2018–2025 and enriched with empirical human success rates. Using structural and thematic clustering, we provided a fine-grained view of question types and knowledge domains, and evaluated 14 recent open-source LLMs with both automatic metrics and an LLM-as-a-Judge approach.

Our results show that the strongest open models, such as `Qwen3-235B-A22B-Thinking` and `DeepSeek-R1`, approach but do not surpass the average human team performance. Large-scale reasoning-enabled architectures consistently outperformed their non-reasoning counterparts, particularly in domains like technology, ancient world, psychology, and nature, while smaller dense models lagged behind across categories. At the same time, omission and wordplay-based questions remained difficult for all systems, underscoring persistent weaknesses in handling associative reasoning and linguistic creativity.

The inclusion of human answer rates allowed us to directly compare model accuracy with human performance. Although correlations between model and human difficulty patterns were statistically significant ($r \approx 0.19$, $p < 10^{-22}$), they were weak, suggesting that humans and models rely on different problem-solving strategies. Qualitative examples further confirmed that models excel more often at fact recall than at reconstructing hidden logic.

Our *What? Where? When?* benchmark is substantially harder than prior Russian quiz datasets. Under the same EM metric, the best result on our data is EM = 0.255 for `Qwen3-235B-A22B-Thinking` (EM = 0.222 for `DeepSeek-V3-0324`), whereas on *CheGeKa* (MERA) `DeepSeek-V3-0324` reaches EM = 0.442; proprietary `Gemini 1.5 Pro` and `Claude 3.7 Sonnet` achieve EM = 0.534 and 0.526, and the human benchmark stands at EM = 0.645 [1]. For metric alignment, we compare EM to EM (MERA reports token-wise F1 and EM), and we avoid contrasting F1 with our judge-based Accuracy. For context, the strongest models judge-based Accuracy on our benchmark is 32.4%.

These findings highlight both the progress of modern open LLMs and their current limitations in intellectual quiz-style reasoning. Future work may expand the dataset, explore interactive multi-agent approaches, and integrate richer evaluation of reasoning traces, bringing automated systems closer to the cognitive style of human quiz players.

References

1. MERA Leaderboard. `https://mera.a-ai.ru/en/text/leaderboard`, accessed: 2025-09-08

2. What? Where? When? `https://en.wikipedia.org/wiki/What%3F_Where%3F_When%3F`, accessed: 2025-09-08

3. Aßenmacher, M., Karrlein, L., Schiele, P., *et al.*: Introducing wwm-german-18k - can LLMs crack the million? (or win at least 500 euros?). In: Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024). pp. 287–296 (2024), `https://aclanthology.org/2024.icnlsp-1.31/`

4. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Advances in Knowledge Discovery and Data Mining. Lecture Notes

in Computer Science, vol. 7819, pp. 160–172. Springer (2013). https://doi.org/10.1007/978-3-642-37456-2_14

5. Chen, A., Stanovsky, G., Singh, S., *et al.*: Evaluating question answering evaluation. In: Proceedings of the 2nd Workshop on Machine Reading for Question Answering. pp. 119–124 (2019). https://doi.org/10.18653/v1/D19-5817

6. Chi, N., Malchev, T., Kong, R., *et al.*: ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models. In: Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. pp. 113–119 (2024), https://aclanthology.org/2024.sigtyp-1.14/

7. Cobbe, K., Kosaraju, V., Bavarian, M., *et al.*: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021), https://arxiv.org/abs/2110.14168

8. Foster, E.J., Friedlander, K.J., Fine, P.A.: Mastermind and expert mind: A qualitative study of elite quizzers. Journal of Expertise 8(1), 38–71 (2025), https://www.journalofexpertise.org/articles/volume8_issue1/JoE_8_1_Foster_etal.html

9. Grootendorst, M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 (2022), https://arxiv.org/abs/2203.05794

10. Hendrycks, D., Burns, C., Basart, S., *et al.*: Measuring massive multitask language understanding. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=d7KBjmI3GmQ

11. Hu, L., Li, Q., Xie, A., *et al.*: GameArena: Evaluating LLM reasoning through live computer games. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=SeQ8l8xo1r

12. Joshi, M., Choi, E., Weld, D.S., *et al.*: TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 1601–1611 (2017). https://doi.org/10.18653/v1/P17-1147

13. Khan, M.A., Yadav, N., Masud, S., *et al.*: QUENCH: Measuring the gap between Indic and non-Indic contextual general reasoning in LLMs. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 4493–4509 (2025), https://aclanthology.org/2025.coling-main.303/

14. Lifar, M., Protsenko, B., Kupriianenko, D., *et al.*: LlaMa meets Cheburashka: impact of cultural background for LLM quiz reasoning. In: Language Gamification - NeurIPS 2024 Workshop (2024), https://openreview.net/forum?id=xCAzTXumhh

15. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013/

16. McInnes, L., Healy, J., Saul, N., *et al.*: Umap: Uniform manifold approximation and projection for dimension reduction. The Journal of Open Source Software 3(29), 861 (2018), https://joss.theoj.org/papers/10.21105/joss.00861

17. Mikhalkova, E., Khlyupin, A.A.: Russian Jeopardy! Data Set for Question-Answering Systems. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 508–514 (2022), https://aclanthology.org/2022.lrec-1.53/

18. Papineni, K., Roukos, S., Ward, T., *et al.*: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002). https://doi.org/10.3115/1073083.1073135

19. Rodriguez, P., Feng, S., Iyyer, M., *et al.*: Quizbowl: The case for incremental question answering (2021), https://arxiv.org/abs/1904.04792

20. Srivastava, A., Rastogi, A., Rao, A., *et al.*: Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=uyTL5Bvosj

21. Taktasheva, E., Shavrina, T., Fenogenova, A., *et al.*: TAPE: Assessing few-shot Russian language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 2472–2497 (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.183

22. Xian, N., Fan, Y., Zhang, R., *et al.*: An empirical study of evaluating long-form question answering. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1141–1151. SIGIR '25 (2025). https://doi.org/10.1145/3726302.3729895

23. Yang, Z., Qi, P., Zhang, S., *et al.*: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 2369–2380 (2018). https://doi.org/10.18653/v1/D18-1259

24. Zhang, Y., Wang, M., Li, X., *et al.*: TurnBench-MS: A benchmark for evaluating multi-turn, multi-step reasoning in large language models. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 19892–19924. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.findings-emnlp.1084

25. Zheng, L., Chiang, W.L., Sheng, Y., *et al.*: Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc. (2023). https://doi.org/10.5555/3666122.3668142