GreekMMLU: A Native-Sourced Multitask Benchmark for Evaluating Language Models in Greek

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are commonly trained on multilingual corpora that include Greek, yet reliable evaluation benchmarks for Greek—particularly those based on authentic, native-sourced content—remain limited. Existing datasets are often machine-translated from English, failing to capture Greek linguistic and cultural characteristics. We introduce GreekMMLU[1], a native-sourced benchmark for massive multitask language understanding in Greek, comprising 19,876 multiple-choice questions across 44 subject areas, organized under a newly defined subject taxonomy and annotated with educational difficulty levels spanning primary to professional examinations. All questions are sourced or authored in Greek from academic, professional, and governmental exams. We publicly release 15,391 samples and reserve 4,485 samples for a private leaderboard to enable robust and contamination-resistant evaluation. Evaluations of over 80 open- and closed-source LLMs reveal substantial performance gaps between frontier and open-weight models, as well as between Greek-adapted models and general multilingual ones. Finally, we provide a systematic analysis of factors influencing performance—including model scale, adaptation, and prompting—and derive insights for improving LLM capabilities in Greek.

1 Introduction

Large language models have achieved strong performance across a wide range of natural language understanding and reasoning tasks, largely driven by large-scale training on multilingual corpora (Brown et al., 2020; Grattafiori et al., 2024; Üstün et al., 2024). Contemporary models are designed to support dozens or even hundreds of languages, a choice driven by the need to maximize data scale and leverage cross-lingual transfer to improve general reasoning capabilities.

[1]Dataset link will be provided upon acceptance.

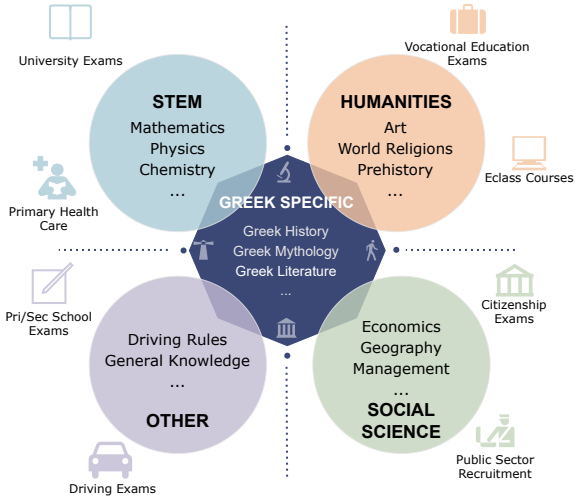

Figure 1: GreekMMLU task overview.

Furthermore, recent research has increasingly emphasized extending LLM capabilities to lower-resource languages (Voukoutis et al., 2024; Roussis et al., 2025; Martins et al., 2025; Shang et al., 2025a,b). In practice, while Greek is often present in the long tail of web-scale training corpora (Grattafiori et al., 2024), it is rarely prioritized as a core language, resulting in significantly lower representation compared to major European languages. Despite this incidental exposure, relatively little work systematically reports LLM performance on Greek, largely due to the absence of large-scale, native-language evaluation benchmarks. Existing evaluations are commonly based on machine-translated datasets originally designed for English (Xuan et al., 2025; Voukoutis et al., 2024), which fail to capture the linguistic and cultural characteristics of authentic Greek language use.

A widely adopted benchmark for assessing broad knowledge and reasoning is the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2020), which evaluates models

across diverse subjects including STEM, social science and humanities. While MMLU has become a cornerstone of LLM evaluation, it is inherently grounded in English and the educational and cultural context of the United States. Extending MMLU to other languages through translation introduces well-known limitations, including translationese, semantic drift, altered difficulty calibration, and the preservation of source-language cultural priors (Singh et al., 2025; Artetxe et al., 2020; Li et al., 2024). As a result, translated benchmarks conflate native language understanding with cross-lingual transfer and often provide a misleading assessment of model capabilities.

These limitations are particularly salient for Greek. Modern Greek exhibits rich morphology, flexible word order, and complex negation, all of which are difficult to evaluate through translated data. Furthermore, many knowledge-intensive domains—such as history, law, civics, and professional certification—are tightly coupled to national curricula, legal frameworks, and cultural context. Evaluating such domains using translated English benchmarks obscures important gaps in localized knowledge and overestimates real-world utility.

To address these issues, we introduce GreekMMLU, the first large-scale, native-sourced benchmark for evaluating massive multitask language understanding in Greek. GreekMMLU consists of 19,876 Multiple-Choice Questions (MCQ) across 44 manually defined subject areas spanning STEM, humanities, social sciences, and other domains, drawn from authentic academic, professional, and governmental examinations. All materials were collected exclusively from sources explicitly released under open-access licenses or educational reuse terms, ensuring ethical data use and legal compliance. The benchmark covers a wide range of educational and difficulty levels, including primary school, secondary school, university, and professional level, and includes eight Greek Specific subject areas requiring localized cultural knowledge. All questions are originally sourced or authored in Greek, preserving linguistic nuance and culturally grounded content.

Our contributions are summarized as follows:

• We introduce **GreekMMLU**, the first large-scale, fully native-sourced benchmark. It comprises 19,876 multiple-choice questions, organized under a **carefully defined subject taxonomy** with 44 subjects and **systematically annotated with educational difficulty levels** ranging from primary education to professional examinations.

• We conduct a large-scale evaluation of **80+ open- and closed-source LLMs** on GreekMMLU, revealing clear and consistent trends: closed-source frontier models substantially outperform open-weight alternatives, model scale strongly correlates with performance and general-purpose multilingual models exhibit pronounced weaknesses on Greek-specific and culturally grounded subject areas.

• We provide an in-depth analysis of factors influencing performance on native Greek language understanding, examining the effects of model scale, instruction tuning, prompting strategies, subject domains, and educational levels. Our findings show that Greek-adapted training leads to significant and systematic gains, particularly on tasks requiring culturally grounded knowledge.

2 Related Work

Language Models in Greek Greek remains underrepresented in large-scale language modeling, as most multilingual LLMs include it only implicitly and without explicit tokenizer design or data balancing, resulting in weaker performance compared to high-resource languages (Chowdhery et al., 2023; Touvron et al., 2023). While models such as Qwen (Yang et al., 2025) and Gemma (Team et al., 2025) can perform Greek tasks, they do not explicitly report Greek training data and performance. Earlier multilingual models like BLOOMZ and mT0 (Muennighoff et al., 2023) explicitly included Greek, but only at a very limited scale (around 0.03%). More recent efforts have addressed this gap through targeted pretraining. EuroLLM (Martins et al., 2025) increased coverage of European languages, including Greek. (Voukoutis et al., 2024) introduced *Meltemi*, the first openly released Greek-centric LLM, trained with large-scale Greek corpora and a Greek-aware tokenizer, achieving substantial gains on Greek benchmarks. Building on this work, (Roussis et al., 2025) proposed *LLaMA-Krikri*, further expanding Greek coverage through increased Greek data. Beyond general-purpose models, domain-specific efforts such as *Plutus-8B* demonstrate the benefits of Greek-adapted training in specialized settings (Peng et al., 2025). Overall, these works highlight the importance of explicit Greek-focused training for robust Greek language understanding.

LLM Evaluation and Multilingual Benchmarks General knowledge and reasoning in LLMs are

Σύμφωνα με την ελληνική μυθολογία, ποιες είναι οι θεότητες των Τεχνών και των Γραμμάτων;

Α. Νύμφες
Β. Χάριτες
Γ. Ερινύες
Δ. Μούσες

According to Greek mythology, which deities are associated with the Arts and Letters?

A. Nymphs
B. Charites
C. Erinyes
D. Muses

Figure 2: Example of a mythology question from GreekMMLU. **Left** shows the content structure derived from native sources and **right** is the English translation. The bold options represent the correct answer keys.

commonly evaluated using multitask benchmarks such as MMLU (Hendrycks et al., 2020), along with datasets like ARC (Clark et al., 2018) and HellaSwag (Zellers et al., 2019). While effective for tracking architectural and scaling progress, benchmarks like MMLU are predominantly English-centric, limiting their ability to assess culturally grounded language understanding. Multilingual extensions often rely on machine translation, including multilingual variants of MMLU (Xuan et al., 2025), but prior work has shown that translated benchmarks suffer from translationese, semantic drift, altered difficulty calibration, and inherited cultural priors, leading to validity concerns (Singh et al., 2025; Artetxe et al., 2020; Li et al., 2024). To address these issues, native-sourced benchmarks such as CMMLU for Chinese (Li et al., 2024), ArabicMMLU for Arabic (Koto et al., 2024), and TurkishMMLU for Turkish (Yüksel et al., 2024) have been proposed, highlighting the need for locally grounded multilingual evaluation.

Greek LLM Evaluation Evaluation of Greek language capabilities has traditionally relied on multilingual benchmarks such as XNLI (Conneau et al., 2018), XTREME (Hu et al., 2020), XQuAD (Artetxe et al., 2020), MASSIVE (FitzGerald et al., 2023), and FLORES-101 (Goyal et al., 2022), where Greek is included as a target language to assess cross-lingual generalization. More recent Greek-centric efforts, such as Meltemi (Voukoutis et al., 2024) and KriKri (Roussis et al., 2025), largely resorted to machine-translated versions of English benchmarks like HellaSwag and MMLU. Several Greek-specific benchmarks enable a more targeted evaluation. GreekSUM (Evdaimon et al., 2024) introduced the first large-scale abstractive summarization dataset for Greek news. Earlier resources such as elNER (Bartziokas et al., 2020), support named entity recognition in Greek, and recent domain-specific benchmarks like Plutus-ben (Peng et al., 2025) and GreekBarBench (Chlapanis et al., 2025) extend evaluation to financial and legal reasoning tasks. However, these efforts are typically task-specific and limited in scale, reflecting a need for dedicated benchmarks that better capture the linguistic and cultural specific challenges in Greek.

3 The GreekMMLU Dataset

3.1 Overview

GreekMMLU is a **native-sourced benchmark** for evaluating massive multitask language understanding in Greek, composed exclusively of original Greek questions drawn from real-world educational and professional assessments. All questions follow an MCQ format as illustrated in Figure 2, with a variable number of answer options (2–4) and exactly one correct answer, reflecting the diversity of Greek national examinations. Each question is annotated with a difficulty level corresponding to its educational context.

A major effort in creating GreekMMLU involved the systematic structuring of heterogeneous raw exam material. We designed a **custom subject taxonomy** and **carefully assigned each task to an educational difficulty level**, enabling consistent analysis across domains and degrees of specialization. The resulting benchmark spans 44 subject areas, organized into four high-level categories—**STEM, Humanities, Social Sciences, and Other** as shown in Figure 1, covering educational levels from **primary school to university and professional examinations**.

Beyond general academic content, GreekMMLU includes a dedicated subset of **Greek-specific tasks** that require explicitly Greek linguistic and cultural knowledge, such as Greek History, Greek Literature, Greek Mythology, Greek Traditions, and the Modern Greek Language. These tasks are tightly coupled to the local cultural context and cannot be

Group	Subjects
Humanities	Art (S, U, Pr), Greek History (P, S, Pr), Greek Literature (S, U), Greek Mythology (P, S, U), Law (S, Pr), Prehistory (P), World History (P, S), World Religions (S, U)
STEM	Agriculture (U, Pr), Biology (P, S), Chemistry (P, U), Civil Engineering (Pr), Clinical Knowledge (Pr), Computer Networks & Security (U), Computer Science (U, Pr), Electrical Engineering (U, Pr), Mathematics (P, U), Medicine (U, Pr), Physics (P, U, Pr)
Social Sciences	Economics (U, Pr), Education (U, Pr), Geography (P, S), Government and Politics (P, S), Greek Traditions (S, Pr), Management (U, Pr), Modern Greek Language (P, S),Accounting (Pr)
Other	Driving Rules (NA), General Knowledge (S, Pr)

Table 1: Subject areas in GreekMMLU. "P", "S", "U", "Pr", and "NA" indicate availability in primary school, secondary school, university, professional, and not available categories, respectively.

reliably assessed through translated benchmarks, highlighting the importance of native-sourced evaluation.

The dataset is divided into a **public, open-source subset** released for research use and a **private subset** reserved for a leaderboard [2] to support more robust and contamination-resistant evaluation. Representative examples of the dataset are provided in Appendix A.

3.2 Data Collection and Curation

We conducted an exhaustive survey of publicly available Greek testing platforms to compile a corpus of questions and solutions. Our sourcing strategy targeted authoritative bodies, identifying a wide spectrum of standardized examinations ranging from primary education to professional licensure. We systematically crawled and ingested data from these repositories, developing custom extraction pipelines to handle heterogeneous file formats—including structured web interfaces, PDF archives, and DOCX documents. These formats typically reside outside the scope of standard web-crawling pipelines (e.g., Common Crawl) used for LLM pre-training, thereby minimizing the risk of data contamination. This process yielded a raw corpus of diverse subject matter, ensuring the benchmark captures the breadth of the Greek educational and professional curriculum.

Given the predominance of PDF documents in the raw corpus, we utilized *PyMuPDF4LLM*[3] for structure-aware text extraction. For legacy scanned files, we applied Tesseract OCR[4]. To mitigate extraction artifacts—particularly in malformed Greek text and mathematical notations—we implemented an LLM-assisted correction phase utilizing Claude 3.5 Sonnet and Haiku. The models were prompted to preserve the original semantic intent verbatim and identify toxic content, ensuring that no generative question creation occurred during the process.

Following this automated restoration, the dataset underwent strict Unicode canonicalization and punctuation standardization (e.g., correcting the Greek question mark ';'). Finally, the curated dataset—after filtering and masking any personal identifying information and meaningless ids—was validated by a dedicated team of **five native Greek-speaking experts** holding graduate-level academic qualifications. This team manually reviewed the question–answer pairs to verify linguistic fidelity and filter out remaining processing artifacts, ensuring the benchmark adheres to the rigorous standards of authentic Greek examinations.

3.3 Quality Control

To ensure the integrity of the benchmark, we performed a rigorous human evaluation of the final dataset. Our 3 Greek-speaking graduate- and professor-level experts randomly selected **5%** of the samples from every individual task to conduct a manual verification of both the question stems and the ground-truth answers. This process focused on identifying OCR errors, semantic drift, or incorrect labeling. We estimated the noise level (proportion of QA samples considered low quality or wrong) in the dataset to be approximately **2.5%**, with any identified errors corrected prior to release.

3.4 Statistics and Analysis

GreekMMLU consists of **19,876** multiple-choice questions across **44 subjects**, organized into **four supercategories** (STEM, Humanities, Social Sciences, and Other) and annotated with **four difficulty levels** corresponding to *Primary*, *Secondary*, *University*, and *Professional* education. We split about 23% from each subject to create two subsets. The **public subset contains 15,391 questions** and

[2] Link will be provided upon acceptance

[3] https://github.com/pymupdf/pymupdf4llm

[4] https://github.com/tesseract-ocr/tesseract

Group	# Questions	# Chars	
		Question	Answer
STEM	5753	97.7	20.0
Humanities	2693	93.6	35.8
Social Science	4937	297.8	17.8
Other	1788	87.9	38.4
Primary School	4525	67.9	16.3
Secondary School	1133	1000.7	14.1
Professional	7061	110.8	26.4
University	794	90.6	34.3
NA	1658	88.7	39.2

Table 2: Average question and answer length (in characters) for each education group and subject area in GreekMMLU.

Αυτό είναι μια ερώτηση [**SUBJECT**]. Επιλέξτε τη σωστή απάντηση!
Ερώτηση: [**QUESTION**]
[**OPTION**]
Απάντηση:

This is a [**SUBJECT**] question. Select the correct answer!
Question: [**QUESTION**]
[**OPTION**]
Answer:

Figure 3: Prompt templates in Greek and English.

is used for all experiments reported in this work, while the **private subset contains 4,485 questions** and is reserved for a future evaluation leaderboard. A detailed statistical distribution can be found in Table 2 and Appendix B.

As shown in Table 2, most questions are drawn from professional examinations, followed by primary, secondary, and university-level questions, with an additional NA category of 1.7K questions not aligned to a specific educational tier. Average question and answer length generally increases with educational level; secondary school items form a notable exception, with longer prompts but relatively short answers, reflecting the inclusion of citizenship and civic education tasks that combine context-rich descriptions with concise answer options. Across subject areas, the dataset is broadly balanced, with STEM, humanities, and social sciences each comprising 2.7K–5.8K questions; social science questions are longer on average due to text-based, situational prompts, while STEM questions tend to be more concise and formula-driven.

4 Experiments and Results

4.1 Experimental Setup

We integrate GreekMMLU into the lm-evaluation-harness framework (Gao et al., 2024) to ensure a standardized and reproducible evaluation environment. All models are evaluated under both zero-shot and five-shot settings. We show the detailed setup in Appendix C.

Evaluation Strategy. Evaluation depends on access to model outputs. For open-weights models, we use a rank-based multiple-choice setup, selecting the answer with the highest token log-likelihood. For closed-source API models, we use free-form generation, prompting models to output the answer key directly and extracting the predicted Greek label with regular expressions (Li et al., 2024).

Prompting Protocol. Models are evaluated under both zero-shot and five-shot prompting using prompts written entirely in Greek. Each prompt includes a subject-specific instruction, the question stem, and the labeled answer options, following Greek examination conventions with answer labels A, B, Γ, and Δ. In the five-shot setting, five representative examples from the task's development set are prepended to the prompt. The prompt templates are illustrated in Figure 3.

Model Selection. We conduct a comprehensive analysis of diverse model families on the GreekMMLU benchmark. Our study covers three categories. We first establish broad capabilities using **General-Purpose LLMs**, including the Qwen (Yang et al., 2025), Llama (Grattafiori et al., 2024), and Gemma-3 (Team et al., 2025), GLM-4 (GLM et al., 2024) families. We compare these against **Greek-Adapted LLMs**—such as Meltemi (Voukoutis et al., 2024), Krikri (Roussis et al., 2025), and Plutus (Peng et al., 2025)—which are selected to quantify the benefits of native-language specialization. Additionally, we evaluate **Multilingual European LLMs**, represented by the EuroLLM family (Martins et al., 2025) alongside multilingual models like Aya-101 (Üstün et al., 2024), BLOOMZ (Muennighoff et al., 2023), and mT0 (Muennighoff et al., 2023). Finally, we assessed frontier closed-source models like ChatGPT and Gemini, while a **Random Baseline** is included to establish a theoretical lower bound.

Model	STEM	Humanities	Social Sci.	Other	Average	Greek-specific
General-Purpose LLMs						
GPT-5.2†	86.75	90.25	89.64	86.80	88.30	91.83
GPT-4o †	84.48	90.02	89.06	86.80	87.19	91.34
Gemini 3 Flash†	91.62	93.24	93.29	91.67	92.46	93.88
Qwen2.5-7B	52.16	56.78	58.74	53.19	55.31	60.05
Qwen2.5-7B-Instruct†	59.46	58.67	61.26	58.95	59.94	62.92
Qwen2.5-14B	63.76	65.01	65.43	59.73	64.06	67.08
Qwen2.5-14B-Instruct†	64.75	66.25	69.33	62.64	66.36	72.21
Qwen2.5-32B	72.24	71.68	73.69	70.30	72.45	75.82
Qwen2.5-32B-Instruct†	71.61	71.36	75.69	69.85	72.84	78.36
Qwen2.5-72B	77.16	78.48	80.38	76.85	78.47	82.43
Qwen2.5-72B-Instruct†	78.01	79.22	80.95	77.07	79.13	83.14
Qwen3-30B	71.65	69.47	74.65	58.67	70.89	76.50
Qwen3-30B-Instruct†	78.01	75.03	78.53	76.45	77.59	80.49
Llama-2-7b-hf	36.33	36.05	33.84	36.80	35.45	32.90
Llama-2-7b-chat-hf†	36.73	34.94	34.30	33.89	35.26	33.83
Llama-3.1-8B	51.36	57.33	54.23	49.50	53.03	53.93
Llama-3.1-8B-Instruct†	56.01	63.13	62.20	56.94	59.36	63.91
Llama-3.1-70B	71.79	78.99	77.89	74.38	75.32	80.90
Llama-3.2-1B	37.98	36.46	34.48	37.30	36.42	33.36
Llama-3.2-1B-Instruct†	37.44	37.24	36.85	35.46	36.97	35.30
Llama-3.2-3B	42.20	42.07	43.28	44.80	42.88	41.34
Llama-3.2-3B-Instruct†	42.92	42.85	45.54	44.13	43.99	45.68
Llama-3.3-70B-Instruct†	79.04	82.48	83.24	79.42	81.09	86.34
Gemma-3-4B-pt	49.82	52.87	53.36	50.45	51.61	54.23
Gemma-3-4b-it†	59.41	60.78	65.11	61.47	61.90	67.16
Gemma-3-12B-pt	72.36	74.44	77.10	71.70	74.29	78.88
Gemma-3-12b-it†	72.15	75.03	78.30	71.98	74.75	80.68
Gemma-3-27B-pt	77.07	79.91	79.52	76.45	78.29	81.53
Gemma-3-27B-it†	77.33	79.86	81.15	75.62	78.87	83.58
Aya-101†	52.62	53.75	62.16	57.05	56.73	59.86
BLOOMZ-7b1†	34.16	32.64	30.50	32.44	35.26	29.56
mT0-xxl†	54.86	53.33	64.73	57.94	58.55	62.68
GLM-4-9b	62.73	63.82	67.52	62.86	64.62	67.38
GLM-4-9b-chat†	61.08	64.60	67.63	65.10	64.41	68.11
Greek and European LLMs						
Llama-Krikri-8B-Base	58.61	69.06	64.80	61.69	62.70	67.62
Llama-Krikri-8B-Instruct†	61.12	70.16	69.57	63.26	65.70	73.42
Meltemi-7B-v1.5	51.61	55.49	53.25	51.79	52.78	55.63
Meltemi-7B-Instruct-v1.5†	56.58	64.51	63.67	60.29	60.70	65.38
Plutus-8B-instruct†	60.28	69.93	68.85	62.92	65.06	72.54
EuroLLM-1.7B	34.17	33.93	30.23	32.38	32.51	29.29
EuroLLM-1.7B-Instruct†	29.67	30.80	28.36	31.88	29.62	29.54
EuroLLM-9B	61.57	71.26	68.74	64.77	65.92	72.51
EuroLLM-9B-Instruct†	63.13	73.15	71.27	66.39	67.88	75.38
EuroLLM-22B	66.33	75.26	72.80	67.90	70.12	76.26
EuroLLM-22B-Instruct-2512†	68.87	74.90	73.64	69.57	71.53	77.54
Random Baseline	32.33	28.77	31.86	32.62	30.42	31.59

Table 3: Overall zero-shot performance of different models on the GreekMMLU benchmark. Accuracy (%) is reported. The *Greek-specific* column includes an average of History, Traditions, and Mythology subsets. Models marked with† are instruction-tuned.

4.2 Main Results

The overall zero-shot performance on part of our evaluated models is summarized in Tables 3. We show all the evaluated models with both zero-shot and five-shot settings in Appendix D.

Model performance on GreekMMLU spans a wide spectrum, ranging from near-random accuracy for weak models to strong results for state-of-the-art models. Small and lightly adapted models have performance near the random baseline, while mid-sized models (3-20B) typically achieve moderate performance in the 55–70% range. A clear separation emerges at the top end, where closed-source frontier models substantially outperform all open-weight alternatives. For example, *Gemini 3 Flash* reaches an average accuracy of 92.46%, while *GPT-5.2* and *GPT-4o* achieve 88.30% and 87.19%, respectively, consistently excelling in all subjects. In contrast, the strongest open-weight models—such as *Llama-3.3-70B-Instruct* and *Qwen2.5-72B-Instruct*—peak at 81.09% and 79.13% average accuracy, leaving a substantial gap to the best models. This persistent performance margin reflects the advantages of closed-source models, including broader exposure to Greek data during training in addition to large-scale optimization and advanced alignment.

Instruction Tuning Effects. Across nearly all model families, instruction-tuned variants consistently outperform their corresponding base models at similar parameter scales. This improvement is observed across subject categories and is particularly pronounced on Greek-specific tasks. Instruction tuning appears to improve alignment with the multiple-choice evaluation format, Greek prompt structure, and answer selection conventions used in GreekMMLU.

Greek- and European-Centric Models. Models explicitly adapted to Greek or European languages consistently achieve higher accuracy than general multilingual baselines of comparable size. Greek-centric models like LLaMA-Krikri-8B demonstrate notable gains, especially on Greek-specific subjects. Similarly, European-centric models outperform globally trained multilingual models, indicating that regional linguistic focus contributes meaningfully to Greek language understanding.

Performance Across Educational Levels. Figure 4 shows that accuracy is generally higher on primary and secondary school questions and lower

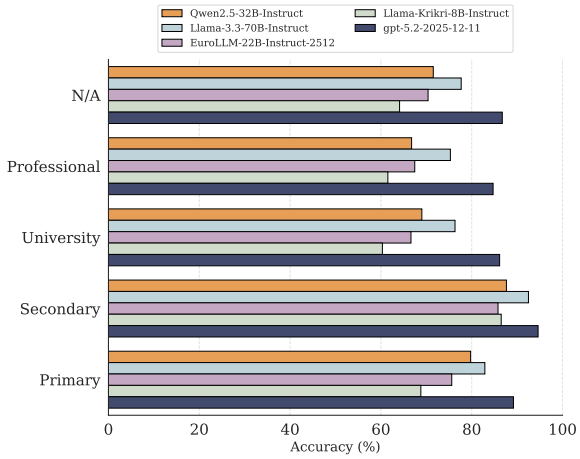

Figure 4: Models performance on different task levels.

on university and professional-level tasks. This pattern appears across model families and scales, including both open-weight and closed-source models. Performance on N/A-level questions typically lies between primary/secondary and university/professional levels. Overall, the reduced accuracy at higher levels reflects the increased complexity, specialized knowledge of university and professional examination questions.

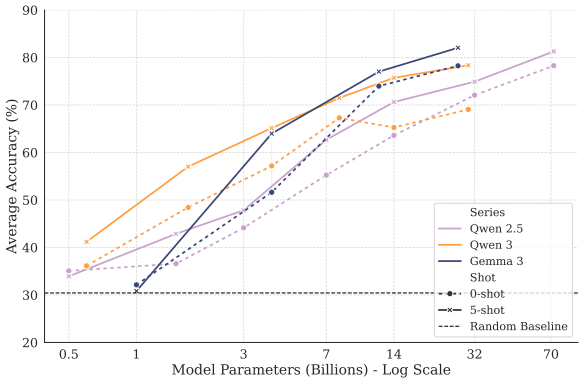

Figure 5: Scaling behavior of average accuracy with respect to model size (in billions of parameters) under zero-shot and five-shot prompting.

4.3 Analysis

Model Scale Effects. Figure 5 shows a clear relationship between model size and performance on GreekMMLU. Very small models (below approximately 2B parameters) generally can't solve these tasks, with accuracy close to the random baseline across all model families. As model size increases, performance improves consistently. Larger models achieve higher accuracy, reflecting better answer selection and more effective handling of longer and

more complex question contexts in Greek. This trend is observed across all evaluated model series, with steady gains as parameters increase, indicating that larger models are better equipped to handle the linguistic and knowledge demands of the benchmark.

Subject-Level and Cultural Differences. Model performance varies across subject domains, with most models achieving higher accuracy in humanities and social sciences than in STEM, reflecting the greater reasoning and technical demands of STEM questions in GreekMMLU. Greek-specific subjects (e.g., history, traditions, mythology) are consistently more challenging than globally shared domains, particularly for general-purpose multilingual models. In contrast, Greek- and European-centric models of comparable size perform better on these culturally grounded tasks, indicating stronger coverage of localized knowledge. Additional analysis is provided in Appendix F.

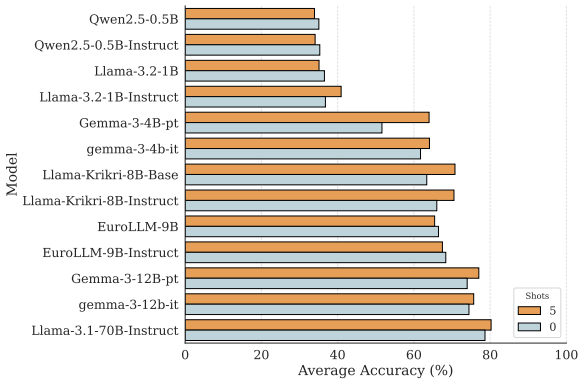

Figure 6: Comparison of average accuracy across models under zero-shot and five-shot prompting.

Zero-Shot vs. Five-Shot Performance. Figure 6 shows that five-shot prompting does not improve performance for models less than 2B, which remain close to the random baseline. In contrast, larger models consistently benefit from additional in-context examples. The improvement is most pronounced for Gemma and Llama families, where five-shot prompting yields noticeable accuracy boosts. Overall, the effectiveness of five-shot prompting is positive and can provide meaningful gains for mid- and large-scale ones.

Calibration Analysis Our analysis of subject-level calibration reveals notable differences across model families (Figure 7). The Greek-specialized

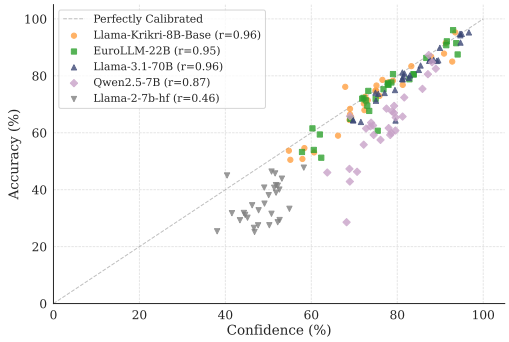

Figure 7: Calibration behavior across models on GreekMMLU.

Llama-Krikri-8B shows strong alignment between confidence and accuracy (r = 0.96), while generic multilingual models such as Qwen-2.5-7B are less well calibrated (r = 0.87). Earlier-generation models (e.g., Llama-2-7B) exhibit pronounced miscalibration (r = 0.46). Although larger models (e.g., Llama-3.1-70B) reduce calibration errors, language-specific training remains the primary driver of reliable calibration. We provide extended calibration analysis in Appendix E.

Correlation Between Question Length and Model Confidence We further experiment with the correlation between the question length and model prediction confidence in Appendix G across different model families and sizes. We found that the model confidence has little or no correlation with question length.

5 Conclusion

We introduced GreekMMLU, the first large-scale, native-sourced benchmark for evaluating massive multitask language understanding in Greek. GreekMMLU comprises 19,876 multiple-choice questions across 44 subjects and multiple educational levels, enabling linguistically and culturally grounded evaluation beyond machine-translated benchmarks. Our comprehensive evaluation highlights substantial performance gaps across model families and demonstrates the benefits of instruction tuning and Greek-adapted training, particularly on Greek-specific domains. We hope GreekMMLU encourages future models to place greater emphasis on native Greek capability and supports the development of more authentic, culturally grounded Greek language models.

Limitations

GreekMMLU is limited to multiple-choice questions drawn from formal educational and professional settings, enabling standardized evaluation but not capturing open-ended generation, interactive reasoning, or informal language use. Although all content is natively sourced in Greek, the benchmark primarily reflects standard Modern Greek as used in official curricula, with limited coverage of regional dialects and colloquial registers, and consists exclusively of text-based inputs without multimodal content. Finally, as with other large-scale benchmarks based on naturally occurring real-world data, potential overlap with the training corpora of some models cannot be entirely ruled out.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Nikos Bartziokas, Thanassis Mavropoulos, and Constantine Kotropoulos. 2020. Datasets and Performance Metrics for Greek Named Entity Recognition. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020)*, SETN 2020, pages 160–167, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Odysseas S Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. Greekbarbench: A challenging benchmark for free-text legal reasoning and citations. *arXiv preprint arXiv:2505.17267*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Iakovos Evdaimon, Hadi Abdine, Christos Xypolopoulos, Stamatis Outsios, Michalis Vazirgiannis, and Giorgos Stamou. 2024. GreekBART: The first pretrained Greek sequence-to-sequence model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7949–7962, Torino, Italia. ELRA and ICCL.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, and 1 others. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640.

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.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025. Plutus: Benchmarking large language models in low-resource greek finance. *arXiv preprint arXiv:2502.18772*.

Dimitris Roussis, Leon Voukoutis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2025. Krikri: Advancing open large language models for greek. *arXiv preprint arXiv:2505.13772*.

Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, Preslav Nakov, Michalis Vazirgiannis, and Eric P. Xing. 2025a. Nile-chat: Egyptian language models for Arabic and Latin scripts. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 306–322, Suzhou, China. Association for Computational Linguistics.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025b. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 9–30, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, and 1 others. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939.

Leon Voukoutis, Dimitris Roussis, Georgios Paraskevopoulos, Sokratis Sofianopoulos, Prokopis Prokopidis, Vassilis Papavasileiou, Athanasios Katsamanis, Stelios Piperidis, and Vassilis Katsouros. 2024. Meltemi: The first open large language model for greek. *arXiv preprint arXiv:2407.20743*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, and 1 others. 2025. Mmluprox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schütze. 2024. Turk-

ishmmlu: Measuring massive multitask language understanding in turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

A GreekMMLU Tasks and Examples

Task	Tested Concepts	Supercategory	# Q
Accounting	Accounting, balance sheets, microeconomics, institutions...	Social Sciences	189
Agriculture Professional	Circular bioeconomy, social economy, agri-food systems...	STEM	344
Agriculture University	Smart agriculture, circular bioeconomy, agri-food systems...	STEM	75
Art Professional	Applied arts, digital design, materials, fashion ...	Humanities	605
Art Secondary School	Greek cinema, Greek music, cultural figures, arts history..	Humanities	28
Art University	Music theory, rhythm, Greek tradition, ethnomusicology...	Humanities	17
Biology	Basic biology, microorganisms, human body, animal biology..	STEM	87
Chemistry	Thermodynamics, physical chemistry, phase equilibria, surface chemistry...	STEM	86
Civil Engineering	Building systems, materials, safety, mechanics.	STEM	774
Clinical Knowledge	Clinical basics, anatomy and physiology, nursing care, dermatology..	STEM	462
Computer Networks and Security	Packet-switched networks, architectures, protocol layers, Internet...	STEM	68
Computer Science Professional	Computing fundamentals, networking, software, digital systems...	STEM	259
Computer Science University	Computer systems, networking, data analysis, interactive technologies.	STEM	109
Driving Rules	Traffic regulations, road safety, traffic signs, driver responsibilities.	Other	1663
Economics Professional	Macroeconomics, international trade, public finance, labor markets.	Social Sciences	201
Economics University	Strategy, digital business, innovation, technology, costs and pricing.	Social Sciences	91
Education Professional	Child development, creativity, play-based learning, pedagogy...	Social Sciences	258
Education University	Educational research, data analysis, academic writing.	Social Sciences	46
Electrical Engineering	Electrical circuits, sensors, automotive systems, protection devices.	STEM	512
General Knowledge	Safety procedures, natural phenomena, first aid.	Other	135
Geography Primary School	Physical and human geography, cartography, regions...	Social Sciences	365
Geography Secondary School	Physical and human geography, cartography...	Social Sciences	67
Government and Politics Primary School	Civics, democracy, institutions, governance, rights and duties.	Social Sciences	285
Government and Politics Secondary School	Political institutions, citizenship, rights, public administration...	Social Sciences	80
Greek History Primary School	Greek history from antiquity to modern times, key events and figures.	Humanities	469
Greek History Professional	Modern Greek political and constitutional history	Humanities	122
Greek History Secondary School	Byzantine Empire and Modern Greek State	Humanities	98
Greek Literature	Authors, Literary Movements and Major Works (Modern and Classical)	Humanities	19
Greek Mythology	Creation myths, heroic cycles, myth in culture...	Humanities	243
Greek Traditions	Food Safety, Baking Science and Culinary Operations	Social Sciences	381
Law	Administrative Law, EU Law and Sports Regulations	Humanities	899
Management Professional	Management, Economics and Business Operations	Social Sciences	646
Management University	Human Resource Management	Social Sciences	30
Mathematics	Mathematical reasoning, algebra, geometry, basic statistics...	STEM	1118
Medicine Professional	Professional-level medical knowledge and clinical reasoning.	STEM	295
Medicine University	System-level human physiology and functional integration of organs.	STEM	77
Modern Greek Language Primary School	Basic grammar, morphology, syntax, orthography..	Social Sciences	1483
Modern Greek Language Secondary School	Vocabulary, comprehension, meaning, language use..	Social Sciences	885
Physics Primary School	Basic physics, matter, energy, measurements, everyday phenomena.	STEM	432
Physics Professional	Applied physics and engineering	STEM	1059
Physics University	Scientific reasoning, physics–biology concepts, thermodynamics, optics...	STEM	76
Prehistory	Cycladic, Minoan, and Mycenaean civilizations.	Humanities	68
World History	Enlightenment, reformation, renaissance, revolutionary movements...	Humanities	25
World Religions	Orthodox Christian hymnology.	Humanities	160
Total			**15,391**

Table 4: Summary of the 44 subjects in the GreekMMLU public dataset. # Q indicates the total number of questions for each task.

Subject	Question	Choices
STEM	Ποια σώματα, όταν δέχονται το ίδιο φως (π.χ. από τον Ήλιο), απορροφούν περισσότερη ενέργεια·	Α. Τα σκουρόχρωμα σώματα Β. Τα ανοιχτόχρωμα σώματα Γ. Τα διαφανή σώματα Δ. Τα μεταλλικά σώματα
	Which bodies, when receiving the same light (e.g. from the Sun), absorb more energy?	A. Dark-colored bodies B. Light-colored bodies C. Transparent bodies D. Metallic bodies
Humanities	Ποιος σκηνοθέτησε την ταινία «Κυνόδοντας», η οποία υπήρξε το 2011 υποψήφια για Όσκαρ Καλύτερης Ξενόγλωσσης Ταινίας·	Α. Παντελής Βούλγαρης Β. Γιώργος Λάνθιμος Γ. Κώστας Γαβράς Δ. Θεόδωρος Αγγελόπουλος
	Who directed the film Dogtooth, which was nominated in 2011 for the Academy Award for Best Foreign Language Film?	A. Pantelis Voulgaris B. Yorgos Lanthimos C. Costa-Gavras D. Theodoros Angelopoulos
Social Sciences	Ποιος είναι ο πρώτος φορέας κοινωνικοποίησης του ατόμου·	Α. Το σχολείο Α Β. Η οικογένεια Γ. Οι φίλοι/συνομήλικοι Δ. Οι επαγγελματικές σχέσεις
	What is the primary agent of socialization of an individual?	A. School B. Family C. Friends/peers D. Professional relationships
Other	Η Ανώνυμη Εταιρεία (Α.Ε.) είναι εμπορική εταιρεία:	Α. όταν ενεργεί εμπορικές πράξεις. Β. όταν ο σκοπός της είναι εμπορικός. Γ. ανεξάρτητα από τον σκοπό της. Δ. Η Ανώνυμη Εταιρεία (Α.Ε.) δεν είναι εμπορική εταιρεία.
	A Société Anonyme (S.A. / public limited company) is considered a commercial company:	A. when it carries out commercial acts. B. when its purpose is commercial. C. regardless of its purpose. D. A Société Anonyme is not a commercial company.
Greek-specific	Ποιος ήταν ο Αίολος·	Α. Ο βασιλιάς των Λαιστρυγόνων Β. Ένας σύντροφος του Οδυσσέα Γ. Ο θεός των ανέμων Δ. Ο πατέρας της Κίρκης
	Who was Aeolus?	A. The king of the Laestrygonians B. A companion of Odysseus C. The god of the winds D. The father of Circe

Table 5: Examples from GreekMMLU with their corresponding English translations across different subjects, where the bold items indicate the correct choices.

B Statistics of GreekMMLU

As reported in Table 6, token-length statistics were obtained using `tiktoken` with the `cl100k_base` encoding. For each question, we computed the number of tokens in the question text (reported as Avg. Q Tokens) and the total number of tokens across all answer choices (reported as Avg. C Tokens), where choices were concatenated using a single space separator. These values were then averaged across all questions within each subject group to summarize token distribution patterns across the dataset.

As shown in Figure 8, the distributional characteristics of text length vary across subject categories in the GreekMMLU benchmark. The left panel reports question-length distributions, indicating that most subjects exhibit median lengths below 200 characters. Notable exceptions include domains such as Modern Greek Language (Secondary School), which display substantially longer inputs, primarily due to the inclusion of extended reading-comprehension passages. The right panel presents answer-length distributions, which are generally more compact; however, technical and professional domains, including Physics and Law, are characterized by longer answer options, reflecting the increased precision and explanatory detail required in these fields. Overall, this variation in sequence length underscores the benchmark's ability to evaluate model performance across both short factual queries and longer, context-dependent inputs.

C Implementation Details

All evaluations were conducted using the lm-evaluation-harness framework (Gao et al., 2024) (version 0.4.9.1). The evaluation setup follows a unified configuration derived from the GreekMMLU benchmark, ensuring consistent assessment across all subject domains.

Models were evaluated on the GreekMMLU benchmark that we implemented based on the MMLU format, which consists of subject-specific multiple-choice questions covering STEM, Humanities, Social Sciences, and Other domains. Evaluation was performed under standardized zero-shot and five-shot prompting conditions.

All evaluations for open-source models relied on the default deterministic behavior of the `lm-evaluation-harness` for multiple-choice tasks. Model predictions were obtained via log-likelihood comparison over answer options rather than generative sampling.

For closed-source models accessed via external APIs, evaluation was performed using a generation-based multiple-choice protocol. Generation was conducted under deterministic settings, with sampling disabled (`do_sample=false`) and temperature set to 0.0.

Model outputs were post-processed using a standardized parsing procedure that extracts the first valid answer symbol, supporting both Greek (Α, Β, Γ, Δ) and Latin (A, B, C, D) representations. When Latin characters were produced, they were deterministically mapped to their Greek equivalents. The extracted prediction was then compared against the answer label to compute accuracy. This procedure supports a variable number of answer choices and was applied uniformly across all evaluated models.

Experiments were conducted on NVIDIA GPUs, primarily using RTX A6000. For larger models exceeding 30B parameters, evaluations were performed on NVIDIA A100 GPUs to accommodate increased memory and compute requirements.

Group	Tasks	#Q	Questions per Task			Avg. Tokens	
			Avg.	Max.	Min.	Question	Choices
STEM	16	5833	364.56	1118	68	86.65	72.60
Humanities	12	2753	229.42	899	17	83.85	130.99
Social Science	14	5007	357.64	1483	30	258.37	68.39
Other	2	1798	899.00	1663	135	79.02	139.14
All	44	15391	349.80	1663	17	141.12	89.45

Table 6: Statistics of the GreekMMLU

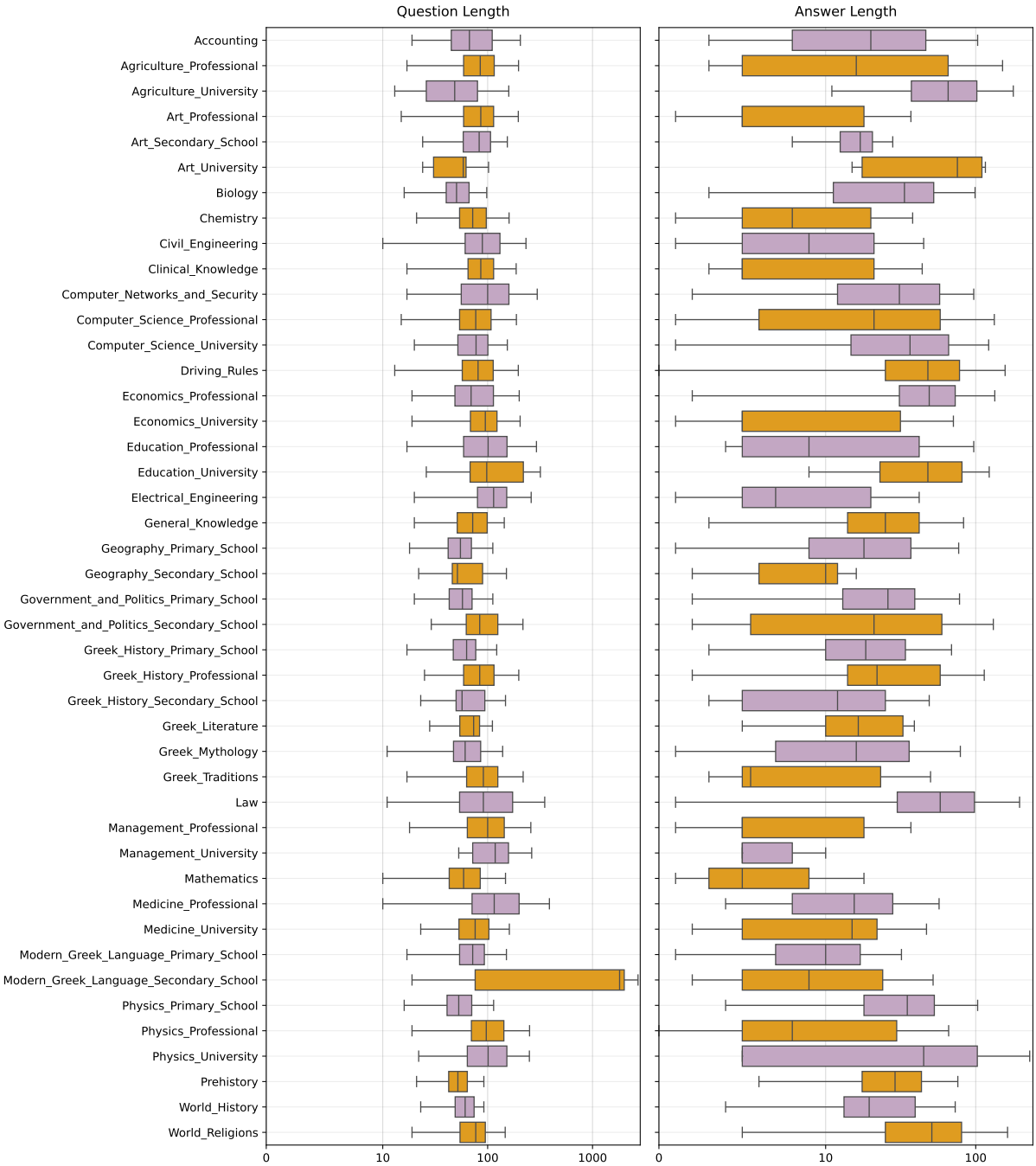

Figure 8: Distribution of character lengths for questions and answer choices across all GreekMMLU subjects.

D Experiment results

Model	STEM	Humanities	Social Sci.	Other	Average	Greek-specific
General-Purpose LLMs						
Qwen2.5-0.5B	35.86	31.49	31.46	37.02	33.79	30.38
Qwen2.5-1.5B	44.41	43.95	42.71	40.04	43.22	42.05
Qwen2.5-3B	46.13	46.67	48.07	50.84	47.46	46.17
Qwen2.5-7B	62.54	62.30	64.18	61.35	62.96	65.19
Qwen2.5-14B	69.74	72.28	72.41	67.90	70.85	75.16
Qwen2.5-32B	75.44	74.53	76.54	72.65	75.37	77.70
Qwen2.5-72B	81.59	82.90	82.05	78.30	81.56	84.40
Qwen3-0.6B	43.35	38.44	40.48	42.28	41.49	38.22
Qwen3-1.7B	59.88	53.29	57.54	56.66	57.72	56.72
Qwen3-4B	69.41	61.15	67.13	63.81	66.75	66.83
Qwen3-8B	74.07	68.92	73.07	69.02	72.37	74.51
Qwen3-14B	77.68	72.37	78.61	73.21	76.73	79.75
Qwen3-30B	81.07	75.45	79.51	76.62	79.18	80.66
Llama-2-7b-hf	37.23	35.68	33.35	32.72	35.08	32.49
Llama-3-8B	61.48	65.79	65.99	63.81	64.00	67.81
Llama-3.1-70B	77.14	82.67	82.71	78.36	80.08	85.77
Llama-3.1-8B	61.08	63.68	64.97	61.97	62.96	65.96
Llama-3.2-1B	35.69	33.89	33.91	37.14	34.96	32.40
Llama-3.2-3B	46.91	49.98	48.32	49.94	48.22	45.52
Gemma-3-1B-pt	31.55	30.53	29.66	31.60	30.73	29.34
Gemma-3-4B-pt	62.02	65.38	66.27	62.19	64.05	68.31
Gemma-3-12B-pt	76.66	78.71	76.46	76.45	76.86	77.87
Gemma-3-27B-pt	81.51	83.03	82.79	80.82	82.10	84.92
XGLM-1.7B	36.03	33.93	31.55	35.01	34.00	30.27
XGLM-2.9B	35.76	35.26	31.93	34.34	34.14	32.76
XGLM-4.5B	36.22	34.11	31.84	35.35	34.24	31.04
XGLM-7.5B	33.63	33.20	31.11	37.47	33.12	29.92
GLM-4-9B	67.39	69.89	73.14	67.95	69.88	72.98
Greek and European LLMs						
Llama-Krikri-8B-Base	65.62	76.14	74.28	67.39	70.45	78.83
Meltemi-7B-v1	59.57	64.87	50.87	64.71	57.81	48.72
Meltemi-7B-v1.5	59.53	67.45	63.12	66.33	62.76	65.71
EuroLLM-1.7B	33.51	31.45	30.76	36.86	32.62	30.90
EuroLLM-9B	59.24	69.93	66.98	66.61	64.43	71.58
EuroLLM-22B	70.97	77.20	76.11	70.58	73.67	80.82
Random Baseline	32.33	28.77	31.86	32.62	30.42	31.59

Table 7: Overall five-shot performance of **base** LLMs on the GreekMMLU benchmark. Accuracy (%) is reported. The *Greek-specific* column includes an average of History, Traditions, and Mythology subsets.

Model	STEM	Humanities	Social Sci.	Other	Average	Greek-specific
General-Purpose LLMs						
GPT-5.2	86.75	90.25	89.64	86.80	88.30	91.83
GPT-4o	84.48	90.02	89.06	86.80	87.19	91.34
Gemini 3 Flash	91.62	93.24	93.29	91.67	92.46	93.88
Qwen2.5-0.5B-Instruct	35.56	32.83	30.96	37.36	33.73	29.75
Qwen2.5-1.5B-Instruct	47.02	46.02	46.62	44.69	46.46	45.71
Qwen2.5-3B-Instruct	51.36	52.69	54.35	55.54	53.12	54.84
Qwen2.5-7B-Instruct	64.12	63.77	64.33	63.37	64.06	66.42
Qwen2.5-14B-Instruct	69.58	70.67	71.51	68.57	70.31	74.26
Qwen2.5-32B-Instruct	72.25	73.84	77.62	72.15	75.53	79.75
Qwen2.5-72B-Instruct	80.15	81.66	82.18	78.13	80.86	84.32
Qwen3-4B-Instruct-2507	70.48	62.39	68.07	66.33	67.97	68.61
Qwen3-30B-Instruct	81.56	77.15	80.49	77.91	80.11	81.94
Llama-2-7b-chat-hf	38.97	39.63	35.33	32.66	37.01	35.08
Llama-3-8B-Instruct	61.41	65.70	67.61	63.87	64.54	69.92
Llama-3.1-8B-Instruct	60.89	67.45	67.59	61.91	64.36	69.84
Llama-3.1-70B-Instruct	78.55	81.93	82.35	78.08	80.34	85.33
Llama-3.2-1B-Instruct	40.73	40.69	39.01	43.68	40.45	36.23
Llama-3.2-3B-Instruct	47.58	49.66	51.68	53.52	50.05	51.15
Llama-3.3-70B-Instruct	79.04	82.48	83.24	79.42	81.09	86.34
Mistral-7B-Instruct-v0.3	49.71	52.14	52.52	53.13	51.47	53.11
Gemma-3-1B-it	45.70	46.48	44.82	45.97	45.53	43.06
Gemma-3-4B-it	62.61	63.59	67.19	62.64	64.40	68.36
Gemma-3-12B-it	74.05	75.45	78.81	74.16	75.97	80.44
Gemma-3-27B-it	79.96	81.75	82.14	78.58	80.84	84.62
Aya-101	54.37	54.71	64.78	57.94	58.59	62.62
Aya-expanse-8b	61.55	66.67	69.02	64.37	65.30	69.78
BLOOMZ-1b1	35.93	34.48	31.92	37.30	34.44	30.93
BLOOMZ-1b7	34.02	32.92	31.42	37.19	33.30	30.03
BLOOMZ-7b1	32.33	32.05	30.91	30.98	31.62	29.89
mT0-large	34.43	32.64	31.53	37.75	33.52	29.04
mT0-xl	41.47	40.83	44.23	45.25	42.82	40.52
mT0-xxl	48.95	43.72	55.58	53.69	51.14	55.36
GLM-4-9B-chat	66.92	66.71	71.60	67.56	68.65	71.58
Greek and European LLMs						
Llama-Krikri-8B-Instruct	65.65	75.68	73.93	67.06	70.23	77.95
Meltemi-7B-Instruct-v1.5	58.40	66.85	67.59	63.53	63.52	69.04
Plutus-8B-instruct	65.97	74.99	73.51	67.90	70.20	76.97
EuroLLM-1.7B-Instruct	33.11	31.54	30.78	36.30	32.42	31.34
EuroLLM-9B-Instruct	60.96	72.18	69.97	67.67	66.60	74.54
EuroLLM-22B-Instruct-2512	71.23	78.94	76.57	72.37	74.39	81.34
Random Baseline	32.33	28.77	31.86	32.62	30.42	31.59

Table 8: Overall five-shot performance of **instruction-tuned** LLMs on the GreekMMLU benchmark. Accuracy (%) is reported. The *Greek-specific* column includes an average of History, Traditions, and Mythology subsets.

Model	STEM	Humanities	Social Sci.	Other	Average	Greek-specific
General-Purpose LLMs						
Qwen2.5-0.5B	36.49	33.43	33.82	36.52	35.09	31.91
Qwen2.5-1.5B	37.27	35.26	34.48	39.65	36.26	31.89
Qwen2.5-3B	46.50	44.51	45.02	39.77	44.89	44.75
Qwen2.5-7B	52.16	56.78	58.74	53.19	55.31	60.05
Qwen2.5-14B	63.76	65.01	65.43	59.73	64.06	67.08
Qwen2.5-32B	72.24	71.68	73.69	70.30	72.45	75.82
Qwen2.5-72B	77.16	78.48	80.38	76.85	78.47	82.43
Qwen3-0.6B	38.00	34.62	34.41	37.47	36.16	32.43
Qwen3-1.7B	50.29	44.55	49.37	49.05	48.99	47.19
Qwen3-4B	64.68	58.30	64.00	60.51	63.03	64.18
Qwen3-8B	69.55	63.59	69.04	66.16	68.11	69.86
Qwen3-14B	62.52	61.98	67.83	68.90	65.10	66.61
Qwen3-30B	71.65	69.47	74.65	58.67	70.89	76.50
Llama-2-7b-hf	36.33	36.05	33.84	36.80	35.45	32.90
Llama-3-8B	52.23	57.15	55.64	48.83	53.76	55.38
Llama-3.1-8B	51.36	57.33	54.23	49.50	53.03	53.93
Llama-3.1-70B	71.79	78.99	77.89	74.38	75.32	80.90
Llama-3.2-1B	37.98	36.46	34.48	37.30	36.42	33.36
Llama-3.2-3B	42.20	42.07	43.28	44.80	42.88	41.34
Gemma-3-1B-pt	33.90	32.83	30.12	31.71	32.13	29.70
Gemma-3-4B-pt	49.82	52.87	53.36	50.45	51.61	54.23
Gemma-3-12B-pt	72.36	74.44	77.10	71.70	74.29	78.88
Gemma-3-27B-pt	77.07	79.91	79.52	76.45	78.29	81.53
XGLM-1.7B	37.01	34.99	33.09	36.74	35.28	31.86
XGLM-2.9B	35.65	35.91	32.83	32.38	34.29	32.13
XGLM-4.5B	37.08	35.59	32.91	37.47	35.41	31.50
XGLM-7.5B	36.42	34.71	31.48	33.33	34.03	30.55
GLM-4-9B	62.73	63.82	67.52	62.86	64.62	67.38
Greek and European LLMs						
Llama-Krikri-8B-Base	58.61	69.06	64.80	61.69	62.70	67.62
Meltemi-7B-v1	52.49	56.83	43.46	55.54	50.23	40.68
Meltemi-7B-v1.5	51.61	55.49	53.25	51.79	52.78	55.63
EuroLLM-1.7B	34.17	33.93	30.23	32.38	32.51	29.29
EuroLLM-9B	61.57	71.26	68.74	64.77	65.92	72.51
EuroLLM-22B	66.33	75.26	72.80	67.90	70.12	76.26
Random Baseline	32.33	28.77	31.86	32.62	30.42	31.59

Table 9: Overall zero-shot performance of **base** LLMs on the GreekMMLU benchmark. Accuracy (%) is reported. The *Greek-specific* column includes an average of History, Traditions, and Mythology subsets.

Model	STEM	Humanities	Social Sci.	Other	Average	Greek-specific
General-Purpose LLMs						
GPT-5.2	85.17	88.37	88.98	85.85	87.08	90.93
GPT-4o	83.17	88.55	88.27	85.57	86.06	91.20
Gemini 3 Flash	91.40	92.74	93.18	92.56	92.37	93.66
Qwen2.5-0.5B-Instruct	36.50	33.93	33.88	37.02	35.25	32.08
Qwen2.5-1.5B-Instruct	33.93	34.21	33.88	33.95	33.95	35.74
Qwen2.5-3B-Instruct	50.62	50.11	50.58	47.15	50.12	51.17
Qwen2.5-7B-Instruct	59.46	58.67	61.26	58.95	59.94	62.92
Qwen2.5-14B-Instruct	64.75	66.25	69.33	62.64	66.36	72.21
Qwen2.5-32B-Instruct	71.61	71.36	75.69	69.85	72.84	78.36
Qwen2.5-72B-Instruct	78.01	79.22	80.95	77.07	79.13	83.14
Qwen3-4B-Instruct-2507	64.61	57.93	66.18	62.75	64.00	66.20
Qwen3-30B-Instruct	78.01	75.03	78.53	76.45	77.59	80.49
Llama-2-7b-chat-hf	36.73	34.94	34.30	33.89	35.26	33.83
Llama-3-8B-Instruct	58.35	63.13	63.52	59.62	61.04	64.45
Llama-3.1-8B-Instruct	56.01	63.13	62.20	56.94	59.36	63.91
Llama-3.1-70B-Instruct	75.30	82.48	81.69	75.17	78.61	85.60
Llama-3.2-1B-Instruct	37.44	37.24	36.85	35.46	36.97	35.30
Llama-3.2-3B-Instruct	42.92	42.85	45.54	44.13	43.99	45.68
Llama-3.3-70B-Instruct	76.24	82.16	81.89	76.62	79.16	85.38
Mistral-7B-Instruct-v0.3	47.92	49.15	51.13	48.32	49.30	51.67
Mistral-Small-24B-Instruct-2501	67.16	71.72	74.79	66.39	70.47	78.06
Gemma-3-1B-it	46.20	44.55	44.01	46.14	45.17	41.67
Gemma-3-4B-it	59.41	60.78	65.11	61.47	61.90	67.16
Gemma-3-12B-it	72.15	75.03	78.30	71.98	74.75	80.68
Gemma-3-27B-it	77.33	79.86	81.15	75.62	78.87	83.58
Aya-101	52.62	53.75	62.16	57.05	56.73	59.86
Aya-expanse-8b	59.83	65.43	67.52	62.75	63.74	68.31
BLOOMZ-1b1	37.16	34.48	33.82	37.02	35.56	32.19
BLOOMZ-1b7	37.18	34.30	33.14	36.74	35.26	31.69
BLOOMZ-7b1	34.16	32.64	30.50	32.44	35.26	29.56
mT0-large	38.48	36.83	36.99	37.92	37.64	34.89
mT0-xl	46.18	48.18	54.43	51.85	50.10	52.49
mT0-xxl	54.86	53.33	64.73	57.94	58.55	62.68
GLM-4-9B-chat	61.08	64.60	67.63	65.10	64.41	68.11
Greek and European LLMs						
Llama-Krikri-8B-Instruct	61.12	70.16	69.57	63.26	65.70	73.42
Meltemi-7B-Instruct-v1.5	56.58	64.51	63.67	60.29	60.70	65.38
Plutus-8B-instruct	60.28	69.93	68.85	62.92	65.06	72.54
EuroLLM-1.7B-Instruct	29.67	30.80	28.36	31.88	29.62	29.54
EuroLLM-9B-Instruct	63.13	73.15	71.27	66.39	67.88	75.38
EuroLLM-22B-Instruct-2512	68.87	74.90	73.64	69.57	71.53	77.54
Random Baseline	32.33	28.77	31.86	32.62	30.42	31.59

Table 10: Overall zero-shot performance of **instruction-tuned** LLMs on the GreekMMLU benchmark. Accuracy (%) is reported. The *Greek-specific* column includes an average of History, Traditions, and Mythology subsets.

E Extended Calibration

We evaluated the calibration of a wider set of 11 models in the 5-shot setting (Table 11). The results highlight three distinct behaviors:

Model	Pearson r
Llama-Krikri-8B-Base	0.96
EuroLLM-22B	0.95
Llama-3.1-70B	0.95
Llama-3.1-8B	0.95
Qwen3-30B	0.93
EuroLLM-9B	0.93
Qwen2.5-72B	0.92
Llama-3.3-70B	0.91
Qwen2.5-32B	0.88
Qwen2.5-7B	0.87
Llama-2-7b-hf	0.46

Table 11: Extended calibration analysis ranked by Pearson correlation coefficient (r) in the 5-shot setting.

We evaluated the calibration of a wider set of 11 models in the 5-shot setting (Table 11). The results highlight three distinct behaviors. The top-performing models, including the Greek-specific Llama-Krikri-8B and the Llama 3.1 family, fall into a highly calibrated cluster, exhibiting the highest correlation ($r \geq 0.95$) and accurately reflecting their true probability of correctness. A second group, comprising the Qwen 2.5/3 family, shows moderate calibration with strong but slightly lower correlation coefficients ($r \approx 0.87 - 0.93$). Finally, the older Llama-2-7b baseline remains poorly calibrated with significantly lower correlation ($r = 0.46$)

F Subject-Level Performance Distribution

Figures 9 and 10 illustrate the distribution of accuracy scores across the 44 subjects of the GreekMMLU benchmark in zero-shot and five-shot settings, respectively, offering granular insights into model robustness beyond aggregate metrics. The analysis consistently reveals a distinct relationship between model specialization, scale, and cross-domain consistency across both prompting strategies. Notably, the Greek-centric `Llama-Krikri-8B` demonstrates a significant upward shift in its performance distribution relative to generic models of comparable size, such as `Llama-3.1-8B` and `EuroLLM-9B`. Its elevated performance baseline indicates a resilience against catastrophic failure modes on linguistically or culturally complex subjects, whereas generic counterparts frequently degrade to near-random accuracy in these areas.

Furthermore, the data highlights the stabilizing effect of model scale. The largest evaluated model, `Llama-3.1-70B`, exhibits the tightest clustering of subject scores within the upper quartile in both settings. This suggests that massive parameterization functions as a stabilizing factor, ensuring consistent competency across niche domains where smaller architectures struggle. In contrast, smaller models like `Qwen2.5-0.5B` display extreme vertical variance; while they occasionally achieve parity on simpler tasks, they lack the generalization capabilities required to maintain robust performance across the full breadth of the academic curriculum.

As illustrated in Figures 11 and 12, a subject-wise breakdown reveals heterogeneous performance patterns across the benchmark. Accuracy levels differ notably between disciplinary groups, with non-technical domains generally exhibiting narrower variance across models, while technical subjects display wider performance dispersion. Topics involving culturally grounded knowledge, including Greek historical and mythological content, introduce additional variability, particularly among general multilingual models. In contrast, models incorporating regionally focused training data tend to show more stable behavior across these subjects.

G Impact of Question Length on Model Confidence

To investigate whether model confidence is a byproduct of input verbosity rather than semantic certainty, we analyze the correlation between question length (measured in characters) and average confidence scores. As illustrated in Figures 13 and 14, we observe no meaningful correlation between question length and model confidence for either the specialized Llama-Krikri-8B or the generic baselines (e.g., Llama-3.1-70B). Pearson correlation coefficients remain consistently close to zero across prompting strategies, indicating that the models' uncertainty estimates are robust to variations in input length and are not driven by superficial properties of the prompt, with Llama-based models showing a marginally higher sensitivity to ques-

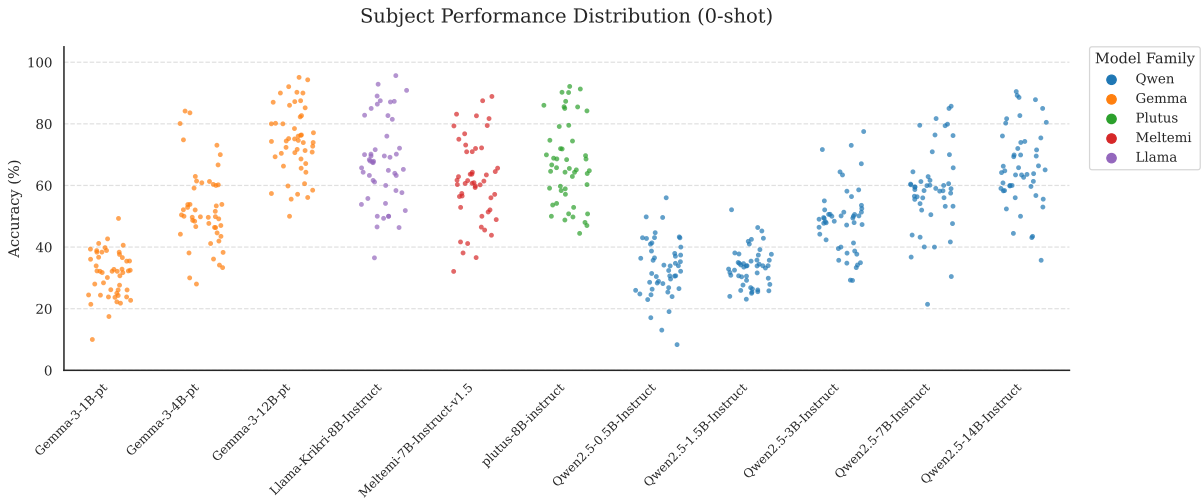

Figure 9: Subject performance distribution (5-shot)

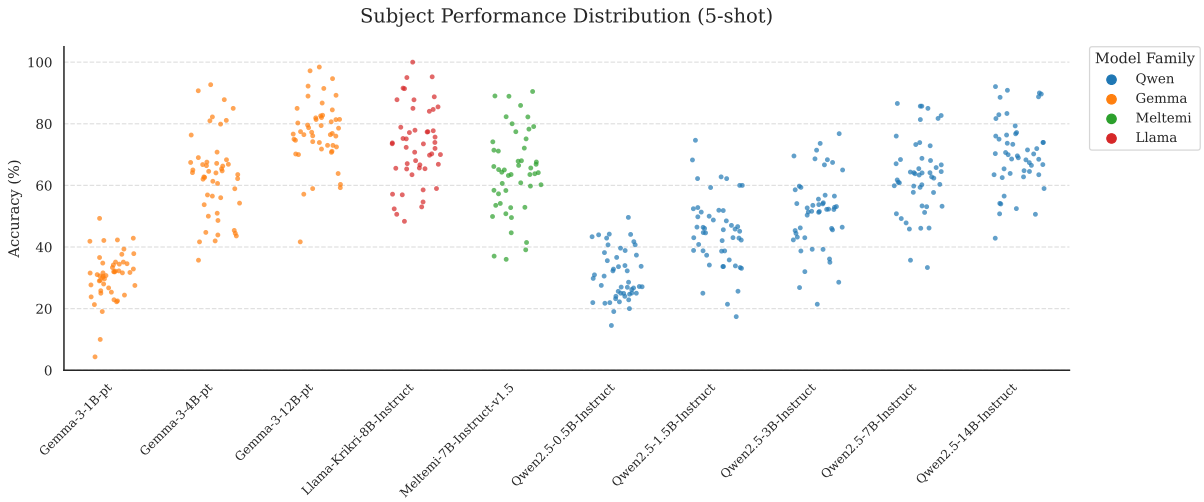

Figure 10: Subject performance distribution (0-shot)

tion length.

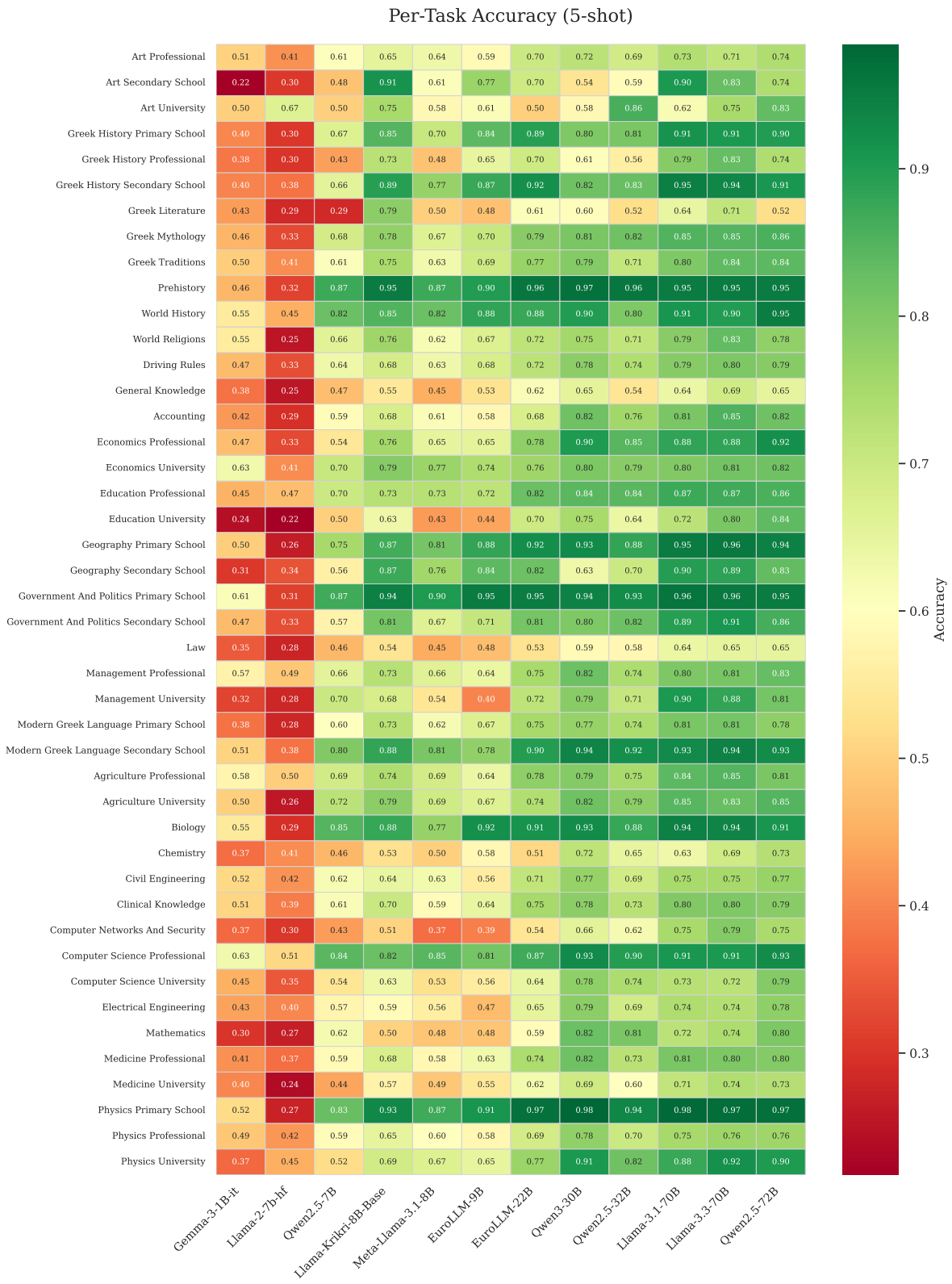

Figure 11: Subject-level accuracy heatmap under five-shot prompting, showing performance variation across GreekMMLU subjects and models.

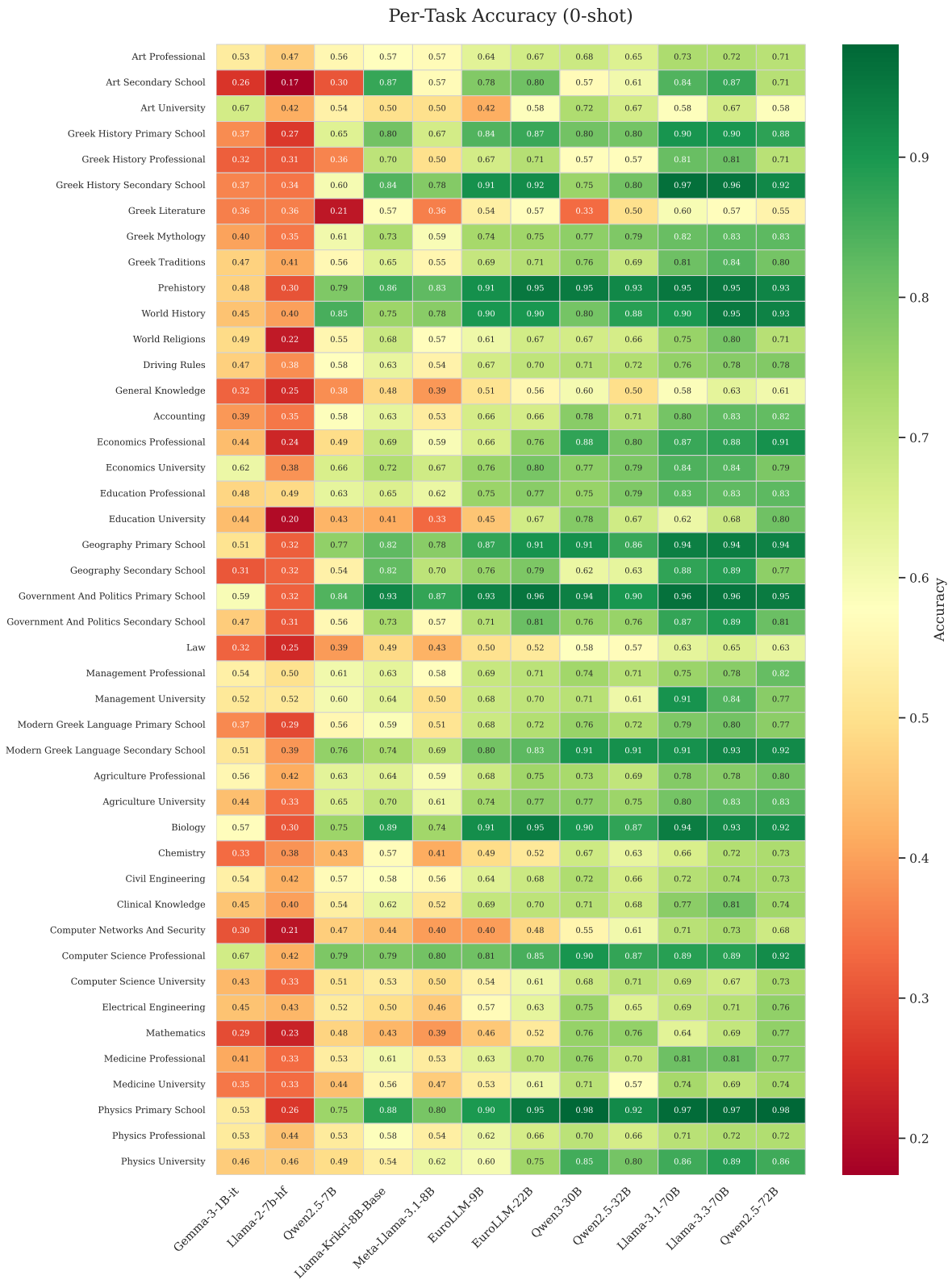

Figure 12: Subject-level accuracy heatmap under zero-shot prompting, showing performance variation across GreekMMLU subjects and models.

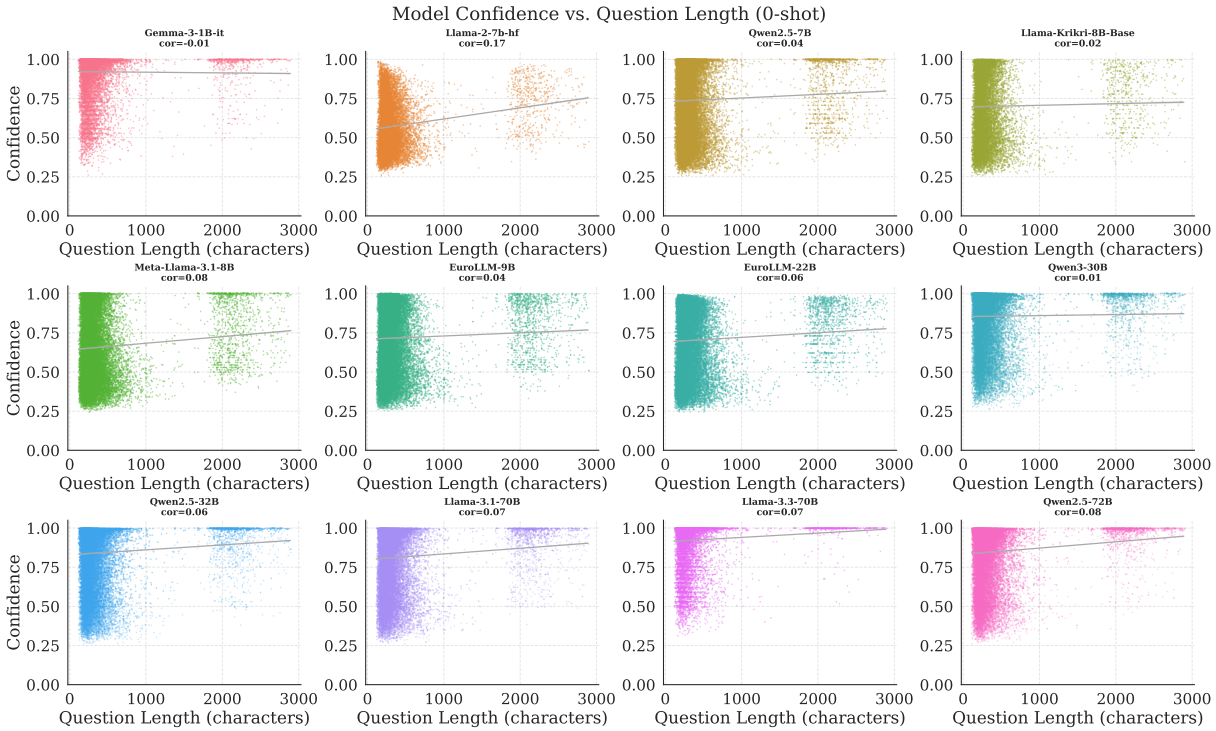

Figure 13: Correlation between Question Length and Confidence (Zero-Shot).

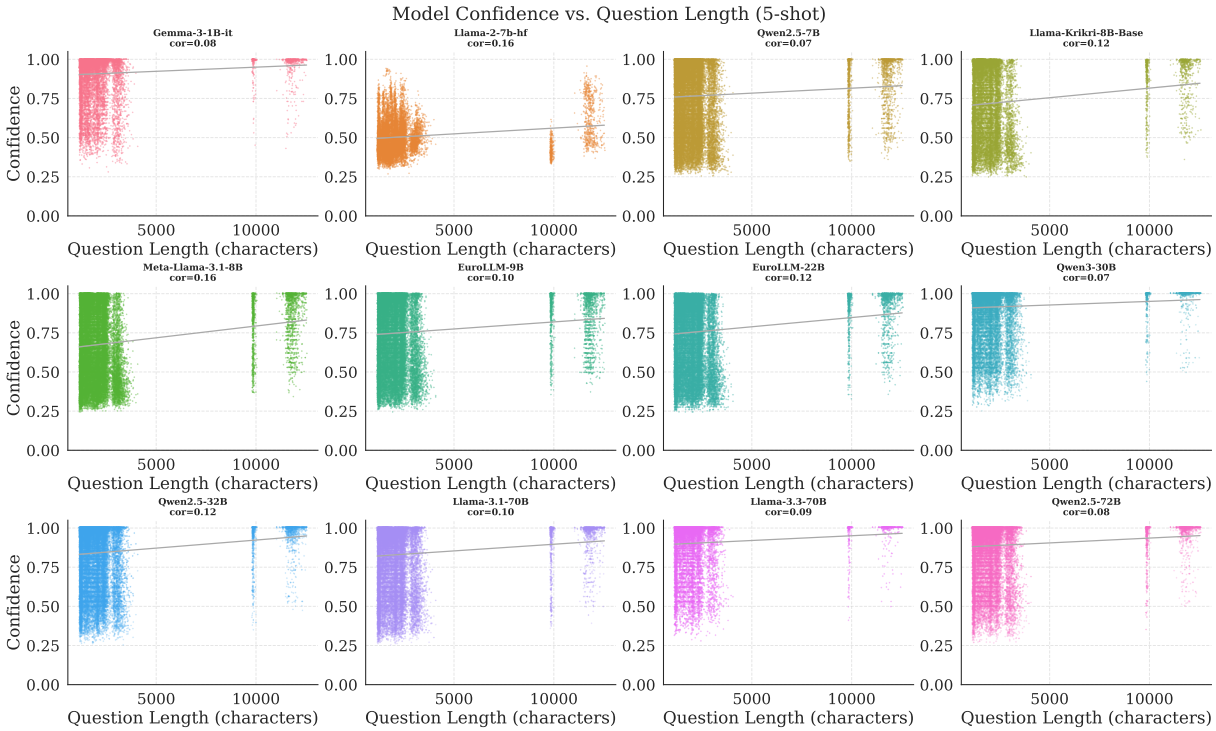

Figure 14: Correlation between Question Length and Confidence (Five-Shot).