This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

EuropeMedQA: A Multilingual, Multimodal Medical Examination Dataset for Language Model Evaluation

Francesco Andrea Causio[1,2], Olivia Riccomi[2], Vittorio De Vita[1,2], Federico Felizzi[2], Michele Ferramola[8], Alessandro Tosi[2], Antonio Cristiano[1,2], Lorenzo De Mori[2,3], Chiara Battipaglia[2], Melissa Sawaya[2], Luigi De Angelis[2,4], Marcello Di Pumpo[1,2], Alessandra Piscitelli[2,9], Pietro Eric Risuleo[1,2], Alessia Longo[7], Giulia Vojvodic[2,5], Mariapia Vassalli[2,11], Bianca Destro Castaniti[2,10], Nicolò Scarsi[1,2,6], and Manuel Del Medico[1,2]

[1] Section of Hygiene, University Department of Life Sciences and Public Health, Università Cattolica del Sacro Cuore, Rome, Italy
[2] Società Italiana Intelligenza Artificiale in Medicina (SIIAM), Rome, Italy
[3] Department of Mental Health, Addiction Treatment Unit, ASL RM 4, Bracciano, Italy
[4] Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy
[5] Università Cattolica del Sacro Cuore, Rome, Italy
[6] Department of Medicine and Surgery, University of Perugia, Perugia, Italy
[7] Université Paris Cité, Paris, France
[8] NSBproject, Mantova, Italy
[9] Department of Psychiatry, University of Campania "Luigi Vanvitelli", Naples, Italy
[10] Department of Diagnostic Imaging, Oncological Radiotherapy and Hematology, Università Cattolica del Sacro Cuore, Rome, Italy
[11] Department of Experimental Medicine, Section of Medical Pathophysiology, Food Science and Endocrinology, Sapienza University, Rome, Italy

Corresponding author: Manuel Del Medico, Università Cattolica del Sacro Cuore, Largo Francesco Vito 1, 00168 Rome, Italy
manuel.delmedico01@icatt.it

Abstract. While Large Language Models (LLMs) have demonstrated high proficiency on English-centric medical examinations, their performance often declines when faced with non-English languages and multimodal diagnostic tasks. We present EuropeMedQA, the first multilingual and multimodal medical examination dataset built from official regulatory exams in Italy, France, Spain, and Portugal. We describe a rigorous curation process and an automated translation pipeline for comparative analysis. We evaluated three contemporary LLMs (OpenAI gpt-5-mini, Anthropic Claude Haiku 3.5, and Claude Sonnet 4.5), using a zero-shot, strictly constrained prompting strategy to assess cross-lingual transfer and visual reasoning. EuropeMedQA aims to provide a contamination-mitigating benchmark that reflects the complexity of European clinical practices and fosters the development of more generalizable medical AI.

Keywords: Medical LLM evaluation · Multilingual benchmark · Multimodal medical QA · Cross-lingual robustness · Trustworthy medical AI.

1 Introduction

The rapid adoption of LLMs in medicine has been driven in part by their strong results on established English-centered medical benchmarks [1–5]. Yet these benchmarks do not fully capture real clinical reasoning, multilingual communication, or the country-specific variation in guidelines, epidemiology, and exam structure that shapes medical decision-making across Europe [6, 7]. Current benchmark ecosystems suffer from four recurring limitations: English bias and insufficient geographic diversity, possible contamination from publicly available training material [6, 8], and failure of synthetic question generation to replicate real-world clinical complexity [9].

These limitations are not only methodological; they also affect how confidently model performance can be interpreted. A model that performs well on a widely circulated English benchmark may not generalize to a French multi-answer exam item, a Spanish image-based diagnostic question, or an Italian licensing question written in a distinct regulatory style [10, 11]. EuropeMedQA is therefore positioned not just as another benchmark, but as an attempt to restore ecological validity to medical LLM evaluation in Europe [6, 12].

This study pursues two main objectives. First, it aims to characterize the distribution and intrinsic features of medical examination questions across different European national contexts. Second, it evaluates the performance of contemporary text-based and multimodal LLMs on these questions in both the original language and an English-translated version, in order to assess cross-lingual robustness and the added value of multimodality.

This dual objective gives the work a broader meaning than simple leaderboard comparison. The dataset is both an infrastructure paper and an evaluation study: it documents how the benchmark was built, and it uses the benchmark to test whether current medical LLMs remain reliable when language, format, and image content change.

2 Data Collection and Extraction

EuropeMedQA was designed as a prospective dataset collection and validation project, with emphasis on transparency, reproducibility, and reusability. The dataset was built from official medical licensing or residency-access examinations collected from four countries: Italy, France, Spain, and Portugal. Eligibility required that questions come from central government or officially recognized regulatory sources, be available in their original language and unedited form, and include retrievable images when they were image-based items. Unofficial preparation materials and questions lacking recoverable official images were excluded.

Question extraction followed a predefined protocol supported by a standardized ChatGPT prompt, but the process was not left to automation alone. Six medical doctors fluent in at least one of the source languages performed the extraction, and eight additional researchers independently cross-checked the records for linguistic and clinical accuracy. Discrepancies were resolved by consensus, which strengthens the validity of the resulting benchmark and reduces the likelihood of extraction artifacts.

The curation process included verification of textual alignment with source exams, validation of image–question correspondence, standardization of variable names and answer labels, correction of grammar and punctuation, normalization of categorical fields, repetition of common stems for multi-question clinical cases, and removal of duplicate records. This workflow is methodologically important because benchmark quality in medical AI depends not only on volume, but also on the fidelity of labels, metadata, and image linkage.

3 Dataset Composition

The final benchmark contains 12,340 questions distributed across four national subsets. Italy contributes 7,816 items, Spain 2,410, Portugal 900, and France 1,214. The Italian subset is the largest and combines medical residency admission questions with items from the older national medical licensing examination, which was discontinued in 2020 but retained because of its topical breadth and volume. The French subset is distinctive because it includes multi-correct questions in which one to five answer options may simultaneously be correct.

Table 1. Composition of the EuropeMedQA benchmark by national subset.

Country	Subset code	Total N	Text only	With image	With-image share	Years
Italy	SSM_Q_ITA	7,816	7,756	60	0.8%	2017–2024
Spain	MIR_Q_ESP	2,410	2,083	327	13.6%	2013–2023
Portugal	PNA_Q_POR	900	890	10	1.1%	2019–2024
France	ECN_Q_FRA	1,214	1,070	144	11.9%	2019–2022

The multimodal content is unevenly distributed. Spain and France contain enough image-bearing questions to support meaningful modality-specific analysis, whereas Italy and Portugal include only a small number of image items, so image-stratified accuracy estimates for those countries should be interpreted cautiously. This asymmetry is not a flaw; it mirrors the real structure of the underlying examinations and therefore preserves benchmark realism.

4 Translation Pipeline

To enable cross-lingual comparison, each national subset was translated into English as a preprocessing step using GPT-4o mini with deterministic decoding.

The translation prompt enforced preservation of clinical meaning, numerical values, units, symbols, and answer-option order, while outputs were constrained to structured JSON to prevent field loss or duplication. After translation, the gold answer mapping was checked to ensure that the correct option remained aligned with the translated alternatives.

This translation strategy matters because it isolates language from reasoning. Since translation was performed before inference rather than at runtime, the evaluation tests how the same model behaves when content is phrased in the original language versus a standardized English version. That makes it possible to distinguish language sensitivity from underlying medical knowledge.

5 Evaluation Methodology and Setup

We evaluated three LLMs: OpenAI gpt-5-mini, Anthropic Claude Haiku 3.5 (claude-3-5-haiku-20241022), and Anthropic Claude Sonnet 4.5 through official provider APIs with no local deployment, no self-consistency, no retries, and exactly one inference call per item. The prompt design was deliberately minimal and rigid: the model received a QuestionID, the question text, up to five answer options, and one or more images when indicated by metadata. The system prompt enforced a strict answer format and prohibited explanations, with malformed outputs excluded rather than counted as incorrect.

Before evaluation, answer options were shuffled using a fixed random seed to minimize positional bias and class-imbalance shortcuts. This is a central methodological choice because LLMs can exploit superficial answer distributions in multiple-choice tests if option placement is not controlled. By redistributing the position of correct answers, the benchmark aims to measure medical reasoning more fairly rather than test-taking heuristics.

Two evaluation regimes are used: Italy, Spain and Portugal employ standard single-answer accuracy (a prediction is correct when the chosen letter matches the gold label), while France is scored with Exact-Match Accuracy (EMA), which requires the predicted set of letters to equal the complete set of correct answers. Because EMA is inherently stricter, French scores are lower by construction and should not be directly compared with the single-answer accuracies from the other countries.

6 Results

The study reports a consistent performance hierarchy across countries. GPT-5-mini and Claude Sonnet 4.5 perform at similarly high levels in Italy, Spain, and Portugal, generally in the high-80s to mid-90s, whereas Claude Haiku 3.5 is substantially weaker across all subsets. The French subset is markedly more difficult for all models, with much lower scores due to the multi-answer exact-match rule.

Table 2. Accuracy by country and language variant. France is scored with Exact-Match Accuracy (EMA) and is not directly comparable to the single-answer accuracies of the other subsets.

Country	Variant	GPT-5-mini	Claude Haiku 3.5	Claude Sonnet 4.5
Italy	Original	93.99%	81.40%	94.59%
Italy	Translated	93.01%	82.47%	93.27%
Spain	Original	94.29%	74.60%	93.83%
Spain	Translated	94.37%	77.14%	93.57%
Portugal	Original	88.20%	68.60%	87.73%
Portugal	Translated	88.64%	68.93%	88.91%
France (EMA)	Original	58.13%	36.08%	42.03%
France (EMA)	Translated	56.61%	32.65%	53.26%

The substantive interpretation is straightforward. On single-answer examinations, frontier models show strong and remarkably stable cross-country performance. On the French multi-correct subset, performance drops sharply for everyone, showing that exact combinatorial answer recovery remains substantially harder than selecting one correct option from five.

A model that scores above 94% on Italian or Spanish single-answer questions can still fail badly on multi-answer tasks or on image-dependent items. The French results are especially revealing because they expose weaknesses that would be hidden by a simpler evaluation regime.

Across models, accuracy is generally lower on questions that include images (Fig. 1). For example, GPT-5-mini in Spain drops from 95.98% on text-only questions to 83.06% on image-bearing ones, while Sonnet 4.5 drops from 95.97% to 79.67% in the same subset. Claude Haiku 3.5 shows a similar pattern, with 75.17% on text-only Spanish questions and 69.84% on image-bearing items. In Italy, the image subsets are very small (60 items for GPT-5-mini and Sonnet 4.5, 59 items for Haiku 3.5), so with-image accuracy estimates are informative but not statistically tight.

The French dataset shows the steepest decline: for GPT-5-mini, text-only EMA is 59.42%, falling to 48.61% when images are included; for Haiku 3.5, text-only EMA is 36.68%, dropping to 30.36% with images; for Sonnet 4.5, text-only EMA is 42.67%, with 37.76% on image-bearing items. These values reflect the added difficulty of multi-correct items with visual components, not just the presence of images per se.

Translation into English has little effect on performance in Italy, Spain, and Portugal. For single-answer tasks, the differences between original-language and translated accuracies are small, and the ranking of models remains stable. This suggests that the translation pipeline preserved clinically relevant content adequately and that top-tier models can transfer medical reasoning across these linguistic settings with relatively little degradation.

France is the main exception. Claude Sonnet 4.5 performs substantially better on translated French than on the original-language version, with EMA increasing

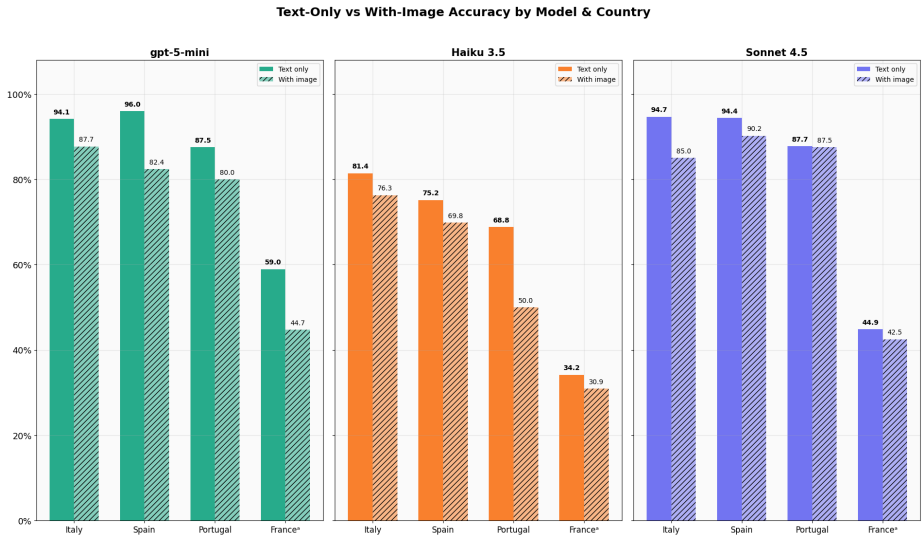

Fig. 1. Text-only versus with-image accuracy for each model across the four national subsets. For every model and country, accuracy on image-bearing items is lower than on text-only items. France ([a]) is scored with Exact-Match Accuracy and is not directly comparable to the single-answer subsets.

from 42.03% to 53.26%. This improvement appears to reflect format sensitivity rather than enhanced medical understanding: in French, the model more often produced verbose text before the expected answer pattern, which led to more parser failures. The translated items seem to impose tighter output formatting, reducing parse errors and improving the measured EMA.

7 Discussion

The clear performance gap between Haiku 3.5 and the two frontier models on every subset suggests that, for content at EuropeMedQA's difficulty level, model capacity remains the dominant factor over language or modality choices: the same Haiku-versus-Sonnet gap is visible on text-only Italian items as on text-plus-image Spanish items. However, the multimodal-versus-text-only comparison must be interpreted carefully. The main practical benefit of multimodal systems is not that they outperform text-only models on every medical question, but that they can exploit information that is genuinely visual and not recoverable from text alone, such as radiographs, pathology slides, skin lesions, or multi-image exam questions requiring integration of several views. Studies in medical visual question answering and imaging consistently suggest that the gain from multimodality is strongest when the question depends on image-derived evidence, cross-modal integration, or spatial pattern recognition rather than on purely textual knowledge.

By contrast, when the image is decorative, redundant, or weakly linked to the answer, multimodality adds little and may even introduce noise. On EuropeMedQA's image-bearing strata, accuracy drops substantially relative to text-only items (for example, for Sonnet 4.5 on Spanish, 96.0% text-only versus 79.7% with image), but the text-only and with-image strata are different question sets. The observed gap therefore conflates two effects: image-bearing questions may be intrinsically harder, and current models may under-exploit the visual evidence, which the present design cannot disentangle. The drop should be read as a property of the harder, image-dependent items rather than as direct evidence that multimodal access degrades performance.

When EuropeMedQA results are compared with other multilingual medical benchmarks such as MedQA and WorldMedQA-V, the French ECN EMA figures (approximately 30–60%) are substantially lower than headline accuracies reported on single-correct USMLE-style tasks [1, 2, 10, 13]. This gap is primarily attributable to scoring rules rather than differences in underlying medical knowledge: EuropeMedQA's exact-match requirement penalizes partial correctness and requires identification of all simultaneously correct options, whereas MedQA and WorldMedQA-V use single-correct multiple-choice formats where selecting any correct option yields full credit, inflating accuracies by roughly 25–30 percentage points relative to exact-match scoring [2, 13]. The pronounced French-specific drop for Sonnet 4.5 (about 11 points from English to French) aligns with cross-lingual heterogeneity documented in WorldMedQA-V, where performance gaps vary substantially by language pair, and suggests that cross-lingual robustness depends on language-specific training exposure, morphological complexity, and output behaviour rather than uniform model properties [13].

The instability of Sonnet 4.5 in French and the moderate EMA levels in ECN underline the need for cautiousness against over-interpreting high accuracy from single-correct, English-centric benchmarks as evidence of clinical readiness. Even the highest EMA observed in the French subset (58.1% for GPT-5-mini) implies that more than 40% of questions are answered incompletely or incorrectly, a rate too high for autonomous clinical or consumer-facing use. Such performance may be acceptable for low-stakes applications such as literature triage, educational tutoring, exam preparation, or clinician-reviewed suggestion generation, but not for unsupervised decision-making in high-stakes settings. Introducing medical AI without proper validation carries documented risks, including data-privacy breaches and algorithmic biases, making it essential to benchmark models on failure modes (partial-credit patterns, formatting sensitivity, language-dependent drift) rather than aggregate accuracy alone. EuropeMedQA contributes to this agenda by providing a multilingual, multimodal, contamination-aware benchmark grounded in real European admission procedures, and by showing that model performance can drift substantially across languages and modalities, even within models from the same provider.

8 Contamination Analysis

To assess whether strong performance might reflect memorization of older public exam material we conducted a year-stratified logistic regression using item correctness as the dependent variable. No significant year trend was found for Spain or Italy, and Portugal showed only a small, non-significant decline. Accuracy remained high even on recent exams such as Spain 2023 and Italy 2024, which argues against contamination as the main explanation for headline performance.

Shuffling removes memorization of the correct letter but cannot fully eliminate memorization of question content. Nevertheless, the absence of a meaningful year effect limits how much contamination can plausibly explain the observed results, and any contribution from memorization appears bounded to a few percentage points.

9 Limitations

The study also has various limitations. First, the benchmark is still exam-based and therefore measures structured test performance rather than full clinical reasoning at the bedside. Second, image-bearing questions are sparse in Italy and Portugal, which limits precision for modality-specific estimates in those subsets. Third, translation may introduce subtle semantic shifts despite the use of deterministic prompts and validation steps.

Additional constraints arise from the evaluation infrastructure. Because the models were accessed via managed cloud APIs, low-level decoding control and full hardware reproducibility were not available. The study also relies on one evaluation run per item, so it does not estimate run-to-run variability under repeated sampling. Finally, the French subset uses a fundamentally stricter scoring rule, which makes cross-country comparison less intuitive even though it is methodologically justified.

10 Conclusion

EuropeMedQA brings coherence to a fragmented evaluation landscape by showing that current frontier models can achieve very strong results on multilingual European medical multiple-choice questions when the task involves a single correct answer, but that their performance becomes less reliable as soon as multimodal reasoning or exact recovery of multiple correct options is required. In this sense, the benchmark shifts the discussion away from the overly simple question of whether LLMs can pass medical exams and toward a more meaningful one: under which linguistic, visual, and scoring conditions they can actually be considered trustworthy. The main contribution of EuropeMedQA is that it goes beyond reporting model scores and shows why those scores vary across countries and task designs, and why the same model may appear highly capable in one context yet considerably less reliable in another. By grounding evaluation in real European medical examinations, the benchmark offers a more realistic

and clinically relevant framework for assessing medical AI in Europe, especially in multilingual settings where translation, formatting, and visual interpretation all influence performance.

Prospects of Application: EuropeMedQA supports safer deployment decisions for medical LLMs in multilingual European settings. Models scoring 90–95% on single-answer subsets may aid education, literature triage, exam preparation, and clinician-supervised drafting, but remain unsuitable as autonomous decision-makers where multiple correct findings, exact formatting, or image-grounded interpretation are required.

Data and Code Availability. The dataset used for all analyses in this paper is publicly available at `https://huggingface.co/datasets/EuropeMedQA/EuropeMedQA` (pinned revision `a803ba6`).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chen, H., Fang, Z., Singla, Y., Dredze, M.: Benchmarking large language models on answering and explaining challenging medical questions. In: Proc. 2025 Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 3563–3599 (2025). https://doi.org/10.18653/v1/2025.naacl-long.182
2. Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H., Szolovits, P.: What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021). https://doi.org/10.3390/app11146421
3. Pal, A., Umapathi, L., Sankarasubbu, M.: MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Proceedings of the Conference on Health, Inference, and Learning (CHIL). Proceedings of Machine Learning Research, vol. 174, pp. 248–260 (2022)
4. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: PubMedQA: A dataset for biomedical research question answering. In: Proc. 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). pp. 2567–2577 (2019). https://doi.org/10.18653/v1/D19-1259
5. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: International Conference on Learning Representations (ICLR) (2021)
6. Alwakeel, M., Nagori, A., Krishnamoorthy, V., Kamaleswaran, R.: Evaluating LLMs in medicine: A call for rigor, transparency. arXiv preprint arXiv:2507.08916 (2025)
7. Felizzi, F., Riccomi, O., Ferramola, M., Causio, F.A., Del Medico, M., De Vita, V., et al.: Are large vision language models truly grounded in medical images? evidence from italian clinical visual question answering. arXiv preprint arXiv:2511.19220 (2025)
8. Riccomi, O., Causio, F.A., De Vita, V., Cristiano, A., Del Medico, M., De Mori, L., et al.: Valutazione one-shot di Mistral7B sul nuovo benchmark EuropeMedQA. Recenti Progressi in Medicina **116**(10), 619–620 (2025). https://doi.org/10.1701/4573.45804
9. Kell, G., Roberts, A., Umansky, S., Khare, Y., Ahmed, N., Patel, N., et al.: RealMedQA: A pilot biomedical question answering dataset containing realistic clinical questions. In: AMIA Annual Symposium Proceedings (2024)
10. Alonso, I., Oronoz, M., Agerri, R.: MedExpQA: Multilingual benchmarking of large language models for medical question answering. Artificial Intelligence in Medicine **155**, 102938 (2024). https://doi.org/10.1016/j.artmed.2024.102938
11. Grzybowski, Ł., Pokrywka, J., Ciesiółka, M., Kaczmarek, J.I., Kubis, M.: Polish-english medical knowledge transfer: A new benchmark and results. arXiv preprint arXiv:2412.00559 (2024)
12. Cheng, N., Li, F.: MulMed: Addressing multiple medical tasks utilizing LLMS. Academic Journal of Computing & Information Science **8**(9), 26–33 (2025). https://doi.org/10.25236/AJCIS.2025.080904
13. Matos, J., Chen, S., Placino, S., Li, Y., Climent Pardo, J., Idan, D., et al.: WorldMedQA-V: A multilingual, multimodal medical examination dataset for multimodal language models evaluation. arXiv preprint arXiv:2410.12722 (2024)