



To What Extent Is LLM Performance on Multiple-Choice Questions Driven by Data Leakage? A Case Study with Contamination-Controlled Spanish Undergraduate Exams

Eva Sánchez Salido¹, Adrián Ghajari¹, Guillermo Marco¹, Julio Gonzalo¹, Jesús Abizanda², Roser Morante¹, Alejandro Benito-Santos¹, Laura Plaza¹, Jorge Carrillo-de-Albornoz¹, Víctor Fresno¹, Enrique Amigó¹, Andrés Fernández García¹

¹UNED Research Center in Natural Language Processing and Information Retrieval
ETSI Informática, UNED - Juan del Rosal, 16 28040 Madrid, Spain

²UNED Barbastro, Huesca, Spain

Correspondence: evasan@lsi.uned.es

Abstract The performance of Large Language Models (LLMs) on multiple-choice university-level exam benchmarks such as MMLU is often reported as highly competitive; however, such results raise persistent concerns regarding contamination of public datasets, English-centric bias, and over-reliance on aggregate accuracy as the primary evaluation signal. In particular, the widespread public availability of evaluation data makes it difficult to disentangle genuine generalization from memorization of seen content, while offering limited insight into models' abilities on culturally grounded assessments beyond English. To address these issues, we introduce LUNES (*Leakage-controlled Undergraduate National Exams of Spain*), a new benchmark of 11,881 multiple-choice questions drawn from official final-year undergraduate exams in Spanish, covering 104 courses across 22 degree programs. The dataset has been rigorously verified to exhibit minimal public web exposure through a combination of automated web search and manual inspection, which enables evaluation under minimal contamination conditions in a non-English, country-specific academic setting. Our results show that (i) LLMs retain strong performance on general knowledge and factual questions, even in the absence of web-accessible training data, suggesting that contamination alone does not explain their success on public benchmarks; (ii) however, their performance degrades substantially on culturally grounded and country-specific content, particularly in domains such as Spanish law, economy, and social structure. Remarkably, models consistently perform better on Anglo-centric content than on Spain-specific material even when answering in Spanish, suggesting that the bottleneck lies in culturally grounded knowledge rather than in language skills per se. A question-level error analysis further reveals that these failures reflect systematic gaps in local institutional, legal, and geographical knowledge, even for high-resource languages such as Spanish, that aggregate metrics systematically obscure.

Keywords: Natural Language Processing, Large Language Models, Evaluation, Exams, Multiple-Choice Questions, Data Contamination, Spanish, Cultural Knowledge Gap.

1 Introduction

Recent advances in Large Language Models (LLMs) have led to remarkable improvements in generalization, reasoning, and task transfer [2, 25]. This progress has increased the demand for robust and diverse evaluation frameworks to better understand both the strengths and limitations of these systems. While early benchmarks focused on single-task evaluations or narrowly defined NLP capabilities—as in GLUE [23], SuperGLUE [22], or MMLU [7]—recent evaluation efforts have moved towards broader, more cognitively demanding tasks spanning multiple domains [20, 12, 21]. However, many of the newer benchmarks rely heavily on synthetic or prompt-engineered data, limiting their ability to reflect the complexity of real-world assessments.

As a reaction, a growing body of work has turned to human-centric benchmarks built from expert-authored questions and official exams, which are expected to better evaluate higher-order cognitive skills. For example, AGIEval [27], GPQA [17], and M3Exam [26] consist of university admission tests, bar exams, and graduate-level materials to assess reasoning and domain expertise in realistic settings. Among these, multiple-choice formats have become especially popular: they allow for scalable, automatic evaluation while still challenging models to reason, infer, and contextualize knowledge meaningfully.

However, major limitations remain. On the one hand, most datasets are publicly available on the web, which introduces the risk of data contamination and can compromise evaluation reliability [18]. On the other hand, most exam-based benchmarks are English-centric, overlooking other languages and cultural domains [10].

In this work, we address two closely related questions concerning the evaluation of Large Language Models. First, we ask *to what extent reported model performance on multiple-choice exam benchmarks is driven by data contamination* stemming from the public availability of evaluation datasets. Second, we investigate *how well LLMs perform on culturally grounded, non-English academic content* once such contamination is explicitly controlled for. To this end, we construct a new benchmark of undergraduate-level multiple-choice exams sourced from a private institutional repository, rigorously verifying the public (un)availability of all questions through automated web search and manual inspection. We then evaluate a diverse set of state-of-the-art LLMs under minimal contamination conditions. Our **main contributions are as follows**: (i) the creation of a **large-scale**, contamination-controlled benchmark in **Spanish**, comparable in size and scope to MMLU and verified to exhibit minimal public web exposure; (ii) a systematic evaluation of LLM performance across academic areas, degree programs, and individual courses under these conditions, complemented by a fine-grained question-level error analysis; and (iii) the identification of persistent weaknesses in **domain-specific and culturally grounded knowledge**. Overall, our results show that while contamination may contribute to performance gains observed in existing benchmarks, it is **not the primary driver of model success**, and substantial limitations remain when models are evaluated on culturally specific academic content.

2 Related Work

The evaluation of LLMs has evolved from simple, skill-specific benchmarks—such as ARC [5], OpenBookQA [13], and RACE [9]—to more comprehensive multitask frameworks like MMLU [7], BIG-Bench [20], and MMLU-Pro [24]. A growing trend favors *human-centric evaluations* based on real-world exam questions to assess reasoning and domain-specific knowledge in realistic settings. Notable examples include AGIEval [27], GPQA [17], M3Exam [26], and Humanity’s Last Exam [3], many of which adopt a multiple-choice format to balance cognitive demand and evaluation scalability.

Another line of work investigates the impact of *data contamination*, i.e., the presence of test data in a model’s training corpus, which can artificially inflate performance [1, 18]. While some approaches attempt to detect contamination through *model-level signals*—e.g., by analyzing the training data or prompting the model directly to recover test instances [8, 14]—others emphasize the need for dataset-level integrity, such as verifying the public availability of benchmark items [11]. Interestingly, some studies suggest that contamination does not always correlate with higher performance [4], highlighting the nuanced nature of its effects.

A further limitation of current benchmarks is their strong bias toward English and their reliance on synthetic or crowd-sourced content. While efforts are emerging to support other languages—such as La

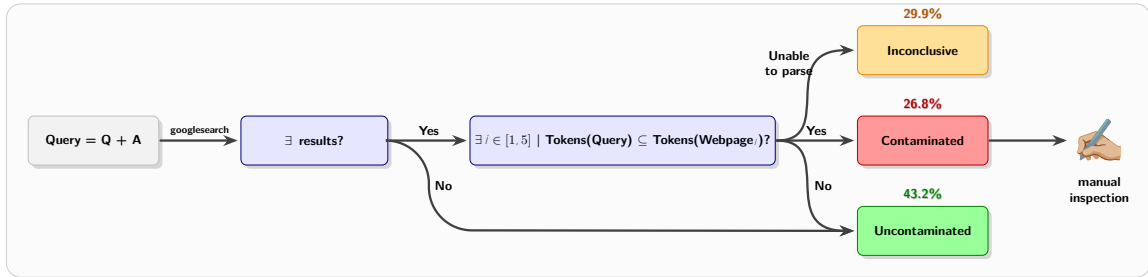


Figure 1: Decision flow for detecting web presence of questions in our dataset.

Leaderboard [6], a leaderboard specifically focused on evaluating LLMs in Spanish—resources based on authentic academic materials remain scarce. Moreover, recent studies have highlighted the challenges LLMs face when reasoning over culturally grounded knowledge. For example, BLEnD [15] evaluates models on everyday facts rooted in diverse linguistic and cultural contexts, exposing performance gaps in non-English and non-Western settings. Similarly, M3Exam [26] reports underperformance on legal and historical questions tied to local contexts. These findings suggest that language coverage alone does not ensure cultural competence, underscoring the need for benchmarks that reflect diverse educational and sociocultural realities.

In summary, we address three key limitations of existing evaluation benchmarks: (i) the risk of data contamination, which we mitigate by sourcing data privately and systematically verifying its absence from web search results; (ii) their English-centric design, which we counter by introducing a large-scale dataset based on academic content in Spanish; and (iii) the lack of cultural grounding, which we tackle by introducing in our dataset real exam questions that reflect Spanish curricula and sociocultural context.

3 Methodology

To address our research question and investigate the impact of data leakage on model performance, we follow a four-step process:

1. **Benchmark construction.** We collect multiple-choice questions from a private institutional repository of final-year undergraduate exams from [university name], whose associated answer keys have never been publicly released. The benchmark is subsequently deduplicated and filtered to remove exact and near-duplicate questions as well as items unsuitable for consistent evaluation (see Section 4).
2. **Web exposure verification.** To assess whether any questions are publicly available online, we implement an automated pipeline that searches each question and its options using a public web search API. A question is flagged as potentially contaminated if its cleaned tokens appear in any of the top-5 retrieved results. Figure 1 illustrates this process.
3. **Manual inspection.** Since web presence does not necessarily imply that the correct answer is available, flagged questions are manually reviewed to determine whether both the question and its correct answer are publicly accessible. This step ensures that only actual cases of meaningful exposure are considered.
4. **Model evaluation.** We evaluate five diverse LLMs (DeepSeek-R1, QwQ-32B, Gemini-2.0-Flash, Phi-4, and LLaMA-3.3-70B) in a standardized zero-shot setting. For each model, we compute Cohen’s Kappa across multiple levels of granularity (academic area, degree program, and course) to assess generalization and domain-specific performance.

This process yields a curated, contamination-controlled benchmark suitable for evaluating LLM performance beyond memorized content.

4 The lunes Dataset

lunes (Leakage-controlled Undergraduate National Exams of Spain) is a curated dataset of 11,881 multiple-choice questions in Spanish, sourced from official third- and fourth-year undergraduate exams at the Universidad Nacional de Educación a Distancia (UNED). It covers 104 courses across 22 degree programs and spans all major academic domains: Sciences, Social Sciences and Law, Arts and Humanities, Engineering, and Health. This dataset builds on uned-access 2024 [19] by introducing more advanced and specialized academic content. To our knowledge, it is the first large-scale benchmark of its kind natively written in Spanish and derived from real university-level assessments.

Data collection The questions were collected via a structured internal API from a private institutional repository of official exams, whose answer keys are not publicly released. All items follow a standardized four-option format and include metadata such as course, degree program, academic area, and year. Academic areas follow the university's official classification.¹

Data distribution Figure 2 shows the distribution of questions per degree program, grouped by academic area. The dataset includes 7,920 questions from Social Sciences and Law, 1,788 from Sciences, 1,352 from Arts and Humanities, 863 from Engineering and Architecture, and 139 from Health Sciences.

Figure 2: Number of questions per degree program.

Deduplication We applied a multi-step filtering and deduplication pipeline to ensure content quality. First, exact duplicates were removed: we tokenized the questions and answers, removed stopwords, and then eliminated exact duplicates based on the cleaned text. Following exact duplicate removal through

¹<https://www.uned.es/universidad/inicio/estudios/grados.html>

text normalization, we addressed near-duplicate questions arising from minor rewordings, answer reordering, or limited option modifications. We compared TF-IDF, Levenshtein distance, and Sentence-BERT [16] embeddings across multiple similarity thresholds, complemented by manual inspection of automatically aggregated cases. Effectiveness was assessed based on whether aggregated pairs aligned with our manual definition of true duplicates i.e., questions differing only by minor wording changes or answer reordering while excluding cases involving variable substitutions, fully reformulated questions, or substantive changes in answer options. Sentence-BERT with a cosine similarity threshold of 0.95 was ultimately selected to prioritize precision over recall, capturing only near-identical or semantically equivalent questions while avoiding over-aggressive filtering of distinct items. Lower thresholds were observed to substantially increase false positives by conflating conceptually related but non-duplicate questions.

Quality control and final selection Lastly, only courses with at least 50 questions were retained to ensure statistical robustness. We also discarded items referencing unavailable visual content (e.g., diagrams or tables) or formats unsuitable for consistent evaluation (e.g., phonetic alphabets). While questions were not individually curated, manual inspection during model evaluation and deduplication revealed only minor artifacts (e.g., numbering), which do not affect comprehension or structure. To preserve its integrity, the dataset will remain private and used exclusively within the ODESIA leaderboard evaluation framework.²

Example questions To illustrate the diversity of courses and reasoning types included in the dataset, Table 1 shows three representative examples from different disciplines.

<p>General Astrophysics</p> <p>Two stars, A and B, have the same radius $R_A = R_B$, but star A is farther from Earth (r_A) than star B (r_B). Which of the following statements is true?</p> <p>A. The relation between their angular radii is $\theta_A > \theta_B$</p> <p>B. The relation between their angular radii is $\theta_A < \theta_B$</p> <p>C. The relation between their angular radii is $\theta_A = \theta_B$</p> <p>D. The angular radius does not depend on distance</p>
<p>Criminal Law II (Law)</p> <p>In the offence of calumny (false accusation), criminal liability shall be extinguished:</p> <p>A. By publishing a retraction in the same medium where the calumny was made, in an identical or similar space, within the deadline set by the sentencing judge or court</p> <p>B. By the pardon of the offended party or their legal representative</p> <p>C. When the perpetrator acknowledges before the judicial authority the falsity or inaccuracy of the accusations and retracts them</p> <p>D. By the publication or dissemination of the conviction judgment, at the expense of the convicted person</p>

Table 1: Example question from the lunes dataset (translated from Spanish for accessibility).

These examples reflect the general type of questions that conform the dataset: domain-specific questions that require factual knowledge, inference, and contextual understanding offering a realistic testbed for evaluating LLMs in Spanish academic contexts.

5 Experimental Setup

5.1 Web contamination estimation

To verify that the lunes dataset is free from meaningful web exposure, we implement the contamination detection pipeline shown in Figure 1. We define potential contamination as the presence of a near-verbatim version of a given question (including its answer options) on publicly accessible web pages. To detect this, we developed an automated procedure that submits each question to a web search engine using the

²https://leaderboard.odesia.uned.es/genai_eval_unedacceso

googlesearch Python package.³ Each search query includes the cleaned version of the question and all its options, with stopwords removed. If all non-stopword tokens appear in any of the top-5 retrieved pages, the question is flagged as potentially contaminated; otherwise, it is considered uncontaminated. In cases where a page cannot be parsed (e.g., due to the format or encoding issues), the result is labeled as inconclusive. This automated process is followed by a multi-step, stratified manual revision to validate the findings and estimate contamination rates more reliably.

5.2 Models

We evaluated five modern language models differing in size and architecture. Gemini 2.0 Flash (Google) is proprietary, with undisclosed size and architecture. The remaining models DeepSeek-R1 (685B; DeepSeek), LLaMA 3.3 70B Instruct (Meta), Qwen's QwQ-32B (Alibaba), and Phi-4 (14B; Microsoft) are open-source. These models range from compact (e.g., Phi-4 with 14 billion parameters) to very large (e.g., DeepSeek-R1 with 685 billion parameters).

Although all models support multilingual input to some extent, their performance on Spanish varies: LLaMA 3.3 and Phi-4 are explicitly multilingual; DeepSeek and QwQ-32B show strong results in English and Chinese; Gemini's language coverage has not been officially documented.

5.3 Inference and prompting strategy

Local inference was conducted using vLLM⁴ on a server with four NVIDIA RTX A5000 GPUs (24 GB VRAM each). LLaMA 3.3, QwQ-32B, and Phi-4 were deployed locally, while Gemini and DeepSeek-R1 were accessed via their respective APIs. All local models used tensor and pipeline parallelism; to accommodate GPU memory limitations, LLaMA 3.3 was quantized using bitsandbytes (default 8-bit quantization).

System prompt (translated from Spanish) You are an expert system for answering exam questions.
User prompt (translated from Spanish) Answer the following question of the subject {} only with the letter of the correct answer. Question: {}

Table 2: System and user prompts used for all models (originally in Spanish, shown here in English for accessibility).

Models were evaluated in a zero-shot setting using a fixed prompt format: a system prompt and a user prompt including the question course, followed by the multiple-choice question with four options labeled A-D. Temperature was set to zero to ensure deterministic outputs. Answers were extracted via regex-based scripts designed to capture the final predicted option. If the model failed to produce a valid letter, the response was marked incorrect. For example, QwQ-32B frequently failed in this regard: although it outputs explicit reasoning mechanisms (e.g., <think> tokens), it often entered unproductive loops and exceeded the context window (8k tokens), yielding no final answer.

5.4 Metrics

To account for the role of random guessing in performance evaluation, we report Cohen's Kappa for all experiments. Kappa corrects for chance agreement setting a random baseline at 0 offering a more informative measure than raw accuracy in balanced multiple-choice settings. Recall that Kappa is calculated as:

³<https://pypi.org/project/googlesearch-python/>

⁴<https://docs.vllm.ai/>

$$\text{Kappa} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}} = \frac{\frac{C}{N} - \frac{1}{M}}{1 - \frac{1}{M}}$$

where the expected accuracy corresponds to that of random guessing, which in our case is 1/4 since questions have 4 options.

6 Results

6.1 Contamination results

Applying the googlesearch pipeline, 8,324 out of 11,881 questions yielded conclusive results, whereas the remaining 3,557 were classified as inconclusive. The most common technical failures leading to inconclusive outcomes included network-related issues (such as connection and read timeouts), server-side retrieval errors, and parsing failures due to malformed or unparseable markup, embedded document formats (e.g., PDFs), dynamically generated content, or character encoding inconsistencies.

Among the conclusive cases, 5,135 showed no web matches, while 3,189 exhibited some degree of web presence. However, web presence alone does not imply usable contamination: a question may appear in search results without its correct answer being accessible or inferable. The following analysis distinguishes between mere textual co-occurrence and meaningful exposure. Figure 3 shows the distribution of contamination results across degree programs, indicating that web presence is fairly evenly distributed.

Figure 3: Distribution of question visibility on the public web based on automated search results, sorted by degree program.

Among the 3,189 flagged cases, 2,811 (88%) came from a single website that hosted only question texts, without answers or explanations. For the remaining 378 questions, we manually reviewed a stratified sample of 287 questions, selected to ensure proportional coverage across courses. Most matches were spurious e.g., token overlap in unrelated documents. Only 18 cases came from informal sources (e.g., student forums) that included answer keys, often embedded in hard-to-extract formats (e.g., interactive elements or bold text). Representative examples illustrating these scenarios are provided in the Appendix.

Since the manual inspection followed a 95% confidence level with 5% margin of error, results can be extrapolated to the full set of 378 and, by extension, to all 3,189 flagged questions. Combining the 3,189 manually validated questions with the 5,135 automatically verified ones, we obtain 8,324 questions over 70% of the dataset free from any usable online presence. Since all conclusively analyzed questions

were found uncontaminated, either via automated detection or manual review, we treat this subset as practically contamination-free. Given that inconclusive cases are evenly distributed across degree programs (Figure 3), this subset is representative of the full dataset. Using Clopper-Pearson exact binomial confidence intervals, we estimate that at least 99.97% of the dataset is free from meaningful web exposure, based on zero observed contaminated samples out of $n = 8,324$, $p = 0.0005$, 95% CI [0.9997; 1.0000]. While conservative, this estimate provides strong support for the reliability of Lunes as a benchmark for evaluating model performance under realistic, minimal-contamination conditions.

6.2 LLM Evaluation

Overall, the evaluation reveals that all models achieve relatively strong performance at an aggregate level, even under minimal-contamination conditions and in a non-English setting. However, this apparent robustness masks substantial heterogeneity across domains, degree programs, and individual courses. While some areas and courses reach high agreement with expert-curated answer keys, others particularly those requiring culturally grounded, country-specific, or highly specialized knowledge exhibit sharp performance drops across models. These patterns motivate a multi-level analysis, as aggregate scores alone fail to capture systematic weaknesses that only emerge at finer levels of granularity.

Evaluation results in terms of Cohen's Kappa, disaggregated by academic area, degree program, and single courses for all the five models considered in our study are supplied below.

6.2.1 Overall performance by model and academic area

Figure 4 displays Cohen's Kappa across academic areas, with 95% bootstrap confidence intervals. Although the models have not seen the correct answers during pre-training, and the questions are posed in Spanish, their performance across all areas of knowledge is consistently at or above a Cohen's Kappa of 0.6 (which corresponds to a raw accuracy of approximately 0.7). DeepSeek-R1 consistently outperforms all other models across domains, followed by Gemini and QwQ though their relative rankings vary slightly in the Sciences. Phi-4 and LLaMA-3.3-70B exhibit lower overall scores.

Figure 4: Bootstrapped Cohen's Kappa (mean ± 95% CI) by area and language model ($n = 1000$ resamples).

At the area level, Arts and Humanities stands out as the domain with the highest overall performance, followed by Sciences. In contrast, Health Sciences emerges as the most challenging area across all models. The wider confidence interval in this latter domain reflects the smaller number of questions in the subset, suggesting that additional data would be needed to draw more robust conclusions. As we will see in the

⁵Although course names are shown in English (e.g., Criminal Law), they are translations of the original Spanish titles and refer specifically to content grounded in the Spanish legal, academic, or cultural context.

following sections, these aggregated results do not capture the full complexity and variation observed across individual courses.

6.2.2 Performance by degree program

Figure 5 presents bootstrapped Cohen's Kappa scores for each model, aggregated by degree program. Programs are sorted by DeepSeek-R1's performance to facilitate comparison across models and disciplines. Each cell displays the model's mean Kappa score with its 95% bootstrapped confidence interval.

Figure 5: Bootstrapped Cohen's Kappa (mean ± 95% CI) per model and degree. Results are sorted by descending DeepSeek's mean results.

In the model-level analysis, DeepSeek-R1 leads across nearly all degrees, with Kappa above 0.9 in Geography and History, and Chemistry, and above 0.7 in most others. Only Mathematics and Law fall below this threshold, though moderately. DeepSeek-R1 is slightly outperformed in Mathematics and IT by Gemini, and in Social Work by both Gemini and QwQ. Gemini and QwQ-32B follow closely, with varying strengths: QwQ surpasses Gemini in Physics, Chemistry, Tourism, Computer Engineering, and Economics. Both models, however, drop considerably in Spanish Literature, Business Administration, Electrical Engineering, and Law where QwQ scores just 0.43. Phi-4 and LLaMA-3.3-70B consistently

yield lower scores. Phi only fails in Law (like QwQ), while LLaMA often falls below 0.5 though it leads in Art History and Social Education.

In the degree-level analysis, Law stands out as the most difficult degree, with three models below 0.5 and the best reaching only 0.65. Mathematics and Electrical Engineering also pose challenges, likely due to abstract reasoning demands, whereas Law's difficulty may stem from its cultural grounding. Other STEM degrees (e.g. Chemistry and Physics), however, rank among the easiest in the benchmark. In fact, no consistent pattern emerges at the degree level: within Social and Legal Sciences, for instance, some degrees perform very well while others rank among the hardest. Sociology and Criminology score high, whereas Economics scores considerably lower. A similar picture holds across the remaining academic areas: in Arts and Humanities, Geography and Art History are quite tractable but Spanish Literature shows the steepest drop in the benchmark; in Engineering, IT performs well but Electrical Engineering proves substantially harder. Since aggregate scores at both the area and degree level mask this heterogeneity, we turn to a finer-grained analysis by individual course, where systematic patterns become more apparent.

6.2.3 Performance by course

Figures 6 and 7 present bootstrapped Cohen's Kappa scores (95% CI) for all 104 courses in the benchmark, grouped by academic area.

Unlike degree-level aggregation, this finer-grained analysis reveals substantial variation across individual courses. For instance, while Geography and History rank high at the degree level, Geography of Spain and its Landscapes performs much worse (DeepSeek-R1: 0.68), with 3 out of 5 models falling below 0.5. A similar trend holds for country-specific courses: Modern History of Spain shows strong results for DeepSeek-R1 (0.88) but failure from two other models; Budget and Public Expenditure in Spain, Spanish and Comparative Economic Policy, and Social Structure of Spain show failure from three or more. Most strikingly, Economic Structure of Spain ranks lowest in the benchmark, with all models scoring below 0.39, whereas History of Modern America tops the list, with all models above 0.73.

Other consistently difficult courses span domains requiring abstract, technical, or specialized reasoning: Criminal Law (three models fail), Environmental Risks in Industry (four fail), Advanced Microeconomics (three fail), and Theory of Money and Banking, where DeepSeek-R1 scores 0.74 but others fall below 0.33.

These results underscore the value of course-level evaluation. Aggregated metrics at the degree or area level may obscure sharp variance in specific topics. The difficulty of these courses likely stems not only from content complexity but also from specialized terminology or contextual reasoning. Spanish academic texts in law, finance, or history may include culturally grounded references or linguistic subtleties that general-purpose models are not equipped to handle, highlighting the need for domain-adapted training and fine-tuning.

Figure 6: Bootstrapped Cohen's Kappa (mean ± 95% CI) per model and course (courses 1-52). Results are sorted by descending DeepSeek-R1 performance. The number of questions per course is shown in grey.

Figure 7: Bootstrapped Cohen's Kappa (mean 95% CI) per model and course (courses 53 104). Results follow the same ordering as in Figure 6. The number of questions per course is shown in grey.

6.3 Performance by cultural category

The results presented so far suggest that performance gaps are not uniformly distributed across domains, but may reflect deeper asymmetries in the cultural grounding of training data. To test this hypothesis directly, we manually classified all 104 courses into three categories based on their subject matter. Spain-specific courses are those whose content is explicitly tied to Spanish law, history, geography, economy, or institutional context (e.g., Modern History of Spain, Spanish Tax System, Geography of Spain and its Landscapes). Anglo-centric courses are those centered on English language and culture (e.g., English Grammar, English for Criminology, Diachrony and Typology of English). All remaining courses, whose content is general and not tied to a specific national or linguistic context, were labeled Universal (e.g., Biochemistry, Thermodynamics I, Advanced Microeconomics). This process yielded 16 Spain-specific courses (2,202 questions), 10 Anglo-centric courses (1,225 questions), and 78 Universal courses (8,454 questions). Figure 8 shows bootstrapped Cohen's Kappa scores (95% CI) for all five models across these three categories. A clear pattern emerges: models consistently achieve their highest performance on Anglo-centric content, followed by Universal courses, with Spain-specific content yielding the lowest scores across all models. This ranking holds regardless of model size or architecture, and suggests that the observed performance degradation is driven by the absence of culturally grounded, country-specific knowledge that is likely underrepresented in model pretraining data, rather than by domain complexity or language alone. The strong performance on Anglo-centric courses, despite being answered in Spanish, further suggests that the bottleneck lies in knowledge rather than in language processing per se.

Figure 8: Bootstrapped Cohen's Kappa scores (95% CI) for all five models across cultural categories.

6.4 Question-level error analysis

To characterize the limitations of current models, we analyze the subset of questions where all evaluated models failed. This subset consists of 750 questions spanning 98 courses.

As a first step, we classified these questions according to whether their content is primarily Spain-related. This classification was based on lexical cues, including courses explicitly tied to Spanish law, geography, institutional structures, or national economic and social systems (e.g., courses containing terms such as Spanish, Iberian, or of Spain). This process identified 15 Spain-related courses, which together account for 206 failed questions.

Figure 9 shows the top 15 courses with the highest number of questions failed by all models, highlighting whether they contain Spain-related content. As shown in the figure, the largest concentration of failures corresponds to Criminal Law II (56 questions), followed by Spanish and Comparative Economic Policy (33 questions). Both courses are strongly grounded in Spain-specific legal and institutional

Figure 9: Top 15 courses with the highest number of questions failed by all models, classified as Spain-specific vs. Universal.

knowledge, and are followed by additional courses related to the Spanish economic, financial, social, or geographical context.

To further characterize these failures, we conducted a manual inspection of representative error cases. While a small number of errors can be attributed to surface-level issues such as strict adherence to textbook definitions that conflict with valid alternative formulations, or transcription and formatting errors in mathematical expressions (e.g., in courses such as Fourier Analysis) the majority of failures reflect deeper limitations related to the absence of country-specific and culturally grounded knowledge. We also identified a small number of potentially ambiguous questions or gold annotation errors; however, these cases are rare and unlikely to affect the overall evaluation trends. Based on manual inspection of failure cases, we identify the following recurrent error categories:

^ Culturally grounded knowledge gaps, with failures concentrated in:

Legal knowledge, particularly in Spanish criminal and administrative law, where correctness depends on fine-grained doctrinal distinctions.

Economic, political, and historical context, especially in courses tied to Spain-specific macroeconomic frameworks, public policy, or historical periodization.

Geographical and territorial knowledge, including region-specific land use, administrative divisions, and physical geography of Spain.

^ Ambiguous or multi-valid questions, where multiple options may appear plausible without access to localized academic conventions.

^ Mathematical or formal notation issues, including transcription errors or formatting artifacts.

^ Answer extraction failures, particularly in models producing verbose reasoning traces (e.g., QwQ-32B).

Overall, these results indicate that aggregate accuracy masks systematic weaknesses in culturally grounded domains, which only emerge through fine-grained, question-level analysis. Table 3 illustrates some representative examples of these failure cases.

<p>Spanish and Comparative Economic Policy (Business Administration)</p> <p>The so-called third Spanish economy was characterized by:</p> <p>A. A growing urbanization process</p> <p>B. An interventionist regime aiming for national self-sufficiency</p> <p>C. A progressive deterioration of the national industrial fabric</p> <p>D. An increase in the active population in the agricultural sector</p>
<p>Geography of Spain and its Landscapes</p> <p>Which land use would most likely dominate in the eastern sector of Sierra Morena (Jaén area)?</p> <p>A. Dehesas used for cattle and pig farming</p> <p>B. Traditional Mediterranean dry farming: olive trees, vineyards, and almond trees</p> <p>C. Use focused on traditional industries, such as local production systems often based on cork oak exploitation in the dehesas</p> <p>D. Forest replanting, increasingly abandoned dehesas, and game hunting</p>

Table 3: Examples of cases where all models fail (translated from Spanish for accessibility). Correct answers are shown in bold.

7 Conclusions

In this paper we introduce *lunes*, a new benchmark of 11,881 curated multiple-choice questions in Spanish, as a methodological element that allows to investigate whether data contamination affects LLMs performance and to help reduce the persistent linguistic and cultural gap in LLM evaluation, which focuses mostly on English. The dataset is sourced from real, private undergraduate exams that are entirely absent from the web, targeting culturally embedded and domain-specific knowledge across a wide range of academic disciplines.

The results of evaluating the state-of-the-art LLMs on *lunes* align with prior findings [4] that suggest that contamination is not a decisive factor in model performance. Overall, performance remains relatively high, particularly at the aggregate level by academic area, indicating model generalization capabilities, even in the absence of question leakage. However, a fine-grained analysis of results by degree and course shows substantial variation across domains: models struggle with courses that demand non-anglo-centric, culturally grounded or language-intensive reasoning, such as Spanish Literature, Law, and Economics.

These results underscore the limitations of works that extract conclusions on the basis of broad metrics, aggregate scores and English-centric datasets. They highlight the need for fine-grained, domain-sensitive, and culturally aware benchmarks to more accurately assess LLM performance in diverse settings. Our findings also point to the importance of leveraging more effective multilingual and domain-adapted training strategies to improve the robustness and cultural inclusivity of language models.

Limitations

Our contamination detection pipeline focuses on identifying near-verbatim public web presence through indexed search results. While this approach is effective for detecting openly accessible leakage, it cannot capture all possible training data pathways. In particular, contamination may still occur through paraphrased versions of questions, documents embedded in non-indexed PDFs, exam-preparation platforms behind authentication walls, or private student-shared materials (e.g., screenshots in closed forums). As such, our verification procedure should be interpreted as a conservative estimate of public exposure rather than an exhaustive audit of all potential training data sources. Nevertheless, by combining automated detection with stratified manual inspection, we ensure that the benchmark is free from meaningful and easily exploitable web contamination, which is the primary concern in benchmark-based evaluation.

Acknowledgments

This work has been funded by the European Union (NextGenerationEU) through the *Recovery, Transformation and Resilience Plan*, by the Ministry of Economic Affairs and Digital Transformation, and by the National Distance Education University (UNED) under cooperation agreement C039-21OT (ODESIA project); and by the Spanish Ministry of Science, Innovation and Universities (project ANNOTATE (PID2024-156022OB-C31), funded by MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+); and TRAIL-LLMs project (AIA2025-163322-C65) in the framework of the coordinated project HumanAIze). Eva Sánchez Salido is supported by an FPI predoctoral fellowship from UNED. However, the views and opinions expressed are solely those of the authors and do not necessarily reflect those of any of the funding agencies or institutions mentioned above, nor can they be held responsible for them.

References

- [1] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [3] Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026.
- [4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1), January 2023.
- [5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018.
- [6] María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, Pedro Reviriego, Leire Rosado Plaza, Alejandro Vaca Serrano, Estrella Vallecillo-Rodríguez, Jorge Vallego, and Irune Zubiaga. La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32482–32524, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021.
- [8] Minhao Jiang, Ken Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Does data contamination make a difference? insights from intentionally contamination pre-training data for language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.
- [9] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical*

- Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [10] Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc., 2024.
- [11] Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An open-source data contamination report for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [12] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification, Outstanding Certification.
- [13] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [14] John X. Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G. Edward Suh, Alexander M. Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize? *arXiv preprint arXiv:2505.24832*, 2025.
- [15] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Victor Gutierrez Basulto, Yazmin Ibanez-Garcia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. BLEnd: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [16] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 EMNLP-IJCNLP*, pages 3982–3992, November 2019.
- [17] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- [18] Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore, December 2023. Association for Computational Linguistics.
- [19] Eva Sánchez Salido, Roser Morante, Julio Gonzalo, Guillermo Marco, Jorge Carrillo-de Albornoz, Laura Plaza, Enrique Amigo, Andrés Fernández García, Alejandro Benito-Santos, Adrián Ghajari Espinosa, and Victor Fresno. Bilingual evaluation of language models on general knowledge in university entrance exams with minimal contamination. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6184–6200, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.
- [20] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska,

- Aitor Lewkowycz, and others. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification.
- [21] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [22] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.
- [23] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.
- [24] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [25] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.
- [26] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: a multilingual, multimodal, multilevel benchmark for examining large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [27] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Appendix

Below we provide representative examples of top web search results used in the contamination analysis, illustrating how web presence was assessed.

Example A: Question-only web presence (no meaningful exposure)

Query: ¿Cuál de los siguientes organismos oficiales de estandarización NO tiene alcance internacional? IEEE. ISO. ITU. ANSI.

Assessment: The question text and answer options are publicly accessible; however, no indication of the correct answer or official solution is provided. This case is therefore classified as free from usable web exposure.

Example B: Web presence with hard-to-extract correct answer

Query: En la búsqueda de nuevos antagonistas de la dopamina, suponga que con la administración conjunta de una nueva droga psicoactiva (5-IT), que activa la liberación de este neurotransmisor, y un

