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 and Svetla Koeva

Institute for Bulgarian Language, Bulgarian Academy of Sciences

CLIB 2026 | 7 – 10 September, Sofia, Bulgaria

Motivation

- LLMs are increasingly used to generate educational content and assessment tasks – but we lack benchmarks that test **understanding and reasoning**, not just factual recall.
- Bulgarian has no specialised benchmark targeting **long-text comprehension**.
- Existing Bulgarian resources (MMLU-bgeval, EXAMS, BgGLUE, MMLU-BG, etc.) are largely **translated/adapted from English** and don't target comprehension processes.
- Traditional benchmarks prioritise factual accuracy or simple logical prediction – **insufficient for measuring human-like deep comprehension**.

Motivation

- **Bulgarian AI evaluation suffers from a shortage of native, cognitively-grounded benchmarks.** There is a need for a benchmark that evaluates information retrieval, interpretation and integration of knowledge, as well as reflection and evaluation on statements – inspired by the PISA reading-literacy framework.

The main goal of **REASONING-BG dataset**:

to shift evaluation focus from surface-level QA to complex interpretative and logical performance in Bulgarian.

The REASONING-BG approach

- **PISA Cognitive Alignment:** inspired by the PISA framework, targeting information retrieval, integration, and reflection.
- **Extended Paragraph Contexts:** utilizes 232 continuous, high-quality offline Bulgarian prose texts optimized for long-text inference.
- **Pedagogical Grounding:** shifts assessment to observable task performance under hierarchical, cognitive-like reading demands.

The Benchmark design

To facilitate the automated generation of assessment items, a specialised dataset of 232 texts with an average length of 3,000 – 4,000 words was compiled.

- 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.
- The collection focuses on continuous text, composed of sentences organised into paragraphs. Materials containing graphics, tables or diagrams were excluded.

The REASONING-BG approach

– a hybrid human-LLM pipeline

1. AI Question Generation
Google DeepMind
Gemma-3-27B-IT

❯

2. Expert Human Selection
Strict Selection & Editing

❯

3. Double-Blind Proofreading
Cross-Annotation & Balance

Automatic Question Generation

AI Question Generation with Google DeepMind Gemma-3-27B-IT, prompted to generate 15 multiple-choice questions (4 options, single correct) per text.

- Prompt 1 targeted logical derivations not explicitly stated.
- Prompt 2 introduced formal propositional logic (e.g., modus ponens)

Human Selection & Editing

- Human annotators selected the best 10 questions per text from the candidate pool.
- Corrected syntax, grammatical errors, and refined word choice to ensure stylistic soundness in Bulgarian.
- Rejected direct text overlap, overly simplistic fact-matching, or ambiguous choices.

Dataset Structure

DATASET UNIT

TEXT

.txt — plain text file

paired with

QUESTIONS *(.tsv)*

- Question text
- Answer options A / B / C / D
- Correct answer

described by metadata

METADATA

JSON

filenameText	filenameQuestions
title	author
source	domain
keywords	numberPar
numberSent	numberWords
validationStatus	

- The dataset is available on: Github HuggingFace

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 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 relations.

Reasoning-BG Spans Nine Scientific Domains

Domain	Texts	Words
Natural Sciences	60	40,462
Science & Technology	42	26,418
Literary Studies	38	26,658
History	37	25,312
Geography	23	19,816
Psychology	13	8,459
Culture	7	5,780
Medicine	7	4,347
Philosophy	5	4,103
Total Dataset	232	161,355

To guarantee functional and stylistic diversity, continuous texts are annotated with their dominant rhetorical structure, mimicking the standardized PISA reading dimensions.

Cross-Annotation

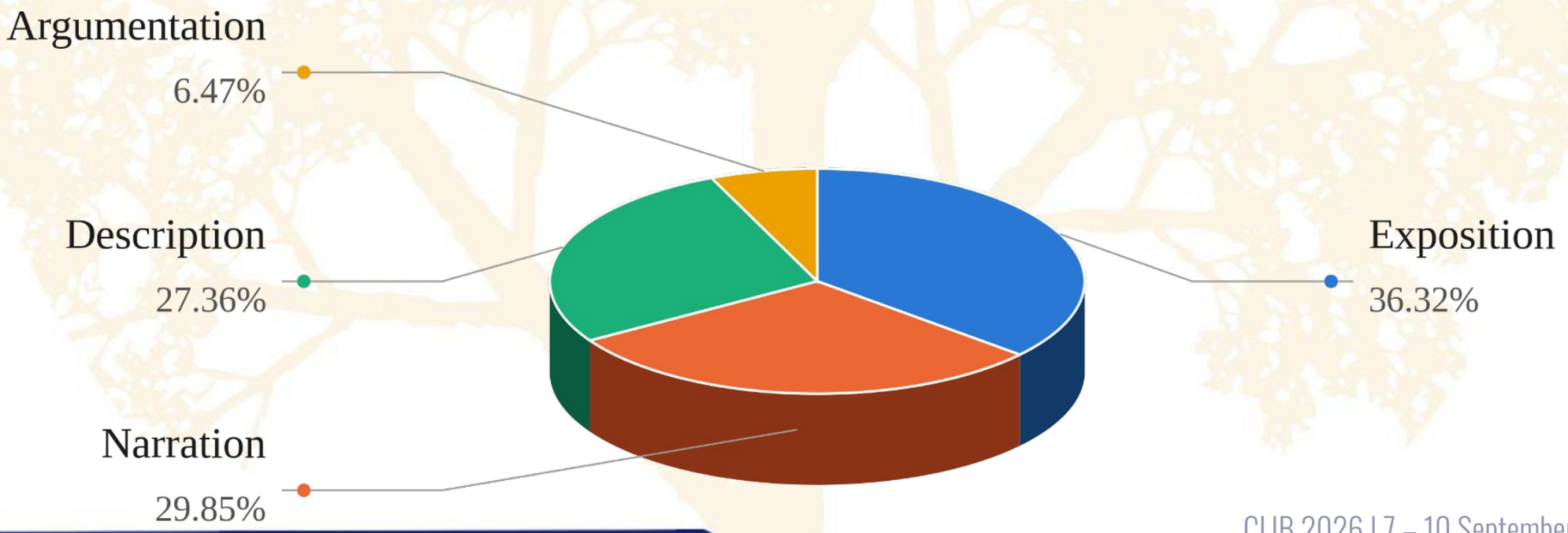
- Exchanged selected question-answer sets among annotators for cross-verification.
- Discussed and aligned inter-annotator disagreement on text interpretation.
- Edited incorrect answers (distractors) to contain partially true or implicit statements to test models.

Text typology and classification

For functional and stylistic diversity, continuous texts are annotated with their dominant rhetorical structure, mimicking PISA reading dimensions.

The texts were automatically mapped to the four categories using an approximate classification based on the dominant rhetorical structure.

Many authentic texts contain mixed structures, so this categorisation should be regarded as an approximate mapping.



Generation of questions using Gemma-3-27B-IT

The prompts used for the creation of multiple choice questions 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.
- 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.) 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.

Generation of questions – human review and validation

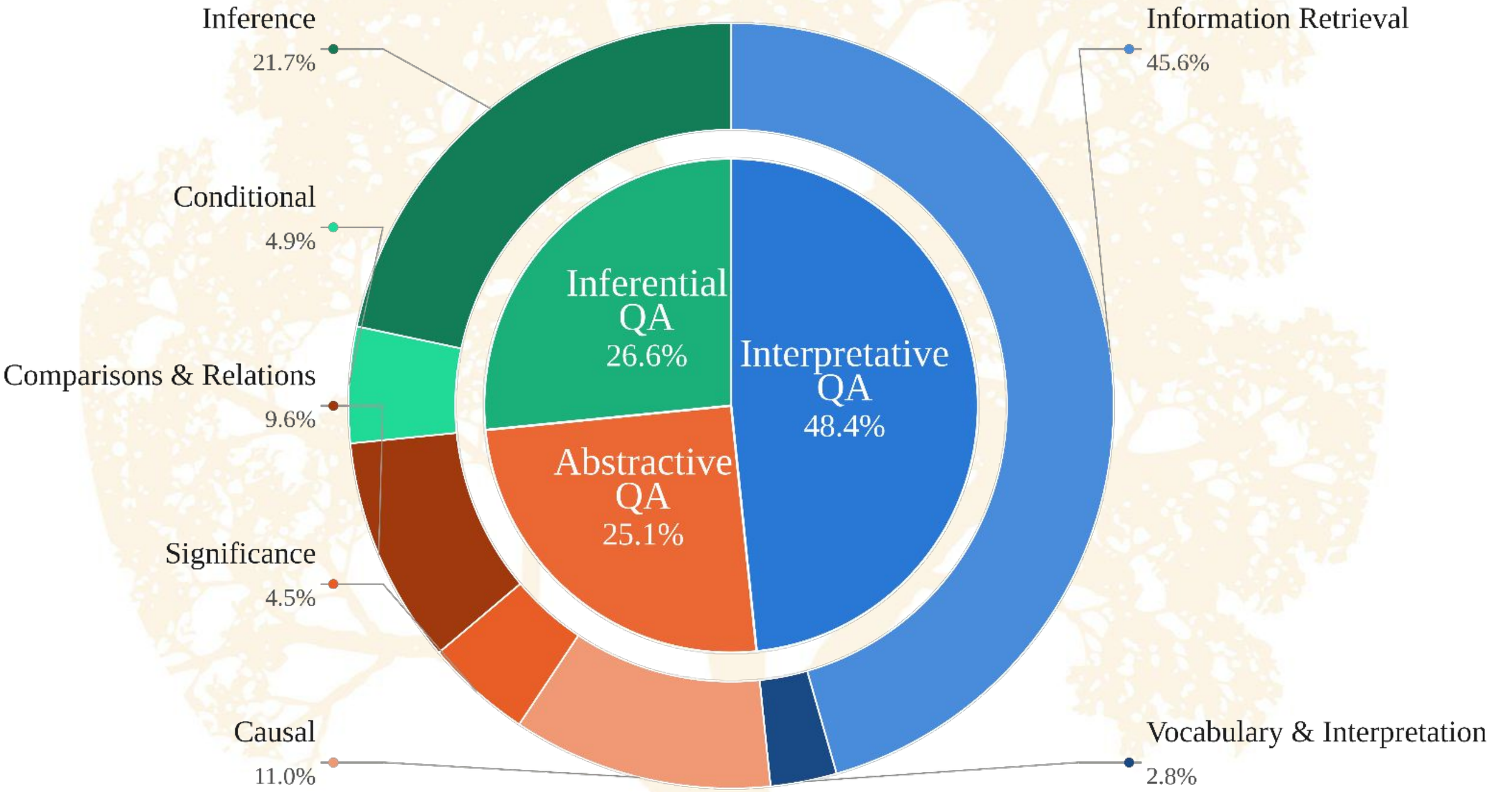
Stage 1: Question selection and editing

- Select 10 questions/text from an initial pool of 30, combining both prompt-generated sets.
- Prioritize difficult, text-grounded questions with one clearly correct answer.
- Reread texts to check factual consistency, linguistic adequacy, and question–answer relevance; edit, adapt, or reject items as needed.

Stage 2: Cross-annotation and refinement

- Exchange question–answer sets and texts for proofreading, refinement, and inter-annotator agreement.
- Validate clarity, correctness, source alignment, non-overlap with the answer, and cognitive level.
- Revise answer options to ensure one correct, text-grounded answer, resolving partial correctness, adding missing answers, and balancing option length/structure to avoid bias.

Question Categories



Question Categories

- The classification is automatic and rule-based.
- The PISA-inspired categories should be interpreted as properties of the assessment tasks rather than as direct measurements of internal cognitive processes.
- The classified feature markers do not strictly adhere to the formal PISA definitions, they served as heuristic cues during rule-based classification and do not exhaustively characterise each subcategory.

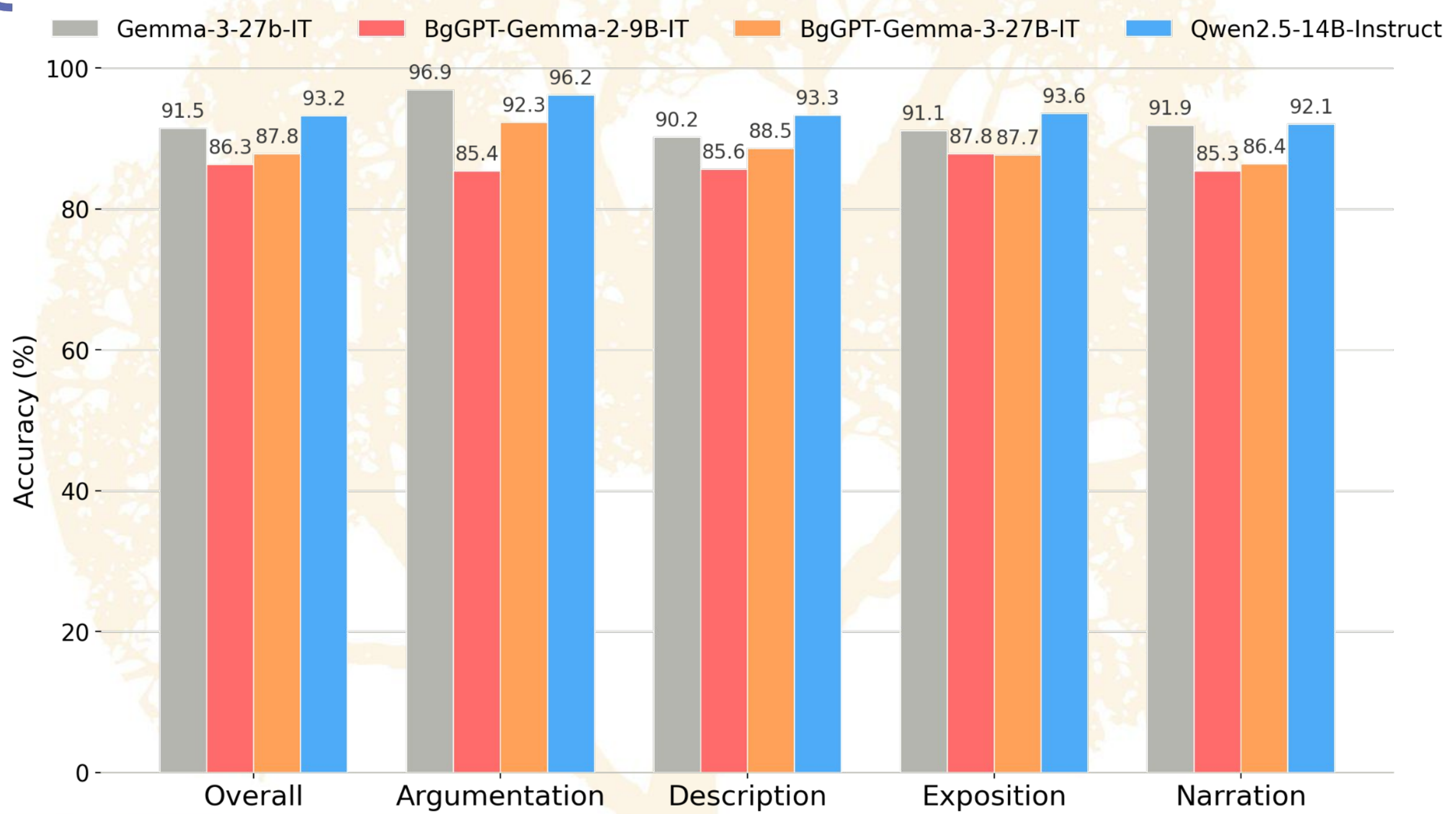
Experiments

We test the performance of several LLMs:

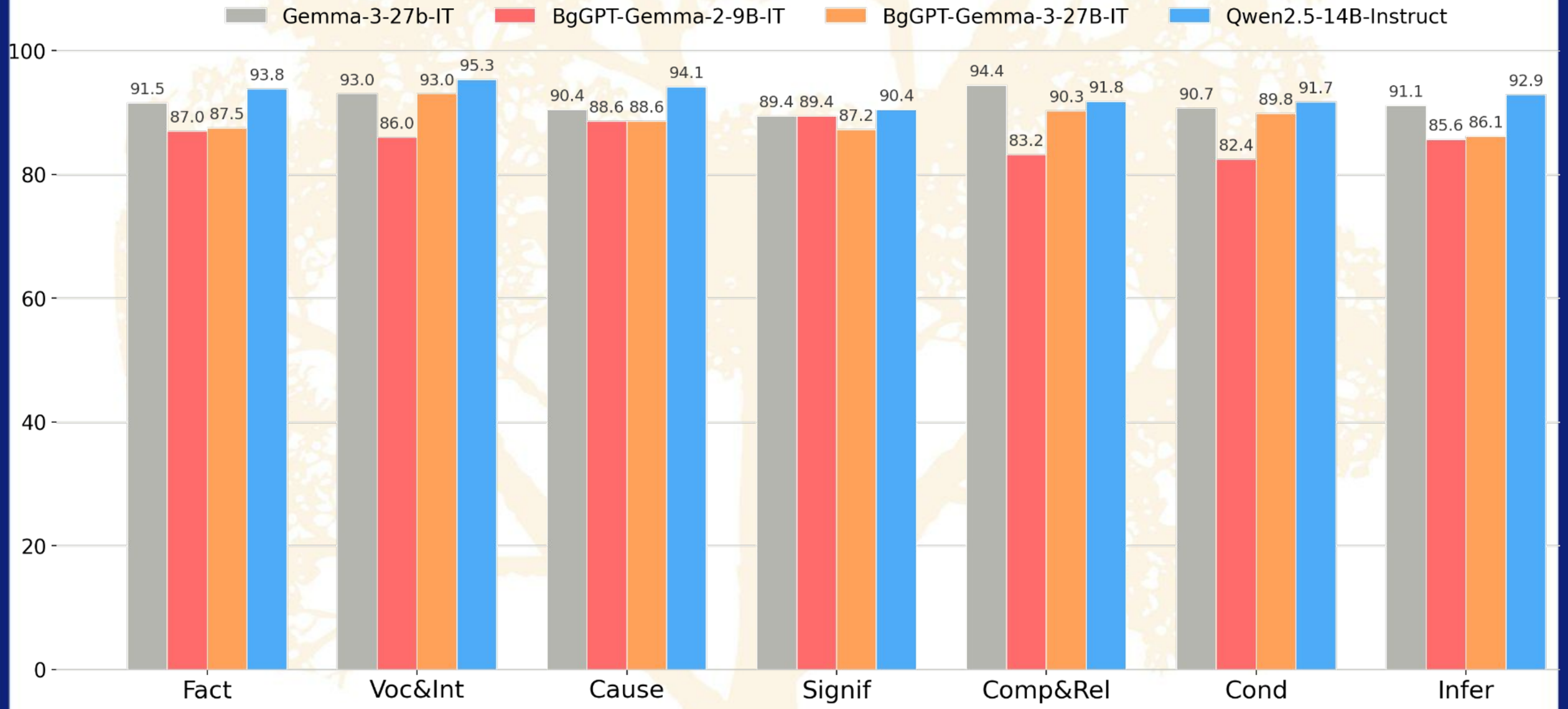
- We show results of **Gemma-3-27B-IT** as a reference only (since it has been used for the initial generation of the questions).
- We use two models specifically focusing on Bulgarian: **BgGPT-Gemma-2-9B-IT-v1.0** and **BgGPT-Gemma-3-27B-IT**, developed by the INSAIT Institute.
- We also include **Qwen2.5-14B-Instruct** as a multilingual model with no specific training on Bulgarian.

All models are obtained from Hugging Face and the tests are performed using NVIDIA A100 GPU (80GB VRAM), fixed temperature of 0.7 across all generations, with binary accuracy evaluation compared with expert human keys.

Experiments



Experiments



Experiments: Discussion

- Highest scores are achieved by **Qwen2.5-14B-Instruct**, consistent with the possibility that multilingual pre-training contributes to performance, but more experiments are needed to confirm this.
- Performance is broadly consistent across text types with some small variance.
- 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 levels of difficulty for LLMs. This finding also suggests 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.

Conclusions and future work

- **Reasoning-BG** provides a foundational resource as a human-validated, LLM-assisted dataset developed 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.
- 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.

Acknowledgments

- Available on GitHub and Hugging Face:
  https://github.com/DCL-IBL/Reasoning-BG-Dataset
  https://huggingface.co/datasets/DCL-IBL/Reasoning-BG-Dataset

- The work is part 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/.

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria

- https://ifgpt.dcl.bas.bg