Reasoning-BG: A Hybrid Benchmark for Long-Text Understanding and Reasoning in Bulgarian

Maria Todorova, Valentina Stefanova, Tsvetana Dimitrova, Mihaela Moskova, Hristina Kukova, Svetlozara Leseva, Ivelina Stoyanova, Dimitar Georgiev, Svetla Koeva

Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences

{maria,valentina,cvetana,mmoskova,hristina}@dcl.bas.bg
{zarka,iva,dgeorgiev,svetla}@dcl.bas.bg

Abstract

The paper introduces the Bulgarian benchmark **Reasoning-BG**, designed to evaluate how accurately large language models (LLMs) solve Bulgarian long-text comprehension tasks involving different comprehension demands. It was developed using a hybrid methodology that combines automated generation of four-option multiple-choice questions by the Gemma-3-27B-IT model, applied to 232 Bulgarian texts, followed by validation, correct-answer determination, and editing by human experts.

The question design, which addresses relatively long texts, is inspired by the Programme for International Student Assessment (PISA) framework, which targets reading-comprehension tasks such as retrieving information, interpreting and integrating ideas, and evaluating and reflecting on content. The benchmark provides a framework for evaluating model-driven, cognitive-like processes, shifting the focus from factual recall to the analysis of complex interpretative and logical performance in Bulgarian.

Keywords: benchmark, reasoning, LLMs, Bulgarian

1 Introduction

The increasing use of large language models (LLMs) to generate (educational) content and assessment tasks has created a need for benchmarks that can evaluate AI systems on their understanding and reasoning over texts with different levels of complexity.

This need is particularly acute for Bulgarian, which currently lacks specialised benchmarks designed to assess LLM understanding and reasoning. Some recent efforts to develop Bulgarian evaluation datasets have been made (see Section 3), but these resources do not specifically target the cognitive processes involved in understanding extended texts. Traditional benchmarks often prioritise factual accuracy or simple logical predictions, and fail to provide the depth required to measure complex comprehension effectively.

This paper addresses a significant gap in Bulgarian NLP resources by introducing **Reasoning-BG**, a benchmark comprising 232 relatively long texts covering a wide range of topics. Each text is accompanied by 10 multiple-choice questions aligned with a cognitive framework that targets key comprehension processes such as information retrieval, interpretation, and evaluation. A distinguishing characteristic of Reasoning-BG is its strong emphasis on content quality: every text and all questions have been carefully crafted and refined by experts to ensure grammatical accuracy, natural word order, and stylistic consistency that faithfully reflects the linguistic characteristics of Bulgarian.

The present study makes both methodological and empirical contributions by advancing the evaluation of text understanding, reasoning, and higher-order cognitive-like processes involved in long-text comprehension by large language models.

- First, a benchmark for LLM evaluation, Reasoning-BG, was developed that is explicitly aligned with the cognitive processes underpinning deep human text comprehension, including the retrieval and integration of information, the establishment of connections between distant textual segments, the construction of local and global inferences, the tracking of causal relationships, and the formation of a coherent mental representation of the text.
- Second, a hybrid human–LLM methodology for benchmark creation is used, combining expert human judgement with the efficiency of large language models. This approach facilitates the creation and validation of high-quality benchmarks while preserving human

oversight of the interpretation and reliability of the content.

- Third, unlike benchmarks primarily designed for fact retrieval or the assessment of short-text understanding, the proposed benchmark is specifically intended to evaluate long-text inferences. It therefore targets more complex levels of comprehension, including the integration of distributed information, the construction of a global text representation, and the generation of inferences across extended textual contexts. In this respect, the study contributes to the development of cognitively grounded approaches to evaluating understanding and reasoning and provides a foundation for future research.

The paper is organised as follows: Section 2 presents the goals and motivation of the research, followed by a review of related work in Section 3. Section 4 details the dataset design, including the selection of the text collection and the automated question generation process using LLMs, as well as the two-stage human review and validation process used to ensure logical consistency and linguistic adequacy. Section 5 outlines the dataset structure, the distribution of texts across domains and text types, and the distribution of questions by category. The experiments evaluating LLMs with the ReasoningBG benchmark are reported in Section 6, followed by a brief discussion of the results in Section 7 and a summary of the findings and directions for future work in Section 8.

2 Goals and motivation

Reading literacy has traditionally been conceptualised as a measure of human comprehension, with its most influential operationalisation provided by the Programme for International Student Assessment (PISA) (OECD, 2019). Rather than measuring factual recall alone, the PISA framework evaluates progressively more complex comprehension processes, including retrieving information, interpreting and integrating ideas, and evaluating and reflecting on textual content (OECD, 2019). These processes closely correspond to cognitive theories of text comprehension, such as Kintsch's construction–integration model, in which readers incrementally construct coherent mental representations by integrating explicit information with inference and prior context (Kintsch, 1998).

Although PISA was developed to assess human reading proficiency, its cognitive organisation is not specific to human learners. The three broad reading-process dimensions describe increasingly demanding forms of reasoning over text and therefore provide a shared conceptual framework for comparing human readers and large language models. Instead of treating reading comprehension as simple question answering, the PISA taxonomy distinguishes between different levels of reasoning, from locating explicitly stated information to integrating evidence distributed across a document and evaluating competing interpretations.

The PISA categories are used here as task design and analysis categories rather than as claims about the cognitive mechanisms employed by LLMs. In particular, successful performance on an inference-oriented item does not establish that a model performs inference in the same way as a human reader. The benchmark therefore evaluates observable task performance under different comprehension demands, while leaving the underlying mechanisms of model behaviour as an open empirical question.

The benchmark organises questions according to progressively more demanding comprehension processes, enabling systematic comparison across language models and, potentially, with human readers using the same cognitive framework. Rather than serving as a direct replication of the PISA assessment, the benchmark transfers its theoretical organisation to the evaluation of LLM reasoning, providing a common reference point for analysing similarities and differences between human and artificial text comprehension.

While PISA operationalises the cognitive processes of retrieving information, interpreting and integrating, and evaluating and reflecting on content (OECD, 2019) to differentiate human proficiency levels, this benchmark adapts this hierarchical approach to evaluate LLM performance on text-comprehension tasks organised according to PISA-inspired cognitive categories in the context of Bulgarian.

The benchmark is designed to include tasks that go beyond direct information retrieval, such as interpretation, causal reasoning, comparison, conditional reasoning, and text-based inference. These labels describe the requirements imposed by the questions and should not be taken as evidence that the models employ formal logical reasoning or human-like cognitive processes.

By mapping AI performance onto pedagogically grounded categories, the benchmark provides a robust framework for investigating whether performance varies systematically across tasks that place different demands on text comprehension, or instead reflects similar surface-level or statistical strategies across different question types.

3 Related work

Recent advances in LLMs have demonstrated substantial progress in common-sense reasoning and mathematical tasks (Chen et al., 2024; Wang et al., 2026a,b). While established benchmarks such as SQuAD (Rajpurkar et al., 2016), BIG-bench (Srivastava et al., 2023), and HellaSwag (Zellers et al., 2019) are widely used to evaluate machine reading comprehension and logical prediction, they often lack systematic alignment with structured pedagogical frameworks such as PISA (OECD, 2019). There have been only limited efforts to create examination datasets and benchmarks, for example RACE (Lai et al., 2017), that resemble a standardised school examination setting suitable for use as an LLM benchmark.

Existing evaluation tools such as TruthfulQA (Lin et al., 2022) focus primarily on factual accuracy rather than deep comprehension of extended texts. A complementary line of work focuses on educational assessment. Resources such as PISA-Bench (Haller et al., 2025), semi-automatic scoring frameworks such as Omethi (Zehner et al., 2025), and AI-based scoring methodologies (Okubo et al., 2023; Šindelář et al., 2026) have been developed to address the complexities of large-scale educational assessments and reduce human bias in scoring. Unlike general NLP benchmarks, these resources are grounded in educational theory; however, they focus primarily on assessment methodology and scoring rather than on constructing native-language benchmarks for evaluating the reading comprehension abilities of large language models.

The evaluation of reading comprehension in LLMs has evolved considerably beyond simple fact extraction towards deeper forms of understanding. The focus on comprehension of longer texts has led to the development of the QuALITY benchmark (Pang et al., 2022), with multiple-choice questions over passages of roughly 5,000 tokens; NarrativeQA (Kočiský et al., 2018), with questions on full novels and film scripts; and SCROLLS and ZeroSCROLLS (Shaham et al., 2022, 2023), which aggregate summarisation, question answering, and natural language inference tasks across domains ranging from literature to legal texts. Alongside text length, there has been growing interest in multi-step and multi-document reasoning, for example the LongBench benchmark (Bai et al., 2024).

For the Bulgarian language, there have been several evaluation datasets, for example MMLU-bgeval, Winogrande-bgeval, and GSM8k-bgeval (Alexandrov et al., 2024), EXAMS (Hardalov et al., 2020), BgGLUE (Hardalov et al., 2023), and MMLU-X (Thellmann et al., 2024). While these resources are crucial for assessing general knowledge, common sense, and mathematical reasoning in a Bulgarian context, they are largely (automatically) translated and adapted from English or other language benchmarks and do not specifically target language understanding with respect to hierarchical cognitive processes (Kintsch, 1998; OECD, 2019).

An important recent addition is MMLU-BG (Koeva et al., 2026), a native Bulgarian benchmark for large-scale multitask language understanding. The benchmark presented in this paper builds on this line of research by shifting the evaluation focus from multitask knowledge assessment to document-level reasoning, grounded in well-established cognitive and educational frameworks. Together, these resources constitute important infrastructure for Bulgarian LLM evaluation. However, none is specifically designed to assess reading comprehension through hierarchical cognitive processes over extended texts.

Recent advances in LLMs have also enabled the scalable generation of educational reading-comprehension datasets. Zou et al. (2025) introduce DocBench, a benchmark for end-to-end document-reading systems constructed through a hybrid pipeline combining human-written and synthetically generated questions over multi-page documents. Säuberli and Clematide (2024) propose an evaluation protocol, including the *text informativity* metric, for assessing the quality of LLM-generated multiple-choice items, demonstrating that current language models can generate educational questions of acceptable quality for PISA-style assessment. At the same time, several studies have identified limitations of LLM-generated reading-comprehension data. Neeman et al. (2024) show that language models frequently rely on internalised parametric knowledge rather than on the supplied document when the two conflict, po-

tentially compromising the validity of reading-comprehension evaluation. Similarly, Jain et al. (2025) demonstrate that although GPT-4o and o1 can estimate question difficulty in a manner correlated with Item Response Theory parameters, their performance remains insufficient for fully automated assessment design.

These findings highlight the limitations of fully automatic benchmark construction. LLM-generated questions may suffer from factual inconsistencies, ambiguity, and weak cognitive validity, allowing models to answer correctly through superficial lexical matching rather than genuine comprehension (Lin et al., 2022; Srivastava et al., 2023). Consequently, expert human validation remains an essential component of educational benchmark development (Zehner et al., 2025). Hybrid LLM–human pipelines combine the scalability of language models with the methodological control provided by domain experts, ensuring both linguistic quality and pedagogical validity (Haller et al., 2025).

Taken together, existing resources address complementary aspects of reading comprehension, educational assessment, multilingual evaluation, and automatic dataset construction. However, to our knowledge, there is currently no benchmark that simultaneously provides (i) a native Bulgarian resource, (ii) document-level reading comprehension over extended texts, (iii) explicit alignment with the cognitive processes underlying the PISA framework, and (iv) a hybrid LLM–human annotation methodology designed specifically for evaluating the reading-comprehension capabilities of large language models. The benchmark presented in this work is intended to fill this gap.

4 Benchmark design

4.1 Collection of texts

To facilitate the automated generation of assessment items, a specialised dataset of 232 texts was compiled. The collection focuses exclusively on text data drawn from a single source continuous text, composed of sentences organised into paragraphs. Materials containing graphics, tables or diagrams were excluded.

The initial phase of data collection yielded over 300 candidate texts, with an average length of 3,000–4,000 words. To ensure data integrity and avoid prior exposure to LLMs, only offline materials not indexed in public digital repositories were selected.

All sources were converted into a machine-readable format through scanning, OCR processing and manual correction to remove noise and ensure accuracy. The texts were orthographically normalised for consistent spelling and punctuation and stored in plain-text format for computational use. Sections that were overly long or complex were shortened while preserving semantic integrity. After manual verification, the corpus was refined to 232 texts. Approximately one third of the initial samples were excluded due to:

1. Lexical complexity: Texts containing an excessive density of specialised technical terminology were omitted.

2. Structural and technical constraints: A subset of materials was excluded due to high technical complexity in the OCR recovery process or structural irregularities that compromised standardisation of the dataset.

3. Length optimisation: Texts that significantly deviated from the average word count or were deemed excessively long for standard assessment windows were removed to ensure homogeneity of the dataset.

4.2 Text typology and classification

To ensure diversity in discourse characteristics while maintaining a coherent domain, the dataset was organised according to the primary communicative purpose and source of the texts. The categorisation reflects the intended use of the documents rather than a predefined educational taxonomy and supports the construction of questions targeting different comprehension skills.

The corpus comprises two broad groups:

- Informational texts: Sourced from scientific articles and biographies, these texts are specifically designed to support questions targeting factual accuracy and temporal reasoning.

- Popular science texts: Derived from magazines and encyclopaedias, these connected prose pieces facilitate questions focused on cause-and-effect relationships and explanatory understanding.

Although this functional categorisation guided corpus construction, it does not directly correspond to internationally established reading assessment

frameworks. Therefore, for analysis and comparison purposes, all texts were also mapped to some of the general PISA text categories. The PISA framework distinguishes texts according to their dominant rhetorical organisation rather than their source or topic, namely *Description*, *Narration*, *Exposition*, *Argumentation*, *Instruction*, *Transaction*. This secondary classification enables comparisons with previous reading comprehension studies and provides a standardised description of the discourse structures represented in the corpus.

As the present benchmark focuses exclusively on continuous prose intended for complex reading comprehension, texts whose primary function is procedural communication (*Instruction*) or interpersonal exchange (*Transaction*) were excluded from the dataset. The remaining texts were automatically mapped to the four applicable categories using an approximate classification based on the dominant rhetorical structure. The resulting distribution is: *Exposition* (36.32%), *Narration* (29.85%), *Description* (27.36%) and *Argumentation* (6.47%).

The assignment of texts to four categories reflects the dominant discourse function of each document. Many authentic texts contain mixed structures, so this categorisation should be regarded as an approximate mapping.

4.3 Generation of questions using large language models

The test questions were generated automatically using **Gemma-3-27B-IT**[1] (Google DeepMind, 2025), a high-performance, instruction-tuned 27B Transformer from the third-generation Gemma family. The model supports long-context processing, multilingual input, and hybrid-attention mechanisms, which enable efficient handling of extended texts. Its instruction-following capabilities and ability to generate coherent, context-aware outputs over long documents make it particularly suitable for automatic question generation. The model also provides structured outputs suitable for multi-step workflows. Despite its scale, the model is optimised for efficient deployment on modern GPUs and is released under the Gemma Terms of Use, which allow commercial application under defined safety constraints.

[1] https://huggingface.co/google/gemma-3-2B-it

4.4 Prompt design

Two prompts were used for the creation of multiple choice questions. Both prompts instruct the model to generate 15 questions with four possible answers, labeled A), B), C), D), with only one correct answer. The answers should not be explicitly stated in the text but should be logically derived from it.

The first prompt instructs the model to generate questions whose answers are not stated directly in the text, but can be logically inferred from it. The questions and answers generated by this type of prompt required minor editorial changes, such as improving the wording, refining misleading answers, removing ambiguities, or addressing instances where more than one answer was correct.

In the second prompt, the instruction is expanded to explicitly include rules from propositional logic (e.g. “modus ponens”, “modus tollens”, hypothetical syllogism, constructive dilemma, etc.), with the expectation that the model will generate questions requiring formal logical reasoning. In practice, however, the generated questions often adopt a logical “If … then …” structure and frequently introduce premises, conditions, or conclusions that are not supported by the original text. Many of these questions are based on hallucinations rather than valid logical inferences from the source text. As a result, only a limited number of questions generated by the second prompt were selected.

The generation process yielded two sets of comprehension questions for each text, resulting from Prompt 1 and Prompt 2.

4.5 Human review and validation

The automatic creation of the question set is followed by a two-stage human review process to ensure consistency, quality, and alignment with the source texts.

4.5.1 Stage 1: Question selection and editing

In the first stage, annotators review the automatically generated questions and select 10 items per text from an initial pool of 30, covering both general types of generated questions. Annotators are instructed to combine the results from the two prompt-generated sets and to choose 10 questions. The goal is to select difficult questions whose answers follow logically from the text, with only one correct option among the proposed answers.

The process involves careful rereading of the source texts to verify factual consistency, linguistic adequacy, and the relevance of questions and

answers; questions are edited, adapted, or rejected where necessary.

During evaluation and selection, preference is given to questions that (i) require drawing actual conclusions from the text, (ii) contain exactly one correct answer that can be derived from the source text, and (iii) demonstrate an appropriate level of difficulty. Since the first prompt consistently generates a higher proportion of questions meeting these criteria, most of the final questions originate from it, while only a small number of questions generated by the second prompt are retained after careful review and, where necessary, minor adjustments.

The automatically generated questions are edited to ensure clarity, accuracy, and readability. This includes correcting grammatical errors, spelling, and word order; improving lexical choice; and removing redundant or irrelevant elements. Questions are rejected if they are not inferable from the text, are excessively narrow or overly broad, duplicate the focus of other questions, or contain answers explicitly stated in the text without requiring interpretation.

4.5.2 Stage 2: Cross-annotation and refinement

In the second stage, the selected question–answer sets and corresponding texts are exchanged among annotators for further proofreading and refinement. The process includes inter-annotator agreement discussions to ensure consistent interpretation and evaluation criteria. Across both stages, all materials are reviewed and revised against the following validation criteria: clarity, correctness, alignment with the source text, absence of literal overlap with the answer, and appropriate cognitive level.

The four answer options for each question are revised to ensure that each question contained only one clearly correct answer, grounded in the source text. The editing process includes:

- resolving cases where multiple options appear partially correct,
- adding a correct answer when it is missing,
- balancing answer length and structure to prevent bias (e.g. avoiding overly detailed correct answers).

In some cases, annotators fully develop both the questions and all four answer options to meet the required quality standards. The incorrect answers are edited to include implicit information and partially true statements, testing models' ability to distinguish between accurate, misleading, and incomplete information.

5 Dataset

5.1 Classification

The overall observation from comparing the questions generated by the two prompts is that those derived from Prompt 1 are primarily grounded in explicit facts from the text and are more directly applicable and suitable for operational use. In contrast, the questions resulting from Prompt 2 require substantial restructuring and refinement, as they often introduce overly elaborate formulations or abstract criteria that are not immediately aligned with the intended assessment format.

To facilitate systematic analysis, the questions were broadly grouped into PISA-inspired task categories developed for this benchmark (Table 1).

The questions were classified into seven categories using a rule-based scheme applied to the question text in a fixed order, with Information retrieval questions serving as the default when no other pattern matches.

- **Information retrieval** — the default category for questions that retrieve information explicitly stated in the text, typically beginning with question words and not matching any higher-priority pattern.
- **Vocabulary and interpretation** — questions that ask about the meaning of a word, phrase, or abstract concept as used in the text.
- **Causal** — questions about reasons, causes, effects, or consequences.
- **Significance** — questions about the role, purpose, importance, or function of something.
- **Comparisons and relations** — questions that ask the reader to compare, contrast, or identify relationships between entities or concepts.
- **Conditional** — hypothetical questions that present a counterfactual or "if X, then what" scenario.
- **Inference** — questions that require reaching a conclusion not explicitly stated in the text, or evaluating a claim against the text.

Category	Features	Subcategories
Interpretative QA	who, when, where	Information retrieval, Vocabulary and interpretation
Abstractive QA	why, how, how do X and Y differ	Causal, Significance, Comparisons and relations
Inferential QA	if... then	Conditional, Inference

Table 1: Classification of question types by categories

These categories are vague and therefore assigned only approximately, and they do not strictly adhere to the formal PISA definitions. In particular, the feature markers shown in Table 1 served as heuristic cues during rule-based classification and do not exhaustively characterise each subcategory; inference questions, for instance, do not need to exhibit conditional syntax to qualify. The PISA-inspired categories should therefore be interpreted as properties of the assessment tasks rather than as direct measurements of internal cognitive processes.

5.2 Dataset structure

The dataset consists of 232 texts paired with a set of questions. The texts are stored in plain text (.txt) format, and the questions are in tab separated values (.tsv) format; each question contains: the question itself, the answers (A, B, C, D), and the correct answer (marked by the corresponding letter).

The following metadata presented in JSON format are stored for each text: **filenameText**, **filenameQuestions**, **title**, **author**, **source**, **domain**, **keywords**, **numberPar**, **numberSent**, and **numberWords** (number of paragraphs, sentences, and words, respectively), **validationStatus**.

Table 2 shows the statistics for the dataset and the texts' distribution across domains. The distribution across question categories is presented in Table 3.

6 Experiments

6.1 Description of selected LLMs

The LLMs used in the experiments span a range of model sizes and architectures, enabling meaningful comparison across models, particularly those that explicitly include or target Bulgarian during training.

Domain	#Texts	#Sents	#Words
Natural Sciences	60	2,535	40,462
Science & Technology	42	1,511	26,418
Literary Studies	38	1,292	26,658
History	37	1,279	25,312
Geography	23	1,108	19,816
Psychology	13	474	8,459
Culture	7	313	5,780
Medicine	7	253	4,347
Philosophy	5	206	4,103
Total	**232**	**8,971**	**161,355**

Table 2: Domain distribution of the dataset

Question Category	N	%	Avg/text
Information retrieval	1,057	45.6	4.56
Vocabulary and interpretation	64	2.8	0.28
Interpretative	1,121	48.3	4.83
Causal	255	11.0	1.10
Significance	104	4.5	0.45
Comparisons and relations	223	9.6	0.96
Abstractive	582	25.1	2.51
Conditional	114	4.9	0.49
Inference	503	21.7	2.17
Inferential	617	26.6	2.66
Total	**2,320**	**100.0**	**10.00**

Table 3: Distribution of the question according to question categories

Gemma-4-31B-IT[2] is an instruction-tuned large language model (released in 2026) with 31 billion parameters and a 256K token context window, supporting over 140 languages, including but not targeting Bulgarian.

BgGPT-Gemma-2-9B-IT-v1.0[3] is a Bulgarian-targeted instruction-tuned LLM (released in 2024) with 9 billion parameters, developed by INSAIT and continually pretrained on approximately 85 billion Bulgarian tokens on top of Google's Gemma 2 base model.

[2] https://huggingface.co/google/gemma-4-31b-it

[3] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-2-9B-IT-v1.0

BgGPT-Gemma-3-27B-IT[4] is a Bulgarian-targeted instruction-tuned LLM (released in 2025) with 27 billion parameters and a 131K token context window, developed by INSAIT through continual pretraining on Bulgarian data on top of Google's Gemma 3 base model.

Qwen2.5-14B-Instruct[5] is an instruction-tuned LLM (released in 2024) with 14 billion parameters and a 128K-token context window, developed by Alibaba Cloud and supporting more than 29 languages, without mentioning Bulgarian in its training data.

6.2 Evaluation environment and metrics

The evaluation is conducted on an NVIDIA A100 GPU with 80 GB of VRAM. The temperature is fixed at 0.7 across all models. The software is written in Python and is publicly available.[6] The language models are all programmatically downloaded from HuggingFace.

Model performance is evaluated by comparing each prediction with the manually determined answer in the dataset. The evaluation metric is binary accuracy, calculated as the proportion of instances in which the model's decision matches the human answer, out of the total number of answers.

6.3 Results

Testing large language models on the benchmark reveals pronounced performance differences across models.

Gemma-3-27B-IT, which is used for the generation of the dataset, achieves 91.5% accuracy across all text types. It is used in the analysis for comparative purposes.

The low scores of the newer **Gemma-4-31b-IT** are due to its failure to process the instruction format used in the evaluation experiments, which led to repeated response formatting failures (repetition of the question, repeatedly listing the answers, looping, hallucinations, etc.). This calls for a reconsideration of the experimental protocols in future applications.

BgGPT-Gemma-3-27B-IT also achieves high accuracy (87.84%), representing a negligible improvement on the previous, smaller version, **BgGPT-Gemma-2-9B-IT**, which scored 86.33%. The lower score of BgGPT-Gemma-3 relative to Gemma-3 motivates further controlled experiments investigating the effect of Bulgarian continual pretraining.

The strongest performer is **Qwen2.5-14B-Instruct** with 93.22%, despite having fewer parameters and only including Bulgarian as part of its multilingual support. Qwen2.5 was trained with an emphasis on instruction following and reasoning (Qwen Team, 2025), and its competitive performance on the multiple-choice task, despite the absence of Bulgarian-targeted training, may reflect strong general instruction-following calibration rather than Bulgarian language understanding.

It is worth noting that model performance is quite consistent across text types. Most models perform best on argumentation texts, except for BgGPT-Gemma-2-9B-IT-v1.0, which scores slightly better on exposition texts.

7 Discussion

The highest scores are achieved by the multilingual model Qwen2.5.The observed performance differences are consistent with the possibility that multilingual pre-training contributes to performance, but the present comparison does not permit causal attribution because model architecture, size, pretraining data, instruction tuning and language-specific training vary simultaneously.

Performance is broadly consistent across text types, with some models showing their strongest results on Argumentation texts and others showing their weakest on Narration. This may reflect the nature of the question types associated with each genre: Argumentation texts prompt more Inference questions, which require the model to evaluate claims against the text – a task that aligns well with the instruction-following training of LLMs. Narration texts, by contrast, tend towards information retrieval and causal questions anchored in specific narrative details, where a model lacking deep familiarity with Bulgarian cultural and historical context may be at a disadvantage.

Across question categories, variation within models is small, suggesting that the distinctions in cognitive demand encoded in the taxonomy (Causal, Inference, etc.) do not translate straightforwardly into differing levels of difficulty for LLMs. This contrasts with findings from human reading comprehension studies, where question type is a re-

[4] https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-3-27B-IT

[5] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[6] https://github.com/DCL-IBL/BG-Benchmark-Evaluation

Model	Type of text				
	Overall	Argumentation	Description	Exposition	Narration
Gemma-4-31B-IT	33.37	36.15	36.00	32.33	31.55
BgGPT-Gemma-2-9B-IT	86.33	85.38	85.64	87.81	85.34
BgGPT-Gemma-3-27B-IT	87.84	92.31	88.55	87.67	86.38
Qwen2.5-14B-Instruct	**93.22**	**96.15**	**93.27**	**93.56**	**92.07**

Table 4: Models' accuracy by text type. The highest observed accuracy for each text type (vertical maximum) is shown in bold. The results for the text type with the best performance for each model (horizontal maximum) are underlined.

Model	Question type						
	Fact	Voc&Int	Cause	Signif	Comp&Rel	Cond	Infer
Gemma-4-31B-IT	34.2	34.9	30.6	31.9	35.2	29.6	33.3
BgGPT-Gemma-2-9B-IT	87.0	86.0	88.6	89.4	83.2	82.4	85.6
BgGPT-Gemma-3-27B-IT	87.5	93.0	88.6	87.2	90.3	89.8	86.1
Qwen2.5-14B-Instruct	**93.8**	**95.3**	**94.1**	**90.4**	**91.8**	**91.7**	**92.9**

Table 5: Accuracy per question category. The model performing best on each question type (vertical highest) is marked in bold. The question category on which each model achieves its highest accuracy (horizontal highest) is underlined.

liable predictor of difficulty (OECD, 2019; Day and suk Park, 2005). One possible explanation is that the models may rely on broadly similar patterns of text–question–option association across different question types, rather than showing clearly distinguishable performance patterns across the reasoning demands represented in the benchmark. However, the present results do not allow us to determine the underlying mechanisms responsible for these patterns.

The dataset is available on GitHub: https://github.com/DCL-IBL/Reasoning-BG-Dataset and HuggingFace: https://huggingface.co/datasets/DCL-IBL/Reasoning-BG-Dataset.

8 Conclusions and future work

Reasoning-BG is a benchmark for measuring how accurately LLMs solve Bulgarian long-text comprehension tasks designed around different, PISA-inspired comprehension demands. It addresses a critical gap in Bulgarian AI benchmarks by shifting the focus from simple factual recall to more demanding forms of text comprehension, including interpretation, integration, and inference.

The main contribution of this work is the development of a human-validated, LLM-assisted dataset through a hybrid human–AI pipeline that combines the scalability of large language models such as Gemma-3-27B-IT with two-stage evaluation by human experts.

This dataset provides a foundation for future research on human–AI comparison and educational assessment tools in the Bulgarian context. There is considerable potential to expand it by adding more texts and designing increasingly complex reasoning tasks.

In addition, the dataset could be used to train LLMs to compile resources for teaching reading literacy and exam preparation.

Error analysis can, on the one hand, help improve dataset quality, particularly by comparing incorrect outputs from multiple models, and, on the other hand, be used to explore differences in the answer-generation behaviour of diverse models and in their approaches to different types of questions. In future work, more complex reasoning tasks can also be designed to test the capabilities of LLMs.

Acknowledgments

The present study is carried out within the framework of the project Infrastructure for Fine-tuning Pre-trained Large Language Models, Grant Agreement No. ПВУ–55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

References

Anton Alexandrov, Veselin Raychev, Dimitar I. Dimitrov, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024. BgGPT 1.0: Extending English-centric LLMs to other languages. ArXiv: 2412.10893.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, Ishaan Singh Rawal, Cheston Tan, Dongkyu Choi, Bo Xiong, and Bo Ai. 2024. LLM-based multi-hop question answering with knowledge graph integration in evolving environments. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14438–14451, Miami, Florida, USA. Association for Computational Linguistics.

Richard R. Day and Jeong suk Park. 2005. Developing reading comprehension questions. *Reading in a Foreign Language*, 17(1):60–73.

Google DeepMind. 2025. Gemma 3 27B-IT. `https://huggingface.co/google/gemma-3-27b-it`.

Patrick Haller, Fabio Barth, Jonas Golde, Georg Rehm, and Alan Akbik. 2025. PISA-Bench: The PISA Index as a multilingual and multimodal metric for the evaluation of vision-language models. ArXiv: 2510.24792.

Momchil Hardalov, Pepa Atanasova, Todor Mihaylov, Galia Angelova, Kiril Simov, Petya Osenova, Veselin Stoyanov, Ivan Koychev, Preslav Nakov, and Dragomir Radev. 2023. bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8733–8759, Toronto, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yoshee Jain, John Hollander, Amber He, Sunny Tang, Liang Zhang, and John Sabatini. 2025. Exploring the potential of large language models for estimating the reading comprehension question difficulty. In *Adaptive Instructional Systems*, pages 202–213. Springer Nature Switzerland.

Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Svetla Koeva, Ivelina Stoyanova, Dimiter Georgiev, Svetlozara Leseva, Valentina Stefanova, Maria Todorova, Tsvetana Dimitrova, Hristina Kukova, Mihaela Moskova, and Tinko Tinchev. 2026. Bulgarian Massive Multitask Language Understanding Benchmark. In *Proceedings of the 2026 Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ori Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2024. LLMs' reading comprehension is affected by parametric knowledge and struggles with hypothetical statements. ArXiv: 2404.06283.

OECD. 2019. *PISA 2018 Assessment and Analytical Framework: Reading, Mathematics and Science*. OECD Publishing, Paris, France. ISBN: 978-92-64-94031-4.

Tomoya Okubo, Wayne Houlden, Paul Montuoro, Nate Reinertsen, Chi Sum Tse, and Tanja Bastianić. 2023. *AI Scoring for International Large-Scale Assessments Using a Deep Learning Model and Multilingual Data*, volume 287 of *OECD Education Working Papers Series*. OECD Publishing, Paris, France.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. 2022. QuALITY: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358, Seattle, United States. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. ArXiv: 2412.15115.

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.

Andreas Säuberli and Simon Clematide. 2024. Automatic generation and evaluation of reading comprehension test items with large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics. ArXiv: 2404.07720.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore. Association for Computational Linguistics.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardised comparison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pavel Šindelář, Dávid Slivka, Christopher Bouma, Filip Prášil, and Ondřej Bojar. 2026. Training data generation for context-dependent rubric-based short answer grading. ArXiv: 2603.28537.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. ArXiv: 2206.04615.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards cross-lingual LLM evaluation for European languages. ArXiv: 2410.08928.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, and Yang Yu. 2026a. A survey on large language models for mathematical reasoning. *ACM Comput. Surv.*, 58(8).

Yi Wang, Yang Hu, and Jiajun Nie. 2026b. Knowledge graph-based chain-of-thought reasoning framework for complex questions. In *Proceedings of the International Conference on Artificial Intelligence and Advanced Autonomous Systems (AIAAS 2026)*, page 47–54, New York, NY, USA. Association for Computing Machinery.

Fabian Zehner, Hyo Jeong Shin, Emily Kerzabi, Andrea Horbach, Sebastian Gombert, Frank Goldhammer, Torsten Zesch, and Nico Andersen. 2025. Down the cascades of Omethi: Hierarchical automatic scoring in large-scale assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 660–671, Vienna, Austria. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2025. DocBench: A benchmark for evaluating LLM-based document reading systems. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 359–373, Albuquerque, New Mexico, USA. Association for Computational Linguistics.