IMB: An Italian Medical Benchmark for Question Answering

Antonio Romano[1,2], Giuseppe Riccio[1,2,*], Mariano Barone[1,2], Marco Postiglione[3] and Vincenzo Moscato[1,2]

[1]University of Naples Federico II, Department of Electrical Engineering and Information Technology (DIETI), Via Claudio, 21 - 80125 - Naples, Italy
[2]Consorzio Interuniversitario Nazionale per l'Informatica (CINI) - ITEM National Lab, Complesso Universitario Monte S.Angelo, Naples, Italy
[3]Northwestern University, Department of Computer Science, McCormick School of Engineering and Applied Science, 2233 Tech Dr, Evanston, IL 60208, United States

Abstract

Online medical forums have long served as vital platforms where patients seek professional healthcare advice, generating vast amounts of valuable knowledge. However, the informal nature and linguistic complexity of forum interactions pose significant challenges for automated question answering systems, especially when dealing with non-English languages. We present two comprehensive Italian medical benchmarks: **IMB-QA**, containing 782,644 patient-doctor conversations from 77 medical categories, and **IMB-MCQA**, comprising 25,862 multiple-choice questions from medical specialty examinations. We demonstrate how Large Language Models (LLMs) can be leveraged to improve the clarity and consistency of medical forum data while retaining their original meaning and conversational style, and compare a variety of LLM architectures on both open and multiple-choice question answering tasks. Our experiments with Retrieval Augmented Generation (RAG) and domain-specific fine-tuning reveal that specialized adaptation strategies can outperform larger, general-purpose models in medical question answering tasks. These findings suggest that effective medical AI systems may benefit more from domain expertise and efficient information retrieval than from increased model scale. We release both datasets and evaluation frameworks in our GitHub repository to support further research on multilingual medical question answering: https://github.com/PRAISELab-PicusLab/IMB.

Keywords

Healthcare NLP, Medical QA Dataset, Generative AI, Large Language Models

1. Introduction

Since the early days of the Internet, online medical forums have facilitated direct, valuable interactions between patients and healthcare professionals, creating an accessible space for medical advice and support. While these platforms serve as vital resources for medical guidance, they present unique challenges for Natural Language Processing (NLP) systems, particularly in Question Answering (QA) tasks. Unlike traditional medical texts, these conversations are characterized by colloquial language, implicit medical knowledge, and cultural nuances that current QA systems struggle to interpret accurately. Existing biomedical QA research has primarily focused on structured, English-language content, leveraging pre-trained models like BERT [1], RoBERTa [2], and BioBERT [3]. While these models have shown promising results on standard QA benchmarks [4], [5], [6], they are predominantly trained on formal medical literature and standardized exam questions [7]. This creates a significant gap between model capabilities and real-world medical communication needs, particularly in non-English contexts. To address these challenges, we introduce two complementary datasets: **IMB-QA** (Italian Medical Benchmark for Question Answering), a comprehensive collection of 782,644 real-world medical conversations across 77 medical categories from Italian online forums MedicItalia[1] and Dica33[2]; and **IMB-MCQA** (Italian Medical Benchmark for Multiple Choice Question Answering), containing 25,862 multiple-choice questions and answers from medical specialty admission exams collected from the simulator CompitoInClasse.org[3]. Both datasets have been carefully curated, with **IMB-QA** specifically enhanced through LLM-based methodologies to ensure quality and anonymity while preserving the authentic nature of patient-doctor interactions.

Our work goes beyond data contribution through extensive experimentation with state-of-the-art language

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
antonio.romano5@unina.it (A. Romano); giuseppe.riccio3@unina.it (G. Riccio); mariano.barone@unina.it (M. Barone); marco.postiglione@northwestern.edu (M. Postiglione); vincenzo.moscato@unina.it (V. Moscato)
https://github.com/LaErre9 (A. Romano); https://github.com/giuseppericcio (G. Riccio); https://github.com/csmariano (M. Barone); http://wpage.unina.it/vmoscato/ (V. Moscato)
0009-0000-5377-5051 (A. Romano); 0009-0002-8613-1126 (G. Riccio); 0009-0004-0744-2386 (M. Barone); 0000-0001-6092-940X (M. Postiglione); 0000-0002-0754-7696 (V. Moscato)

[1]https://www.medicitalia.it/
[2]https://www.dica33.it/
[3]https://www.compitoinclasse.org/

models. We conduct a systematic evaluation of various LLM architectures, comparing models of different sizes and training backgrounds, with particular attention to those specialized in biomedical domains. Through this analysis, we explore the two standard approaches to enhance medical QA performance: Retrieval Augmented Generation (RAG) and in-domain fine-tuning. Our experiments with RAG demonstrate significant improvements in response accuracy and completeness, while our fine-tuning studies reveal the potential of task adaptation even for smaller models. The dual nature of our datasets — spanning both informal forum discussions and formal medical examinations — provides a unique opportunity to assess model performance across different types of medical communication. Our findings challenge conventional assumptions about model size and generalization, suggesting that targeted task adaptation and retrieval-based approaches may be more crucial for medical QA than raw model scale.

Table 1
Comparison of QA and MCQA datasets from prior literature and our proposed **IMB** datasets.

Type	Dataset	# Q/A	Language
QA	BiQA [9]	>7.4K	English
	HealthQA [10]	>7.5K	English
	EPIC-QA [11]	45	English
	COVID-QA [12]	>2K	English
	CliCR [13]	>100K	English
	LiveQA-Med [14]	738	Multilingual
	PubMedQA [15]	>212K	English
	emrQA [16]	>455K	English
	webMedQA [17]	>63K	English
	BioASQ [18]	>3.2K	English
	IMB-QA (Ours)	>782K	Italian
MCQA	HEAD-QA [19]	>6.8K	Spanish
	MedMCQA [20]	>194K	English
	cMedQA [15]	>54K	Chinese
	ChiMed [21]	>24.9K	Chinese
	MEDQA [15]	>61K	English-Chinese
	QA4-MRE [22]	>1.5K	Multilingual
	IMB-MCQA (Ours)	>25K	Italian

2. Related work

In Question Answering (QA), models are typically provided with a relevant text from which they must extract answers. However, in real-world applications, manually curating such texts is impractical due to the high cost of obtaining annotated contexts. This challenge has driven the development of Open-Domain QA (OpenQA), where models must autonomously retrieve and understand relevant information to generate accurate responses [8]. In the biomedical domain, numerous datasets have been introduced to advance QA, particularly in high-resource languages such as English (as shown in Table 1). However, resources for other linguistic domains—especially Italian—remain scarce, limiting the development and evaluation of multilingual biomedical QA models.

Open-Domain and MRC Biomedical QA Several datasets support OpenQA and Machine Reading Comprehension (MRC) in the biomedical field. BiQA [9] compiles questions from online forums (e.g., Stack Exchange, Reddit) and links them to PubMed articles, though the accuracy of this linking remains largely unverified. HealthQA [10] consists of manually curated medical questions with answers sourced from patient information websites, yet it lacks a systematic quality assessment. BioRead [23] and its extended version, BioMRC [24], annotate texts using Unified Medical Language System (UMLS) concepts, enhancing knowledge representation but focusing more on structured information extraction rather than OpenQA. The COVID-19 pandemic and the creation of specialized datasets such as EPIC-QA [11] and COVID-QA [12], which compile question-answer pairs from pandemic-related literature. However, their long-term relevance is inherently limited to this specific context. CliCR [13] employs cloze-style questions derived from clinical case reports to assess comprehension and inference abilities, yet its scope is restricted to a narrow set of medical conditions. Although most biomedical QA datasets are available only in English, some efforts have targeted other languages. LiveQA-Med [14] provides a small set of 634 annotated medical question-answer pairs, but its test set (104 questions) is too limited for robust evaluation. MEDQA [15], built from medical board exams in English and Chinese, does not clearly specify the balance between languages or the translation quality. WebMedQA [17], derived from Chinese health consultancy platforms, reflects real-world medical inquiries, though its reliability depends on the moderation of user-generated content.

Multiple Choice QA Several datasets focus on multiple-choice QA (MCQA) for biomedical applications. HEAD-QA [19] and MedMCQA [20] assess domain knowledge and reasoning skills but lack coverage for Italian. PubMedQA presents a distinct format where article titles serve as binary-answer questions, though it does not address complex inferential reasoning. While ChiMed [21] and cMedQA [15] provide Chinese-language biomedical MCQA datasets, Italian biomedical QA resources remain virtually nonexistent. QA4-MRE [22] attempted to introduce multilingual medical reading comprehension, yet its dataset was limited in both scale and scope. To address this gap, we introduce a large-scale Italian biomedical QA dataset, consisting of 782,644 question-answer pairs spanning 77 medical categories, alongside an Italian biomedical MCQA dataset with 25,862 multiple-

Table 2
Example of a question-answer pair from **IMB-QA**.

IMB-QA	
ID	14571
URL	https://www.medicitalia.it/consulti/psicoterapia/23549-psichiatri-per-fare-psicoterapia-necessitano-di-specializzazione.html
Question	**ITA:** Psichiatri: per fare psicoterapia necessitano di specializzazione? Gentili dottori, uno psichiatra è automaticamente abilitato a fare lo psicoterapeuta, o deve, come gli psicologi che intendono anche essere psicoterapeuti, fare i 4 anni di specializzazione in psicoterapia? Grazie. **ENG:** Psychiatrists: do they need specialization to practice psychotherapy? Dear doctors, is a psychiatrist automatically allowed to practice psychotherapy, or does he need, like psychologists who also wish to be psychotherapists, to undergo 4 years of specialization in psychotherapy? Thank you.
Reformulated Answer	**ITA:** Dopo la specializzazione, uno psichiatra può iscriversi all'albo degli psicoterapeuti senza dover completare un'ulteriore specializzazione. Sebbene alcuni psichiatri scelgano di proseguire con studi supplementari in psicoterapia, tale formazione non è obbligatoria. Gli psicologi, invece, devono frequentare un corso di specializzazione per diventare psicoterapeuti. **ENG:** After specialization, a psychiatrist can register with the list of psychotherapists without needing further specialization. While some psychiatrists choose to pursue additional studies in psychotherapy, this training is not mandatory. Psychologists, however, must attend a specialization course to become psychotherapists.
Category	**ITA:** Psicoterapia - **ENG:** Psychotherapy
Macro Category	**ITA:** Salute Mentale - **ENG:** Mental Health

choice questions across 60 categories. Compared to existing datasets, our corpus is significantly larger and more diverse, enhancing both domain-specific knowledge extraction and OpenQA capabilities. Furthermore, we employ advanced post-processing techniques to improve answer accuracy and applicability in medical information retrieval tasks.

3. IMB Dataset

The IMB dataset consists of two structured subsets: **IMB-QA**, which focuses on unstructured, patient-driven medical inquiries and professional responses, and **IMB-MCQA**, which contains structured multiple-choice questions designed for evaluating domain-specific medical knowledge. The **IMB-QA** dataset captures natural, patient-driven inquiries and professional responses, reflecting real-world medical concerns and interactions (refer to Table 2 for an example).

In contrast, the **IMB-MCQA** dataset consists of structured multiple-choice questions derived from medical specialization exam simulators, providing a controlled environment for evaluating domain-specific knowledge (an example is shown in Table 3).

Table 3
Example of a multiple-choice question from **IMB-MCQA**.

IMB-MCQA	
ID	121
Category	**ITA:** Dermatologia e venereologia **ENG:** Dermatology and Venereology
Question	**ITA:** Dermatite da contatto: quale delle affermazioni sottoriportate è corretta? **ENG:** Dermatitis: which of the following statements is correct?
Answer A	**ITA:** È una genodermatosi **ENG:** It is a genodermatosis
Answer B	**ITA:** È più frequente negli individui di razza nera **ENG:** It is more common in individuals of African descent
Answer C	**ITA:** È causata spesso dall'uso di cosmetici **ENG:** It is often caused by the use of cosmetics
Answer D	**ITA:** Si realizza al 1° contatto con l'allergene **ENG:** It occurs at the first contact with the allergen
Answer E	**ITA:** Tutte le precedenti **ENG:** All of the above
Percentage Correct	49%
Correct Answer	**ITA:** È causata spesso dall'uso di cosmetici **ENG:** It is often caused by the use of cosmetics

3.1. Data Collection

The **IMB-QA** dataset was constructed by collecting questions and answers from two Italian medical forums: MedicItalia and Dica33. These public platforms facilitate interactions between users and certified healthcare professionals. The selection of these forums was guided by qualitative reliability criteria, including verification of medical credentials and assessment of response quality. The data extraction process was conducted through automated retrieval of publicly available information. To enhance compliance with GDPR requirements, an anonymization procedure was applied to remove Personally Identifiable Information (PII). However, we acknowledge that ensuring complete anonymization is inherently challenging, especially in medical contexts where indirect re-identification risks may persist. Future iterations of the dataset will incorporate additional validation steps to assess and improve the effectiveness of the anonymization process. The dataset covers a broad spectrum of common clinical conditions, supporting its medical representativeness. Each sample consists of the following components: A *question* formulated by a user, representing a real medical concern and assigned to a specific medical category; An *answer* provided by a certified healthcare professional, reformulated when necessary to improve clarity and coherence while ensuring the anonymization of personal data; Additional *metadata*, including the *corresponding medical category*, the *macro-category*, and, where applicable, the *URL* of the original source.

The **IMB-MCQA** dataset, on the other hand, was constructed by collecting multiple-choice questions from Italian medical specialization exam simulator CompitoInClasse.org. Each sample consists of the following components: A *question* related to a specific clinical topic, selected from official simulators that provide access to past examination questions; The *multiple-choice answers* associated with the question, including one correct answer validated by domain experts; The *medical category* of the question, identifying the relevant medical field (e.g., physiology, cardiology, etc.); The *percentage of correct answers*, calculated based on responses from a substantial number of candidates who have used the simulator, with a minimum response threshold to ensure reliability.

3.2. Data preprocessing methods

The **IMB-QA** dataset was built from Italian medical forums, collecting 782,644 patient questions and certified professional answers across 77 categories (up to July 2024), capturing real-world interactions.

The **IMB-MCQA** dataset was compiled from official Italian medical specialization exams through 2024 and includes 25,862 multiple-choice questions across 60 clinical fields, each with 4–5 options. As typical with unstructured sources, both datasets had inconsistencies, redundancies, and PII. A multi-stage preprocessing pipeline improved their quality and NLP usability. Summary statistics are in Table 4.

Table 4
Overall statistics for **IMB-QA** and **IMB-MCQA**.

Statistic	IMB-QA	IMB-MCQA
# Questions and Answers	782,644	25,862
# Categories	77	60
Last Update	July 2024	July 2024
Tot. Answer Tokens	40,370,381	9,321
Unique Answer Vocab.	154,837	1,234
Tot. Question Tokens	137,129,435	282,239
Unique Question Vocab.	1,397,929	19,214
Unique Total Vocab.	1,552,766	20,448
Avg. Answer Length	352.05	9.3
Max. Answer Length	9,817	21
Avg. Question Length	1,056.77	10.91
Max. Question Length	13,390	124

3.2.1. Preprocessing for IMB-QA

Data cleaning Incomplete/truncated questions were removed, doctor signatures and timestamps stripped, and minor inconsistencies fixed, preserving meaning.

Text Normalization, Answer Reformulation, and Data Anonymization These operations were carried out using Llama3-Med42-8B [25], a Large Language Model (LLM) specialized in the medical domain and adapted for multilingual tasks. The model underwent a *prompt engineering* phase to enhance the clarity, coherence, and grammatical accuracy of the responses while preserving an adequate level of fidelity to medical information. User-submitted questions were retained in their original form to preserve the natural variability and authenticity of real-world patient inputs. In contrast, doctors' responses were reformulated according to three main criteria: (i) removal of redundancies and colloquial language, (ii) stylistic consistency across responses, and (iii) improved readability for more effective processing by NLP models. To address anonymization, we utilized Italian_NER_XXL [26], a NER model specifically trained in Italian. This model successfully identified PII, such as names of patients and doctors, cities, online resources, email addresses, healthcare facilities, and other identifiers that could enable re-identification. The identified PII underwent an anonymization procedure using the same LLM employed for reformulation, which preserved sentence semantics while substituting terms with generic medical context-appropriate alternatives. The effectiveness of anonymization was evaluated by calculating the percentage of PII — detected using the same NER model as in the anonymization phase — in the initial, reformulated, and anonymized responses on a subset of approximately 2163 responses equally selected from all medical categories in the dataset. Initially, 27% of answers contained PII; reformulation reduced this to 7%, and ultimately, anonymization decreased the presence of PII to just 1%.

Table 5
Macro-categories and number of related questions in **IMB-QA**.

Category	N.o Questions
Urology, andrology and male health	110,052
Gastroenterology and digestive health	104,449
Mental health	103,893
General Medicine and General Surgery	87,789
Ophthalmology, otolaryngology, dentistry and pneumology	83,710
Cardiology, circulatory system and hematology	81,232
Gynecology and female health	65,792
Orthopedics and musculoskeletal system	50,283
Dermatology, allergies and aesthetics	49,288
Neurology	46,704

Data Categorization To group questions into broader semantic fields, unsupervised topic modeling via BERTopic [27] was applied. Sentence embeddings were generated with "paraphrase-multilingual-MiniLM-L12-v2" [28], reduced via UMAP [29], and clustered using HDBSCAN [30]. This enabled flexible, interpretable macro-categorization without enforcing rigid class definitions. Final groupings are reported in Table 5.

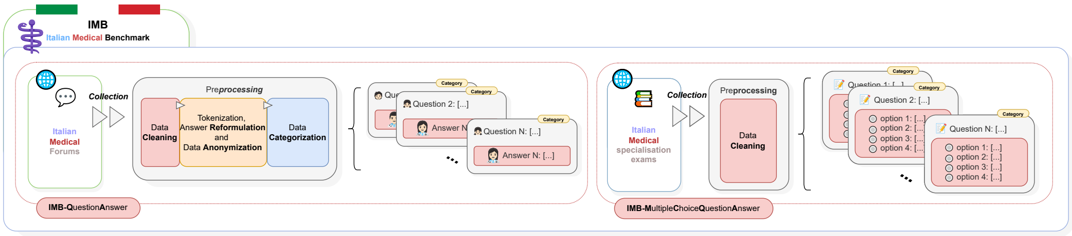

Figure 1: Workflow for the construction of the Italian Medical Benchmark (IMB), consisting of open-ended question-answer pairs (IMB-QA) and multiple-choice question-answer assessments (IMB-MCQA).

3.2.2. Preprocessing for IMB-MCQA

As this dataset was already in a clean, structured exam format, preprocessing mainly involved organizing entries and ensuring consistent formatting. No major cleaning or reformulation was necessary. The workflow is summarized in Figure 1.

3.3. Data Analysis

3.3.1. Diversity of Questions

Clinical medicine covers a broad range of topics, reflected in the question types within the **IMB** dataset. To assess this variety, a qualitative analysis was conducted on a random sample of 102 questions from **IMB-QA** and **IMB-MCQA**. Given the complexity of accurately classifying questions as **fact-based** or **case-based** through automated methods, manual categorization was chosen. **Fact-based** questions focus on specific medical knowledge and clear reasoning, such as “Which condition is linked to persistent fatigue?”. **Case-based** questions, instead, present a patient’s symptoms or medical background, requiring multi-step reasoning for diagnosis, treatment decisions, or prognosis, such as assessing a patient with chest pain. The analysis indicates that **IMB-QA** is predominantly composed of **case-based** questions, where patients describe symptoms and seek medical guidance, requiring models to perform complex reasoning. Although **IMB-MCQA** mainly consists of **fact-based** questions, as it evaluates medical knowledge for specialization exams, it also includes a considerable number of **case-based** inquiries. This dual function highlights the dataset’s role in assessing both factual knowledge and clinical decision-making, with **IMB-QA** emphasizing patient narratives and **IMB-MCQA** blending factual recall with clinical reasoning.

3.3.2. Need for Domain-Specific Expertise

To evaluate the datasets’ complexity, we assessed question difficulty. In **IMB-QA**, a sample of 2,500 questions was analyzed using a difficulty index based on length,

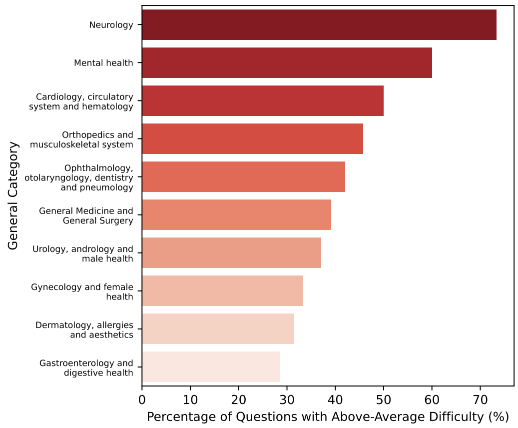

Figure 2: Percentage of questions with above-average difficulty by macro-category in **IMB-QA**. The score refers to the percentage of questions in each category that were classified as above-average in difficulty, based on our difficulty index

terminology, and syntax. 39.24% were above-average in difficulty, with Neurology exceeding 70%, indicating high specialization demands (Figure 2).

In **IMB-MCQA**, difficulty was estimated from participant accuracy. Categories like ”Thermal Medicine” (80.12%), ”Ophthalmology” (72.86%), ”Neurosurgery” (71.30%), and ”Nuclear Medicine” (66.95%) showed high complexity (Figure 3).

These results confirm that both datasets require advanced clinical knowledge, making them valuable for training models in specialized medical reasoning.

3.3.3. Diversity of Categories

The **IMB** dataset shows uneven category distribution, affecting model performance across specialties. **IMB-QA** (Figure 4) overrepresents areas like ”Gastroenterology”, ”Cardiology”, and ”Urology”, while fields like ”Sleep Medicine” and ”Pediatric Surgery” are underrepresented. This may lead to imbalanced model capabilities. **IMB-**

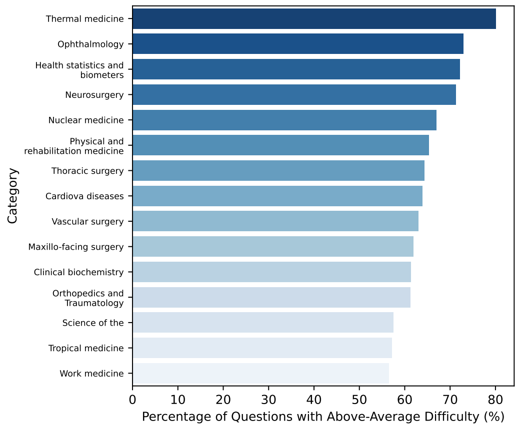

Figure 3: Percentage of questions with above-average difficulty by category in **IMB-MCQA**. The score refers to the percentage of questions in each category that were classified as above-average in difficulty, based on our difficulty index

MCQA (Figure 5) shows a more uniform distribution, with most categories having ~350 questions, except "General Medicine" (~5,000), reducing but not eliminating coverage gaps in niche fields.

3.3.4. Presence of Information Noise and Ambiguity in Responses

Challenges in the **IMB** dataset include noise and ambiguity. In **IMB-QA**, informal forum responses often contain contextual or generic advice, sometimes prioritizing in-person consultation over definitive answers. These traits, while realistic, introduce variability. Preprocessing helped filter irrelevant elements and standardize responses. In **IMB-MCQA**, ambiguity stems from distractors designed to assess reasoning, with some questions allowing multiple valid interpretations. Such complexity enhances the dataset's value in training models to manage uncertainty and emulate clinical decision-making.

4. Applications

4.1. Benchmarking Large Language Models

Evaluating LLMs on domain-specific datasets is essential to measure their suitability for fields like medicine, where precise understanding is required [31]. Despite advancements in general-purpose knowledge, performance in non-English clinical contexts remains limited [32]. **IMB-QA** and **IMB-MCQA** enable benchmarking in Italian for both open-ended and multiple-choice medical QA, capturing language-specific features, technical terminology,

Table 6
Language models benchmarked in our experiments.

Model	Size	Fine-tuned	Language
Mistral-7B-Instruct-v0.3	7B	No	English
LLaMa-3.1-70B-Instruct	70B	No	English
LLaMa-3.1-8B-Instruct	8B	No	English
LLaMa-3.2-3B-Instruct	3B	No	English
Gemma-2-9b-it	9B	No	English
BioMistral-7B	7B	Yes	English
Bio-Medical-Llama-3-8B	8B	Yes	English
Maestrale-Chat-v0.4	7B	Yes	Italian
LLaMAntino 3-8B	8B	Yes	Italian
Velvet-14B	14B	No	Italian

and clinical nuances.

We evaluate open-ended QA using BERTScore [33] with the multilingual model `bert-base-multilingual-cased`, chosen for its cross-lingual semantic similarity capabilities and its widespread adoption in multilingual NLP benchmarks. For MCQA tasks, we report standard accuracy. This dual evaluation highlights LLM strengths and limitations in Italian clinical applications.

4.2. Medical Question Answering

Medical QA demands models that handle informal, complex queries without hallucinating [34, 35]. We apply **Retrieval-Augmented Generation** (RAG) using a separate knowledge base of 100k anonymized *IMB-QA* answers, explicitly excluding evaluation samples to avoid data leakage. Relevant contexts are retrieved via dense embeddings generated with `all-MiniLM-L6-v2`[4] and indexed using FAISS [36]. We retrieve the top-5 most similar passages, which are then prepended to the query. This ensures factual grounding while maintaining separation between retrieved context and target answers. Although we did not perform a separate retriever evaluation, the overall gain in BERTScore (Table 7) confirms the added value of retrieval. The process is formalized as:

$$A = \text{LLM}(Q, R(Q, D)) \quad (1)$$

where Q is the query, D the dataset, and R the retrieval function. Table 7 shows RAG improves BERTScore Precision across all categories.

4.3. Fine-tuning

Fine-tuning improves domain alignment for LLMs, especially in non-English medical contexts [37, 38]. Using **IMB-QA**, we fine-tune Small Language Models (SLMs) like Llama-3.2-1B, Gemma-2-2b-it, and Qwen2.5-1.5B [39], leveraging [CLS]/[SEP] token strategies, cross-entropy loss, and Curriculum Learning [40] via the Unsloth[41] library. This approach aims to enhance output

[4]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

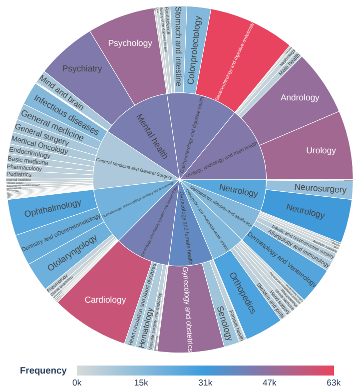

Figure 4: Distribution of macro-categories in **IMB-QA**.

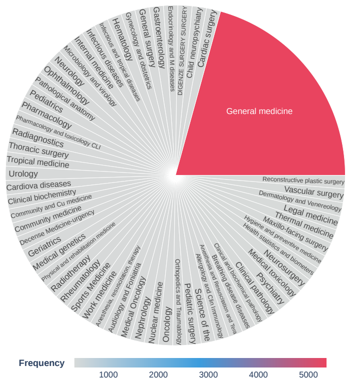

Figure 5: Distribution of categories in **IMB-MCQA**.

Table 7
BERTScore Precision: gemma-2-9b-it with and without RAG on **IMB-QA**.

Category	w/o RAG	RAG	Δ%
Cardiology, hematology	0.632	**0.672**	6.33%
Dermatology, aesthetics	0.636	**0.678**	6.60%
Gastroenterology	0.638	**0.679**	6.42%
General medicine	0.636	**0.674**	5.97%
Gynecology	0.630	**0.671**	6.51%
Mental health	0.636	**0.677**	6.45%
ENT, ophthalmology	0.647	**0.685**	5.87%
Orthopedics	0.628	**0.669**	6.52%
Urology, andrology	0.638	**0.679**	6.42%
Neurology	0.653	**0.706**	8.12%

accuracy and reduce hallucinations while ensuring efficient deployment in clinical environments. Although formal hallucination metrics are not reported, results in Table 8 show that fine-tuning on **IMB-QA** leads to modest improvements across several metrics, particularly in BERTScore and BLEU. Gains are model-dependent and not uniform across all scores: for instance, METEOR slightly decreases in some cases. Nonetheless, the overall trend supports the effectiveness of task-specific adaptation in improving answer quality in Italian medical QA.

5. Experiments

5.1. Experimental Setup

Experiments were conducted on Google Colab Pro using an NVIDIA T4 GPU and Intel Xeon CPU. Due to hardware constraints, the evaluation focused on the most complex categories, as defined in Section 3.3.2. For **IMB-QA**, ~2,000 instances were sampled per category, except for the "Neurology" category, which includes only 998 instances. In the case of **IMB-MCQA**, the full set of instances for each category was used. Models were implemented with Hugging Face Transformers and fine-tuned using the Unsloth library, leveraging mixed precision (fp16) to optimize memory and convergence speed. Each model was fine-tuned for 6 epochs using the Cross Entropy loss function and a fixed learning rate of $2.97e^{-4}$.

5.2. Benchmarking LLMs & SLMs Results

IMB-MCQA offers a robust benchmark for clinical QA in multiple-choice format, evaluated using accuracy. As shown in Figure 6, models with more than 8B parameters achieve nearly 85% accuracy, outperforming smaller models, which struggle with domain-specific reasoning. These trends align with prior analyses of category difficulty, where questions involving underrepresented or cognitively complex fields proved more challenging even for advanced LLMs.

5.3. Medical QA Results

IMB-QA allows assessment of open-ended medical QA, where semantic accuracy is paramount. In Figure 7, `gemma-2-9b-it` outperforms larger models, likely due to its multilingual training. Despite its smaller size, it achieves competitive BERTScore Precision (up to 0.638), suggesting high semantic alignment. This metric is more informative than fluency-based ones in clinical settings, where accurate, relevant answers are crucial.

5.4. Fine-tuning SLMs Results

We fine-tuned several SLMs, including `Llama-3.2-3B`, on **IMB-QA** using an 80/20 train/eval split and leveraging Unsloth library. As shown in Table 8, fine-tuned models generally showed modest improvements over base versions, although gains varied across metrics and models,

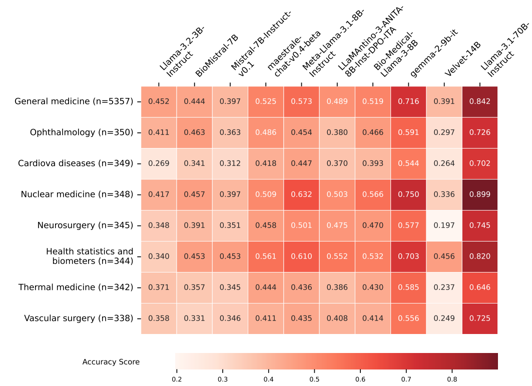

Figure 6: LLM benchmark on **IMB-MCQA**.

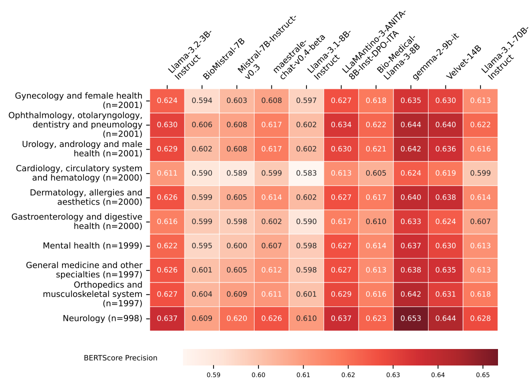

Figure 7: LLM benchmark on **IMB-QA**.

Table 8
Comparison between fine-tuned and non-fine-tuned models on **IMB-QA**.

Model	Fine-Tuned	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	BERTScore P	BERTScore R	BERTScore F1
Llama-3.2-1B-Instruct	Yes	0.2857	0.0572	0.1998	0.0309	0.1682	0.7107	0.6860	0.6976
	No	0.2315	0.0445	0.1552	0.0148	0.2137	0.6186	0.6680	0.6423
gemma-2-2b-it	Yes	0.2673	0.0586	0.1890	0.0336	0.1617	0.7098	0.6775	0.6926
	No	0.2932	0.0511	0.1918	0.0228	0.2055	0.6783	0.6870	0.6821
Llama-3.2-3B-Instruct	Yes	0.2994	0.0642	0.1995	0.0424	0.1952	0.7031	0.6924	0.6972
	No	0.2523	0.0509	0.1607	0.0213	0.2310	0.6332	0.6830	0.6569
Qwen2.5-1.5B-Instruct	Yes	0.2628	0.0438	0.1761	0.0201	0.1571	0.7049	0.6859	0.6948
	No	0.1141	0.0180	0.0756	0.0103	0.1283	0.6021	0.6617	0.6302

with some showing performance drops in specific scores such as METEOR. This confirms that task adaptation improves answer quality and contextual understanding, even for compact models, making them well-suited for clinical applications.

6. Conclusion & Future Work

In this work, we introduced IMB, the first Italian dataset for medical question-answering, which includes both open-ended (QA) and multiple-choice (MCQA) questions. The dataset, sourced from medical forums and exam simulators, provides a valuable resource for the development of advanced NLP models. Our qualitative and quantitative analysis highlighted a diverse range of medical specialties, while also revealing challenges related to question difficulty and clinical complexity. Initial experiments with state-of-the-art language models demonstrated that these models struggle with clinically complex Italian questions but perform relatively well on multiple-choice questions. Future work will focus on expanding the dataset by incorporating additional medical specialties and languages (such as English), improving category balancing, and implementing advanced filtering techniques to reduce informational noise. Furthermore, we will explore strategies for adapting language models to improve their ability to understand and reason effectively about medical content.

Limitations **IMB** has several limitations, including an imbalance in specialty representation. Fields such as "Gastroenterology" and "Cardiology" are overrepresented, while others, such as "Sleep Medicine" and "Pediatric Surgery", have limited coverage. This imbalance may affect model generalization. We will address this issue through data balancing techniques, such as oversampling and weighted training strategies. Another limitation arises from informational noise, as the questions were automatically collected from public sources, which may include irrelevant or ambiguous details. We plan to tackle this challenge by employing semantic filtering and human verification methods. Additionally, ambiguity in responses, particularly in the **IMB-MCQA** dataset, poses a challenge, which we aim to overcome through disambiguation techniques and more precise annotation strategies.

Ethical and Legal Considerations Our dataset has been developed using content sourced information from publicly accessible Italian medical sites (MedicItalia, Dica33) as well as a medical exam simulator (CompitoInClasse.org). The dataset is intended exclusively for academic research with non-commercial objectives, adhering to legal guidelines regarding GDPR compliance, data anonymization, and research-related copyright exemptions as outlined in Italian and EU legislation. To mitigate any legal and ethical challenges, and based on consultations with legal experts, we implemented several measures: **(1) Anonymization:** All identifying details (e.g. names, contact details, emails) were removed or altered with the help of automated scripts and LLM-supported redaction, conforming to GDPR's tenets of data minimization and protection. **(2) Textual Transformation:** While we provide links to the original source of each data sample, the raw questions and answers were linguistically restructured and polished, involving grammatical adjustments, simplification, and content refinement with the aid of LLMs and manual oversight. **(3) Scientific Scope:** This data serves strictly educational, illustrative, and scientific purposes as permitted under Article 89 of the GDPR and Article 70 of the Italian Copyright Law, which allows non-commercial research data usage under specified conditions. For this reason, the dataset is distributed under a *Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0)* license. This license strictly restricts usage to non-commercial research, prohibits redistribution of altered versions, and mandates proper author attribution.

Acknowledgments

This work was conducted with the financial support of (1) the PNRR MUR project PE0000013-FAIR and (2) the Italian ministry of economic development, via the ICARUS (Intelligent Contract Automation for Rethinking User Services) project (CUP: B69J23000270005).

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019) –. URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[3] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinform. 36 (2020) 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682. doi:10.1093/BIOINFORMATICS/BTZ682.

[4] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://aclanthology.org/D16-1264/. doi:10.18653/v1/D16-1264.

[5] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, NeurIPS, Vancouver, BC, Canada, 2019, pp. 5754–5764. URL: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

[6] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, S. Petrov, Natural questions: A benchmark for question answering research, Transactions of the Association for Computational Linguistics 7 (2019) 452–466. URL: https://aclanthology.org/Q19-1026/. doi:10.1162/tacl_a_00276.

[7] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artières, A. N. Ngomo, N. Heino, É. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition, BMC Bioinform. 16 (2015) 138:1–138:28. URL: https://doi.org/10.1186/s12859-015-0564-6. doi:10.1186/S12859-015-0564-6.

[8] D. Wang, Q. Huang, M. Jackson, J. Gao, Retrieve what you need: A mutual learning framework for open-domain question answering, Trans. Assoc. Comput. Linguistics 12 (2024) 247–263. URL: https://doi.org/10.1162/tacl_a_00646. doi:10.1162/TACL_A_00646.

[9] A. Lamurias, D. Sousa, F. M. Couto, Generat-

ing biomedical question answering corpora from q&a forums, IEEE Access 8 (2020) 161042–161051. doi:10.1109/ACCESS.2020.3020868.

[10] M. Zhu, A. Ahuja, W. Wei, C. K. Reddy, A hierarchical attention retrieval model for healthcare question answering, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2472–2482. URL: https://doi.org/10.1145/3308558.3313699. doi:10.1145/3308558.3313699.

[11] M. A. Weinzierl, S. M. Harabagiu, The university of texas at dallas hltri's participation in EPIC-QA: searching for entailed questions revealing novel answer nuggets, CoRR abs/2112.13946 (2021) –. URL: https://arxiv.org/abs/2112.13946. arXiv:2112.13946.

[12] T. Möller, A. Reina, R. Jayakumar, M. Pietsch, COVID-QA: A question answering dataset for COVID-19, in: ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID), ACL, Online, 2020, pp. –. URL: https://openreview.net/forum?id=JENSKEEzsoU.

[13] S. Šuster, W. Daelemans, CliCR: a dataset of clinical case reports for machine reading comprehension, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1551–1563. URL: https://aclanthology.org/N18-1140/. doi:10.18653/v1/N18-1140.

[14] A. B. Abacha, E. Agichtein, Y. Pinter, D. Demner-Fushman, Overview of the medical question answering task at TREC 2017 liveqa, in: E. M. Voorhees, A. Ellis (Eds.), Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017, volume 500-324 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA, 2017, pp. –. URL: https://trec.nist.gov/pubs/trec26/papers/Overview-QA.pdf.

[15] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL: https://arxiv.org/abs/2009.13081. arXiv:2009.13081.

[16] A. Pampari, P. Raghavan, J. J. Liang, J. Peng, emrqa: A large corpus for question answering on electronic medical records, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2357–2368. URL: https://doi.org/10.18653/v1/d18-1258. doi:10.18653/V1/D18-1258.

[17] J. He, M. Fu, M. Tu, Applying deep matching networks to chinese medical question answering: a study and a dataset, BMC Medical Informatics Decis. Mak. 19-S (2019) 91–100. URL: https://doi.org/10.1186/s12911-019-0761-8. doi:10.1186/S12911-019-0761-8.

[18] A. Nentidis, A. Krithara, K. Bougiatiotis, M. Krallinger, C. R. Penagos, M. Villegas, G. Paliouras, Overview of bioasq 2020: The eighth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, volume 12260 of *Lecture Notes in Computer Science*, Springer, Thessaloniki, Greece, 2020, pp. 194–214. URL: https://doi.org/10.1007/978-3-030-58219-7_16. doi:10.1007/978-3-030-58219-7_16.

[19] D. Vilares, C. Gómez-Rodríguez, HEAD-QA: A healthcare dataset for complex reasoning, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 960–966. URL: https://aclanthology.org/P19-1092/. doi:10.18653/v1/P19-1092.

[20] A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: G. Flores, G. H. Chen, T. J. Pollard, J. C. Ho, T. Naumann (Eds.), Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event, volume 174 of *Proceedings of Machine Learning Research*, PMLR, Online, 2022, pp. 248–260. URL: https://proceedings.mlr.press/v174/pal22a.html.

[21] Y. Tian, W. Ma, F. Xia, Y. Song, ChiMed: A Chinese medical corpus for question answering, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 250–260. URL: https://aclanthology.org/W19-5027/. doi:10.18653/v1/W19-5027.

[22] A. Peñas, E. H. Hovy, P. Forner, Á. Rodrigo, R. F. E. Sutcliffe, R. Morante, QA4MRE 2011-2013: Overview of question answering for machine reading evaluation, in: P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Information

Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, volume 8138 of *Lecture Notes in Computer Science*, Springer, Valencia, Spain, 2013, pp. 303–320. URL: https://doi.org/10.1007/978-3-642-40802-1_29. doi:10.1007/978-3-642-40802-1_29.

[23] D. Pappas, I. Androutsopoulos, H. Papageorgiou, Bioread: A new dataset for biomedical reading comprehension, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA), Miyazaki, Japan, 2018, pp. –. URL: http://www.lrec-conf.org/proceedings/lrec2018/summaries/795.html.

[24] D. Pappas, P. Stavropoulos, I. Androutsopoulos, R. McDonald, BioMRC: A dataset for biomedical machine reading comprehension, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2020, pp. 140–149. URL: https://aclanthology.org/2020.bionlp-1.15/. doi:10.18653/v1/2020.bionlp-1.15.

[25] C. Christophe, P. K. Kanithi, T. Raha, S. Khan, M. A. Pimentel, Med42-v2: A suite of clinical llms, CoRR abs/2408.06142 (2024) –. URL: https://doi.org/10.48550/arXiv.2408.06142. doi:10.48550/ARXIV.2408.06142. arXiv:2408.06142.

[26] DeepMount00, Italian_ner_xxl, https://huggingface.co/DeepMount00/Italian_NER_XXL, 2024.

[27] M. Grootendorst, Bertopic: Neural topic modeling with a class-based TF-IDF procedure, CoRR abs/2203.05794 (2022) –. URL: https://doi.org/10.48550/arXiv.2203.05794. doi:10.48550/ARXIV.2203.05794. arXiv:2203.05794.

[28] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3980–3990. URL: https://doi.org/10.18653/v1/D19-1410. doi:10.18653/V1/D19-1410.

[29] L. McInnes, J. Healy, UMAP: uniform manifold approximation and projection for dimension reduction, CoRR abs/1802.03426 (2018) –. URL: http://arxiv.org/abs/1802.03426. arXiv:1802.03426.

[30] M. F. Rahman, W. Liu, S. B. Suhaim, S. Thirumuruganathan, N. Zhang, G. Das, HDBSCAN: density based clustering over location based services, CoRR abs/1602.03730 (2016) –. URL: http://arxiv.org/abs/1602.03730. arXiv:1602.03730.

[31] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, M. L. Li, Benchmarking large language models on cmexam - A comprehensive chinese medical exam dataset, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, NeurIPS, New Orleans, LA, USA, 2023, pp. –. URL: http://papers.nips.cc/paper_files/paper/2023/hash/a48ad12d588c597f4725a8b84af647b5-Abstract-Datasets_and_Benchmarks.html.

[32] Y. Jin, M. Chandra, G. Verma, Y. Hu, M. D. Choudhury, S. Kumar, Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries, in: T. Chua, C. Ngo, R. Kumar, H. W. Lauw, R. K. Lee (Eds.), Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024, ACM, Singapore, 2024, pp. 2627–2638. URL: https://doi.org/10.1145/3589334.3645643. doi:10.1145/3589334.3645643.

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, Addis Ababa, Ethiopia, 2020, pp. –. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[34] S. Zhang, X. Zhang, H. Wang, J. Cheng, P. Li, Z. Ding, Chinese medical question answer matching using end-to-end character-level multi-scale cnns, Applied Sciences 7 (2017) 767.

[35] N. Yagnik, J. Jhaveri, V. Sharma, G. Pila, A. Ben, J. Shang, Medlm: Exploring language models for medical question answering systems, CoRR abs/2401.11389 (2024) –. URL: https://doi.org/10.48550/arXiv.2401.11389. doi:10.48550/ARXIV.2401.11389. arXiv:2401.11389.

[36] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, arXiv preprint arXiv:2401.08281 (2024). arXiv:2401.08281.

[37] H. Tran, Z. Yang, Z. Yao, H. Yu, Bioinstruct: instruction tuning of large language models for biomedical natural language processing, J. Am. Medical Informatics Assoc. 31 (2024) 1821–1832.

URL: https://doi.org/10.1093/jamia/ocae122. doi:10.1093/JAMIA/OCAE122.
[38] F. J. Dorfner, A. Dada, F. Busch, M. R. Makowski, T. Han, D. Truhn, J. Kleesiek, M. Sushil, J. Lammert, L. C. Adams, K. K. Bressem, Biomedical large languages models seem not to be superior to generalist models on unseen medical data, CoRR abs/2408.13833 (2024) –. URL: https://doi.org/10.48550/arXiv.2408.13833. doi:10.48550/ARXIV.2408.13833. arXiv:2408.13833.
[39] C. Van Nguyen, X. Shen, R. Aponte, Y. Xia, S. Basu, Z. Hu, J. Chen, M. Parmar, S. Kunapuli, J. Barrow, et al., A survey of small language models, arXiv preprint arXiv:2410.20011 (2024).
[40] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 41–48. URL: https://doi.org/10.1145/1553374.1553380. doi:10.1145/1553374.1553380.
[41] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) and DeepL Write / DeepL Translate in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.