Bulgarian Massive Multitask Language Understanding Benchmark

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

Institute for Bulgarian Language, Bulgarian Academy of Sciences | svetla@dcl.bas.bg

Language Resources and Evaluation Conference ◆ Palma de Mallorca, Spain ◆ 11-16 May 2026

Motivation and objectives

The main issues related to low-resource languages such as Bulgarian:

- **Most LLM benchmarks target English**; very few evaluation resources for Bulgarian: e.g., bgGLUE, BgGPT Evaluation Suite.
- **A growing need for more evaluation benchmarks** as existing ones quickly become outdated due to model overfitting.

Why MMLU?

- **Domain scope and universality** – academic subjects in STEM, Social Sciences, Humanities, Professional Skills.
- **Standardised evaluation** – 4-options multiple-choice format.
- **Multilinguality** – multilingual versions, e.g. Global MMLU, MMLU-X.
- **Practical relevance** – smaller models, common in low-resource settings, score far below saturation.
- **Comparison of human-translated and machine-translated benchmarks.**

MMLU-BG is a large-scale benchmark in Bulgarian designed to assess the general knowledge and multi-domain versatility of LLMs, identifying model strengths and weaknesses across different knowledge areas.

Structure of MMLU-BG

Domain	# subjects	# questions	# words
STEM	19	3,606	159.6K
Humanities	15	4,134	338.4K
Social Sciences	6	1,808	76.8K
Professional	16	4,550	232.6K
Total	56	14,093	807.3K

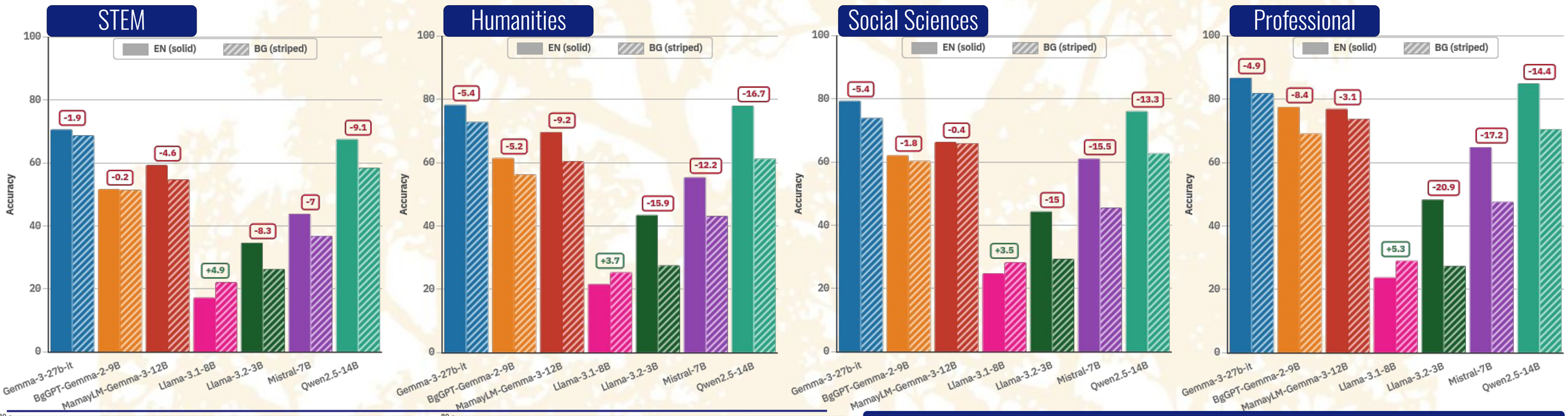
Development protocol of MMLU-BG

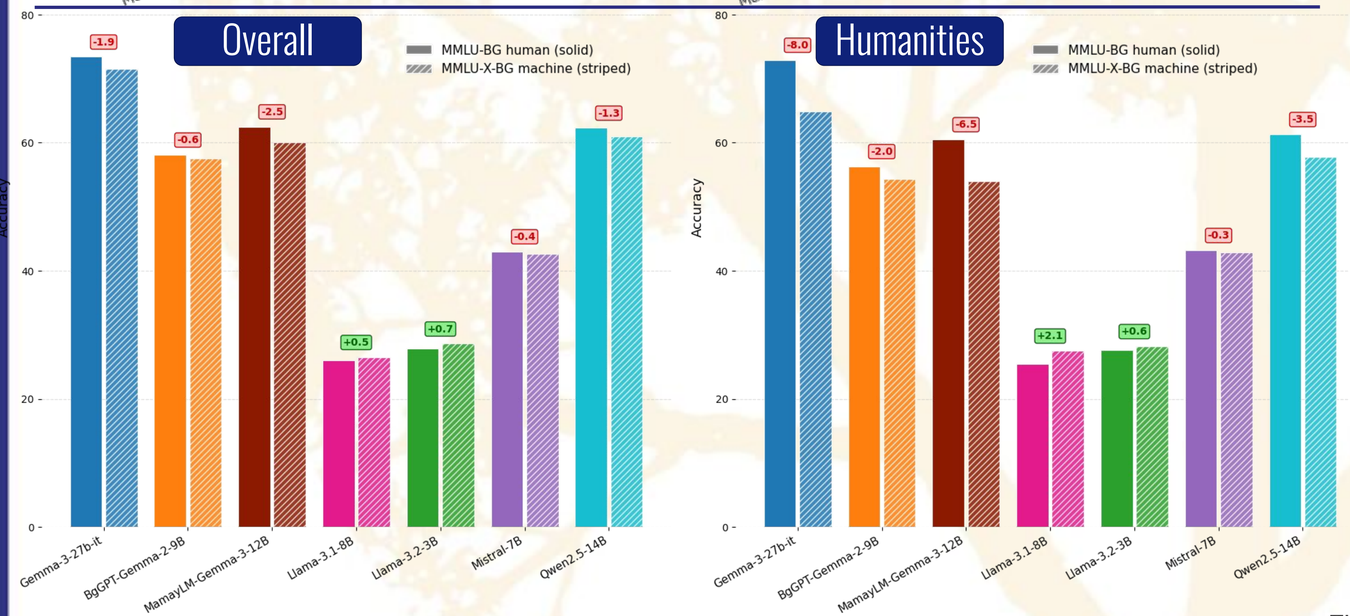
The main issues related to low-resource languages such as Bulgarian:

- **Human expert translation by Bulgarian language specialists.** Experts consulted research articles, textbooks, terminology dictionaries and encyclopaedias.
- **Terminology consistency** was checked semi-automatically to ensure each unique term was translated uniformly.
- **All subjects were cross-validated by a second expert.** Any errors or inconsistencies in spelling, punctuation, or grammar were corrected directly. Issues involving unclear meaning or factual errors were returned to the first expert for a final decision.

Experiments: (1) Comparing 7 LLMs on MMLU (EN) vs. MMLU-BG; (2) MMLU-BG (Human) vs. MMLU-BG-X (MT)

NVIDIA A100 GPU (80 GB VRAM); temp = 0.7. One-shot evaluation. Prompt: question + 4 options (A/B/C/D). Metric: accuracy. Unified pipeline for 7 LLMs.





Conclusions and future work

- Benchmarks focusing on STEM-related capabilities consistently show stronger correlations with human judgements across languages; tasks requiring understanding of linguistic nuances and cultural contexts are more sensitive to translation.
- The next step is to expand MMLU-BG with language- and culture-specific knowledge, as well as data for more advanced evaluation tasks.

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

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria