

SQARB: A Benchmark of Sophisticated Question Answering and Reasoning in Belarusian

Vladislav Poritski¹, Oksana Volchek¹, Maksim Aparovich²,
Volha Harytskaya¹, Pavel Smrz²

¹ Independent researcher, Vilnius, Lithuania

² Brno University of Technology, Brno, Czech Republic

Correspondence: belarusianglue@gmail.com

Abstract

As large language models saturate existing benchmarks, language- and culture-specific tasks remain a productive source of problems that are challenging for current models yet solvable by humans. We present SQARB (Sophisticated Question Answering and Reasoning in Belarusian), the first dedicated, natively authored question answering resource for Belarusian, a low-resource East Slavic language. SQARB contains 500 short-answer questions manually written by domain experts, based on doctoral dissertations defended in Belarus between 1994 and 2023, covering five topics: the Belarusian language, literature, folklore, art, and history. The questions are designed to be culture-specific, hard to retrieve online, ever-green, and to require a mix of factual knowledge and reasoning. The dataset is split into an easier development set and a harder test set, both filtered adversarially against locally deployed models. We evaluate near-frontier models without tool use on SQARB and find a large gap between proprietary and open-weight models, with the best-performing model (Gemini 3.1 Pro) achieving 42.3% accuracy on the test set. Models consistently exhibit overconfidence and perform relatively better on questions requiring plausible reasoning, rather than pure factual recall. We release SQARB under gated access with a canary string.¹

1 Introduction

The steady progress of large language models (LLMs) has caused many benchmarks to saturate quickly. Consequently, much of the recent research focuses on identifying tasks that still reveal a significant gap between state-of-the-art model capabilities and human performance. This gap has been observed in non-linguistic domains, such as spatial and visual reasoning in the ARC-AGI benchmarks (Chollet, 2019; Chollet et al., 2025; ARC Prize

Foundation, 2026). Another niche for finding this gap lies in language- and culture-specific tasks.

The pursuit of harder tasks is important not as an end in itself, but rather as an instrument to systematically diagnose what modern multilingual LLMs can and cannot do. Because LLM generation quality heavily relies on the volume of training and fine-tuning data, lesser-resourced languages offer a convenient testing ground for such diagnostics. In particular, while pure linguistic competence in these languages is gradually improving, the intersection of reasoning and culture-specific knowledge remains largely unmapped. By exploring tasks that range from factual retrieval and cultural trivia to plausible reasoning, it is possible to better understand model limitations and failure modes in absence of massive training corpora.

Our focus in the present contribution is Belarusian, a low-resource East Slavic language. Initial evidence regarding LLM capabilities in this language has been provided by the Belarusian-GLUE benchmark (Aparovich et al., 2025) and language-specific tasks in OpenAI Evals (Poritski et al., 2026). However, while these evaluation sets are currently approaching saturation, not much is known so far about the performance of LLMs on Belarusian culture-specific knowledge and reasoning. There is a clear need for a novel, culturally grounded evaluation in this domain.

To address this need, we introduce SQARB² (Sophisticated Question Answering and Reasoning in Belarusian), a challenging closed-book question answering benchmark requiring short, free-form answers. The benchmark contains 500 questions, split into an easier 30-question development set and a harder 470-question test set, and ranging across five domains: Belarusian language, literature, folklore, art, and history. To mitigate the risk of pre-training data contamination, the questions were

¹<https://hf.co/bel-11ms/SQARB>

²Pronounced like the Belarusian word *скапо́* ‘treasure’.

derived from doctoral dissertations (1994–2023) that are legally treated as unpublished manuscripts in Belarus and aren’t easily accessible online. To ensure high difficulty while avoiding live data leaks via proprietary APIs, the questions were adversarially filtered against a suite of locally hosted open-weight models (up to 31B parameters). Using this novel benchmark, we evaluate the performance of near-frontier proprietary and open-weight LLMs against a human baseline, relying entirely on manual grading to guarantee evaluation accuracy.

The rest of the paper is structured as follows. Section 2 places our contribution in the context of related work. Section 3 describes the construction and adversarial filtering of SQARB in detail. Section 4 showcases the results of human and LLM evaluations. Finally, new findings and directions for future work are discussed in the Conclusion.

2 Related Work

From Reading Comprehension to Knowledge in Weights. The modern task of open-domain question answering initially gained traction as reading comprehension (Rajpurkar et al., 2016), where models extracted contiguous answer spans from guaranteed evidence text. Subsequent benchmarks progressively relaxed these assumptions by introducing unanswerable questions (Rajpurkar et al., 2018; Kwiatkowski et al., 2019) and requiring multi-hop reasoning across multiple documents (Welbl et al., 2018; Yang et al., 2018). Datasets like WikiQA (Yang et al., 2015) and MS MARCO (Nguyen et al., 2016) pushed towards realistic search query evaluations, while SimpleQuestions (Bordes et al., 2015) replaced textual evidence with knowledge-graph triples. However, as transformer-based foundational models began compressing massive amounts of world knowledge into their parameters, the paradigm shifted. Open-domain QA is now frequently framed as zero-shot generation without provided evidence, testing the model’s parametric memory rather than its reading comprehension.

Adversarial Difficulty and Evaluation Format. With LLMs quickly saturating multiple-choice benchmarks, both traditional, like MMLU (Hendrycks et al., 2021), and highly specialized, like GPQA (Rein et al., 2023), the post-ChatGPT reality demands harder evaluation sets. Recent benchmarks, such as SimpleQA (Wei et al., 2024), challenge models through adversarial filtering. HLE – Humanity’s Last Exam (Phan et al., 2025)

specifically pre-screened questions, keeping only those that frontier models failed to answer. Inspired by HLE, SQARB employs an adversarial filtering step; however, in order to prevent premature leakage of our dataset via queries to proprietary API providers, we restrict this filtering to locally hosted open-weight models.

In modern free-form QA, grading often relies on LLM-as-a-judge (Kamalloo et al., 2023, 2024; Badshah and Sajjad, 2025) using a three-way scheme: correct / incorrect / not attempted (Wei et al., 2024). While this offers a trade-off between human effort and accuracy at scale, our benchmark targets maximum diagnostic precision over a small set of questions, so instead of using LLM evaluators we grade all answers manually.

Mitigating Data Contamination. In culturally-aware evaluation, datasets based on trivia games, such as TriviaQA (Joshi et al., 2017), Quizbowl (Rodriguez et al., 2019), or the Russian CheGeKa (Mikhalkova and Khlyupin, 2022), are particularly vulnerable to data contamination, as their source material is publicly accessible online. Standard ways to prevent test data from leaking into the pre-training corpora include canary GUIDs (BIG-bench authors, 2023) and gated-access release format (Phan et al., 2025). SQARB introduces an additional barrier: sourcing questions from unpublished doctoral dissertations hosted by the National Library of Belarus. These scanned documents reside behind an authwall requiring in-person registration, and much of their content is absent from public datasets like CommonCrawl. To preserve integrity upon publication, we release SQARB under gated access and with a canary string extending the standard BIG-bench GUID.

Cultural Grounding vs. Reasoning Paradigms. Culture-specific benchmarks consistently reveal sharp performance drops for non-Western regions. This holds across answer formats: multiple-choice (Tanwar et al., 2025; Chiu et al., 2025), short-answer (Myung et al., 2024), and long-form (Arora et al., 2025), but these evaluations largely focus on factual recall. Conversely, benchmarks based on linguistic puzzles (Bean et al., 2024; Sánchez et al., 2024; Chi et al., 2024; Goyal and Dan, 2025) target logical and symbolic reasoning in low-resource settings, while intentionally avoiding external factual knowledge. SQARB combines elements of both paradigms: although all instances require at least some degree of culture-specific knowledge to solve,

many of them can be solved through plausible reasoning, without recalling obscure facts.

Multilingual and Slavic QA. Multilingual QA has evolved from translating English resources (Artetxe et al., 2020; Lewis et al., 2020) towards natively collected data. TyDi QA (Clark et al., 2020) covers 11 typologically diverse languages; exam-based datasets, such as EXAMS (Hardalov et al., 2020), INCLUDE (Romanou et al., 2024), cover a wider range of languages but spread thinly across them. Dedicated single-language benchmarks for Tigrinya (Gaim et al., 2023), Amharic (Taffa et al., 2024), Kazakh (Umbet et al., 2025) etc., as well as multilingual NativQA (Hasan et al., 2025) all share the premise that community-specific knowledge requires community-authored questions.

While Slavic languages have seen coverage – including such datasets as Russian SberQuAD (Efimov et al., 2019) and Polish PoQuAD (Tuora et al., 2023) for reading comprehension, Russian RuBQ (Korablinov and Braslavski, 2020) for knowledge-base QA, Polish open-domain PolQA (Rybak et al., 2024), Ukrainian exam-based ZNO-Eval (Syromiatnikov et al., 2024), and multilingual CUS-QA (Libovický et al., 2025) probing regional knowledge, Belarusian remains relatively neglected:

- A subset of QALD-9-Plus (Perevalov et al., 2022), crowd-translated into Belarusian, provides Wikidata-dependent answers whose correctness can degrade in the future.
- INCLUDE (Romanou et al., 2024) contains several hundred instances in Belarusian with documented quality issues.³
- Global PIQA (Chang et al., 2025) addresses physical-commonsense reasoning in Belarusian, not culturally-grounded knowledge.
- BelarusianGLUE (Aparovich et al., 2025) targets language understanding tasks but does not feature question answering.

To our knowledge, SQARB is the first dedicated, natively authored, manually evaluated, and contamination-resistant QA resource for the Belarusian language.

3 Dataset Description

Desiderata. We designed SQARB around a set of seven desiderata that, taken together, characterize a good question for our benchmark. Each question should be:

- (i) answerable by a human;
- (ii) not easily searchable online;
- (iii) hard for locally hosted LLMs;
- (iv) culture-specific;
- (v) solvable through knowledge and/or reasoning;
- (vi) reasonably short, with a short exact answer;
- (vii) evergreen, i.e. its correct answer should not change in the future.

The rest of this section explains how the questions are sourced and authored, and how the construction procedure satisfies these criteria.

Sources. The questions are derived from doctoral dissertations written in Belarusian between 1994 and 2023. These dissertations are available in digital form to registered readers of the National Library of Belarus, both on-premises and online.⁴ In Belarus, dissertations are treated as unpublished manuscripts and are therefore not easy to access: in particular, they are not available globally through ProQuest or similar databases, although a significant fraction of their content has been published in the authors’ research papers and monographs. This restricted availability reduces the scope for potential contamination. We deliberately ignore dissertations defended before 1994, since very few of them are digitized, as well as those defended after 2023, whose coverage in the library’s database was sparse at the time of our access. As an order-of-magnitude estimate, about 20K doctoral dissertations (including both PhD and DrHab) have been defended in Belarus since it gained independence in 1991; of these, about 1K were written in Belarusian and the rest in Russian. Our work therefore draws on approximately a 5% subset of the total research output of Belarusian scholars and scientists.

Question Authoring. Questions were contributed by four experts (all of whom are authors of this paper), with approximately 90% questions written by two experts with a background in philology. The primary topics are the Belarusian language, literature, folklore, art, and history. A question may combine a primary and a secondary topic, e.g. language and literature, or art and history.

Because the dissertations are ordered by library catalogue codes, their sampling is not truly random. To avoid over- or under-representing any period, we subsampled the dissertations by year of defence. To write a question, an expert skimmed the dissertation text until finding one or more statements they found

³<https://hf.co/datasets/CohereLabs/include-base-44/discussions/5>

⁴<https://elib.nlb.by>

suitable for formulating a question.

Several categories of dissertations were skipped. Although there have been attempts to write dissertations in Belarusian outside the humanities – in mathematics, engineering, medicine, and other fields – such dissertations are very rare and clearly not culture-specific, so we produced no questions from them. We eventually also skipped most dissertations on pedagogy, as they proved particularly hard to extract statements worth asking about. Further reasons for failing to produce a question include a dissertation being thematically unrelated to Belarus, being fully published and thus easy to find online, or simply yielding facts that are hard to phrase as a question.

We normally followed the source’s spelling of proper names, and acknowledge any discrepancies in the comments accompanying a question. Occasionally the answer is not stated verbatim in the source, or the question relies on additional facts verified by us; such cases are likewise acknowledged in the comments. The factual correctness of the source’s claims is taken for granted. We attempted to strike a reasonable balance between the uniqueness of the answer and the difficulty of the question (more detail makes the answer more unique but the question easier). Across the dataset, each distinct answer occurs only once.

Adversarial Filtering. Filtering was deliberately restricted to local models, in order to reduce the attack surface for potential contamination. We used role prompts in both English – mimicking HLE – and Belarusian (see Appendix A), and all model answers were graded manually.

Filtering proceeded in two stages. In the first stage, we ran a set of thirteen local models released in 2024 and 2025, each under 15B parameters.⁵ Every model was queried in two runs, one with the English prompt and one with the Belarusian prompt, at zero temperature and with a budget of 4096 generated tokens. A question was dropped if multiple models answered it correctly or showed significant progress toward the correct answer – specifically, more than one correct answer and/or more than three partial answers across the thirteen models in either run. From the questions that survived this

⁵EuroLLM-9B-Instruct, gemma-2-9b-it, gemma-3-12b-it, Llama-3.1-8B-Instruct, Llama-PLLuM-8B-instruct, Ministral-8B-Instruct-2410, Mistral-Nemo-Instruct-2407, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, salamandra-7b-instruct, Vikhr-Llama3.1-8B-Instruct-R-21-09-24, Vikhr-Nemo-12B-Instruct-R-21-09-24, Vikhr-YandexGPT-5-Lite-8B-it.

stage, we randomly sampled a topic-balanced set of 30 questions and designated it as the development set, used for human baseline evaluation and for a thorough analysis of model performance.

In the second stage, we used two strong local models released in 2026, Qwen 3.5 27B and Gemma 4 31B, running both with Ollama⁶ in 4-bit quantization without tool use, generating up to 16384 tokens. Each model was queried three times with the English prompt and three times with the Belarusian prompt, using the default nonzero temperature, i.e. non-deterministic decoding. A question was dropped if at least one correct answer was produced in these runs. This stage was not applied to the development set. All remaining questions formed the 470-question test set. To reach a round number, we had to compensate for the filtered-out instances by writing a few additional questions, sufficiently hard to survive both stages.

The dataset construction process is summarized in Figure 1 and can be described as a funnel:

- 953 dissertations were available to us.
- We attempted to write questions based on 711 of them, and successfully produced at least one question from 639 dissertations.
- After filtering, 500 questions remain – one per dissertation.

The eventual distribution of sources by year of defence is shown in Figure 2.

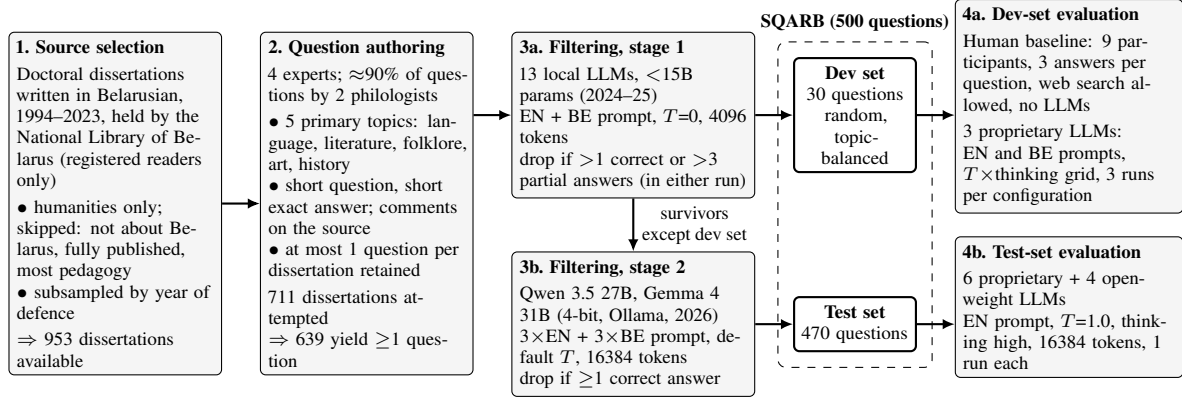
Question Types and Topics. The question types are diverse, spanning exact factual knowledge, plausible reasoning, needle-in-a-haystack retrieval, etc. Realistic examples are provided in Table 6 in Appendix C. We estimate that in $\approx 32\%$ of all questions (159/500), reasoning may help arrive at the correct answer without knowing the fact.

The distribution of questions by primary topic is shown in Table 1. History of Belarus is the most frequent primary topic, followed by Belarusian literature and language.

Topic	No. questions	
	dev.	test
Art	6	46
Folklore	4	34
History	10	162
Language	6	100
Literature	4	128
Total:	30	470

Table 1: Question distribution by primary topic.

⁶<https://github.com/ollama/ollama>



Every LLM and human answer is graded manually (correct/partial/incorrect); all LLM runs are without tool use.

Figure 1: Overview of the SQARB pipeline: source selection, question authoring, two-stage adversarial filtering against locally hosted open-weight LLMs, and evaluation. The development set is sampled after the first filtering stage and bypasses the second one; a few additional questions that survived both stages were written to bring the total to 500.

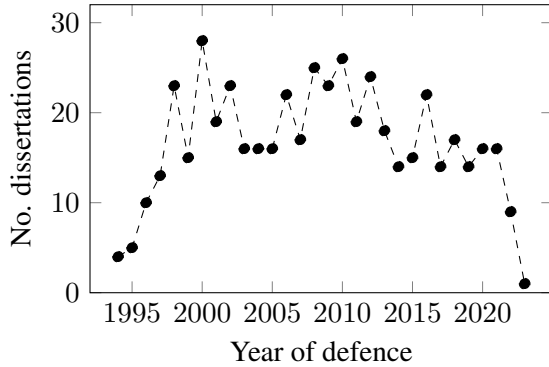


Figure 2: Distribution of sources by year.

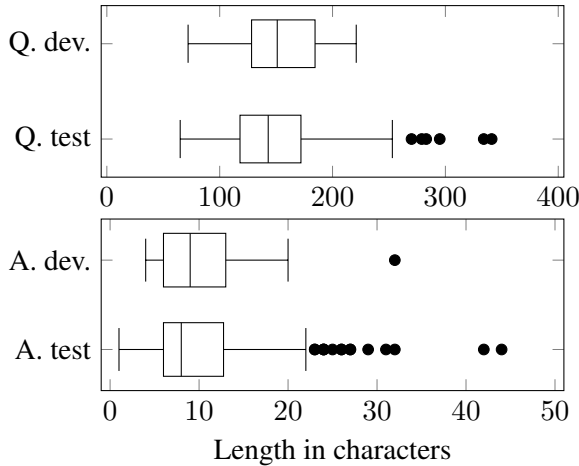


Figure 3: Distribution of questions and answers by length in characters in the development and test sets.

Length Statistics. The distributions of question and answer lengths are shown in Figure 3. The questions are moderately short, typically ranging between 100–200 characters, and the gold answers

are succinct, with a median length of 8–9 characters, which amounts to 1–2 words. A two-sample Kolmogorov–Smirnov test reveals no statistically significant difference between the length distributions of the development and test sets for either questions ($D = 0.151$, $p = 0.541$) or answers ($D = 0.098$, $p = 0.950$).

Answer Types. The distribution of questions by entity type in the answer is shown in Table 2. Persons, most of them real and some fictional (e.g., book characters), are the most frequent type of named entities in the answers. Locations and titles (including the titles of literary and artistic works, periodical names, other organization or project names) are also common, along with linguistic units. Purely numerical answers, such as years and quantities, make up a small minority. Almost 20% of the answers that don’t neatly fit into any category are grouped as “miscellanea”.

Type	No. questions dev.	test
Person	7	128
Location	5	60
Title	3	73
Word/phrase	7	93
Year	1	11
Number	0	15
Misc.	7	90
Total:	30	470

Table 2: Question distribution by answer type.

Meeting the Desiderata. We now revisit the desiderata and describe how each is satisfied.

- (i) *Answerability* holds by construction: each question can be answered by the author of the respective dissertation.
- (ii) For *searchability*, we verified that the answers are hard – though not necessarily impossible – to find online with a search engine.⁷
- (iii) *Difficulty for local LLMs* is enforced by the filtering procedure described above.
- (iv) *Cultural specificity* is reflected in the distribution of primary topics, reported above.
- (v) The required *mix of knowledge and reasoning* follows from the diversity of question types (see Table 6 in Appendix C).
- (vi) *Brevity* of both questions and answers is documented in the length statistics above.
- (vii) Finally, the questions are *evergreen* modulo possible future changes in the territory of Belarus and in the standard variety of the Belarusian language, which could affect $\approx 6\%$ of the questions (29/500).

Biases. The dataset carries several biases that should be kept in mind when interpreting results.

- (1) Not absolutely all of the dissertations written in Belarusian by Belarusian researchers between 1994 and 2023 were available to us.
- (2) The dataset reflects the research community’s focus on particular topics, including, after 2020, the increased risk associated with researching certain sensitive topics.
- (3) Our own selection favors the humanities.
- (4) Subsampling by year and not-quite-random ordering of the dissertations introduce further skew.
- (5) Finally, each expert applied an individual, intuitive notion of what constitutes a question-worthy fact.

4 Experiments

4.1 Human Baseline

While 100% correct human answers is a theoretical upper bound, it is also important to estimate how average subjects perform on the development set of SQARB. To that end, through a combination of community outreach and personal contacts we recruited nine unpaid participants, all but one 20–39 years old, with at least B.A. degree; 7/9 self-reported to be fluent speakers of Belarusian.

We partitioned the development set three times: each partition splits the 30 questions into three folds of 10. Every fold is roughly balanced by topic, and folds belonging to different partitions overlap in at most 5 questions. Each participant

was prompted to answer a single fold of questions through an online interface, within a one-hour time limit. Participants were allowed to use any tools except LLMs. Web search was permitted, and we did not specifically ban Google search, since the experiment took place in December 2025, when its AI mode was still an early-stage product. The partition design yields 3 answers per question.

Every answer was manually labeled as correct, partial, or incorrect (including not attempted). Individual scores range between 1/10 and 7/10 correct answers. At least one correct answer was collected for 22/30 questions, and a further 2 questions received at least one partial answer; 6 questions were left without any correct or partial answer. The full table of labels is provided in Appendix B.

Because each question was answered by multiple participants, it is possible to quantify their agreement. Not counting empty answers, average pairwise percent agreement between the answers is 0.22, and Krippendorff’s alpha (Hayes and Krippendorff, 2007) calculated on the labels is 0.154. These low scores indicate large discrepancies between participants, suggesting that a team of humans would perform stronger than each of them individually.

4.2 Development Set Evaluation

Later, the development set was processed with three near-frontier proprietary LLMs:

- Gemini 3.5 Flash via Gemini Enterprise Agent Platform in batch inference mode;
- Claude Sonnet 4.6 via the Converse API of AWS Bedrock;
- GPT-5.4 via the OpenAI Platform in batch inference mode.

We submitted the questions with prompts in English and Belarusian (see Appendix A) and explored the following parameter space, as far as each model’s settings allow:

- temperature (T): 0.0 / 0.5 / 1.0;
- thinking: disabled / low / medium / high.

Gemini Flash doesn’t let the user disable thinking, thus 18 configurations remain. Claude Sonnet requires $T=1.0$ in thinking mode, leaving 12 configurations. In GPT-5.4, temperature is deprecated (always $T=1.0$), so there are 8 configurations to consider. To estimate how stable the results are, 3 runs per configuration were performed.

No model was given access to tools, in particular web search, whereas the human participants were. This asymmetry is deliberate. The human base-

⁷Typically Startpage: <https://www.startpage.com>

line asks whether a non-specialist with ordinary resources can answer the questions at all; since the questions are drawn from specialist dissertations, this cannot be tested without search. The model evaluation asks how much of the required language- and culture-specific knowledge is already stored in the weights, and therefore uses the same closed-book regime as the adversarial filtering, so that difficulty is defined consistently for the filter and for the evaluated models. The resulting scores are a lower bound on LLM performance and should not be read as a direct comparison with the human baseline.

Counts of correct answers, averaged over runs and prompt languages, are presented in Table 3.

Gemini Flash				
$T \setminus \text{Think.}$	no	low	med.	high
0.0	–	16.0	17.3	17.3
0.5	–	15.7	16.0	17.5
1.0	–	16.5	18.3	18.3

Claude Sonnet				
$T \setminus \text{Think.}$	no	low	med.	high
0.0	8.2	–	–	–
0.5	7.5	–	–	–
1.0	9.0	8.2	7.8	8.3

GPT-5.4				
$T \setminus \text{Think.}$	no	low	med.	high
1.0	7.8	12.8	13.5	14.0

Table 3: Development set evaluation results: average numbers of correct answers, out of 30 questions.

$T=1.0$, thinking=high is usually the best configuration, except Claude Sonnet in which it is the second best. With disabled thinking, the performance of GPT-5.4 degrades, while the other two models are not particularly sensitive to the choice of thinking level. The use of English / Belarusian language in the prompt impacts the average scores only slightly (Gemini Flash: 16.9 / 17.1 correct answers; Claude Sonnet: 8.2 / 8.2; GPT-5.4: 12.0 / 12.1). We observed high inter-run variance of scores: in their respective best configurations, the three models yield 15–22, 8–10, and 10–17 correct answers to the development set questions.

4.3 Test Set Evaluation

Six proprietary models and four open-weight models, made publicly available no later than June 1st, 2026, were evaluated on the 470-question test set:

- Gemini 3.5 Flash and Gemini 3.1 Pro;
- Claude Sonnet 4.6 and Claude Opus 4.8;
- GPT-5.4 and GPT-5.5;
- DeepSeek V4 Pro, MiMo-V2.5-Pro, MiniMax

M3 (all in 8-bit quantization) and Kimi 2.6 (in 4-bit quantization) – via OpenRouter.⁸

Based on HLE’s findings (Phan et al., 2026, p. 1143), token quota per question was set to 16384 tokens, including thinking. In all models, temperature was set to 1.0. In proprietary models, thinking level was set to high. Guided by the development set evaluation, we also tried disabling thinking in both Claudes – this resulted in slightly (Sonnet) or significantly (Opus) degraded accuracy scores, which are not reported here. In open-weight models, we kept the default thinking level unchanged.

Same as in the development set, here we evaluated all models without tool use. We used the English version of the prompt, as the choice of the instruction language was found to impact the results only slightly, while in weaker models English instructions may provide more stable behavior.

Due to budget constraints, only a single run has been performed for each model. API errors and requests that exceeded the token quota were retried up to 3 times. The outputs have then been manually labeled as correct, partial, or incorrect (including not attempted). The results are shown in Table 4.

Gemini 3.1 Pro is the top competitor with 42.3% correct answers, followed by GPT-5.5 with 36.2%. The results of proprietary model families are hinting at their specialization: Gemini Pro and Gemini Flash are stronger than GPT-5.5 and GPT-5.4 respectively, which are in turn stronger than their contemporary versions of Claude Opus and Claude Sonnet. Among the open-weight models, only DeepSeek reaches the Claude Sonnet level; the other three models, advertised to be best suited for agentic workflows, don’t offer comparable quality. In particular, Kimi frequently exceeds the token quota before emitting the answer.

Questions that involve elements of reasoning, while being less demanding on factual knowledge, are easier for all models than purely factual questions, as the accuracy scores in Table 4 show. The accuracy gain on reasoning vs. factual questions is somewhat higher in weaker models.

On average, wrong answers are preceded by longer chain of thought. This is especially true for stronger models, such as Gemini Pro or GPT-5.5 that spend over 2x more tokens on wrong answers than correct ones. Claude Opus is an outlier, both in terms of how sparingly it generates output tokens and how small is the difference in length.

⁸<https://openrouter.ai>

Model	Released	Answers				Accuracy, %		Spent, M tok. In / Out	Avg. out tok.		
		+	±	−	∅	Overall	Fact. / Reas.		+	±	−
Gemini 3.5 Flash	2026.05.19	162	34	274	0	34.5	28.8 / 46.9	0.066 / 1.774	2319	3249	4702
Gemini 3.1 Pro	2026.02.19	199	35	236	0	42.3	33.4 / 61.9	0.066 / 2.514	3122	5446	7214
Claude Sonnet 4.6	2026.02.17	85	32	347	6	18.1	14.9 / 25.2	0.080 / 1.430	1959	2148	3160
Claude Opus 4.8	2026.05.28	140	31	298	1	29.8	22.6 / 45.6	0.090 / 0.347	685	710	713
GPT-5.4	2026.03.05	138	31	300	1	29.4	21.7 / 46.3	0.065 / 1.846	2238	4144	4640
GPT-5.5	2026.04.23	170	19	277	4	36.2	28.2 / 53.7	0.065 / 2.517	3117	4060	6657
DeepSeek V4 Pro	2026.04.24	95	30	345	0	20.2	15.8 / 29.9	0.071 / 1.365	1877	2599	3214
Kimi 2.6	2026.04.20	31	21	369	49	6.6	4.6 / 10.9	0.074 / 4.711	8849	9304	9318
MiMo-V2.5-Pro	2026.04.27	33	22	405	10	7.0	5.0 / 11.6	0.081 / 1.612	2912	2085	3225
MiniMax M3	2026.06.01	39	32	397	2	8.3	5.0 / 15.6	0.148 / 2.481	5583	4970	5217

Table 4: Test set evaluation results. The answers are labeled correct (+), partial (±), or incorrect (−); some answers are missing (∅), due to exceeded token quota. Accuracy score is the ratio of correct answer count to the test set size. Accuracies on purely factual questions and questions involving reasoning are reported separately. Also reported are the total counts of consumed tokens (in million) and average output token counts stratified by the answer label.

Gemini Flash	45.7	55.9	36.4	31.0	25.0
Gemini Pro	52.2	55.9	45.7	43.0	30.5
Claude Sonnet	19.6	26.5	21.6	14.0	14.1
Claude Opus	19.6	52.9	30.9	33.0	23.4
GPT-5.4	32.6	44.1	31.5	28.0	22.7
GPT-5.5	28.3	55.9	37.0	36.0	32.8
DeepSeek	19.6	38.2	20.4	22.0	14.1
Kimi	10.9	14.7	5.6	6.0	4.7
MiMo	15.2	11.8	5.6	9.0	3.1
MiniMax	10.9	11.8	7.4	11.0	5.5
	Art	Folk.	Hist.	Lang.	Lit.

Figure 4: Test set accuracy percentages by primary topic of the question (Art, Folklore, History, Language, Literature). Darker cells denote higher accuracy.

Gemini Flash	42.2	38.3	26.0	31.2	18.2	26.7	34.4
Gemini Pro	46.9	41.7	37.0	40.9	45.5	20.0	45.6
Claude Sonnet	21.1	20.0	13.7	14.0	27.3	13.3	20.0
Claude Opus	33.6	26.7	20.6	31.2	27.3	26.7	33.3
GPT-5.4	30.5	33.3	21.9	30.1	36.4	26.7	30.0
GPT-5.5	39.8	36.7	32.9	34.4	27.3	26.7	37.8
DeepSeek	21.1	18.3	12.3	21.5	18.2	13.3	26.7
Kimi	8.6	6.7	1.4	5.4	0.0	6.7	10.0
MiMo	6.3	5.0	5.5	10.8	0.0	6.7	7.8
MiniMax	6.3	6.7	8.2	11.8	0.0	6.7	10.0
	Pers.	Loc.	Title	Word/phr.	Year	Number	Misc.

Figure 5: Test set accuracy percentages by entity type in the answer (Person, Location, Title, Word/phrase, Year, Number, Miscellaneous). Darker cells denote higher accuracy.

Figures 4 and 5 show each model’s accuracy on the subsets of test set questions by primary topic and answer type. None of the models achieves 60% accuracy on any of the topics. Folklore is consistently the easiest topic, while literature is often the hardest one. With respect to the answer type, there are considerable differences in model performance. Persons and miscellanea tend to be systematically easier than titles and numbers.

Because our prompt is based on HLE’s prompt (Phan et al., 2025, p. 24) that asks the model to return a confidence score together with each answer, we followed the approach of Wei et al. (2024) to construct calibration curves, shown in Figure 6.

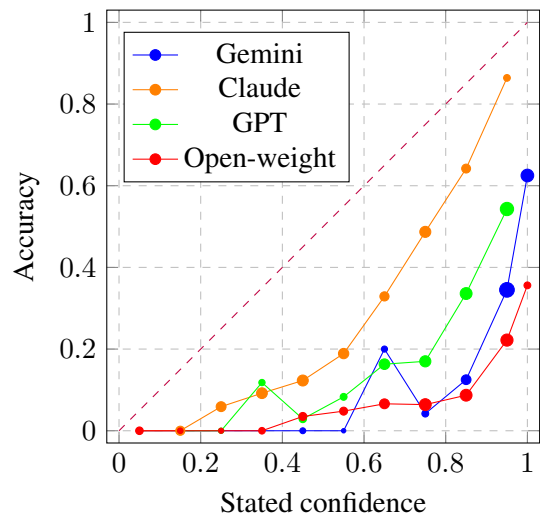


Figure 6: Test set calibration of the evaluated models based on their stated confidences. Models are grouped by family (Gemini, Claude, GPT) or type (open-weight). Marker sizes are proportional to answer counts in each of the confidence bins: 0–9%, ..., 90–99%, 100%.

Of the models we evaluated, the Claude fam-

ily is clearly the most well-calibrated, though still overconfident. Other proprietary and open-weight models are even more overconfident: in particular, only 2.7% answers of Gemini models have confidence less than 50%, while 82.5% answers have confidence 90–100%.

As one further note, we observed that longer questions tend to elicit more accurate answers, but this effect might as well be caused by confounding variables, such as uneven distribution of question lengths across topics or answer types.

5 Conclusion

We presented SQARB, a natively-authored question answering benchmark for Belarusian, and demonstrated that, as of mid-2026, it is challenging for near-frontier proprietary LLMs in no-tools configuration, although their performance is now comparable to what a team of three fluent speakers with access to web search could deliver under a time limit. There is a gap in performance between proprietary and open-weight models. In the domain of Belarusian humanities, all evaluated models exhibit imperfect calibration, limited factual recall but relatively stronger reasoning capacity.

We publish SQARB⁹ in “gated access” format, with a canary string that is a superset of the BIG-bench canary string, so that contamination could be reliably detected. Together with the dataset, we release the full set of model answers and their manually assigned labels, to facilitate LLM-as-a-judge automatic grading.

A recent study (Akhtar et al., 2026) found that expert curation and adversarial design may improve benchmark robustness to saturation. Both of these ingredients are present in SQARB. Even though frontier models can be projected to saturate it rather quickly, especially in agentic settings with access to web search and browsing, we anticipate that for smaller open-weight multilingual models, SQARB could remain a useful benchmark for a longer time, assuming that the underlying corpus of dissertations isn’t made available for model training.

Limitations

Our study has several limitations. First, the frontier of LLM performance is a moving target, so we weren’t able to evaluate the currently most recent generation of models (including Claude Fable 5.1, GPT-6 Astra, Kimi K3) due to budget and

availability constraints. Second, as the strongest proprietary models obscure their reasoning traces (providing only summaries or non-human-readable signatures), we weren’t able to perform a cross-model analysis of reasoning patterns. Third, evaluating models without tool use provides only a lower-bound estimate of LLM performance; allowing tool use, in particular web search or multilingual retrieval-augmented generation, would yield a complementary upper-bound estimate. The latter is of special interest, since Ranaldi et al. (2026) report that cross-lingual retrieval with translation of the retrieved documents into a pivot language benefits low-resource languages the most. Finally, meta-prompting (Ye et al., 2024), i.e. generating the prompts with the models themselves and analyzing the impact of the prompt on model performance, is also left to future work.

Acknowledgments

The authors are grateful to the nine participants of the human baseline evaluation. Computational resources for running the locally hosted models during the adversarial filtering stage were provided by Brno University of Technology.

References

- Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, and 18 others. 2026. [When AI benchmarks plateau: A systematic study of benchmark saturation](#). *Preprint*, arXiv:2602.16763.
- Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. [BelarusianGLUE: Towards a natural language understanding benchmark for Belarusian](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527, Vienna, Austria. Association for Computational Linguistics.
- ARC Prize Foundation. 2026. [ARC-AGI-3: A new challenge for frontier agentic intelligence](#). *Preprint*, arXiv:2603.24621.
- 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:*

⁹<https://hf.co/bel-11ms/SQARB>

- Long Papers*), pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. [On the cross-lingual transferability of mono-lingual representations](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.
- Sher Badshah and Hassan Sajjad. 2025. [Reference-guided verdict: LLMs-as-judges in automatic evaluation of free-form QA](#). In *Proceedings of the 9th Widening NLP Workshop*, pages 251–267, Suzhou, China. Association for Computational Linguistics.
- Andrew Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A., Ryan Chi, Scott Hale, and Hannah Rose Kirk. 2024. [LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 26224–26237. Curran Associates, Inc.
- BIG-bench authors. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Transactions on Machine Learning Research*.
- Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. [Large-scale simple question answering with memory networks](#). *Preprint*, arXiv:1506.02075.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanahun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, and 361 others. 2025. [Global PIQA: Evaluating commonsense reasoning across 100+ languages and cultures](#). *Preprint*, arXiv:2510.24081.
- Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. [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*, pages 113–119, St. Julian’s, Malta. 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.
- Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. [ARC-AGI-2: A new challenge for frontier ai reasoning systems](#). *Preprint*, arXiv:2505.11831.
- François Chollet. 2019. [On the measure of intelligence](#). *Preprint*, arXiv:1911.01547.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages](#). *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2019. [SberQuAD – Russian reading comprehension dataset: Description and analysis](#). *Preprint*, arXiv:1912.09723.
- Fitsum Gaim, Wonsuk Yang, Hanchchol Park, and Jong Park. 2023. [Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada. Association for Computational Linguistics.
- Satyam Goyal and Soham Dan. 2025. [IOL-BENCH: Benchmarking LLMs on linguistic reasoning](#). *Preprint*, arXiv:2501.04249.
- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. [EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.
- Md. Arif Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. [NativQA: Multilingual culturally-aligned natural query for LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.
- Andrew F. Hayes and Klaus Krippendorff. 2007. [Answering the call for a standard reliability measure for coding data](#). *Communication Methods and Measures*, 1(1):77–89.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.
- Ehsan Kamalloo, Shivani Upadhyay, and Jimmy Lin. 2024. [Towards robust QA evaluation via open LLMs](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, pages 2811–2816, New York, NY, USA. Association for Computing Machinery.
- Vladislav Korablinov and Pavel Braslavski. 2020. [RuBQ: A Russian Dataset for Question Answering over Wikidata](#), pages 97–110. Springer International Publishing.
- Tom Kwiakowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Patrick Lewis, Barlas Oguz, Rutu Rinott, Sebastian Riedel, and Holger Schwenk. 2020. [MLQA: Evaluating cross-lingual extractive question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.
- Jindřich Libovický, Jindřich Helcl, Andrei-Alexandru Manea, and Gianluca Vico. 2025. [CUS-QA: Local-knowledge-oriented open-ended question answering dataset](#). *Preprint*, arXiv:2507.22752.
- Elena Mikhalkova and Alexander A. Khlyupin. 2022. [Russian Jeopardy! data set for question-answering systems](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 508–514, Marseille, France. European Language Resources Association.
- 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, and 3 others. 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.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. [MS MARCO: A human generated machine reading comprehension dataset](#). *Preprint*, arXiv:1611.09268.
- Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022. [QALD-9-plus: A multilingual dataset for question answering over DBpedia and Wikidata translated by native speakers](#). In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 229–234.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1136 others. 2025. [Humanity’s last exam](#). *Preprint*, arXiv:2501.14249.
- Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 1081 others. 2026. [A benchmark of expert-level academic questions to assess AI capabilities](#). *Nature*, 649(8099):1139–1146.
- Vladislav Poritski, Oksana Volchek, Maksim Aparovich, Volha Harytskaya, and Pavel Smrz. 2026. [Tracking the evolution of LLM capabilities for Belarusian with OpenAI Evals](#). In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 378–387, Rabat, Morocco. Association for Computational Linguistics.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2026. [Multilingual retrieval-augmented generation for knowledge-intensive question answering task](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 697–716, Rabat, Morocco. Association for Computational Linguistics.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Di-rani, Julian Michael, and Samuel R. Bowman. 2023.

- GPQA: A graduate-level google-proof Q&A benchmark. *Preprint*, arXiv:2311.12022.
- Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2019. [Quizowl: The case for incremental question answering](#). *Preprint*, arXiv:1904.04792.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. [INCLUDE: Evaluating multilingual language understanding with regional knowledge](#). *Preprint*, arXiv:2411.19799.
- Piotr Rybak, Piotr Przybyła, and Maciej Ogrodniczuk. 2024. [PolQA: Polish question answering dataset](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12846–12855, Torino, Italia. ELRA and ICCL.
- Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. [Linguini: A benchmark for language-agnostic linguistic reasoning](#). *Preprint*, arXiv:2409.12126.
- Mykyta V. Syromiatnikov, Victoria M. Ruvinskaya, and Anastasiya S. Troynina. 2024. [ZNO-Eval: Benchmarking reasoning capabilities of large language models in Ukrainian](#). *Informatics. Culture. Technology*, 1(1):186–191.
- Tilahun Abedissa Taffa, Ricardo Usbeck, and Yaregal Assabie. 2024. [Low resource question answering: An Amharic benchmarking dataset](#). In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 124–132, Torino, Italia. ELRA and ICCL.
- Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. [Do you know about my nation? Investigating multilingual language models’ cultural literacy through factual knowledge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14956–14979, Suzhou, China. Association for Computational Linguistics.
- Ryszard Tuora, Aleksandra Zwierzchowska, Natalia Zawadzka-Paluckta, Cezary Klamra, and Łukasz Kobyliński. 2023. [PoQuAD - the Polish question answering dataset - description and analysis](#). In *Proceedings of the 12th Knowledge Capture Conference 2023, K-CAP ’23*, pages 105–113, New York, NY, USA. Association for Computing Machinery.
- Sanzhar Umbet, Sanzhar Murzakhmetov, Beksultan Sagyndyk, Kirill Yakunin, Timur Akishev, and Pavel Zubitski. 2025. [KazBench-KK: A cultural-knowledge benchmark for Kazakh](#). In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 38–57, Vienna, Austria. Association for Computational Linguistics.
- Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. [Measuring short-form factuality in large language models](#). *Preprint*, arXiv:2411.04368.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [Constructing datasets for multi-hop reading comprehension across documents](#). *Trans. of the Association for Computational Linguistics*, 6:287–302.
- Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. [WikiQA: A challenge dataset for open-domain question answering](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. [Prompt engineering a prompt engineer](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts

These prompts, based on the HLE prompt with an addition of an expert persona, have been used for question filtering (Section 3) and development set evaluation (Section 4.2). The English version of the prompt has been used for test set evaluation (Section 4.3).

A.1 English

You're an expert in Belarusian history, culture, language, and literature. You will be asked a question on one of these topics. Your response should be in the following format:

Explanation: {your explanation for your final answer}

Exact Answer: {your succinct, final answer}

Confidence: {your confidence score between 0% and 100% for your answer}

A.2 Belarusian

Вы спецыяліст па гісторыі і культуры Беларусі, беларускай мове і літаратуры. Вам будзе задана пытанне на адну з гэтых тэм. Ваш адказ павінен мець наступны фармат:

Тлумачэнне: {вашы развагі і тлумачэнні да канчатковага адказу}

Дакладны адказ: {ваш канчатковы адказ без лішняй інфармацыі}

Упэўненасць: {паказчык вашай упэўненасці ў адказе, ад 0% да 100%}

B Human Baseline Evaluation Details

Q. id	i1	i2	i3	i4	i5	i6	i7	i8	i9	Best
A1		−			+		+			+
A2			−	−				−		−
A3	−					−			+	+
A4		−			+				−	+
A5			+	±				−		+
A6	−					+	−			+
F1		±				+	+			+
F2	+			−				−		+
F3		+			−			+		+
F4			+		−				−	+
H1			−	−					−	−
H2	−					+			−	+
H3		−		−					−	−
H4			−		+			−		+
H5			+			−			−	+
H6	−					−		−		−
H7		−		+			+			+
H8		−				−	−			−
H9	+				+			+		+
H10			+		+		−			+
La1	±			±				±		±
La2		±			±			−		±
La3			+			−	+			+
La4	−				−		+			+
La5		−		−					−	−
La6		−		−			+			+
Li1	+					+			+	+
Li2			+			+	−			+
Li3	+				+			−		+
Li4			+	−					+	+
No. +	4	1	7	1	6	5	6	2	3	22
No. ±	1	2	0	2	1	0	0	1	0	2

Table 5: Detailed results of the human baseline evaluation. Each of the nine informants (i1 to i9) answered 10 development set questions. Question topics are encoded in their identifiers: art (A), folklore (F), history (H), language (La), literature (Li).

C Sample Questions

Type	Topics	Question	Answer
Factual knowledge	Art, history	<i>Назавіце горад на тэрыторыі Беларусі, у якім была ўзведзена першая ў Рэчы Паспалітай барочная пабудова.</i> ‘Name the city in Belarus where the first Baroque building in the Polish-Lithuanian Commonwealth was built.’	<i>Нясвіж</i> ‘Nyasvizh’
Culture-specific trivia	Folklore	<i>Адгадайце беларускую загадку: «Сам пан у зелён, яго дзеці ў чырвані».</i> ‘Guess the Belarusian riddle: “The lord himself is in green, his children are in red”.’	<i>бурак</i> ‘beetroot’
Plausible reasoning	History	<i>У які дзень 2008 года ў Магілёўскім аддзеле ЗАГС бралі шлюб ажно 70 пар? Дайце адказ у фармаце YYYY.MM.DD.</i> ‘On what day in 2008 as many as 70 couples married at the Mahilyow registry office? Give your answer in the format YYYY.MM.DD.’	2008.08.08
Needle in a haystack	Language, literature	<i>Які назойнік-люксонім у творах Якуба Коласа спалучаецца з прыметнікамі «халодны», «цёмны», «зялёны», «сіні», «агністы» і часам сустракаецца ў форме множнага ліку?</i> ‘What luxonym noun in the works of Yakub Kolas co-occurs with the adjectives “cold”, “dim”, “green”, “blue”, “fiery” and sometimes occurs in the plural form?’	<i>бляск</i> ‘gleam, glitter’

Table 6: Sample questions and answers. The examples in the table are similar to the actual instances but easier, as these were correctly answered by at least one local LLM.