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
Sofia, Bulgaria
svetla@dcl.bas.bg, iva@dcl.bas.bg, dgeorgiev@dcl.bas.bg, zarka@dcl.bas.bg,
valentina@dcl.bas.bg, maria@dcl.bas.bg, cvetana@dcl.bas.bg, hristina@dcl.bas.bg,
mmoskova@dcl.bas.bg, tinko@dcl.bas.bg

Abstract

Assessing the broad general knowledge of Large Language Models (LLMs) across multiple domains in Bulgarian remains challenging due to the limited availability of Bulgarian evaluation benchmarks. To address this gap, we introduce the **Bulgarian Massive Multitask Language Understanding benchmark** (**MMLU-BG**), designed to evaluate whether LLMs possess generalised knowledge capabilities beyond simple text prediction in Bulgarian. This paper presents the structure, the development protocol, and the scope of the MMLU-BG benchmark. It is tested in comparison with the original MMLU for English across seven LLMs selected according to specific criteria. The experiments demonstrate that the MMLU-BG benchmark assesses multi-domain versatility and highlights the models' strengths and weaknesses across different subject areas.

1. Introduction

The development of language processing tools is always linked to the evaluation of their precision and efficiency. This also applies to Large Language Models (LLMs), with all their peculiarities and the challenges that thorough evaluation may reveal. Publicly available benchmarks are often created for English; moreover, it has long been noted that evaluation benchmarks become outdated very quickly, as models may overfit to them. Recent surveys identify issues such as inflated scores caused by data contamination, as well as cultural and linguistic biases (Ni et al., 2025); others highlight the necessity of benchmarks that reflect real-world applications and present challenges (Mohammadi et al., 2025).

The cited surveys explored hundreds of benchmarks. In this context, benchmarks for Bulgarian are few. The specially designed benchmark for Bulgarian that is freely available is **bgGLUE**: Bulgarian General Language Understanding Evaluation (Hardalov et al., 2023). Another publicly available benchmark is **BgGPT Evaluation Suite** (INSAIT Institute, 2023).

The **MMLU** benchmark (**Massive Multitask Language Understanding**) is widely regarded as one of the most important and standard benchmarks for evaluating LLMs (Hendrycks et al., 2021). Its significance arises from several key factors, including the evaluation of logical reasoning, problem-solving, and factual recall, rather than simple pattern matching.

Given the scarcity of Bulgarian benchmarks, measuring broad general knowledge across domains is challenging. To address this gap, we developed **Bulgarian Massive Multitask Language Understanding**, or **MMLU-BG**) to provide datasets for evaluating whether an LLM possesses generalised knowledge beyond simple text prediction in Bulgarian. Moreover, unlike many efforts to develop LLM benchmarks that rely on machine translation, MMLU-BG is the result of human expert translation.

As MMLU is already available in many languages, a human-translated Bulgarian MMLU will provide a reliable test set for cross-lingual comparison of LLM reasoning across different subjects. In contrast to some other benchmarks that emphasise only one type of knowledge (e.g., mathematical reasoning or general knowledge), and following the design of MMLU, MMLU-BG assesses multi-domain versatility, identifies the strengths and weaknesses of models across knowledge areas, and helps to understand their capabilities.

The paper is organised as follows. Section 2 provides a brief overview of the most important evaluation benchmarks used to assess LLMs. Section 2.3 describes the key features of the Massive Multitask Language Understanding (MMLU) benchmark and the results of its application in evaluating various LLMs. The motivation for developing the Bulgarian dataset MMLU-BG is presented in Section 3. Section 4 outlines the structure and scope of MMLU-BG, as well as the development protocol. The relevance and utility of MMLU-BG are demonstrated through several experiments in Section 5, which presents the selected LLMs, the experimental setup, the results, and a discussion of the outcomes. Finally, Section 6 presents the conclusions and our plans for the future development and expansion of MMLU-BG.

The **main contributions** of this work are as follows:

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4658–4672
11–16 May 2026.

- First, we have developed a large-scale benchmark in Bulgarian, the **MMLU-BG**, for evaluating LLMs, aimed at assessing general knowledge and multi-domain versatility. This benchmark enables the identification of model strengths and weaknesses across different knowledge areas, providing valuable insight into model capabilities. Human experts invested significant effort in translating and adapting MMLU into Bulgarian to create a linguistically authentic evaluation resource aligned with terminology already accepted in Bulgarian.
- Second, we demonstrate experimentally that **MMLU-BG** can be successfully applied to evaluate models across various subjects, knowledge levels, languages, and LLM architectures.

The **MMLU-BG** is distributed openly under Creative Commons Attribution 4.0 International License (CC BY), which facilitates the replication of experiments.

2. Benchmarks for LLMs evaluation

The creation of robust evaluation benchmark, that can measure progress in the field has received considerable attention. The ability to evaluate allows comparison of various strategies and guides further research and improvement.

Three basic types of evaluation tasks for LLMs are defined: language generation, knowledge utilisation, and complex reasoning (Zhao et al., 2023, 29-34). Different surveys emphasise various classifications, descriptions, and features of evaluation benchmarks (Chang et al., 2024; Ni et al., 2025). Here, we briefly present some of the most important benchmarks and evaluation frameworks that assess broad reasoning, knowledge, and multitask abilities, which is also the purpose of MMLU.

2.1. SuperGLUE

GLUE benchmark (Wang et al., 2018) is a framework for evaluating general-purpose language understanding technologies. GLUE comprises nine language understanding tasks based on existing public datasets. It also includes private test data, an evaluation server, a single-number aggregate metric, and an additional diagnostic set.

SuperGLUE (Wang et al., 2019) is a benchmark designed to provide a more demanding test of general-purpose language understanding for English. SuperGLUE provides a single performance metric and an analysis toolkit, maintaining GLUE's overall design while increasing task difficulty (Wang et al., 2019).

It includes eight diverse tasks: BoolQ, CB, COPA, MultiRC, ReCoRD, RTE, WiC, and WSC, covering question answering, entailment, causal reasoning, coreference resolution, complex reading comprehension with reasoning and word-sense disambiguation.

For several languages – Chinese (Xu et al., 2020), Korean (Park et al., 2021), Indonesian (Koto et al., 2020), Russian (Shavrina et al., 2020), and Bulgarian among others (Hardalov et al., 2023) – similar benchmarks have been created. Significant efforts were made to adhere to the original GLUE's guiding principles, covering a wide range of domains and tasks and incorporating language-specific characteristics.

2.2. BIG-bench

The **BIG-bench** (Beyond the Imitation Game Benchmark) (Srivastava et al., 2023) is a benchmark designed to evaluate LLMs using challenging tasks. BIG-bench comprises over 200 tasks and human baseline performance data collected through crowdsourcing. These baselines are used to compare model and human performance against gold labels provided by the original creator of each task. The tasks are diverse, covering areas such as linguistics, childhood development, mathematics, common-sense reasoning, biology, physics, social bias, software development, and more. The benchmark tasks are novel, span a wide range of topics and languages, and are intentionally designed to be difficult for pretrained models without task-specific fine-tuning. The BIG-bench GitHub repository (Google, 2023) contains the tasks, a benchmark API that supports task evaluation on publicly available models, and evaluation results. In addition, BIG-bench Lite is available as a condensed, canonical subset of tasks, allowing quicker evaluation than the full benchmark.

2.3. Massive Multitask Language Understanding

The **MMLU** (**Massive Multitask Language Understanding**) benchmark was introduced by Hendrycks et al. (2021) and covers 57 diverse subjects across multiple domains and difficulty levels. The MMLU benchmark contains 15,908 multiple-choice questions designed to evaluate models in zero-shot and few-shot prompting, making it significantly more challenging at the time than its predecessors, such as GLUE and SuperGLUE.

The questions are divided into a few-shot Development set, a Validation set, and a Test set. The few-shot Development set contains exactly five questions per subject. For each subject, at least 100 test questions are provided. The Validation

set, which may be used for selecting hyperparameters, consists of 1,540 questions across all subjects. The subjects are divided into four main categories: 19 subjects in **STEM** (abstract algebra, astronomy, school physics, etc.), 16 subjects in **Humanities** (history, philosophy, religion, art, etc.), 6 subjects in **Social Sciences** (economics, politics, sociology, psychology, culture), and 16 subjects in the category **Professional skils**, which can be further divided into business, medicine, management, and professional skills.

Early experiments with MMLU showed that LLMs such as GPT-3 (175B parameters) achieved an accuracy of 43.9%, far below the estimated human-expert level of 89.8%, while smaller models performed worse (25%) (Hendrycks et al., 2021).

Moreover, LLMs exhibited varied performance across domains, particularly at college and professional levels. Over the years, MMLU has become established as a standard for evaluating general language understanding by LLMs, and the benchmark is now widely used for LLM evaluation.

In recent years, LLMs have demonstrated significantly improved performance on MMLU. GPT-4 achieved 86.4% accuracy on MMLU in a 5-shot setting, excelling particularly in humanities and STEM topics (Achiam et al., 2023). Gemini 1.5 Pro achieved 83.7% accuracy with standard 5-shot evaluation, while Gemini Ultra reached 90.0% using chain-of-thought reasoning combined with uncertainty routing (Gemini et al., 2023).

Open-source models also showed competitive performance, with Llama 3 70B Instruct scoring approximately 79–82% (Dubey et al., 2024), and Phi-3-medium (14B parameters), with a notably smaller size, achieving 78–82% (Abdin et al., 2024).

Frontier models consistently clustered around 86–89% accuracy on MMLU. Claude 3 Opus achieved 88.2% on MMLU (Anthropic, 2024). Gemini 2.5 Pro achieved 89.8% on Global MMLU (Gemini, 2025). DeepSeek-V3 demonstrated strong performance with 88.3% on MMLU, as did Llama 3.3 70B with 88.5% (Vidhya, 2025). The saturation effect, together with exceeding the reported human expert accuracy of 89.8% (Hendrycks et al., 2021, 3), motivated the development of more challenging variants of the MMLU benchmark.

2.4. Key Features of Benchmarks

Benchmarking practices are heavily concentrated on a relatively small number of datasets for a particular task, and many benchmark datasets are used to evaluate progress on tasks for which the data were not originally intended (Koch et al., 2021).

A developing theme in the Artificial Intelligence Index Report 2025 is the observed performance saturation across many popular benchmarks (Maslej et al., 2025, 14), which has prompted the development of new benchmarks such as Humanity's Last Exam, a multimodal academic test with broad subject coverage (Center for AI Safety et al., 2026).

Overall, most existing LLM benchmarks have several limitations: many are restricted to English, are narrow in scope, quickly reach saturation as model performance improves, and often rely on data created by non-experts or artificial data.

The limited number of existing benchmarks for Bulgarian covers only a narrow range of domains and general knowledge, as the data are collected from social media posts, Wikipedia, and news articles.

Furthermore, most benchmarks are created or translated automatically. In contrast to common practice, we rely on human translation, adaptation, and evaluation of benchmarks. We aim to investigate LLM performance in more specialised subject areas, as well as for specific tasks requiring evaluation across a broader range of subjects.

2.5. Related Work

MMLU-Pro Wang et al. (2024) substantially increased the difficulty by expanding answer choices from four to ten, thereby reducing the probability of correct guesses by random selection. The MMLU-Pro benchmark comprises 12,032 questions that emphasise reasoning rather than simple knowledge retrieval. Questions are sourced from high-quality datasets, including the original MMLU (with trivial questions removed), STEM websites, TheoremQA, and SciBench. Experimental results showed that MMLU-Pro causes a significant accuracy drop of 16–33% compared to MMLU across all tested models and demonstrates greater stability under prompt variations.

Xuan et al. (2025) extended MMLU-Pro into a multilingual benchmark (excluding Bulgarian), **MMLU-ProX**, addressing the need to evaluate LLMs across different languages and cultural contexts. MMLU-ProX introduces additional complexity through multi-hop reasoning requirements and longer-context questions that test advanced reasoning capabilities. MMLU-ProX specifically targets the evaluation of state-of-the-art LLMs by incorporating questions that require integrating information from multiple sources and performing complex logical operations. The benchmark represents a significant step towards understanding how well language models can handle sophisticated reasoning tasks across linguistic boundaries and reflects the requirements of multilingual AI applications.

Singh et al. (2024) addressed one of the main limitations of the original MMLU – its monolingual nature – by developing **Global MMLU**, which extends the original benchmark to multiple languages

(excluding Bulgarian) and incorporates cultural differences in knowledge representation and question formulation. The authors showed that translating questions often fails to capture culturally specific knowledge and may introduce biases that disadvantage models trained on data from other languages and cultures. Their work involved creating culturally adapted versions of questions and adding region-specific content that reflects diverse educational systems and cultural contexts. This investigation revealed significant performance differences between models when evaluated on culturally adapted versus directly translated questions, highlighting that benchmark design can favour or disadvantage certain linguistic and cultural backgrounds.

The only known MMLU dataset to include Bulgarian is the multilingual **MMLU-X** dataset (Thellmann et al., 2024). MMLU-X extends the original MMLU benchmark to multiple European languages through automated translation pipelines. The work showed that cross-lingual performance varies considerably across languages, highlighting that evaluation benchmarks built solely through automatic translation may not fully capture the linguistic diversity and complexity of European languages. In section 5.3 we perform a comparative analysis between the results on MMLU-BG and the automatically derived Bulgarian counterpart of MMLU-X.

To address the challenge of translating benchmarks into other languages while efficiently incorporating their unique linguistic features into the evaluated tasks, an automated translation framework was introduced (Yukhymenko et al., 2026). The aim is to enable machine translation of datasets and benchmarks with minimal manual supervision and maximum configurability, offering various methods to balance costs and time investment according to language performance in selected machine translation models. As a result, several benchmarks, including MMLU, were translated into Eastern and Southern European languages such as Ukrainian, Romanian, Slovak, Lithuanian, Bulgarian, Turkish, Greek, and Estonian. At the beginning of March 2026, this version of the Bulgarian MMLU machine-translated benchmark had not been released publicly.

3. Motivation for Bulgarian MMLU

Among the widely discussed concerns regarding the use of the MMLU benchmark are the high percentage of identified errors, the quality of translations – especially in low-resource languages, which may amplify errors and ambiguities – and the amount of culturally dependent content, which may disadvantage models for low-resource languages rather than test general reasoning ability. Gema et al. (2025) found that approximately 6.5% of questions in MMLU contain flaws, ambiguities, or factual errors, setting an artificial ceiling on model performance scores. The **MMLU-Redux** benchmark is proposed as a cleaned variant addressing these documented problems.

The decision to develop **MMLU-BG** as the evaluation benchmark for Bulgarian is motivated by the following key features:

1. Domain scope and universality: MMLU covers 57 subjects spanning universal academic knowledge (mathematics, physics, chemistry, biology, law, humanities), making it one of the benchmarks with the widest coverage in terms of knowledge areas. This allows assessment of LLM performance across various fields and helps identify the most challenging ones.

2. Standardised evaluation protocol: MMLU's multiple-choice format provides a consistent evaluation methodology, enabling objective comparison without requiring language-specific human expert evaluation or complex scoring. Moreover, this aligns with widely used human testing systems and facilitates further extensions of the benchmark.

3. Multilinguality: Global MMLU has been extended to cover 41 languages, and MML-X to 20 languages, providing standardised evaluation in diverse linguistic contexts, including low-resource languages. This enables direct comparison of model performance across various languages, including Bulgarian.

4. Multicultural extensions: Multilingual variants within the Global MMLU initiative explicitly flag items that carry cultural or political biases across languages, making it possible to identify when models perform poorly on low-resource languages due to inadequate cultural adaptation, and to build on that knowledge. This facilitates further improvements to the Bulgarian MMLU benchmark.

5. Practical implications: While proprietary and larger models are beginning to reach saturation on MMLU and are close to or surpass the estimated human-expert 89.8% accuracy, the performance of smaller models, which are typically used in low-resource language scenarios, is significantly lower. Moreover, there are very few evaluation benchmarks for Bulgarian, especially standardised ones, which help to evaluate the language-specific performance of LLMs.

6. Comparing human-translated and machine-translated benchmarks: The intuition is that advances in LLMs will reduce the differences between human and machine translation; however, there are relatively few benchmarks developed by human experts, which are based on existing English benchmarks, and the comparison between human and machine translation has not been extensively investigated. Moreover, human translation of benchmarks such as MMLU needs

to be adapted for language-specific terminology, for which LLMs may not have been sufficiently exposed during pretraining

7. Extensions: Existing extensions of MMLU provide the opportunity to further develop the Bulgarian evaluation dataset for more advanced evaluation tasks, such as reasoning and domain-specific tasks. Additionally, a possible direction is to develop a new benchmak based on the methodology of MMLU.

In summary, the development of MMLU in Bulgarian provides a reference point for evaluating LLM capabilities and analysing the specifics of language-dependent benchmarks. This will facilitate the creation of more advanced and comprehensive benchmarks for Bulgarian in the future.

4. Bulgarian MMLU

The **Bulgarian Massive Multitask Language Understanding** benchmark, or **MMLU-BG** benchmark, is based on the original **MMLU** dataset and aims to provide a new resource for evaluating state-of-the-art LLMs and comparatively ranking their performance across different subjects in Bulgarian.

4.1. Bulgarian MMLU Structure

MMLU-BG follows the structure and data organisation of the original MMLU dataset (Hendrycks et al., 2021). Each subject in MMLU-BG consists of three parts: Development, Validation, and Test, and is intended to facilitate few-shot experiments and reliably evaluate LLMs.

The current version of MMLU-BG includes **56 subjects** out of the original 57 (see Table 3). The dataset covers both general knowledge topics and language- and culture-specific ones, but excludes the large highly specialised subset of Professional Law (1,709 questions).

The total number of questions in Bulgarian is **14,093** distributed as follows: **280** questions in the Development set, **1,348** questions in the Validation set, and **12,470** questions in the Test set. The Bulgarian MMLU comprises a total of **807,294 words** of running text.

4.2. MMLU-BG Development Protocol

MMLU-BG is a human translation of the original dataset from English into Bulgarian.

A 2025 investigation based on over 2,000 multilingual benchmarks shows that the majority (61.4%) are created in their original languages and are not translated. Human translations account for 13.2% of benchmarks, representing the highest quality but also the most resource-intensive approach, while machine translation covers the remaining 25.4% (Wu et al., 2025). This distribution highlights both the prevalence of native-language benchmark development and the increasing role of various machine translation technologies in multilingual benchmark creation. Moreover, the authors show that simply translating English benchmarks is insufficient for robust evaluation. For example, localised benchmarks such as CMMLU for Chinese (Li et al., 2024) show a substantially higher correlation with human judgements (0.68) than translated equivalents (0.47 and 0.49), highlighting the need for culturally and linguistically authentic evaluation resources. In line with these conclusions, the development of the MMLU-BG focuses on high-quality human translation using terminology already accepted in Bulgarian scientific literature.

To ensure the Bulgarian language quality and terminology consistency of the translation, the following protocol for the creation of MMLU-BG has been established:

- **MMLU subjects are translated** by seven Bulgarian language experts, four of whom have high proficiency in English to Bulgarian translation. Each expert familiarises themselves as much as possible with the specific subject and its specialised terminology in both English and Bulgarian. To this end, the experts consult review articles, textbooks, terminology dictionaries, and encyclopaedias. For difficult cases, the question is transformed into a statement with the correct answer, and this string is searched for on the internet.
- **Errors previously identified** in the original MMLU Gema et al. (2025), along with some newly discovered errors, are eliminated in both English and Bulgarian as follows: **Bad Question Clarity**, concerning clarity, grammar, and sufficiency of information, and **Bad Options Clarity**, concerning unclear, similar, or irrelevant answers to the question, are either corrected or, in most cases, deleted. **Questions with No Correct Answer** are deleted. **Multiple Correct Answers** are revised so that only one correct answer remains. In cases of **Wrong Ground Truth**, when the correct answer differs from the ground truth provided in MMLU, the correct answer is given if possible; otherwise, the question is deleted. When eliminating questions, the total of 100 questions in the Test set and 5 questions in the Development set is maintained by moving a question from the Validation set to replace the deleted question if necessary. In the resulting dataset, MMLU and MMLU-BG have exactly the same number of subjects and questions.
- **Other minor changes** to the original English formulations are made solely to ensure language consistency, correct spelling, punctua-

tion, and grammatical accuracy in both questions and answers.

- **The same term or string** within a subject was translated consistently unless there were specific reasons for a different choice. Each unique answer in English, along with the questions in which it was nested, is automatically aligned with the Bulgarian translations and the corresponding cell in the table where they are placed. The experts check whether different translations are appropriate for specific reasons, for example, if the question requires defining a synonym of a term; if not, only one translation is retained in the dataset.

- **Each subject is checked** and validated by a second expert. Any errors or inconsistencies in spelling, punctuation, or grammar are corrected directly. Issues involving unclear meaning or incorrect options are marked and returned to the first expert for a final decision. A final reading by an editor has not yet been completed due to the large volume of data. Cross-validation ensures consistent application of quality criteria.

5. Experiments

Our experiments aim to determine whether **MMLU-BG** is a suitable and viable benchmark for evaluating the capabilities of contemporary LLMs, with a particular focus on Bulgarian. We are especially interested in conducting meaningful comparative analyses between Bulgarian and English, as well as between expert-translated and machine-translated benchmarks, and in making domain-specific observations and analyses.

In other words, we seek to determine whether the MMLU-BG benchmark is a reliable tool for re-evaluating models already tested with MMLU and its derivatives, as well as for testing newly developed models whose pretraining datasets include Bulgarian or which are further fine-tuned for conversations in Bulgarian. This objective informs our selection of models: models from two major LLM families whose pretraining datasets include Bulgarian; models for which Bulgarian is not listed as a supported language in the pretraining data; models specifically designed for Bulgarian; and a control model specifically designed for Ukrainian.

5.1. Description of Selected LLMs

Table 1 lists the models that are used in the Experiments with the number of parameters, whether English and Bulgarian are explicitly mentioned in pretraining, and the year of release.

Model	Params	Languages	Year
Gemma-3-27b-it	27b	EN, BG	2025
BgGPT-Gemma-2-9B-IT	9b	EN, BG	2024
MamayLM-Gemma-3-12B-IT	12b	EN, BG, UK	2025
Llama-3.1-8B-Instruct	8b	EN	2024
Llama-3.2-3B-Instruct	3b	EN	2024
Mistral-7B-Instruct-v0.3	7b	EN	2024
Qwen2.5-14B-Instruct	14b	EN	2024

Table 1: Comparison across models. Languages: BG – Bulgarian, EN – English, UK – Ukrainian.

Gemma-3-27B-IT (Google, 2025) is an instruction-tuned LLM (released in 2025) with 27 billion parameters and a 128K token context window, supporting over 140 languages, including Bulgarian. Open weights are available for both pretrained and instruction-tuned variants. Gemma-3-27B-IT reportedly achieves 67.5% accuracy on the MMLU-Pro benchmark (Gemma Team, 2025).

BgGPT-Gemma-2-9B-IT-v1.0 (INSAIT Institute, 2024) was built on Google's Gemma 2 9B open-weight models (both instruction-tuned and base variants) in 2024. It underwent continued pretraining on approximately 100 billion tokens (85 billion in Bulgarian) using the Branch-and-Merge strategy (Alexandrov et al., 2024b). No reports have been found regarding its performance on MMLU or its variants.

MamayLM-Gemma-3-12B-IT-v1.0 (INSAIT Institute, 2025) was built on Gemma 3 12B open-weight models (both instruction-tuned and base variants) in 2025. It underwent continued pretraining on a large, prefiltered Ukrainian dataset using a combination of data mixing and model merging techniques to enhance Ukrainian language and culture. No detailed, publicly released evaluation results with MMLU (or its variants) have been found.

Llama-3.1-8B-Instruct (Meta, 2024a) is an open-weight, instruction-tuned model with 8 billion parameters, optimised through supervised fine-tuning and reinforcement learning with human feedback (released in 2024). Llama-3.1-8B-Instruct supports a 128K-token context window. It was pretrained on a large multilingual corpus of approximately 15 trillion tokens; however, Bulgarian is not listed among the supported languages. The model achieves 68.21% accuracy on MMLU (5-shot) and 44.3% accuracy on MMLU-Pro, as reported by Meta.

Llama-3.2-3B-Instruct (Meta, 2024b) is an open-weight model released in 2024, developed through pruning and distillation from larger Llama 3.1 models, and optimised for specific tasks using supervised fine-tuning and reinforcement learning with human feedback. The model supports a 128K-token context window and can process up to 128K tokens in a single prompt. Its pretraining corpus contains up to 9 trillion tokens from public multilingual data sources; however, Bulgarian is not listed among the officially supported languages. In

Meta's public materials, the 3B model is reported to achieve approximately 63% accuracy on MMLU (5-shot).

Mistral-7B-Instruct-v0.3 (Mistral AI, 2024) is an open-weight model released in 2024. It was fine-tuned on publicly available instruction datasets focusing on dialogue and instruction-following tasks; however, there is no explicit mention of Bulgarian. The model supports a context window of up to 32K tokens. This version builds on the Mistral-7B-v0.3 base model, which uses grouped-query attention and sliding window attention mechanisms for efficient inference. The model achieves approximately 62.6% accuracy on the MMLU (5-shot) benchmark.

Qwen2.5-14B-Instruct (Alibaba Cloud, 2024) is an open-weight model with 14 billion parameters, released in 2024. It was pretrained on up to 18 trillion tokens and supports more than 29 languages; however, Bulgarian is not officially listed among them. The model uses an optimised transformer architecture. It supports a context window of up to 128K tokens and can generate up to 8K tokens. The model achieves 79.7% on the standard MMLU (5-shot) benchmark.

5.2. Evaluation Environment

The evaluation is conducted on an NVIDIA A100 GPU with 80 GB of VRAM. The temperature is set to the default value of 0.7. The software is written in Python and is publicly available.

The language models are all programmatically downloaded from Hugging Face. They are then loaded with PyTorch and quantised to 4 bits and 8 bits respectively using the `bitsandbytes` Python library.

The evaluation of a given model for a given subject is performed as follows. First, each question from the Test set is separated into its own prompt. The prompts do not share any context. Each prompt asks the model to provide the correct answer to a given question. An example is then shown using only the first question from the Development set, with its answers labelled by Latin letters ('A', 'B', 'C', and 'D'), followed by the correct answer. The real question is then given, with its possible answers also labelled by Latin letters.

We use a one-shot design, which is empirically determined after several experiments. An additional motivation is that instruction-following models are known to generalise from very few examples, and one-shot evaluation tests whether the model can learn the pattern from a single example.

The model is expected to return exactly one answer as an uppercase Latin letter. This is checked by first limiting the model to generate at most 20 tokens, then splitting the output text on whitespace, treating each substring as a word, and finally comparing those words against the available answer labels. The possible labels for every question are exactly 'A', 'B', 'C', and 'D'. For example, if the model produces a response containing both 'A' and 'B', the answer is considered incorrect. If it contains only 'B', but the correct answer is 'A', the answer is also incorrect. The model output is considered correct (for correct answer 'A') only if it contains 'A' and none of 'B', 'C', or 'D'. In Bulgarian, the labels are the same as in English, again using capital Latin letters 'A', 'B', 'C', or 'D'.

We evaluate model performance using accuracy by comparing each model prediction to the ground truth and scoring it as correct or incorrect depending on whether it matches the ground truth. The evaluation metric is binary accuracy, calculated as the proportion of instances where the model's decision matches the ground truth answer out of the total number of answers.

5.3. Evaluation Results

In general, the scores are lower for Bulgarian but not significantly (an overall drop of 5.3 for 4-bit quantisation and 5.4 for 8-bit quantisation). The results show only a minor difference between 4- and 8-bit quantisations (see Tables 4, 5, 6, and 7). The implementation uses optimised quantisation techniques, such as those provided by the `bitsandbytes` library, leading to the assumption – confirmed by results – that 4-bit quantisation can retain almost the same effective information as 8-bit quantisation.

Table 9 presents the detailed performance of the models (4-bit and 8-bit quantisations) in both English and Bulgarian within each subject category. Figures 1, 2, 3, and 4 show a comparison across subject categories for English and Bulgarian for the seven selected models (4-bit quantisation).

The highest scores are achieved by **Gemma-3-27b-it** for both English and Bulgarian, in both 4-bit and 8-bit quantisation (see Table 2), and across all subjects – STEM, Humanities, Social Sciences, and Professional. This is expected, as the Gemma 3 family is designed for broad, cross-domain capability and multilingual support, and the relatively large 27B model is pretrained with Bulgarian data included in its dataset. The results are consistent with previous reports for English.

Similar high scores are also achieved by **Qwen2.5-14B-Instruct** for English. However, for Bulgarian, the scores are significantly lower (a drop of 13.4 for 4-bit and 10.8 for 8-bit) compared to English, ranging from 6.9 to 16.7 across subjects. This can be explained by the fact that, although the model is multilingual, Bulgarian is not explicitly mentioned in the pretraining dataset, and even if present, the amount is negligible. For both English and Bulgarian, the model performs worse on STEM and considerably better on Social Sciences.

Model	4-bit		8-bit	
	EN	BG	EN	BG
Gemma-3-27b-it	**77.8**	**73.4**	**78.1**	**72.7**
BgGPT-Gemma-2-9B-IT-v1.0	61.3	58.1	62.7	56.1
MamayLM-Gemma-3-12B-IT-v1.0	66.9	62.5	69.2	67.6
Llama-3.1-8B-Instruct	21.8	26.0	31.5	32.8
Llama-3.2-3B-Instruct	42.2	27.9	44.3	29.2
Mistral-7B-Instruct-v0.3	55.5	43.0	56.4	41.4
Qwen2.5-14B-Instruct	75.7	62.3	75.0	64.2

Table 2: Overall average accuracy across models for 4-bit and 8-bit quantisation.

The two models developed by the INSAIT Institute – **BgGPT-Gemma-2-9B-IT**, which is specifically focused on Bulgarian, and **MamayLM-Gemma-3-12B-IT-v1.0**, which is focused on Ukrainian – also demonstrate strong performance in both languages, with a drop of 3.2 for 4-bit and 6.6 for 8-bit quantisation experiments with the first model, and a drop of 4.4 for 4-bit and 1.6 for 8-bit quantisation experiments with the second model, respectively.

Notably, MamayLM-Gemma-3-12B-IT-v1.0, whose primary specialisation is Ukrainian, maintains robust English and multilingual performance inherited from Gemma 3. The 12B model is smaller than large models (e.g., 27B) and may lag in tasks requiring advanced reasoning or knowledge recall. However, it still outperforms the Bulgarian-specialised but smaller BgGPT model, especially in 8-bit experiments. The results suggest that initial pretraining with data for a given language and the model size has a greater impact on behaviour than additional pretraining and fine- or instruction-tuning: MamayLM-Gemma-3-12B-IT-v1.0, pretrained and instruction-tuned for Ukrainian, outperforms BgGPT-Gemma-2-9B-IT-v1.0, which is pretrained and instruction-tuned for Bulgarian.

Mistral-7B-Instruct-v0.3 achieves results in English comparable to those of the BgGPT-Gemma-2-9B-IT-v1.0 model, but shows a significant decline in performance for Bulgarian (a drop of 12.5 for 4-bit quantisation and 15.0 for 8-bit quantisation). This may be due to the smaller model size and, more importantly, to the uncertain representation of Bulgarian in the pretraining data, as the language distribution in it has not been publicly reported.

The low results for Bulgarian obtained by the two models from the Llama family – **Llama-3.1-8B-Instruct** and **Llama-3.2-3B-Instruct** – are likely due to the limited representation of Bulgarian in their pretraining datasets. Regarding the lower English results compared to those reported by Meta, two factors may explain the difference. First, Meta provides a standardised evaluation setup for Llama-3.1 models using deterministic five-shot prompting for benchmarks such as MMLU. Second, in our experiments, we apply a unified evaluation pipeline across all models. Unexpectedly, the results of Llama-3.1-8B-Instruct for English are lower than for Bulgarian, as many outputs display a failure mode by generating questions or statements instead of answers.

Due to its tailored training, BgGPT-Gemma-2-9B-IT-v1.0 often outperforms larger general-purpose models (e.g. Qwen 2.5 72B, Llama 3.1 70B) on Bulgarian benchmarks (Alexandrov et al., 2024a). Despite its specialisation, the model retains much of its English and cross-lingual capability inherited from the Gemma-2 base, making it useful in bilingual and multilingual contexts. The model remains competent in reasoning, mathematics, knowledge tasks, and general-purpose language tasks, not only in Bulgarian.

The comparison between the evaluation of 4-bit quantisation models using the MMLU-X-BG machine translation (Table 8) and human expert translation (Table 5) in MMLU-BG is very interesting.

Although the overall results for MMLU-BG are higher than those for MMLU-X-BG, there are some categories and models where the machine translation achieves slightly better results. The explanation for these observations may be both linguistically and computationally grounded: some terminology may be translated correctly automatically, as certain terms in English have Bulgarian equivalents as transcriptions; the machine translation may preserve sentence structure between the source and target languages; similar subword or token patterns may occur in both source and translated data; and some machine-translated texts may be present in the pretraining data.

6. Conclusion and Future Work

In this paper, we present **MMLU-BG**, a large-scale benchmark in Bulgarian for evaluating LLMs, designed to assess general knowledge and multi-domain versatility. MMLU-BG follows the subjects and structure of MMLU and has been developed through human translation and adaptation. Its purpose is to evaluate the strengths and weaknesses of LLMs across different knowledge domains. Despite the ongoing discussion on the saturation effect of LLMs on existing benchmarks, MMLU-BG can be used to evaluate new models among the rapidly growing number of LLMs.

Additionally, **MMLU-BG** aims to provide a linguistically authentic evaluation resource aligned with terminology already accepted in Bulgarian. Benchmarks focusing on STEM-related capabilities have consistently shown stronger correlations with human judgements across languages, while tasks requiring understanding of specific linguistic nuances or cultural contexts may be more sensitive to

translation errors, resulting in weaker correlations with human judgements (Wu et al., 2025). In line with this, we believe that expert translation is better aligned with human judgements, especially for certain subjects.

A next step in our research is to use existing extensions of MMLU to further expand and improve the MMLU-BG benchmark for more advanced evaluation tasks, such as reasoning and domain-specific assessments.

Additionally, as the original MMLU lacks Bulgarian language- and culture-specific knowledge and may be subject to English cultural biases due to its content, our efforts are also aimed at addressing these potential shortcomings. In its current form, as a translation and adaptation of the original MMLU, **MMLU-BG** provides a viable basis for evaluating English language- and culture-specific features, which may inform subsequent efforts to expand the dataset with subjects and questions specific to Bulgarian language and culture. Therefore, an immediate goal is to extend the dataset with other subjects widely covered in the Bulgarian educational system or related to traditional Bulgarian culture.

Another direction for future research inspired by the current study is to analyse the sensitivity of MMLU-BG to the increasing capabilities of LLMs (including smaller models), and to address the need for more advanced evaluation benchmarks intended for specific subjects and tasks.

7. Acknowledgments

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

8. Ethics Statement

This research involves the compilation of an MMLU-based benchmark in Bulgarian. The work uses the translation of a preexisting and widely known dataset and does not involve collecting human subject data or generating potentially harmful content.

All LLMs evaluated in this study are open source or publicly accessible, and their use complies with the respective licences.

9. Limitations

The benchmark MMLU-BG is a translation and adaptation of the original MMLU. It may inherit potential errors and biases – both linguistic and cultural – that could affect performance evaluation, as discussed in the paper. Additionally, it lacks language- and culture-specific knowledge necessary for a representative evaluation focused on Bulgarian.

The evaluation is also limited to seven LLMs, which may not be sufficient to establish comprehensive performance baselines across the full range of available LLMs for Bulgarian. However, the selected models represent a diverse range of sizes, architectures, and training data languages, enabling meaningful comparative analysis across different model families and approaches.

Any observations and experience gained from the experiments presented here will support the development of a more reliable evaluation benchmark.

10. Bibliographical References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

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

Anton Alexandrov, Veselin Raychev, Mark Niklas Müller, Ce Zhang, Martin Vechev, and Kristina Toutanova. 2024b. Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17167–17186, Miami, Florida, USA. Association for Computational Linguistics.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Report.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. 2026. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and

Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are We Done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Google DeepMind Gemini. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Gemma Team. 2025. Gemma 3 Technical Report. *Google DeepMind*.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, pages 1–27. OpenReview.net.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. *arXiv preprint arXiv:2112.01716*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. *The AI Index 2025 Annual Report*. Stanford, CA.

Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. 2025. Evaluation and Benchmarking of LLM Agents: A Survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6129–6139, New York, NY, USA. Association for Computing Machinery.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. A Survey on Large Language Model Benchmarks. *arXiv preprint arXiv:2508.15361*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean Language Understanding Evaluation. *arXiv preprint arXiv:2105.09680*.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre FT Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. *arXiv preprint arXiv:2412.03304*.

Aarohi Srivastava et al. 2023. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Transactions on Machine Learning Research*.

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 preprint arXiv:2410.08928*.

Analytics Vidhya. 2025. DeepSeek-V3 vs GPT-4o vs Llama 3.3 70B – Let the Best AI Model Win.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Track on Datasets and Benchmarks*.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. *arXiv preprint arXiv:2504.15521*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2026. Recovered in Translation: Efficient Pipeline for Automated Translation of Benchmarks and Datasets. *arXiv preprint arXiv:2602.22207*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*.

11. Language Resource References

Alibaba Cloud. 2024. *Qwen2.5-14B-Instruct*. HuggingFace. PID https://huggingface.co/Qwen/Qwen2.5-14B-Instruct.

Google. 2023. *BIG-bench*. HuggingFace. PID https://github.com/google/BIG-bench.

Google. 2025. *Gemma-3-27B-IT*. HuggingFace. PID https://huggingface.co/google/gemma-3-27b-it.

INSAIT Institute. 2023. *BgGPT Evaluation Suite*. HuggingFace. PID https://github.com/insait-institute/lm-evaluation-harness-bg.

INSAIT Institute. 2024. *BgGPT-Gemma-2-9B-IT-v1.0*. HuggingFace. PID https://huggingface.co/INSAIT-Institute/BgGPT-Gemma-2-9B-IT-v1.0.

INSAIT Institute. 2025. *MamayLM-Gemma-3-12B-IT-v1.0*. HuggingFace. PID https://huggingface.co/INSAIT-Institute/MamayLM-Gemma-3-12B-IT-v1.0.

Meta. 2024a. *Llama-3.1-8B-Instruct*. HuggingFace. PID https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Meta. 2024b. *Llama-3.2-3B-Instruct*. HuggingFace. PID https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

Mistral AI. 2024. *Mistral-7B-Instruct-v0.3*. HuggingFace. PID https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Wang et al. 2019. *General Language Understanding Evaluation*. PID https://super.gluebenchmark.com/.

Appendix

Subject	Test	Val	Dev	#Qs	#Words
STEM (19 subjects, 3,606 questions, 159.6K words)					
Abstract Algebra	100	5	11	116	3.8K
Anatomy	135	5	14	154	4.7K
Astronomy	152	5	16	173	8.3K
College Biology	144	5	16	165	7.4K
College Chemistry	100	5	8	113	4.3K
College Computer Science	100	5	11	116	8.2K
College Mathematics	100	5	11	116	5.3K
College Physics	102	5	11	118	5.5K
Conceptual Physics	235	5	26	266	7.0K
Elementary Mathematics	378	5	41	424	13.5K
High School Biology	310	5	32	347	17.6K
High School Chemistry	203	5	22	230	10.1K
High School Computer Science	100	5	9	114	7.8K
High School Mathematics	270	5	29	304	9.8K
High School Physics	151	5	17	173	10.5K
High School Statistics	216	5	23	244	19.1K
Machine Learning	112	5	11	128	6.1K
Medical Genetics	100	5	11	116	3.4K
Virology	166	5	18	189	7.4K
Humanities (15 subjects, 4,134 questions, 338.4K words)					
Formal Logic	126	5	14	145	9.4K
High School European History	165	5	18	188	46.7K
High School World History	236	5	26	267	54.4K
High School Geography	198	5	22	225	6.4K
High School Government and Politics	193	5	21	219	10.6K
High School US History	204	5	22	231	51.9K
Logical Fallacies	163	5	18	186	7.2K
Moral Disputes	346	5	38	389	17.3K
Moral Scenarios	895	5	100	1000	84.3K
Philosophy	311	5	34	350	12.6K
Prehistory	324	5	35	364	14.8K
World Religions	171	5	19	195	3.6K
Global Facts	100	5	10	115	3.7K
International Law	119	5	12	136	9.8K
Jurisprudence	108	5	11	124	5.7K
Social Sciences (6 subjects, 1,808 questions, 76.8K words)					
High School Macroeconomics	390	5	43	438	19.8K
High School Microeconomics	238	5	26	269	13.2K
High School Psychology	545	5	60	610	24.4K
Human Sexuality	131	5	12	148	5.2K
Sociology	200	5	22	227	9.8K
US Foreign Policy	100	5	11	116	4.4K
Professional Skills (16 subjects, 4,550 questions, 232.6K words)					
Business Ethics	100	5	11	116	5.1K
Clinical Knowledge	265	5	29	299	10.9K
College Medicine	172	5	22	199	11.9K
Computer Security	100	5	11	116	5.0K
Econometrics	114	5	12	131	7.8K
Electrical Engineering	145	5	16	166	4.4K
Human Aging	223	5	23	251	8.3K
Management	103	5	11	119	2.9K
Marketing	234	5	25	264	9.4K
Miscellaneous	753	5	75	833	22.0K
Nutrition	306	5	33	344	15.0K
Professional Accounting	282	5	31	318	22.9K
Professional Medicine	272	5	31	308	35.3K
Professional Psychology	610	5	68	683	34.1K
Public Relations	110	5	12	127	5.0K
Security Studies	244	5	27	276	32.5K

Table 3: MMLU-BG by subject with number of questions (Test, Val, Dev, #Qs) and number of words.

Model	STEM	Hum.	Soc.	Prof.	Avrg.
Gemma-3-27b-it	**70.7**	**78.3**	**86.8**	**79.4**	**77.8**
BgGPT-Gemma-2-9B-IT-v1.0	51.7	61.5	77.6	62.2	61.3
MamayLM-Gemma-3-12B-IT-v1.0	59.4	69.7	77.0	66.4	66.9
Llama-3.1-8B-Instruct	17.3	21.7	23.8	24.8	21.8
Llama-3.2-3B-Instruct	34.7	43.5	48.4	44.4	42.2
Mistral-7B-Instruct-v0.3	43.9	55.4	64.9	61.1	55.5
Qwen2.5-14B-Instruct	67.6	78.0	85.0	76.1	75.7

Table 4: Model performance (accuracy) by subject category – MMLU-EN, 4-bit quantisation. Hum. – Humanities; Soc. – Social Sciences; Prof. – Professional Skills; Avrg. – Average.

Model	STEM	Hum.	Soc.	Prof.	Avrg.
Gemma-3-27b-it	**68.8**	**72.9**	**81.9**	**74.0**	**73.4**
BgGPT-Gemma-2-9B-IT-v1.0	51.5	56.3	69.2	60.4	58.1
MamayLM-Gemma-3-12B-IT-v1.0	54.8	60.5	73.9	66.0	62.5
Llama-3.1-8B-Instruct	22.2	25.4	29.1	28.3	26.0
Llama-3.2-3B-Instruct	26.4	27.6	27.5	29.4	27.9
Mistral-7B-Instruct-v0.3	36.9	43.2	47.7	45.6	43.0
Qwen2.5-14B-Instruct	58.5	61.3	70.6	62.8	62.3

Table 5: Model performance (accuracy) by subject category – MMLU-BG, 4-bit quantisation.

Model	STEM	Hum.	Soc.	Prof.	Avrg.
Gemma-3-27b-it	**70.9**	**78.2**	**86.8**	**80.2**	**78.1**
BgGPT-Gemma-2-9B-IT-v1.0	54.4	62.9	77.2	63.3	62.7
MamayLM-Gemma-3-12B-IT-v1.0	60.5	69.5	80.9	71.2	69.2
Llama-3.1-8B-Instruct	28.1	32.3	33.0	32.7	31.5
Llama-3.2-3B-Instruct	36.9	44.2	49.2	48.2	44.3
Mistral-7B-Instruct-v0.3	43.9	57.5	65.4	61.6	56.4
Qwen2.5-14B-Instruct	68.6	77.2	82.7	75.0	75.0

Table 6: Model performance (accuracy) by subject category – MMLU-EN, 8-bit quantisation.

Model	STEM	Hum.	Soc.	Prof.	Avrg.
Gemma-3-27b-it	**69.2**	**71.5**	**82.2**	**72.8**	**72.7**
BgGPT-Gemma-2-9B-IT-v1.0	49.6	54.3	67.3	58.2	56.1
MamayLM-Gemma-3-12B-IT-v1.0	60.1	68.4	77.4	68.7	67.6
Llama-3.1-8B-Instruct	26.7	31.9	40.8	35.1	32.8
Llama-3.2-3B-Instruct	27.1	28.6	28.4	31.6	29.2
Mistral-7B-Instruct-v0.3	34.5	42.9	48.8	42.6	41.4
Qwen2.5-14B-Instruct	61.7	63.9	69.8	64.1	64.2

Table 7: Model performance (accuracy) by subject category – MMLU-BG, 8-bit quantisation

Model	STEM	Hum.	Soc.	Prof.	Avrg.
Gemma-3-27b-it	**68.7**	**64.9**	**82.8**	**75.2**	**71.5**
BgGPT-Gemma-2-9B-IT-v1.0	50.8	54.3	66.7	62.1	57.5
MamayLM-Gemma-3-12B-IT-v1.0	54.8	54.0	73.1	64.4	60.0
Llama-3.1-8B-Instruct	22.4	27.5	29.8	27.6	26.5
Llama-3.2-3B-Instruct	26.8	28.2	28.8	30.4	28.6
Mistral-7B-Instruct-v0.3	35.1	42.9	48.0	46.1	42.6
Qwen2.5-14B-Instruct	58.1	57.8	68.8	63.1	61.0

Table 8: Model performance (accuracy) by subject category – MMLU-X-BG, 4-bit quantisation.

Model	Gemma-3		BgGPT		MamayLM		Llama-3.1		Llama-3.2		Mistral		Qwen-2.5	
Subject	EN	BG	EN	BG	EN	BG	EN	BG	EN	BG	EN	BG	EN	BG
STEM														
Abstract Algebra	49.0	51.0	37.0	29.0	36.0	36.0	14.0	22.0	33.0	21.0	25.0	34.0	60.0	53.0
Anatomy	68.1	61.5	64.4	52.6	65.9	61.5	26.7	21.5	44.4	23.0	50.4	43.0	66.7	51.9
Astronomy	84.9	79.6	67.1	69.7	73.7	55.9	24.3	17.1	41.4	25.7	54.6	37.5	80.3	75.0
College Biology	86.8	82.6	81.9	70.8	85.4	77.1	24.3	31.9	50.7	27.1	61.1	47.2	84.0	68.1
College Chemistry	56.0	60.0	47.0	39.0	52.0	46.0	19.0	21.0	32.0	23.0	42.0	32.0	44.0	48.0
College Computer Science	56.0	55.0	40.0	42.0	56.0	51.0	25.0	27.0	32.0	26.0	41.0	30.0	60.0	44.0
College Mathematics	55.0	51.0	25.0	32.0	40.0	35.0	8.0	21.0	26.0	24.0	25.0	23.0	47.0	49.0
College Physics	59.8	56.9	33.3	32.4	38.2	42.2	10.8	19.6	19.6	25.5	35.3	18.6	48.0	39.2
Conceptual Physics	80.4	75.3	57.4	58.3	63.8	69.8	13.6	23.0	39.1	24.7	44.3	37.9	77.4	62.1
Elementary Mathematics	77.5	78.0	43.1	50.0	53.7	54.8	8.2	18.5	24.1	30.7	35.7	29.6	73.5	66.1
High School Biology	87.1	85.2	78.7	76.1	81.9	70.6	25.2	28.7	36.8	26.1	64.8	56.1	84.8	73.5
High School Chemistry	71.4	71.9	58.6	48.3	64.5	57.1	15.8	19.7	37.4	27.6	39.4	41.9	67.0	54.7
High School Computer Science	84.0	82.0	62.0	65.0	79.0	78.0	25.0	26.0	50.0	34.0	58.0	61.0	82.0	79.0
High School Mathematics	54.1	55.2	27.4	35.6	43.7	43.0	7.0	14.1	30.7	21.1	25.2	23.7	57.4	50.4
High School Physics	62.3	62.3	29.1	45.0	29.1	50.3	17.9	21.2	31.8	25.2	29.1	31.8	54.3	45.0
High School Statistics	70.4	66.2	49.1	48.6	56.5	25.9	25.5	27.8	30.1	26.9	43.1	31.5	67.1	56.5
Machine Learning	59.8	56.2	45.5	40.2	57.1	50.0	17.0	19.6	28.6	25.9	42.0	30.4	58.0	48.2
Medical Genetics	86.0	80.0	78.0	63.0	79.0	81.0	22.0	38.0	53.0	39.0	73.0	46.0	84.0	68.0
Virology	56.6	56.6	45.2	47.6	56.6	48.2	13.9	13.9	35.5	26.5	49.4	41.6	48.8	48.2
Humanities														
Formal Logic	58.7	53.2	47.6	37.3	53.2	52.4	19.8	24.6	27.0	23.0	33.3	34.9	57.1	47.6
High School European History	85.5	80.6	78.8	75.8	77.6	25.5	64.2	55.8	59.4	35.2	61.8	48.5	79.4	76.4
High School World History	89.8	75.4	83.1	68.6	84.7	70.3	53.4	42.4	46.2	31.4	63.1	54.2	88.1	69.1
High School Geography	86.9	83.8	77.8	72.7	81.3	81.3	24.7	18.7	30.8	28.3	67.7	58.6	86.4	73.2
High School Government And Politics	95.3	87.0	88.6	77.7	88.1	78.8	17.6	23.8	50.8	10.9	80.3	45.6	92.7	70.5
High School Us History	88.2	83.3	83.3	77.0	86.8	17.6	67.6	48.5	59.3	33.8	72.1	47.5	89.7	69.1
Logical Fallacies	85.9	81.6	66.9	66.9	74.8	69.3	19.6	25.2	48.5	33.1	61.3	44.8	83.4	73.6
Moral Disputes	80.3	72.8	66.8	57.5	78.0	67.6	14.2	21.4	40.2	24.0	62.1	48.6	78.0	62.7
Moral Scenarios	62.9	62.0	28.3	33.3	43.4	54.6	5.8	7.8	28.5	24.2	30.2	28.3	67.3	46.7
Philosophy	81.0	68.5	66.6	58.5	76.2	66.2	12.9	27.7	53.7	31.8	60.1	50.2	77.8	61.4
Prehistory	87.0	76.2	68.5	58.6	79.3	67.0	14.8	34.9	56.5	32.1	65.4	43.2	84.6	62.0
World Religions	87.1	85.4	82.5	61.4	82.5	71.3	18.1	18.7	67.8	37.4	76.0	53.8	84.8	66.1
Global Facts	49.0	51.0	41.0	34.0	48.0	42.0	3.0	16.0	26.0	23.0	30.0	24.0	54.0	53.0
International Law	85.7	81.5	74.8	70.6	83.2	73.1	26.1	49.6	53.8	29.4	72.3	59.7	86.6	68.9
Jurisprudence	81.5	84.3	71.3	67.6	78.7	73.1	26.9	32.4	39.8	22.2	63.9	46.3	78.7	71.3
Social Sciences														
High School Macroeconomics	81.5	75.9	70.3	60.8	72.8	63.6	12.6	27.7	40.3	22.8	50.0	34.4	80.3	60.8
High School Microeconomics	87.0	81.1	71.8	64.7	77.7	79.8	22.7	30.3	43.7	29.0	51.7	39.5	86.6	74.4
High School Psychology	90.5	87.2	81.8	78.3	76.7	79.8	33.0	34.9	56.9	27.0	75.4	56.0	87.0	76.1
Human Sexuality	85.5	78.6	79.4	61.1	73.3	64.9	16.0	7.6	42.0	26.0	61.8	48.1	82.4	67.9
Sociology	85.0	79.5	82.5	69.0	82.5	72.5	28.5	27.0	53.5	36.0	78.5	53.5	85.0	72.0
Us Foreign Policy	92.0	87.0	84.0	74.0	87.0	83.0	20.0	32.0	44.0	30.0	74.0	62.0	92.0	70.0
Professional Skills														
Business Ethics	82.0	73.0	68.0	66.0	72.0	63.0	9.0	13.0	53.0	27.0	53.0	47.0	71.0	73.0
Clinical Knowledge	80.8	74.7	67.5	57.7	71.3	69.1	12.8	35.5	24.9	23.4	63.0	51.7	81.1	66.4
College Medicine	73.3	66.3	62.8	55.2	68.0	61.6	24.4	23.8	36.6	23.3	58.1	47.1	73.3	58.1
Computer Security	78.0	79.0	72.0	73.0	75.0	69.0	21.0	37.0	48.0	32.0	68.0	57.0	77.0	68.0
Econometrics	56.1	57.9	49.1	40.4	58.8	51.8	17.5	18.4	20.2	23.7	40.4	32.5	63.2	56.1
Electrical Engineering	69.7	70.3	59.3	49.0	60.0	62.1	14.5	19.3	46.2	33.1	48.3	33.1	67.6	56.6
Human Aging	80.3	75.3	66.8	58.3	74.0	69.1	15.7	36.3	49.3	37.2	61.0	46.2	74.9	58.7
Management	84.5	79.6	75.7	66.0	79.6	73.8	19.4	14.6	46.6	35.9	72.8	61.2	81.6	67.0
Marketing	94.0	87.2	83.3	80.8	83.8	81.6	7.7	27.4	50.4	37.6	83.3	65.0	89.3	76.9
Miscellaneous	91.0	83.4	81.1	68.3	86.1	76.1	19.3	17.7	62.9	30.5	77.6	56.0	86.9	66.8
Nutrition	75.5	73.9	69.3	61.1	64.1	69.6	31.7	36.9	47.7	29.4	55.2	44.8	75.2	64.1
Professional Accounting	63.1	56.4	40.1	37.9	43.6	24.1	25.5	29.1	28.0	30.5	39.7	35.5	59.6	45.4
Professional Medicine	74.3	69.1	18.4	57.7	16.9	65.4	40.1	36.8	28.3	26.1	55.1	33.8	64.3	56.6
Professional Psychology	79.7	72.0	46.6	61.0	56.6	65.1	36.6	29.8	41.3	27.5	55.2	32.6	77.4	65.2
Public Relations	70.0	74.5	69.1	67.3	67.3	65.5	18.2	23.6	50.0	24.5	60.0	56.4	64.5	65.5
Security Studies	79.1	72.5	70.1	55.3	79.9	68.9	46.3	45.9	46.3	27.9	56.1	42.6	74.2	56.6

Table 9: Accuracy per subject for 4-bit quantisation. Models: Gemma-3 = Gemma-3-27b-it; BgGPT = BgGPT-Gemma-2-9B-IT-v1.0; MamayLM = MamayLM-Gemma-3-12B-IT-v1.0; Llama-3.1 = Llama-3.1-8B-Instruct; Llama-3.2 = Llama-3.2-3B-Instruct; Mistral = Mistral-7B-Instruct-v0.3; Qwen-2.5 = Qwen2.5-14B-Instruct.

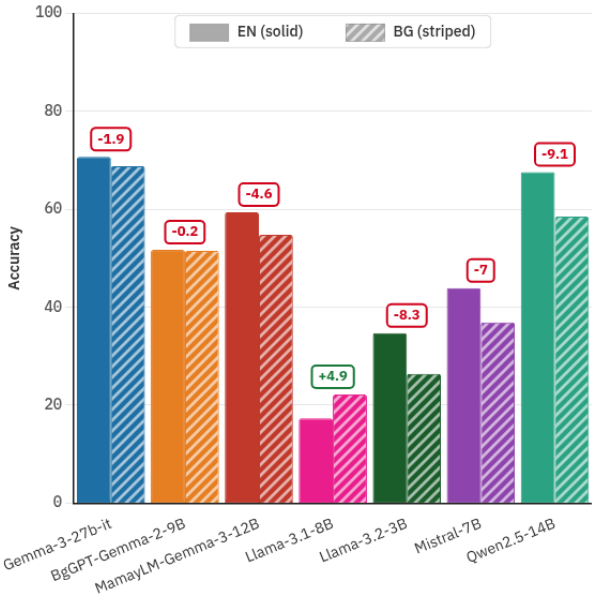

Figure 1: Comparative results by model and by language for STEM subjects – MMLU-EN and MMLU-BG (4-bit quantisation).

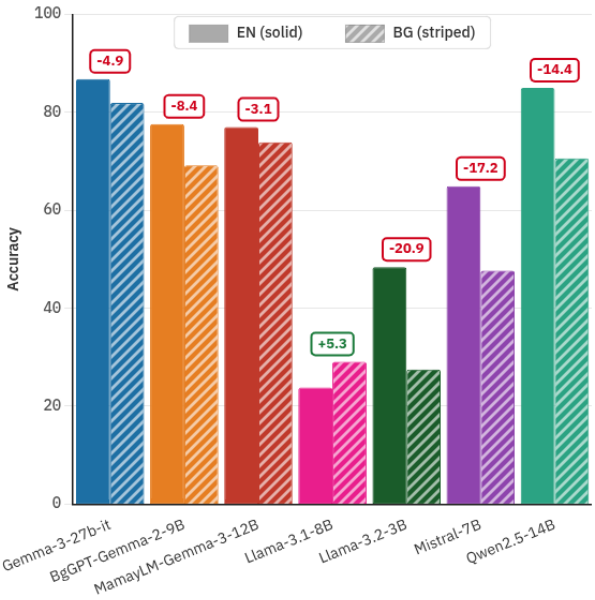

Figure 3: Comparative results by model and by language for Social Sciences subjects – MMLU-EN and MMLU-BG (4-bit quantisation).

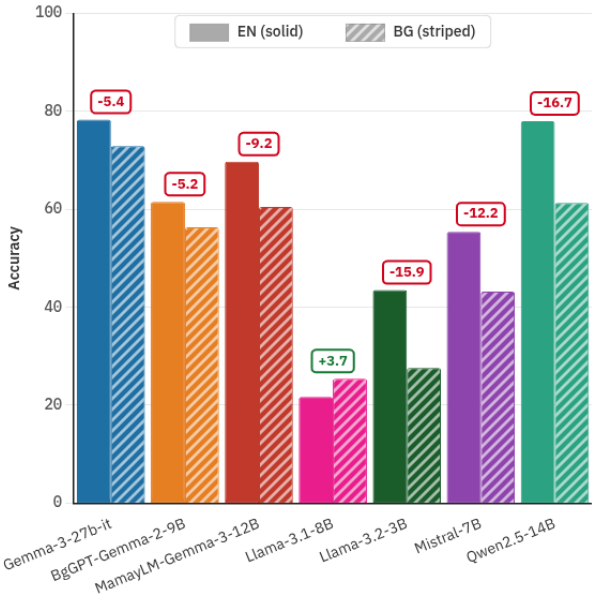

Figure 2: Comparative results by model and by language for Humanities subjects – MMLU-EN and MMLU-BG (4-bit quantisation).

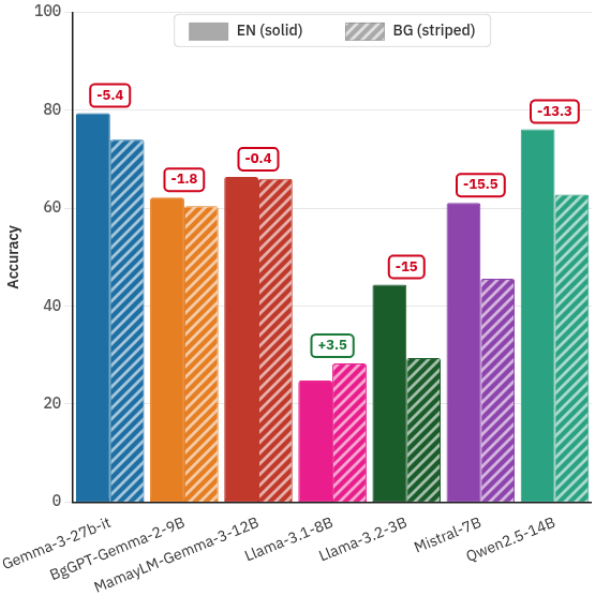

Figure 4: Comparative results by model and by language for Professional subjects – MMLU-EN and MMLU-BG (4-bit quantisation).