

Article

RoCulturaMCQ: Building a Benchmark While Learning Statistics

Denis Iorga ¹, Razvan Muntean ¹, Mihai Masala ¹ , Andreea Dutulescu ¹ , Stefan Ruseti ^{1,*} 
and Mihai Dascalu ^{1,2} 

¹ Computer Science and Engineering Department, National University of Science and Technology Politehnica Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania; denis.iorga@upb.ro (D.I.); razvan.muntean@stud.acs.upb.ro (R.M.); mihai_dan.masala@upb.ro (M.M.); andreea.dutulescu@upb.ro (A.D.); mihai.dascalu@upb.ro (M.D.)

² Academy of Romanian Scientists, Str. Ilfov 3, 050044 Bucharest, Romania

* Correspondence: stefan.ruseti@upb.ro

Abstract

The broad adoption of Large Language Models (LLMs) has increased the need for human-curated datasets that serve as evaluation benchmarks. This need is particularly pronounced for non-English languages and for tasks that are inherently subjective and require multiple human perspectives. One such example is the development of benchmarks designed to assess the cultural awareness of LLMs. Statistics and data science courses offer a potential setting for developing such benchmarks while teaching students to apply LLM evaluation techniques using statistical inference. This paper presents a pilot project in which students in a statistics course within a data science engineering program created culturally diverse multiple-choice questions, generated answers using LLMs, and applied statistical methods to assess model accuracy. Student feedback indicated the project was engaging and useful for learning, while also highlighting a notable reliance on LLMs, particularly for interpreting statistical results. The resulting dataset comprises 1355 multiple-choice questions across 18 categories, including language, social media, and politics. After filtering valid items, the dataset was used to evaluate both closed- and open-source LLMs. Results show that the Gemini (closed-source) and Qwen (open-source) model families achieved the best performance, with improvements linked to model size, reasoning capabilities, and access to search tools. The best closed-source model achieved an accuracy of 97.66%, whereas the best open-source model achieved an accuracy of 79.07%. Qualitative analyses of errors in the filtering procedure and model reasoning process point to possible explanations into the challenges LLMs face when handling culturally specific content. Furthermore, results support a cultural injection hypothesis, whereby cultural knowledge is embedded during pretraining and accessed through instruction tuning. Through this work, we aim to demonstrate how statistics and data science courses can provide productive contexts for developing open-source benchmarks for non-English languages while also enriching students' learning experiences. The dataset is publicly available.



Academic Editors: Colin Layfield and John Abela

Received: 19 June 2026

Revised: 16 July 2026

Accepted: 29 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: benchmark; LLM; culture; evaluation; education; statistics; data science; multiple-choice questions

1. Introduction

Large Language Models (LLMs) represent a major area of ongoing research and development [1,2]. One of the most visible outcomes of this activity is the widespread adoption of tools such as ChatGPT, Claude, and Gemini among the general population.

Other models, including Llama, Qwen, and Gemma, are more commonly used within the data science and machine learning research communities.

LLM evaluation has emerged as one of the most high-impact research directions for ensuring the proper functioning of these technologies [3]. A central activity within this area is the development of benchmarks. A benchmark is “a particular combination of a dataset or sets of datasets [...] and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods” [4]. In general, benchmarks can be thought of as datasets that include instructions for LLMs and an answer assessment protocol for those instructions.

Benchmarks guide LLM development and support the assessment of models’ capabilities and limitations [5]. A recent survey of 283 benchmarks identifies three main categories: general capabilities benchmarks (e.g., natural language understanding, college-level exams, causal reasoning), domain-specific benchmarks (e.g., mathematical problem solving, financial knowledge, code understanding), and target-specific benchmarks (e.g., frequency of hallucinations, planning ability, cultural intelligence) [6].

A direction in recent work related to benchmark development focuses on the construction of culturally grounded benchmarks. A recent survey [7] provides a comprehensive overview of this area. The survey highlights a growing diversity of datasets designed to capture cultural variation across regions and linguistic contexts. However, the survey indicates that Romanian culture has limited representation within existing cultural benchmarking datasets.

Against this background, recent efforts have sought to develop benchmarks for the Romanian language [8–14], with some focusing specifically on cultural knowledge [12,14,15]. However, culture is both multifaceted and constantly evolving; therefore, a holistic evaluation requires multiple perspectives and continuous updating. In this context, the development of Romanian cultural benchmarks is particularly timely and important.

One opportunity for the development of such benchmarks is provided by statistics and data science courses. This aligns with initiatives aimed at integrating LLMs in statistics and data science activities [16–19]. In these settings, students can engage not only with the use of LLMs as learning tools but also with their evaluation and improvement. This is especially relevant because the relationship between LLMs and statistics and data science is bidirectional: LLMs can support data-related work, while statistical and data science perspectives can contribute to improving LLMs, particularly in areas such as trustworthiness, transparency, alignment, evaluation, optimization, and uncertainty quantification [20–24].

From an educational perspective, this creates an opportunity to rethink statistics and data science courses as spaces in which students can develop both technical and critical competencies related to LLMs. When appropriately guided, LLMs can support teaching and learning processes [25], including data visualization, statistical reasoning, and prompt-based learning, particularly for students with limited programming experience [16–18]. At the same time, involving students in benchmark development can foster critical, ethical, and reflective engagement with LLMs, while also contributing to resources that require multiple perspectives and continuous updating.

In this context, the present paper introduces a case study of a pilot project exploring the development of a cultural LLM evaluation benchmark for the Romanian language. The benchmark was developed together with students enrolled in a statistics course within a data science engineering program.

Through this work, we aim to achieve the following three research objectives:

- **RO1:** Provide educators in statistics and data science courses with a case study of a project-based learning activity that aims to introduce students to LLMs evaluation via statistical methods;
- **RO2:** Present an assessment of students' perceptions of the project, as well as the extent and manner in which they used LLMs to support its completion;
- **RO3:** Release a new multiple-choice question (MCQ) benchmark of 1355 questions focused on Romanian culture (RoCulturaMCQ), and report results from a series of analyses conducted using the dataset. The dataset is available at <https://github.com/upb-nlp/RoCulturaMCQ> (accessed on 28 July 2026).

2. Material and Methods

The present case study involved developing a data generation and annotation pipeline as part of a statistics course project, designed both to provide students with a meaningful learning opportunity and to produce a valuable research output. The project was carried out with students enrolled in a statistics course at a Romanian technical university, part of a data science master's program, most of whom held an engineering bachelor's degree. The project included 7 stages that are summarized in Figure 1.

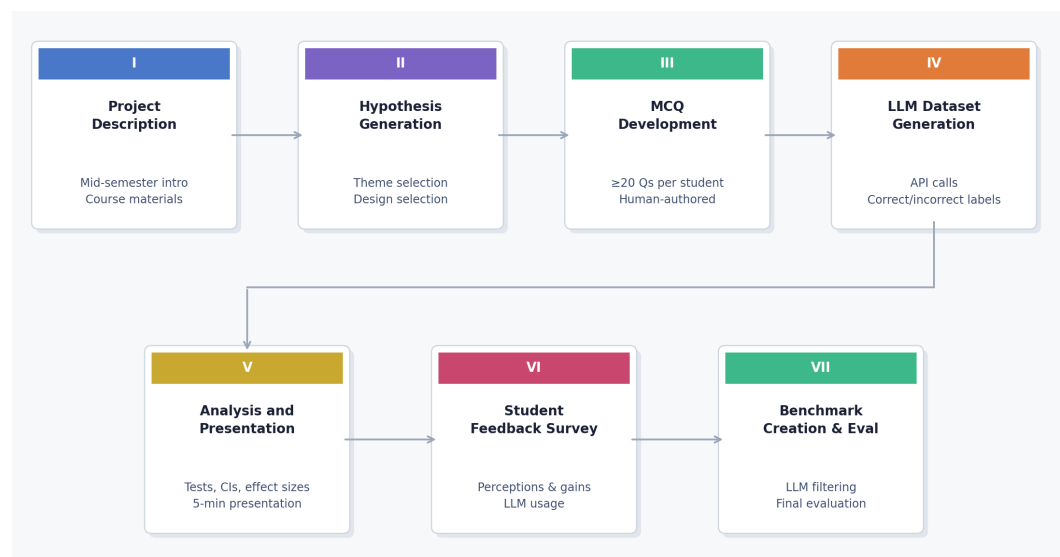


Figure 1. Project stages.

The **project description** phase involved introducing the project to the students. The task was presented midway through the semester during class and through course materials. Besides the course, students could participate in three laboratory sessions dedicated to project completion. The first laboratory focused on theme selection, hypothesis generation, and question development (Stages II and III in Figure 1). The second laboratory focused on refining the questions and using LLMs to generate the dataset (Stages III and IV). The final laboratory focused on applying statistical techniques to analyze the data (Stage V).

The **hypothesis generation** stage required students to formulate both a null and an alternative hypothesis. Students were advised to first select a theme related to national and local culture for question development. Then, they were advised to consider simple accuracy comparisons, either between the answers of two specific LLMs or between prompts to the same questions (within-subjects design), or between the responses of the same LLM to two different types of questions (between-subjects design). All hypotheses were centralized in a Google Sheets file, and individual feedback was provided.

In the **MCQ development** phase, students created at least 20 multiple-choice questions related to the selected theme. Each question included one correct answer and three incor-

rect options (i.e., distractors). Students were instructed not to use LLMs to generate the questions or the answer options. Between 20 and 30 question–answer pairs were collected per student and afterward reviewed in a Google Sheets file. Where possible, individual feedback was provided on the appropriateness of the questions and answers, as well as on their alignment with the proposed hypothesis.

In the **LLM dataset generation** stage, students submitted their question–answer pairs as input to one or more LLMs to obtain model responses. They were advised to use application programming interfaces (APIs) rather than chat interfaces. A Google Colab notebook was provided to the students, containing examples of LLM API usage. For each question, they recorded whether the LLM’s answer was correct or incorrect. This process resulted in a dataset that could be used for testing the initial hypothesis.

During the **analysis and presentation** phase, students tested their hypotheses using the collected data. A Google Colab notebook was provided to the students, containing examples of statistical tests, confidence intervals, and effect size calculations in Python 3.12 using statsmodels v0.14.6 and scipy.stats v1.17.0 libraries. Each student presented their results in a five-minute presentation structured around the hypothesis, dataset, results, and a discussion of the findings.

For the **feedback** stage, an anonymous survey was sent to the students to assess their perceptions of the project. Participation was voluntary, and respondents were informed about the intended use of the collected data for research and publication purposes. The survey focused on (1) perceived enjoyment and usefulness of the project, (2) perceived knowledge gains following the project, and (3) LLM usage during project completion. It included two demographic questions (sex and age), two 5-point Likert scales measuring enjoyment and usefulness, a battery of 11 five-point Likert scales assessing perceived knowledge gains, and a battery of 7 tick-box items capturing LLM usage beyond that required by the project. The released dataset is accompanied by the full set of survey questions.

The **benchmark creation and evaluation** stage involved a series of manual and automated procedures conducted after the end of the course. The following paragraphs describe the procedures used to review the final questions and answers for errors, compile them into a dataset, and evaluate the performance of various LLMs.

Benchmark creation involved reviewing the centralized question–answer pairs using both qualitative and automated methods. Qualitative methods were used to group the items into higher-order themes and identify types of flaws in the dataset (i.e., wrong questions, incorrect correct answers, invalid distractors, cases where no definitive answer exists). The latter analysis allowed the development of an automated method to filter potentially flawed items. The procedure involved using three closed-source LLMs (Gemini Pro Preview 3 with search enabled, GPT 5.2, and Claude 4.6 Opus) to identify flawed items. Items that were flagged by at least two models were removed. Details regarding the prompt used for the three filtering models can be found in Appendix B.

Additionally, 100 items were randomly sampled and manually evaluated to estimate the error rate of the LLM-based filtering approach. An initial human reviewer without access to the LLM-filtering method labels assessed each question and assigned a validity label. A second reviewer then adjudicated cases of disagreement between the LLM-based method and the first reviewer. This latter version of the 100 items dataset was used to estimate the error rates.

Benchmark analysis was inspired by Hsieh et al. [26] and Zhang et al. [27], who recently introduced national-specific cultural benchmarks. The current analysis involved evaluating the answers of a set of closed-source and open-source LLMs to the consolidated question-answers dataset. We evaluated the responses of the following closed-source models: GPT 5.2, Claude 4.6 Opus, and Gemini 3 Pro Preview.

For Gemini 3 Pro Preview, we evaluated the model with and without grounding in Google Search. For GPT 5.2 and Claude 4.6 Opus, we also compared performance under an alternative system prompt that emphasized the Romanian cultural context.

Regarding open-source models, the selection focused on widely known models and LLMs trained on Romanian data. Their performance was evaluated using a logit-based approach, whereby each answer option was scored according to the model's pre-softmax logits, and the option with the highest score was selected as the predicted answer. This evaluation strategy provides a consistent basis for comparing both base and instruction-tuned models, which further enables the investigation of the cultural injection hypothesis explored by Masala et al. [15] and presented in (Section 3.3.3).

For closed-source models, logits are inaccessible, making a logit-based evaluation unfeasible. As such, we use a generative approach to evaluate the performance of closed-source models. Although the original prompt already instructed the models to respond with a single letter corresponding to the correct answer choice, we reinforced this instruction by repeating it in the prompt. The generated responses were manually reviewed and corrected when an output did not match the provided answer choices by removing any additional explanations and extracting only the letter corresponding to the selected answer. Manual intervention was required in one case for GPT 5.2, no cases for Claude 4.6 Opus, 16 cases for Gemini without Google search, and 352 cases for Gemini with Google search option.

For Qwen 3.5, we additionally compared the model with and without reasoning (i.e., the model was prompted to think through the process before giving an answer). Because logit-based evaluation may limit a reasoning model's ability to use its reasoning trace, the model was first allowed to generate the reasoning trace freely. The generated reasoning was then passed to the completion-model variant, which was constrained to output only a single letter. For single-token answer choices, this constrained decoding setup is equivalent to logit-based evaluation.

Except for system prompts embedded in closed-source models, we use the same prompts across all models and apply the chat template whenever available. Open models are evaluated using log-likelihood, so generation parameters such as temperature, top-(k), top-(p), and maximum token length do not affect their reported performance. For open reasoning models, generation is used only to produce the reasoning trace, with a temperature of 0.7. All other hyperparameters, including those for closed-source models, use the default implementation or API settings. The benchmark creation and evaluation workflow is shown in Figure 2, and all scripts are released alongside the dataset to support reproducibility.

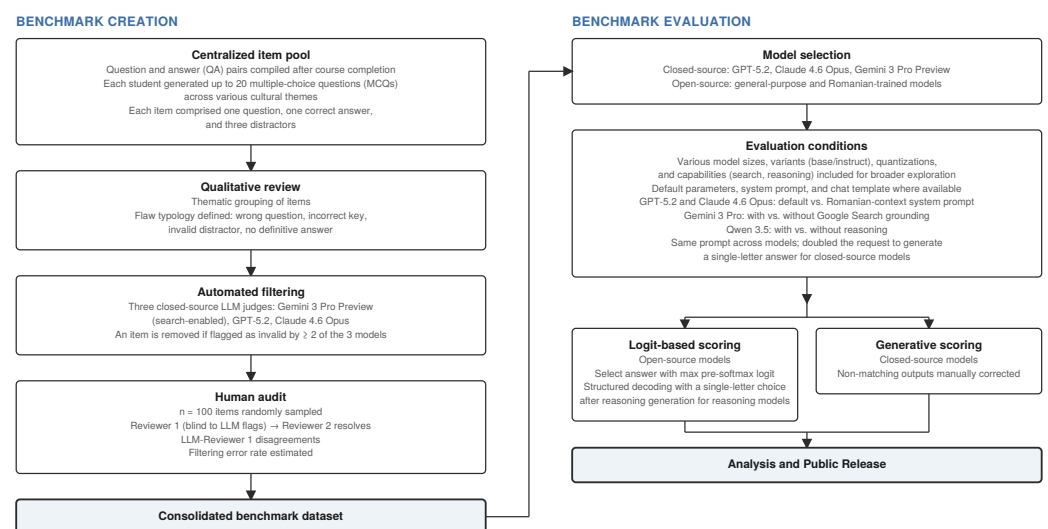


Figure 2. Benchmark creation and evaluation workflow.

3. Results

This section presents the main findings of both the analysis of students' perceptions and the benchmark analysis. The released dataset is accompanied by a table reporting the performance of all tested LLMs, including Wilson confidence intervals.

3.1. Students' Perceptions and LLM Usage

Of the 67 enrolled students, 56 participated in the project. Among these, 42 completed the anonymous feedback survey. The mean age of the respondents was approximately 24 years, with 78.6% identifying as male and 21.4% as female. Students' perceptions of the project were generally positive. The survey recorded a mean enjoyment score of 4.67 out of 5 and a mean perceived usefulness score of 4.52 out of 5 for gaining knowledge about LLM evaluation through statistical methods. The survey also illustrated the extent to which students' knowledge of the main statistical concepts involved in the project had improved relative to their initial level of understanding (Figure 3).

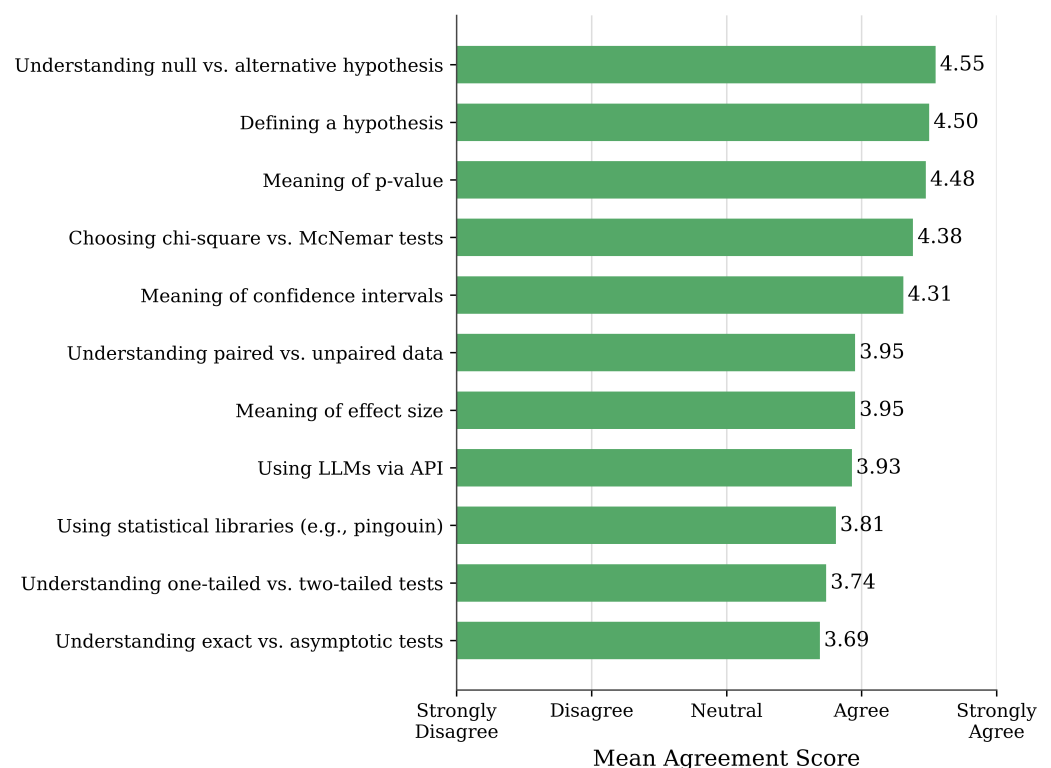


Figure 3. Perceived knowledge gain of concepts required to complete the project.

Students generally agreed that they improved their knowledge of formulating hypotheses, interpreting p -values and confidence intervals, and distinguishing between the uses of the chi-square and McNemar tests. In contrast, responses were more varied regarding their understanding of the differences between paired and unpaired data, the meaning of effect sizes, the use of LLMs via APIs, the use of statistical libraries, the distinction between one-tailed and two-tailed tests, and between exact and asymptotic tests.

The results presented in Figure 3 should be interpreted in the context of an additional factor, beyond the distinction between actual and perceived knowledge gains: the higher level of agreement regarding statistical tests for paired versus unpaired data, compared to the more moderate agreement concerning the underlying distinction between paired and unpaired data itself (which is a prerequisite for selecting the appropriate test), suggests that the terminology or framing of the concepts may have influenced students' responses.

Figure 4 illustrates how LLMs were used in completing the project. As can be observed, commercial solutions such as Gemini, ChatGPT, or Claude were primarily used to interpret results, work with LLM APIs, and implement statistical methods. They were used to a lesser extent to prepare the final presentation, refine benchmark questions, and generate hypotheses or questions for the benchmark.

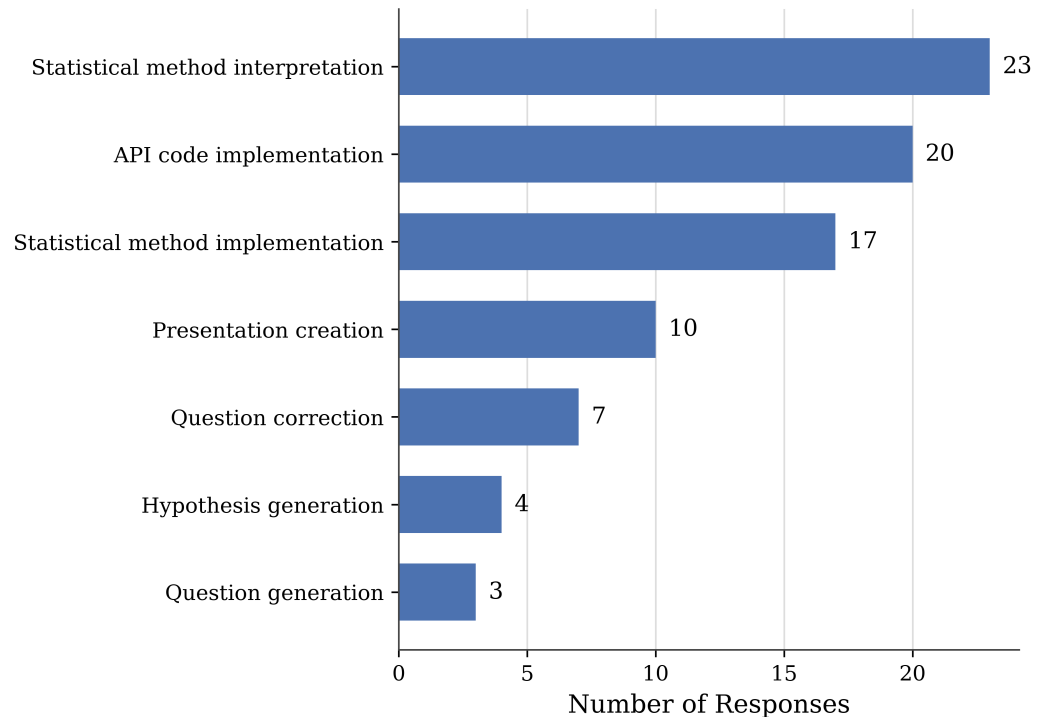


Figure 4. Self-reported LLM usage for project completion.

The results presented in Figure 4 prompted an investigation into the relationship between LLM usage during the project and perceived knowledge gains. A correlational analysis was conducted to examine the links between LLM usage—specifically for implementing API code, implementing statistical methods, and interpreting statistical method results—and students’ self-reported gains in API proficiency and statistical literacy (e.g., understanding statistical libraries, confidence intervals, effect sizes, and p -values). However, only very weak correlations were identified (the lowest p value = 0.318). These findings suggest a lack of evidence that LLM usage to complete the project directly influenced perceived knowledge acquisition, though it remains possible that such tools impact actual rather than perceived learning outcomes. Moreover, the small sample size may have limited the statistical power of the analysis, reducing the likelihood of detecting meaningful associations even if they were present.

3.2. Benchmark Creation

Following project completion, a dataset comprising 1355 question–answer pairs was collected and evaluated using LLMs to identify flawed questions or answers. A total of 328 items were marked as flawed and therefore removed. These consisted of items that were marked as “not valid” by at least two LLMs due to any of the following reasons: wrong questions, invalid correct answers or distractors, and cases where no definitive answer exists.

The final benchmark consists of 1027 question–answer pairs covering 18 categories. As illustrated in Figure 5, there is a notable imbalance in the distribution, with categories ranging from “Expressions & Idioms” (115 questions) to “Trivia” (11 questions).

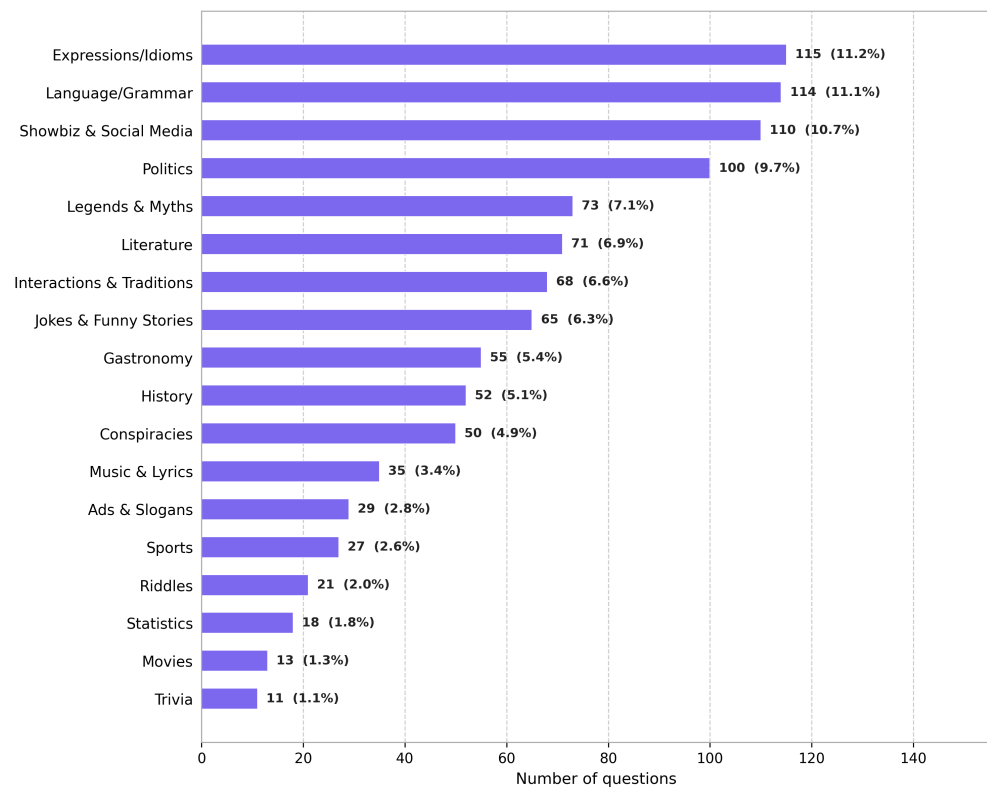


Figure 5. Category distribution of the benchmark.

To evaluate the effectiveness of the LLM-based question filtering method, we manually annotated a sample of 100 question–answer pairs, drawn using stratified random sampling proportional to the prevalence of the valid/not valid classes. The analysis revealed an error rate of 7.1% (95% CI: 3.5–13.9%) for questions labeled valid by at least two LLMs and 36.7% (95% CI: 27.9–46.5%) for those labeled not valid. Inter-annotator reliability analysis resulted in a Cohen’s kappa of 0.59 and a percent agreement of 84%. Based on these point estimates and their 95% confidence intervals, among the 1027 retained questions, approximately 73 may still be invalid despite being marked as valid, with a confidence interval range of about 36 to 143 questions. Conversely, among the 328 removed items, approximately 120 may still be valid, with a confidence interval range of about 92 to 153 questions.

We further analyzed the 100 manually labeled samples to better understand the errors produced by the LLM-based filtering approach. Among the questions that were erroneously marked as valid by the models, discernible causes included unrecognized temporal inconsistencies and incorrect references. For example, in some cases, a question asked about the winner of a television show in a specific year, but the wrong answer was marked as correct and the model failed to detect the error. In another case, a question asked about the current soloist of a band but incorrectly referred to the most popular former soloist, which the model also failed to identify. Similarly, some questions contained incorrect references, such as naming a song “My Half” instead of the correct title, “Your Half,” but these mistakes were not detected by the model.

Among the questions that were erroneously marked as invalid by the models, potential causes included limited recognition of YouTube and TikTok references, uncertainty around political or celebrity quotations, and difficulty evaluating questions that are valid from a Romanian cultural perspective but less widely recognized internationally. Concerning the latter, a representative item concerned the title of the “world’s first female engineer,” a claim that depends on the historical, regional, and disciplinary perspective adopted.

3.3. Benchmark Evaluation

After creating the benchmark, we evaluated a series of open- and closed-source LLMs using it. All model micro and macro accuracies are reported in Table A1. The following sections present the main insights from the evaluation using the micro accuracies. The released dataset is accompanied by a table reporting confidence intervals and category-wise accuracy. These results reveal patterns in LLM behavior and serve as a reference for selecting models for culturally related tasks.

3.3.1. Overall Performance of LLMs

Closed-source models achieved overall better results. Gemini 3 Pro Preview obtained the highest overall accuracy (96.01%), followed by Claude 4.6 Opus (87.63%) and GPT 5.2 (88.12%). For open-source models, Qwen 3.5 with reasoning obtained the highest accuracy (79.07%). Figure 6 illustrates the comparison between these models across categories. As can be observed, Gemini 3 Pro Preview outperforms other models across nearly all categories. The same figure reveals that the Entertainment & Sports cluster—particularly Ads & Slogans (Gemini 3 Pro Preview: 86.2%, GPT 5.2: 72.4%, Claude 4.6 Opus: 62.1%, and Qwen 3.5: 48.3%)—represents the most challenging domain for all models.

We also evaluated the impact of using a system prompt tuned to the Romanian cultural context and of enabling Gemini 3 Pro Preview's embedded search tool. GPT 5.2 registered an under 1% increase in overall performance, whereas Claude 4.6 Opus showed a very slight decrease, suggesting that this adjustment has only a marginal effect on performance. Regarding the use of Gemini 3 Pro Preview's search capability, enabling this option increased the model's overall performance to 97.66%.



Figure 6. Performance overview of best-performing models across categories.

Notably, the use of the search option led to a performance decrease in four categories. Lower scores were observed for Expression/Idioms (−0.87%), History (−1.93%), Jokes & Funny Stories (−1.54%), and Gastronomy (−3.64%). Further investigation into the causes of this phenomenon is required, given the relatively small differences observed and the possible confounding effects of the evaluation methodology and errors in the dataset.

A final experiment relevant to model performance examined the use of reasoning—i.e., allowing the model to think before producing an answer. We observed that this led to an overall performance increase of 6.72% for the Qwen 3.5 model. Here again, we observed a decrease in performance for the Showbiz & Social Media category (0.91%), suggesting a possible overthinking effect.

We further analyzed all the unfiltered questions in the Showbiz & Social Media category for which the reasoning model answered incorrectly while the non-reasoning version produced the correct response. A recurring pattern emerged in the reasoning traces: references to more prominent online personalities, particularly those more extensively represented in written sources, often diverted the model from the correct answer. These competing associations sometimes came to dominate the reasoning process and pulled the model away from the correct answer, even when that answer had been identified correctly. This pattern may also reflect the greater availability of written online content compared with video- and TikTok-centered cultural information. In other cases, the model failed to identify the person referenced in the question, recalled the relevant fact incorrectly, or was constrained by its knowledge cutoff.

3.3.2. Impact of LLM Size and Instruction Tuning

It is generally expected that larger model variants would achieve better performance than smaller variants. This expectation arises because larger models are typically exposed to more data during pretraining (i.e., when the model learns to predict the next token using large-scale internet data, such as websites, books, and news articles). Therefore, observing such a relationship would provide an initial validation of the proposed benchmark.

However, model performance is not determined solely by pretraining. Modern LLMs typically undergo an additional instruction-tuning (or alignment) phase, during which the model learns to follow instructions by learning from examples of human responses to prompts. While pretraining is the stage during which most knowledge acquisition occurs, the alignment stage primarily exposes models to instruction-following behaviors.

Although instruct-tuned models may achieve higher benchmark scores, they can also be more susceptible to evaluation instability and prompt formatting sensitivity [28]. In this context, we examined how benchmark performance varies with model size across different model families, while controlling for the training stage and distinguishing between pretrained and instruction-tuned models.

Figure 7 shows a relationship between model size and cultural knowledge across a broad range of models. Instruct variants generally achieve higher performance, suggesting that instruction tuning improves alignment with the multiple-choice evaluation format, prompts, and answer-selection conventions. Qwen 3 and Qwen 3.5 follow a similar scaling trajectory, but unlike Qwen 3.5, where instruct consistently outperforms the base, Qwen 3 shows the opposite pattern at smaller scales, with the base model matching or surpassing the instruct variant.

Overall, performance improves as model size increases, although the rate of improvement differs across model families. Gemma 3 achieved one of the strongest scaling trends, with performance increasing rapidly as model size grows. Salamandra and EuroLLM models show more conservative growth, with EuroLLM exhibiting a slightly reversed scaling effect for the instruct variant. Finally, Llama-3.1 model scales consistently, though this

trend is largely driven by the inclusion of the 70B model, which is the largest of the models shown in Figure 7.

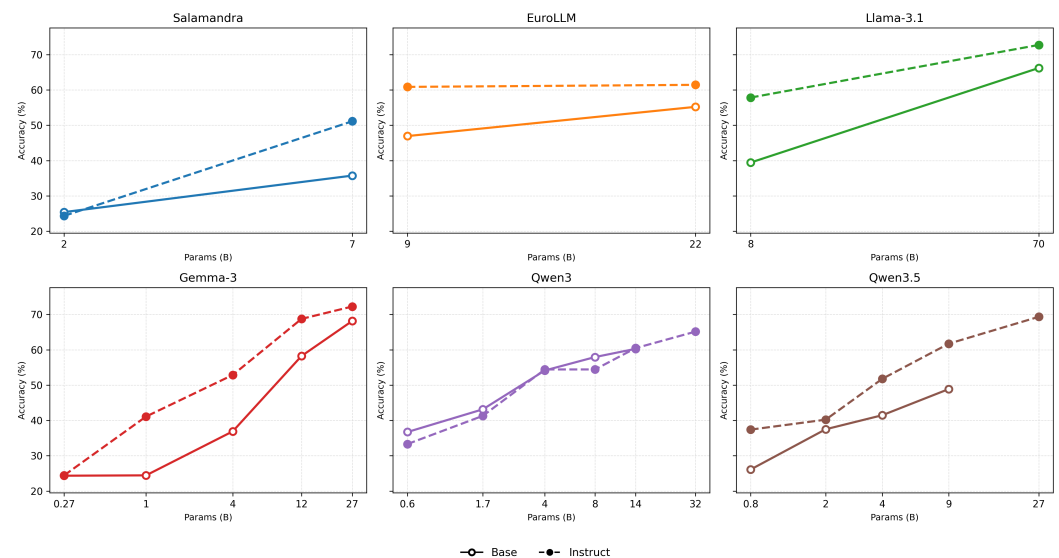


Figure 7. Influence of model size with regard to overall accuracy.

3.3.3. Language Adaptation and Cultural Injection

Because many languages are underrepresented in the data used to train LLMs, such models are often further adapted to better support specific languages. This adaptation can involve continued pretraining (i.e., further training the model on large amounts of text in the target language after the initial training phase) and/or instruction fine-tuning (i.e., training the model with curated prompt–response examples in the target language to better follow human instructions). Such adaptation can improve fluency and the model’s ability to reflect linguistic and cultural nuances.

In Romanian, several models have been adapted through these additional training steps. This raises an important question, also explored by Masala et al. [15]: to what extent do such strategies also introduce or strengthen cultural knowledge within the model? As such, we aimed to use the benchmark in order to answer this question.

The experimental setup focused on the following open-source models: (I) Llama2-7B Base [15] and its instruction-tuned variant; (II) RoLlama2-7B-Base, a version of Llama2-7B that has undergone continued pretraining on Romanian data, and RoLlama2-7B-Instruct-DPO, which further incorporates general instruction fine-tuning and alignment; (III) Mistral-7B-v0.3 Base, along with its original instruction-tuned (III-A) and Romanian fine-tuned (III-B) counterparts.

This choice of models allowed us to assess the contributions of pretraining and instruction tuning in shaping cultural understanding during LLM adaptation to a non-English language. The results of this evaluation are presented in Figure 8, with category (I) shown in blue, (II) in orange, (III-A) in green, and (III-B) in red. Note that for the Mistral models, the performance of pretrained models (PT) is identical, as the same pretrained model was used in both cases (i.e., only instruction tuning was done for that model). This allowed us to isolate and observe the influence of instruction tuning alone.

Additionally, it should be noted that the instruction fine-tuning datasets used to train these models do not contain information specific to Romanian culture. Instead, the data is general in nature and largely derived from translations from English sources. Consequently, adapting the models to Romanian via instruction tuning entailed replacing the original prompt–response datasets with Romanian-translated prompt–response datasets collated

from various sources. Nonetheless, the same instruction fine-tuning datasets were used for both RoLlama2 and RoMistral.

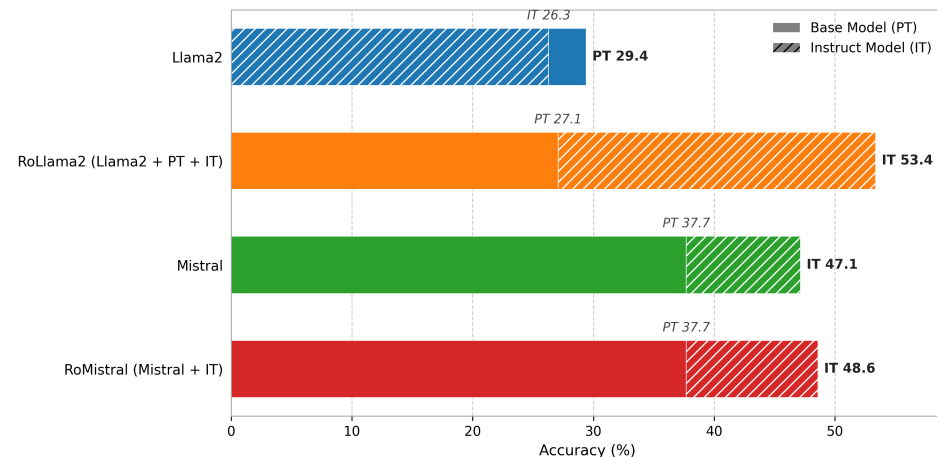


Figure 8. Influence of continued pretraining (CPT) and instruction tuning (IT) with regard to overall accuracy.

Starting with the influence of continued pretraining (i.e., the first two bars in Figure 8), we observe that performing this procedure in Romanian does not always (or at least directly) benefit cultural adaptation. This is reflected in a drop of around 3%, from 29.4% to 27.1%, in the comparison between Llama2 (PT) and RoLlama2 (PT).

It is important to note that for older models like Llama2, the pretraining variants are only completion models, in contrast to the newer training paradigm that adds instruction-like data in the pretraining phase [29,30]. This, combined with the nature of the proposed benchmark, may account for the slight drop in performance, as it deviates from the multiple-choice question-answering format.

However, later in the training life cycle, after instruction tuning, the performance increase is considerable for the language-adapted Llama2 model. The version exposed to Romanian training data via continued pretraining (RoLlama2 IT) performs almost twice as well as its counterpart (Llama2 IT) (53.4% vs. 26.3%). This pattern does not hold for the language-adapted Mistral model (RoMistral IT), which did not undergo continued pretraining and shows only a marginal improvement over the original instruction-tuned version (Mistral IT), with accuracies of 48.6% and 47.1%, respectively.

These results show that starting from a relatively “weak” base model (Llama2), followed by the injection of cultural knowledge and general instruction tuning, results in a highly culturally aware model. Remarkably, this model performs on par with, or even outperforms, newer models built on substantially stronger base architectures, such as RoMistral, RoLlama3, and RoGemma.

This leads us to the conclusion that cultural knowledge is largely encoded and integrated during the pretraining phase, while instruction-tuning—even if it is general—enables the efficient extraction and utilization of this knowledge. These findings are consistent with previous studies [15].

4. Discussion

The results presented in this work indicate that all three research objectives were successfully achieved. First, we provided a concrete case study of a project-based learning activity in statistics and data science courses, illustrating how students can be introduced to the use and evaluation of LLMs via statistical methods (RO1). Second, through the survey analysis, we assessed students’ perceptions of the project, including its perceived usefulness and enjoyment, as well as the extent and manner in which LLMs were used

to support task completion (RO2). Finally, we released the RoCulturaMCQ benchmark, consisting of 1355 multiple-choice questions centered on Romanian culture, and reported a series of analyses evaluating the performance of various LLMs on this dataset (RO3).

Overall, students reported generally positive perceptions of both enjoyment and knowledge gains. However, because knowledge gains were assessed subjectively, it remains possible that the learning objectives were not fully achieved. Future work should therefore incorporate objective assessments of students' understanding of statistical methods for LLM evaluation after completing such projects. Moreover, while one of the project's implicit aims was to teach students how to formulate a hypothesis, design an experiment, and carry it out, the face-to-face presentations revealed that a notable proportion of students iterated on LLM model choice and prompt selection until they could confirm their hypothesis. While this had no impact on the benchmark itself, it poses a problem in terms of achieving the learning objectives for the students. A possible solution to this issue is to introduce a dedicated stage in which experimental conditions are predefined before data collection and analysis. Alternatively, the activity could be reframed as "finding hard-to-answer questions," with such iterations included as part of the project objective.

Aggregating and curating the questions yielded a multiple-choice question (MCQ) benchmark that can assist LLM evaluation on culturally related tasks. Such MCQ benchmarks continue to play an important role in the research ecosystem. They remain valuable for analyzing patterns in LLM behavior, as illustrated by work that uses MCQ-based settings to investigate internal model mechanisms [31]. MCQ datasets are also widely used in alignment and training strategies, where structured question-answer formats facilitate controlled evaluation and optimization [32,33]. Furthermore, they can serve as a foundation for more complex evaluation frameworks, with recent work demonstrating how simple MCQ tasks can be extended into multi-hop or more dynamic reasoning settings [34].

The released benchmark—RoCulturaMCQ—supplements recent efforts to develop cultural evaluation datasets for Romanian. Unlike RoCulturaBench [15], which uses 100 open-ended prompts designed to capture subjective aspects of Romanian culture, RoCulturaMCQ uses multiple-choice questions with a single intended correct answer, making it more suitable for accuracy-based evaluation. RoCulturaMCQ also differs in size, diversity, and format from the dataset introduced in a Romanian history bias study [12], which uses a small set of controversial historical statements to examine cross-lingual bias. Finally, compared with the "Who Wants to Be a Millionaire?" dataset [14], RoCulturaMCQ offers broader category diversity, including domains such as social media or politics. Furthermore, because its questions were not extracted from curated television quiz-show material up to 2019, RoCulturaMCQ also captures a more grassroots and recent view of Romanian cultural reality. However, unlike the "Who Wants to Be a Millionaire?" dataset, RoCulturaMCQ lacks labels related to Romanian-specific versus international content or difficulty level.

The results reported for the "Who Wants to Be a Millionaire?" dataset [14] also provide an opportunity for comparison with those obtained on RoCulturaMCQ. The model rankings are not identical across the two benchmarks, but they are highly similar. Among the overlapping models, Llama-3.3-70B-Instruct achieves the highest score on both RoCulturaMCQ and the Romanian-specific subset of the "Who Wants to Be a Millionaire?" dataset, while Mistral-7B remains the lowest-performing model in both settings. The main differences consist of small rank reversals among mid-performing models, such as Gemma2-9B and EuroLLM-9B, as well as among Romanian-adapted or smaller models. Overall, this suggests that the two benchmarks produce broadly comparable model orderings when assessing Romanian cultural knowledge, despite differences in dataset source, coverage, and evaluation protocol. Furthermore, across models, performance on RoCulturaMCQ tends to be close to, but slightly lower than, performance on the Romanian-specific subset

of the “Who Wants to Be a Millionaire?” dataset, suggesting that the former is slightly more difficult.

Several limitations of the resulting RoCulturaMCQ benchmark should be considered. First, although variability in how students formulated questions was partly controlled by capping the number of items per student (maximum 30), the dataset still reflects contributions from a relatively homogeneous group. All questions were created by students from a single university in one city, which may limit the diversity of perspectives and topics covered. In addition, the distribution of categories is imbalanced, which may bias the evaluation toward more represented topics. Therefore, one direction for future benchmark development is to expand cultural coverage through multi-institutional, multi-annotator, and multi-regional balanced data collection.

It is also worth noting that a small number of questions address topics at the intersection of international and local culture. For example, 20 items in the “Conspiracies” category refer to globally circulating narratives. Moreover, these are not framed as “According to conspiracy theories...” but instead posed as direct questions. Nevertheless, the answer options appear to guide models toward identifying the conspiratorial response. Likewise, 15 items in the “Literature” category draw on widely known international works. These were retained, as such topics are still shaped by and interact with the national cultural context. In this regard, future work should focus on labeling RoCulturaMCQ benchmark items according to whether they reflect Romanian-specific or international content.

Another limitation concerns how students generated their questions. During presentations, at least three strategies were identified: consulting people in their immediate environment, using offline books and materials, and relying on online sources. The latter two approaches may partly explain the strong performance of the Gemini model, as it is developed by Google and may benefit from training data or retrieval mechanisms aligned with the types of online resources students likely use. This is closely linked with the risk that some items may overlap with content present in models’ training data, creating potential data contamination. As such, assessing the benchmark’s exposure to data contamination risk is another important direction for future work.

Furthermore, because the benchmark uses a multiple-choice format, model performance may be affected by random guessing or by option-cueing, whereby answer options provide clues that make questions easier than they would be in an open-ended setting. Accordingly, future work should assess order sensitivity and evaluate model accuracy in an open-ended version of the benchmark (i.e., where answer options are removed and responses are assessed based on their distance from the correct answer).

The filtering method employed prior to the benchmark analysis also represents a limitation of the current results. Of particular concern is the circularity introduced by using the same closed-source LLMs for both filtering and evaluation. This approach may have led some difficult but valid questions to be classified as invalid, thereby potentially inflating the performance of these models and possibly of other models as well. Consistent with this concern, the human annotation of 100 items showed that some questions marked as invalid were in fact valid.

To assess the impact of the filtering method on model performance, we conducted a sensitivity analysis. The results showed that filtering had minimal impact on the relative ranking of models, with a Kendall’s coefficient of concordance of 0.99 between the rankings obtained with and without filtering. Overall, filtering increased model accuracy, suggesting that it successfully removed problematic items. However, this increase was more pronounced for closed-source models, which showed an average accuracy gain of approximately 8%, compared with approximately 4% for open-source models. This pattern suggests that, for closed-source models, the observed improvement may partly reflect the

removal of invalid items, but may also partly result from the use of the same models to filter difficult questions. To support experimenting with alternative filtering methods, we released the full unfiltered dataset, including the models' filtering judgments.

A further limitation concerns differences in the evaluation protocol across open-source and closed-source models. Specifically, open-source models were evaluated using a logit-based approach, whereas closed-source models were evaluated through a generative approach, with manual corrections applied where necessary. This limitation arose from the inability to access logits for closed-source models. Furthermore, for the reasoning-based Qwen 3.5 model, we adapted the evaluation procedure so that the reasoning trace was generated before applying logit-based evaluation. An alternative would have been to evaluate it using a generative approach, similar to the closed-source models, but such an approach would not have allowed for a fair comparison with the open-source models. Accordingly, future work should focus on discerning the differences between logit-based and generative evaluations for the released benchmark.

Another limitation affecting both the evaluation protocol and the filtering method is the absence of a systematic prompt-sensitivity analysis. Future work should therefore examine performance under one-shot and few-shot settings, evaluate a broader range of prompt formulations, and investigate automated prompt optimization and selection techniques.

Considering these limitations, the performance of the closed-source models suggests that Gemini 3 Pro Preview shows promising results in terms of Romanian cultural awareness, especially for information that is readily available on the internet. This conclusion should be interpreted in light of the automated filtering, which may have removed difficult but valid questions (i.e., estimated to 92–153 such questions). Nonetheless, other closed-source models and open-source models still exhibited gaps in this scenario. Our benchmark indicated that Gemini is the strongest closed-source choice for culturally oriented tasks, while Qwen (with reasoning) is the best-performing open-source alternative among the models we tested.

The qualitative analyses of errors in the filtering procedure, together with cases in which the highest-performing reasoning-enabled Qwen 3.5 model produced an incorrect response while its non-reasoning counterpart answered correctly, suggest two important hypotheses. First, LLMs appear to have greater difficulty accessing cultural references that are primarily disseminated through video-based media. Second, the overrepresentation of certain cultural references in written sources can introduce competing associations that divert the model's reasoning and lead it away from the correct answer in multiple-choice cultural questions. However, future work is needed to quantitatively validate these hypotheses.

The comparison between the Llama 2 and Mistral model families offers support for the cultural injection hypothesis. Continued pretraining alone did not improve RoLlama2-Base, whereas subsequent Romanian instruction tuning produced a substantial gain. By contrast, RoMistral, which used the same translated instruction-tuning data but no prior Romanian continued pretraining, showed only a marginal improvement. This suggests that continued pretraining provides much of the underlying cultural knowledge, while instruction tuning mainly helps the model access and express it. However, this conclusion is limited by the absence of a Llama 2 baseline with Romanian instruction tuning but no continued pretraining, and by the lack of instruction-tuning data containing explicit Romanian cultural knowledge. As these experimental setups could not be identified among the available models, future work should focus on constructing and evaluating them.

In conclusion, we believe that such a project benefits both students, who gain experience working with current LLM systems and applying inferential statistical methods, and researchers who can gain and produce insights into the area of LLM evaluation. Fur-

thermore, we believe that this approach creates valuable opportunities for collaboration between the statistics and data science communities and the natural language processing community.

Author Contributions: Conceptualization, D.I.; Methodology, D.I.; Software, M.M. and A.D.; Validation, M.M. and A.D.; Investigation, R.M.; Resources, R.M.; Data curation, D.I. and R.M.; Writing—original draft, D.I., R.M., M.M. and A.D.; Writing—review & editing, S.R. and M.D.; Supervision, D.I. and S.R.; Project administration, M.D.; Funding acquisition, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Project “Romanian Hub for Artificial Intelligence—HRIA”, Smart Growth, Digitization and Financial Instruments Program, 2021–2027, MySMIS no. 351416.

Data Availability Statement: Data and code have been made available at the following URL: <https://github.com/upb-nlp/RoCulturaMCQ> (accessed on 28 July 2026).

Acknowledgments: Generative AI tools (ChatGPT 5.3 and Claude Sonnet 4.6) were used to improve language and readability.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Overall Model Performance

The table below presents the micro- and macro-accuracy scores for the tested models on the filtered dataset. Please note that the same three closed-source models (GPT 5.2, Claude 4.6 Opus, and Gemini 3 Pro Preview with search) were used to filter invalid items, which may have inflated their performance relative to the other models.

Table A1. Overall micro and macro accuracy of the models (sorted by micro acc.).

Model	Micro Acc.	Macro Acc.	Model	Micro Acc.	Macro Acc.	Model	Micro Acc.	Macro Acc.
Gemini 3 pro search	97.66	98.12	RoGemma 2 9b it	58.91	56.88	Gemma 2 2b it	44.69	44.14
Gemini 3 pro	96.01	96.18	Gpt oss 20b medium	58.42	59.75	TildeOpen 30b	43.72	41.60
Gpt 5.2 (sp)	88.12	88.94	Gemma 3 12b pt	58.23	56.57	Qwen3 1.7B base	43.14	41.99
Claude 4.6 opus	87.73	87.40	Gpt oss 20b low	58.13	57.91	Tiny aya base	42.84	42.76
Claude 4.6 opus (sp)	87.63	87.69	Qwen 3 8B base	57.94	55.69	Qwen3.5 4b base	41.48	39.99
Gpt 5.2	87.34	88.62	Llama 3.1 8B it	57.84	57.16	Qwen3 1.7B	41.29	42.01
Qwen 3.5 122b a10b reason.	79.07	82.52	Gemma 2 27b	57.35	58.83	Gemma 3 1b it	41.09	39.73
Llama 3.3 70B it	73.13	74.30	Aya expanse 8b	57.06	55.03	RoLlama 3 8b it	40.99	42.28
Llama 3.1 70B it	72.74	73.56	EuroLLM 22B	55.21	54.08	Qwen 3.5 2b	40.21	40.73
Qwen 3.5 122b a10b	72.35	73.18	Qwen 3 4B it	54.43	54.91	Llama 3.1 8B	39.44	37.30
Gemma 3 27b it	72.25	73.50	Qwen 3 8B	54.43	54.58	Mistral 7B v0.3	37.68	35.97
Aya expanse 32b	71.37	71.05	Qwen 3 4B	54.33	54.15	Qwen 3.5 2b base	37.49	36.71
Qwen 3.5 27b	69.33	67.92	Qwen 3 4B base	54.14	53.86	Qwen 3.5 0.8b	37.39	35.29
Llama 3 70B it	69.13	70.23	Llama 3 8B it	53.94	54.96	Gemma 3 4b pt	36.90	35.38
Gemma 3 12b it	68.74	66.79	RoLlama 2 7b it DPO	53.36	50.95	Qwen 3 0.6B base	36.71	36.98
Gpt oss 120b high	68.55	71.77	RoLlama 3 8b it DPO	53.36	52.27	Llama 3 8B	36.42	34.34
Gemma 3 27b pt	68.16	67.17	RoLlama 3.1 8b it DPO	53.16	51.80	Salamandra 7b	35.74	35.43
Qwen 3.5 9b reason.	68.06	68.92	Qwen 3.5 35b a3b base	53.07	51.19	Mistral 7B v0.1 Romanian	34.27	33.90
Gemma 2 27b it	67.48	67.77	Gemma 3 4b it	52.87	52.30	Qwen3 0.6B	33.30	34.31
Qwen 3.5 35b a3b	67.19	66.18	RoLlama 3.1 8b it	52.00	51.16	Llama 2 7b Romanian glora	30.57	30.08
Gpt oss 120b medium	67.09	69.55	Qwen 3.5 4b	51.80	48.70	Gemma 2 2b	30.57	31.46
Llama 3 70B	66.99	68.58	Salamandra 7b it	51.12	47.53	EuroMoE 2.6B A0.6B it	30.57	31.53
Gpt oss 120b low	66.60	68.96	RoLlama2 7b it	50.54	48.79	Gemma 7b it	30.28	30.09
Llama 3.1 70B	66.21	68.74	Aya 23 8B	49.95	47.73	Llama 2 7b Romanian	29.41	31.12
Aya 23 35B	65.43	65.03	RoGemma 7b it DPO	49.85	47.63	Llama 2 7b hf	29.41	31.02
Qwen 3 32B	65.14	64.72	RoMistral 7b it	49.66	46.94	Llama 3.2 1B it	28.63	27.90
Gemma 2 9b it	65.14	64.51	Qwen3.5 9b base	48.88	44.51	RoLlama 2 7b base	27.07	25.29
Qwen 3 30B A3B it	64.36	63.64	RoGemma 7b it	48.78	48.11	Llama 2 7b chat hf	26.29	26.20
Qwen 3.5 9b	61.73	61.35	RoMistral 7b it DPO	48.59	47.48	Llama 3.2 1B	26.19	25.41
EuroLLM 22B it	61.44	60.06	Mistral 7B it v0.3	47.13	47.87	Qwen 3.5 0.8b base	26.10	24.55
EuroLLM 9B it	60.86	57.95	Gemma 2 9b	47.03	44.55	EuroMoE 2.6B A0.6B	25.80	25.45
Qwen3 14B	60.47	58.13	EuroLLM 9B	46.93	47.16	EuroLLM 1.7B it	25.71	26.39
Qwen3 14B base	60.27	58.89	Tiny aya global	46.06	44.72	Salamandra 2b	25.41	27.13
RoGemma 2 9b it DPO	59.88	60.12	Tiny aya water	46.06	45.70	Gemma 3 1b pt	24.44	24.64
Qwen 3 30B A3B	59.88	60.14	Gemma 7b	44.99	45.37	Salamandra 2b it	24.34	25.12
Gpt oss 20b high	59.20	60.17	LLaMa 3.1 8b	44.79	43.54	EuroLLM 1.7B	23.47	21.44

Appendix B. Prompt for Invalid-Question Filtering

This appendix presents the prompt used to filter invalid questions with the three closed-source models. Each model received the same Romanian-language prompt and was run using its default parameters. An item was classified as invalid if at least two models assigned it any label other than “Valid question and answers.” An English translation of the prompt is provided below for readability.

Original Romanian prompt

Box A1. Romanian prompt.

Ești un evaluator de întrebări tip grilă.
Trebuie să analizezi întrebarea, variantele de răspuns și varianta de răspuns marcată drept corectă din câmpul “answer” și să alegi EXACT una dintre următoarele etichete:

1. “No definitive answer”—Nu se poate determina dacă răspunsul corect este într-adevăr corect sau întrebarea nu are un răspuns clar.
2. “Valid question and answers”—Întrebarea și toate variantele sunt corecte, iar răspunsul marcat este corect.
3. “Invalid distractor”—Cel puțin una dintre variantele greșite ar putea fi corectă.
4. “Wrong correct answer”—Răspunsul marcat ca fiind corect este de fapt greșit.
5. “Wrong question”—Întrebarea este greșită, extrem de vagă, face referire la ceva ce nu există sau nu are sens.

Răspunde DOAR cu eticheta (1, 2, 3, 4 sau 5).

English translation

Box A2. English translation.

You are an evaluator of multiple-choice questions.
You must analyze the question, the answer options, and the answer option marked as correct in the “answer” field, and choose EXACTLY one of the following labels:

1. “No definitive answer”—It cannot be determined whether the correct answer is truly correct, or the question does not have a clear answer.
2. “Valid question and answers”—The question and all answer options are correct, and the marked answer is correct.
3. “Invalid distractor”—At least one of the incorrect options could also be correct.
4. “Wrong correct answer”—The answer marked as correct is actually wrong.
5. “Wrong question”—The question is wrong, extremely vague, refers to something that does not exist, or does not make sense.

Reply ONLY with the label (1, 2, 3, 4, or 5).

References

1. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Dong, Z.; Hou, Y.; Zhang, B.; Min, Y.; Zhang, J.; Liu, P.; et al. A survey of large language models. *Front. Comput. Sci.* **2026**, *20*, 2012627. [\[CrossRef\]](#)
2. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–72. [\[CrossRef\]](#)
3. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [\[CrossRef\]](#)
4. Raji, I.D.; Denton, E.; Bender, E.M.; Hanna, A.; Paullada, A. AI and the Everything in the Whole Wide World Benchmark. In Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), Virtual, 6–14 December 2021.
5. Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; Kochenderfer, M.J. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 21763–21813. [\[CrossRef\]](#)
6. Ni, S.; Chen, G.; Li, S.; Chen, X.; Li, S.; Wang, B.; Wang, Q.; Wang, X.; Zhang, Y.; Fan, L.; et al. A survey on large language model benchmarks. *arXiv* **2025**, arXiv:2508.15361.

7. Pawar, S.; Park, J.; Jin, J.; Arora, A.; Myung, J.; Yadav, S.; Haznitrana, F.G.; Song, I.; Oh, A.; Augenstein, I. Survey of cultural awareness in language models: Text and beyond. *Comput. Linguist.* **2025**, *51*, 907–1004. [\[CrossRef\]](#)
8. Dumitrescu, S.D.; Rebeja, P.; Lorincz, B.; Gaman, M.; Avram, A.; Ilie, M.; Pruteanu, A.; Stan, A.; Rosia, L.; Iacobescu, C.; et al. Liro: Benchmark and leaderboard for romanian language tasks. In Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), Virtual, 6–14 December 2021.
9. Rogoz, A.C.; Ionescu, R.T.; Anghel, A.V.; Antone-Iordache, I.L.; Coniac, S.; Ionescu, A.I. A large-scale benchmark for evaluating large language models on medical question answering in Romanian. *npj Digit. Med.* **2026**, *9*, 268. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Ghinea, D.D.; Corbeanu, A.N.; Dumitran, M.A. RoBiologyDataChoiceQA: A Romanian Dataset for improving Biology understanding of Large Language Models. In Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025); ACL: Stroudsburg, PA, USA, 2025; pp. 551–567.
11. Dumitran, M.; Dumitran, A.; Danila, A.M. GRILE: A Benchmark for Grammar Reasoning and Explanation in Romanian LLMs. In Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing—Natural Language Processing in the Generative AI Era; INCOMA Ltd.: Shoumen, Bulgaria, 2025; pp. 316–324.
12. Cocu, M.I.; Cristia, R.C.; Dumitran, A.M. A Cross-Lingual Analysis of Bias in Large Language Models Using Romanian History. *arXiv* **2025**, arXiv:2510.02362.
13. Cosma, A.; Bucur, A.M.; Radoi, E. Romath: A mathematical reasoning benchmark in romanian. In Proceedings of the 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025); ACL: Stroudsburg, PA, USA, 2025; pp. 95–111.
14. Ganea, A.; Popovici, A.A.; Dumitran, M. A Culturally-Rich Romanian NLP Dataset from “Who Wants to Be a Millionaire?” Videos. In Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing—Natural Language Processing in the Generative AI Era; Angelova, G., Kunilovskaya, M., Escribe, M., Mitkov, R., Eds.; INCOMA Ltd.: Shoumen, Bulgaria, 2025; pp. 378–387.
15. Masala, M.; Ilie-Ablachim, D.; Dima, A.; Corlatescu, D.G.; Zavelca, M.A.; Olaru, O.; Terian, S.M.; Terian, A.; Leordeanu, M.; Velicu, H.; et al. “Vorbești Românește?” A Recipe to Train Powerful Romanian LLMs with English Instructions. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024; Al-Onaizan, Y., Bansal, M., Chen, Y.N., Eds.; ACL: Stroudsburg, PA, USA, 2024; pp. 11632–11647.
16. Levine, R.A. Facilitating a Collaborative Relationship between Generative AI and the Statistics Student. *Am. Stat.* **2026**, *80*, 360–371. [\[CrossRef\]](#)
17. Bien, J.; Mukherjee, G. Generative AI for data science 101: Coding without learning to code. *J. Stat. Data Sci. Educ.* **2025**, *33*, 129–142. [\[CrossRef\]](#)
18. Browning, J.W.; Bustard, J.; Anderson, N.; Galway, L. A data science course utilizing GenAI. In Proceedings of the 2024 IEEE Frontiers in Education Conference (FIE); IEEE: Piscataway, NJ, USA, 2024; pp. 1–7.
19. DeLuca, L.S.; Reinhart, A.; Weinberg, G.; Laudénbach, M.; Miller, S.; Brown, D.W. Developing Students’ Statistical Expertise Through Writing in the Age of AI. *J. Stat. Data Sci. Educ.* **2025**, *33*, 266–278. [\[CrossRef\]](#)
20. Tu, X.; Zou, J.; Su, W.; Zhang, L. What should data science education do with large language models? *Harv. Data Sci. Rev.* **2024**, *6*, 1–28. [\[CrossRef\]](#)
21. Sun, M.; Han, R.; Jiang, B.; Qi, H.; Sun, D.; Yuan, Y.; Huang, J. A survey on large language model-based agents for statistics and data science. *Am. Stat.* **2025**, *80*, 331–344. [\[CrossRef\]](#)
22. Su, W. Do large language models (really) need statistical foundations? *Ann. Appl. Stat.* **2026**, *20*, 724–743. [\[CrossRef\]](#)
23. Dobriban, E. Statistical methods in generative ai. *arXiv* **2025**, arXiv:2509.07054.
24. Ji, W.; Yuan, W.; Getzen, E.; Cho, K.; Jordan, M.I.; Mei, S.; Weston, J.; Su, W.J.; Xu, J.; Zhang, L. An overview of large language models for statisticians. *Am. Stat.* **2026**, 1–49. [\[CrossRef\]](#)
25. Ellis, A.R.; Slade, E. A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education. *J. Stat. Data Sci. Educ.* **2023**, *31*, 128–133. [\[CrossRef\]](#)
26. Hsieh, H.Y.; Liu, S.W.; Meng, C.C.; Chen, C.H.; Lin, S.Y.; Lin, H.J.; Huang, H.H.; Wu, I.C. TaiwanVQA: Benchmarking and Enhancing Cultural Understanding in Vision-Language Models. In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track; NeurIPS: San Diego, CA, USA, 2025.
27. Zhang, Y.; Konomi, M.; Xypolopoulos, C.; Divriotis, K.; Skianis, K.; Nikolentzos, G.; Stamou, G.; Shang, G.; Vazirgiannis, M. GreekMMLU: A Native-Sourced Multitask Benchmark for Evaluating Language Models in Greek. *arXiv* **2026**, arXiv:2602.05150.
28. Luan, H.; Tian, C.; Huan, Z.; Zhang, X.; Chen, K.; Zhang, Z.; Zhou, J. BOSE: A systematic evaluation method optimized for base models. In Proceedings of the Findings of the Association for Computational Linguistics; ACL: Stroudsburg, PA, USA, 2025; pp. 16147–16158.
29. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388.
30. Olmo, T.; Ettinger, A.; Bertsch, A.; Kuehl, B.; Graham, D.; Heineman, D.; Groeneveld, D.; Brahman, F.; Timbers, F.; Ivison, H.; et al. Olmo 3. *arXiv* **2025**, arXiv:2512.13961.

31. Zhao, X.; Choenni, R.; Saxena, R.; Titov, I. Finding culture-sensitive neurons in vision-language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*; ACL: Stroudsburg, PA, USA, 2026; Volume 1, pp. 3366–3381.
32. Elrefai, E.; Khaled, E.; Ehab, A. Star at PalmX 2025: Arabic Cultural Understanding via Targeted Pretraining and Lightweight Fine-tuning. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*; ACL: Stroudsburg, PA, USA, 2025; pp. 825–829.
33. Chatwal, P.; Mishra, S.K. Cultura-Arabica: Probing and Enhancing Arabic Cultural Awareness in Large Language Models via LORA. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*; ACL: Stroudsburg, PA, USA, 2025; pp. 837–842.
34. Permadi, V.A.; Tan, X.; Moosavi, N.S.; Aletras, N. No Shortcuts to Culture: Indonesian Multi-hop Question Answering for Complex Cultural Understanding. *arXiv* **2026**, arXiv:2602.03709.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.