POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Comparative analysis of medical-domain and general-purpose large language models: evaluating specialization in healthcare applications

Tıbbi alan ve genel amaçlı büyük dil modellerinin karşılaştırmalı analizi: sağlık uygulamalarında uzmanlaşmanın değerlendirilmesi

***Yazar(lar) (Author(s))**: Daniel Quillan ROXAS[1], Hakan EMEKCİ[2]*

ORCID[1]: 0009-0000-4484-6751

ORCID[2]: 0000-0002-4074-5600

To cite to this article: Roxas D. Q., and Emekci H., "Comparative Analysis of Medical-Domain and General-Purpose Large Language Models: Evaluating Specialization in Healthcare Applications", ***Journal of Polytechnic***, 29(1):290111:1-7 (2026).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Roxas D. Q. ve Emekci H., "Comparative Analysis of Medical-Domain and General-Purpose Large Language Models: Evaluating Specialization in Healthcare Applications", ***Politeknik Dergisi***, 29(1):290111:1-7 (2026).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1719005

Comparative Analysis of Medical-Domain and General-Purpose Large Language Models: Evaluating Specialization in Healthcare Applications

Highlights

- *General-purpose LLMs consistently outperform medical-specific models on specialized medical tasks*
- *Zero-shot performance often matches or exceeds few-shot results, challenging conventional assumptions*
- *Domain-specific training may not be necessary for high-quality medical language understanding*
- *Architectural improvements may matter more than specialized training data for medical NLP tasks*

Graphical Abstract

This graphical abstract summarizes the evaluation of general-purpose versus medical-domain LLMs, showing the superior performance of general models on specialized medical question-answer datasets.

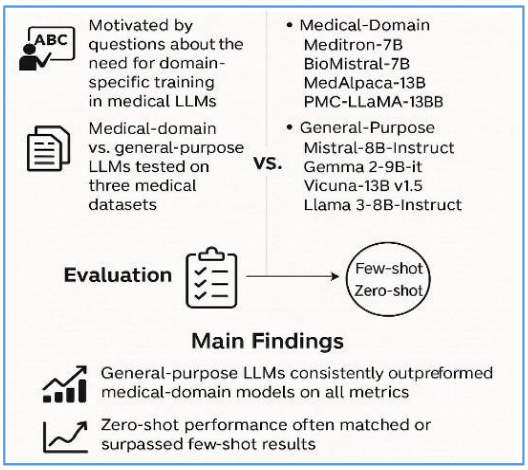

Figure. The summarized process of the LLM comparison

Aim

This study compares medical-domain and general-purpose LLMs to evaluate if specialized training is necessary for healthcare applications.

Design & Methodology

Four medical and four general-purpose LLMs were compared on three medical datasets using zero-shot and few-shot settings. Performance was measured by semantic similarity metrics like BERTScore and SimCSE.

Originality

This study offers a comprehensive comparison of the latest models from as recent as 2024 across multiple datasets, analyzing prompting (zero-shot, few-shot) strategies and performance efficiency.

Findings

General-purpose models consistently outperformed medical-specific models across all metrics. Zero-shot performance often matched or exceeded few-shot results.

Conclusion

Domain-specific pretraining may not be as necessary as previously thought, shown by general models being superior. Future work should focus on improving general architecture rather than creating specialized versions.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Comparative Analysis of Medical- Domain and General-Purpose Large Language Models: Evaluating Specialization in Healthcare Applications

Araştırma Makalesi / Research Article

Daniel Quillan ROXAS[1*], Hakan EMEKCİ[1]

[1]Applied Data Science, TED University, Türkiye

(Geliş/Received : 13.06.2025 ; Kabul/Accepted : 04.12.2025 ; Erken Görünüm/Early View : 02.01.2026)

ABSTRACT

The growing use of Large Language Models (LLMs) in healthcare raises important questions about the need for domain-specific training in medical applications. This study presents a detailed evaluation of medical- domain and general-purpose LLMs using three medical datasets (PubMedQA, BioASQ, and WikiDoc), which contain approximately 11,000 question-answer pairs. We evaluated four medical-domain models (Meditron-7B, BioMistral-7B, MedAlpaca- 13B, and PMC-LLaMA-13B) against four general-purpose instruction-tuned models (Ministral-8B-Instruct, Gemma 2-9B-it, Vicuna-13B v1.5, and Llama 3-8B-Instruct). Across 182,944 prompts in both zero-shot and few-shot settings, our findings show that general-purpose models consistently outperformed their medical-specific counterparts on all evaluation metrics. Specifically, Ministral-8B-Instruct achieved the highest performance in few-shot settings with a BERTScore of 0.613, SimCSE of 0.764, and semantic similarity of 0.684. These scores were significantly higher than those of the best medical model, BioMistral- 7B (0.545, 0.678, and 0.533, respectively). Furthermore, zero-shot performance often matched or surpassed few-shot results, as seen with Llama-3-8B-Instruct achieving a SimCSE score of 0.794. These findings challenge the common assumption that domain-specific pretraining is required for optimal performance in specialized tasks and have major implications for how resources are allocated in healthcare AI development.

Keywords: Large Language Models, Domain-Specific Models, Healthcare AI.

Tıbbi Alan ve Genel Amaçlı Büyük Dil Modellerinin Karşılaştırmalı Analizi: Sağlık Uygulamalarında Uzmanlaşmanın Değerlendirilmesi

ÖZ

Büyük Dil Modellerinin (BDM'ler) sağlık alanında yaygın kullanımı, tıbbi uygulamalar için alana özgü eğitimin gerekli olup olmadığı konusunda temel soruları gündeme getirmiştir. Bu çalışma, yaklaşık 11.000 soru-cevap çiftinden oluşan üç tıbbi veri seti (PubMedQA, BioASQ ve WikiDoc) kullanarak tıbbi alan ve genel amaçlı BDM'lerin kapsamlı bir değerlendirmesini sunmaktadır. Dört tıbbi alan modeli (Meditron-7B, BioMistral-7B, MedAlpaca- 13B ve PMC-LLaMA-13B) ile dört genel amaçlı yönerge ayarlı modeli (Ministral-8B- Instruct, Gemma 2-9B-it, Vicuna-13B v1.5 ve Llama 3-8B-Instruct) hem sıfır-atış hem de az-atış ayarlarında 182.944 istem üzerinde değerlendirdik. Sonuçlarımız, genel amaçlı modellerin tüm değerlendirme metriklerinde tıbbi alan modellerini tutarlı bir şekilde geride bıraktığını göstermektedir. Özellikle, Ministral-8B-Instruct, az-atış ayarlarında 0.613 BERTScore, 0.764 SimCSE ve 0.684 anlamsal benzerlik ile en yüksek genel performansı elde ederek, sırasıyla 0.545, 0.678 ve 0.533 puan alan en iyi tıbbi model (BioMistral-7B) performansını önemli ölçüde aşmıştır. Ayrıca, sıfır-atış performansı genellikle az- atış sonuçlarına eşit veya daha üstün olmuş, Llama-3-8B-Instruct sıfır-atış ayarlarında 0.794 SimCSE puanı elde etmiştir. Bu bulgular, özelleşmiş tıbbi görevlerde optimal performans için alana özgü ön eğitimin gerekli olduğu şeklindeki yaygın varsayımı sorgulamakta ve sağlık AI sistemlerinin geliştirilmesinde kaynak tahsisi açısından önemli çıkarımlar sunmaktadır.

Anahtar Kelimeler: Büyük Dil Modelleri, Alan Özgü Modeller, Sağlıkta Yapay Zeka.

1. INTRODUCTION

Medical-focused Large Language Models (LLMs), such as MedAlpaca and BioMistral, were built on the idea that training them on specialized medical datasets, terminologies, and clinical conversations would improve their performance on healthcare tasks. In medicine, where accuracy directly impacts patient outcomes, the assumption that specialized training yields better results has been a guiding principle.

Recently, however, evidence has emerged to challenge this view. General-purpose models like GPT-4 have shown stronger performance against specialized models such as Med-PaLM (a prompt-tuned version of Flan-PaLM 540B), even without undergoing medical-specific pretraining [1]. This surprising result raises important questions about the need for domain specialization to achieve the best results.

Evaluating LLMs in a medical context presents unique challenges. This study uses a detailed suite of evaluation

*Sorumlu Yazar (Corresponding Author)
e-posta : dquillan.roxas@tedu.edu.tr

metrics, including BERTScore and SimCSE for semantic understanding, while assessing both few-shot and zero-shot capabilities. This approach allows us to evaluate not only the models' handling of medical terminology but also their ability to generate contextually sound information. Our evaluation covers three distinct medical datasets to ensure a broad assessment: PubMedQA for research questions [2], BioASQ for biomedical question answering [3], and WikiDoc for patient-focused information [4].

Through a quantitative analysis of eight similarly-sized LLMs—four medical- specialized and four general-purpose—this study investigates three primary research questions: (1) Does domain-specific training offer a measurable advantage for medical language tasks? (2) In medical contexts, how does zero-shot capability compare to few-shot performance? (3) What are the implications of these findings for the future of medical AI?

Our findings have major implications for the specialized AI ecosystem, particularly for resource allocation and the cost-benefit analysis of developing specialized LLMs. The results challenge established assumptions and suggest that a re-evaluation of development strategies for medical AI may be needed.

2. RELATED WORK

Recent progress in large language models has fueled important discussions about the need for domain-specific training for specialized tasks. This section reviews recent developments in medical and general- purpose LLMs, their evaluation methodologies, and the evolving understanding of few-shot versus zero-shot learning in medical applications.

2.1. Recent Studies in Medical Language Model Evaluation

Table 1 provides an overview of recent studies comparing medical-domain and general-purpose language models, highlighting the datasets, metrics, key findings, and the research gaps our study aims to address.

2.2. Domain-Specific Language Models in Healthcare

The development of domain-specific language models for healthcare has grown considerably, with models like Med-PaLM 2 achieving 85.4% accuracy on MedQA (USMLE), 72.3% on MedMCQA, and 75.0% on PubMedQA, significantly overperforming the original Flan- PaLM which scored only 67.6% on MedQA [5]. These specialized models are typically pretrained on large medical text corpora and then fine-tuned for specific medical tasks. However, recent research suggests the benefits of this specialized pretraining may decrease as model scale and architectural designs improve [6]. While our focus is on language models, the push for specialized medical AI spans multiple modalities, from diagnostic imaging for COVID-19 detection [7] to clinical decision support systems. For instance, Eriç et al. [8] reviewed the use of ChatGPT in medical applications and found it demonstrates competitive performance on medical tasks including the USMLE examination.

2.3. General-Purpose LLMs in Specialized Tasks

While specialized models were in development, general-purpose LLMs also showed impressive capabilities in specialized domains. The work of Rosa et al. [9] offers key insights on model scaling, showing that larger models generalize significantly better to new domains. They noted that "increasing model size results in marginal gains on in-domain test sets, but much larger gains in new domains." This suggests that a model's scale and architectural sophistication could be more important for strong performance than domain-specific training. The development of specialized AI for healthcare extends beyond language models to various clinical applications, including fall risk assessment [10] and diagnostic support systems, reflecting a broader trend toward domain- specific solutions. Additionally, GPT-4's performance varied between its base and production versions, with GPT-4-base achieving 86.1% on MedQA compared to the production version's 81.4%, while maintaining competitive scores of 80.4% on PubMedQA, demonstrating that general models can achieve medical exam performance comparable to specialized models [5].

2.4. Model Selection Rationale

Our model selection was guided by several criteria to ensure a fair and detailed comparison. We prioritized models with parameter counts between 7B and 13B to maintain computational feasibility while working with sufficiently powerful models.

Table 1. Summary of Recent Studies on Medical and General-Purpose Language Models

Study	**Datasets**	**Metrics**	**Key Findings**	**Gaps Addressed by Our Study**
Singhal et al. [5]	USMLE, MedQA	Accuracy, F1	Med-PaLM 2 achieved expert-level performance on medical exams	Limited to exam-style questions with latest general models
Hong et al. [6]	Scientific Papers	Perplexity	Diminishing returns of domain-specific pretraining with scale	Focused on scientific text, not medical Q&A
Eriç et al. [8]	Medical Q&A	Accuracy	Gap between specialized and general models narrowing	Limited model selection; no zero/few-shot comparison
Guo & Hua [29]	MedQuAD	BLEU, ROUGE	Continuous training improves domain performance	Single dataset; older model architectures
Our Study	PubMedQA, BioASQ, WikiDoc	BERTScore, SimCSE, Semantic Similarity	General models outperform medical models; strong zero-shot performance	Comprehensive multi-dataset evaluation with latest models (2024)

The exclusion of Google's Med-PaLM 2, despite its prominence, was a necessary decision for three reasons: (1) it is not openly accessible for standard research, (2) its API requires special permissions that were unavailable to us, and (3) our study design required full control over model deployment to ensure a fair comparison. Focusing on openly available models ensures our findings are reproducible and can be validated by others.

Regarding the timing of model selection, we acknowledge that medical-domain models have not been updated as frequently as general-purpose ones. The medical models we chose (Meditron-7B, BioMistral-7B, MedAlpaca-13B, and PMC-LLaMA-13B) were the most current, openly available options when we conducted our study. This difference in release frequency reflects a practical reality of the field: general-purpose models often benefit from larger development teams and more frequent updates. This gap is relevant to our research; if medical specialization requires significant resources but its products still lag behind general model development, this strengthens the case for re-evaluating whether such specialization remains the most effective strategy.

2.5. Few-Shot Versus Zero-Shot Learning in Medical Applications

The comparison between few-shot and zero- shot learning has grown more complex. Li and Flanigan [11] have challenged traditional ideas by introducing the concept of "task contamination." Their work reveals that model performance on datasets released before their training was complete is significantly higher than on newer datasets. This suggests that what appears to be few-shot learning might actually be a reflection of implicit knowledge gained during pretraining.

2.6. Evaluation Methodologies and Metrics

Evaluation methods for medical language tasks have evolved considerably. Herbold [12] shows that using fine-tuned models to directly predict semantic similarity is more effective than traditional similarity measures, especially in domain-specific applications. This is supported by Koroteev [13], who demonstrates that embedding-based methods like BERTScore have advantages over character-based metrics for capturing semantic meaning. These insights have shaped new evaluation frameworks, like those from Glushkova et al. [14], which combine traditional and domain-specific metrics.

Our evaluation builds on these recent developments while addressing several gaps in the literature. First, while previous studies have compared medical and general models, none have done so across multiple medical datasets using the latest 2024 model architectures. Second, the relationship between model type (specialized vs. general) and prompting strategy (few-shot vs. zero- shot) in medicine remains underexplored, particularly in light of recent findings about task contamination. Finally, our work offers new data on the efficiency-performance trade-offs in medical tasks, an often- overlooked aspect of such studies.

3. METHODOLOGY

3.1. Hyperparameter Selection

To ensure a fair and integral comparison, we used uniform hyperparameters for all models, shot types, and datasets. We randomly selected 100 question-answer pairs to test different temperature settings, aiming for a balance between creativity and factuality. Initial tests with Meditron-7B showed that balanced temperature settings yielded optimal scores on metrics including ROUGE-L [15], BLEU [16], BERTScore [17], and Semantic Similarity [18]. Subsequent tests with Ministral-8B-Instruct confirmed these results, showing only minimal differences between conservative and balanced configurations. Based on this, we used balanced settings for all experiments.

3.2. Dataset Preparation and Processing

We selected three datasets with varied formats to ensure a comprehensive evaluation within our resource constraints. These datasets contain approximately 11,000 unique question-answer pairs, which were processed 182,944 times across the different models and shot types.

Preprocessing involved extracting questions and answers into a uniform format to streamline the workflow. For the few-shot experiments, examples were randomly selected from within each dataset. This method ensures that models were prompted with relevant examples, allowing a fair assessment of their ability to recognize dataset-specific patterns.

3.3. Model Selection and Implementation

All models were accessed from the Hugging Face model hub [19]. The medical domain models were Meditron-7B [20], BioMistral- 7B [21], MedAlpaca-13B [22], and PMC- LLaMA-13B [23]. The general-purpose, instruction-tuned models were Ministral-8B- Instruct [24], Gemma-2-9B-it [25], Vicuna-13B v1.5 [26], and Llama-3-8B-Instruct [27].

We evaluated an equal number of medical and general LLMs, choosing them based on their established reputation and how recently they were released; many (Gemma 2, Ministral 8B, Llama 3 8B) were released or updated in 2024. This allowed for a direct comparison of state-of-the-art architectures against specialized models. All selected models have between 7B and 13B parameters, which strikes a balance between computational efficiency and model capacity. We used 4-bit quantization for efficient inference and deployed all models on RTX 4090 GPUs in a cloud environment.

3.4. Evaluation Protocol

Our metric selection prioritized both efficiency and effectiveness. We chose established automated metrics that offer valuable insights without excessive computational costs.

BERTScore [17] served as a primary metric because of its ability to capture semantic similarity beyond literal word matching, which is vital in medical contexts where meaning must be preserved. We complemented this with other semantic similarity metrics [18] that consider sentence meaning irrespective of structure, an important feature for medical concepts expressed in varied ways.

We also incorporated SimCSE scores [28] to assess semantic similarity beyond direct word overlap, using the unsupervised RoBERTa-base version for its straightforward integration and established role in NLP research. This metric is particularly useful in a field where medical concepts can be expressed with diverse terminology. This methodology provides a robust framework for comparing the models while accounting for dataset diversity.

4. FINDINGS

4.1. General-Purpose Models Demonstrate Superior Performance

Our findings show that general-purpose language models consistently achieved higher scores than medical-specific models across all evaluation metrics, as illustrated in Figure 1. This result challenges the common view that domain-specific training is required for high performance in medical tasks.

In few-shot settings, Ministral-8B-Instruct stood out as the top performer with a BERTScore of 0.613, SimCSE of 0.764, and semantic similarity of 0.684 (detailed in Table 2). By comparison, the leading medical model, BioMistral-7B, scored lower on all three metrics with 0.545, 0.678, and 0.533, respectively.

Table 2. Model Performance by Category and Shot Type

Metric	Zero-shot		Few-shot	
	General	Medical	General	Medical
BERTScore	**0.548**	0.461	**0.521**	0.451
SimCSE	**0.731**	0.638	**0.672**	0.596
Semantic Sim	**0.598**	0.420	**0.473**	0.330

4.2. Zero-Shot Performance Challenges Few- Shot Assumptions

We also found that zero-shot learning performance frequently matched or even surpassed few-shot results, which challenge the conventional wisdom that example-based prompting is always necessary. Figure 2 illustrates this surprising pattern. Notably, Llama-3-8B-Instruct achieved an impressive zero-shot SimCSE score of 0.794 and semantic similarity of 0.680, while Vicuna-13B recorded the highest zero-shot BERTScore of 0.596. These results, shown in Table 3, suggest that high-quality, general-purpose instruction tuning can enable models to perform effectively in new domains without needing specific training examples.

The small performance differences between few-shot and zero-shot learning, along with cases where zero-shot was superior, support the idea that LLMs trained on sufficiently broad data can develop impressive generalization skills.

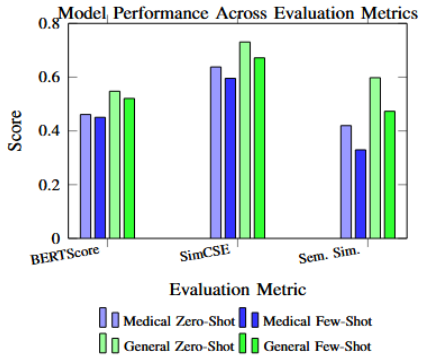

Figure 1. Comparison of model performance across evaluation metrics for medical and general models in zero-shot and few-shot settings.

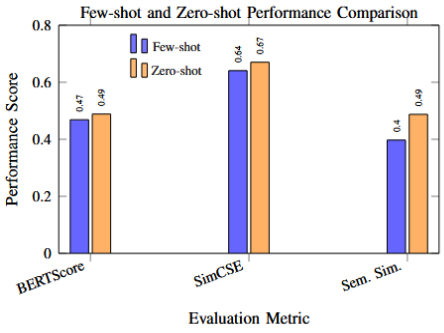

Figure 2. Comparison of few-shot and zero-shot performance across evaluation metrics, demonstrating superior zero-shot performance.

4.3. Performance Analysis by Model Architecture

Table 3 provides a full breakdown of each model's performance. When looking at combined performance, Ministral-8B- Instruct had the highest overall scores (BERTScore: 0.582, SimCSE: 0.729, semantic similarity: 0.616). Vicuna-13B followed, with scores of 0.542, 0.710, and 0.543.

Among medical models, BioMistral-7B showed the strongest results. It is also important to note that parameter count was not a consistent predictor of performance; the 8B-parameter Ministral-8B-Instruct outperformed the 13B-parameter MedAlpaca on all metrics, despite its lack of specialized training.

4.4. Efficiency Analysis

Table 4 shows the processing efficiency of the models in tokens per second. These efficiency differences are visualized in Figure 3, which clearly illustrates the performance gap between few-shot and zero-shot inference speeds across all models. In few-shot settings, BioMistral-7B had the highest throughput at 24.245 tokens/second, while MedAlpaca-13B was the slowest at just 5.077 tokens/second. In zero-shot inference, Llama-3-8B-Instruct showed superior efficiency at 30.462 tokens/second, while Gemma-2-9B-it was the slowest at 12.183 tokens/second.

Table 3. Detailed Model Performance Across Shot Types and Metrics

Model	Category	Few-shot			Zero-shot		
		BERT Score	SimCSE	Sem. Sim.	BERT Score	SimCSE	Sem. Sim.
Ministral-8B-Instruct	General	**0.613**	**0.764**	**0.684**	0.552	0.694	0.547
BioMistral-7B	Medical	0.545	0.678	0.533	0.528	0.681	0.510
Gemma-2-9B-it	General	0.543	0.672	0.487	0.514	0.662	0.496
Vicuna-13B	General	0.488	0.647	0.418	**0.596**	0.774	0.669
PMC-LLaMA-13B	Medical	0.453	0.614	0.328	0.557	0.727	0.615
Llama-3-8B-Instruct	General	0.439	0.604	0.303	0.529	**0.794**	**0.680**
Meditron-7B	Medical	0.415	0.555	0.226	0.428	0.646	0.480
MedAlpaca-13B	Medical	0.390	0.535	0.232	0.331	0.498	0.075

Table 4. Model Processing Efficiency (Tokens per Second)

Model	Category	Few-shot	Zero-shot
BioMistral-7B	Medical	**24.245**	24.355
Ministral-8B-Instruct	General	22.774	22.372
Vicuna-13B	General	16.631	28.254
Llama-3-8B-Instruct	General	13.748	**30.462**
Gemma-2-9B-it	General	10.711	12.183
PMC-LLaMA-13B	Medical	10.384	22.130
Meditron-7B	Medical	9.573	13.139
MedAlpaca-13B	Medical	5.077	23.477

Table 5. Combined Average Performance Metrics

Model	BERTScore	SimCSE	Semantic Sim.
Ministral-8B-Instruct	**0.582**	**0.729**	**0.616**
Vicuna-13B	0.542	0.710	0.543
BioMistral-7B	0.536	0.679	0.521
Gemma-2-9B-it	0.528	0.672	0.492
PMC-LLaMA-13B	0.505	0.671	0.471
Llama-3-8B-Instruct	0.484	0.699	0.492
Meditron-7B	0.421	0.600	0.353
MedAlpaca-13B	0.360	0.517	0.154

These variations in efficiency highlight the importance of considering computational costs alongside performance metrics when choosing models for real-world use. The balance between processing speed, model size, and task performance is a critical factor for any deployment scenario.

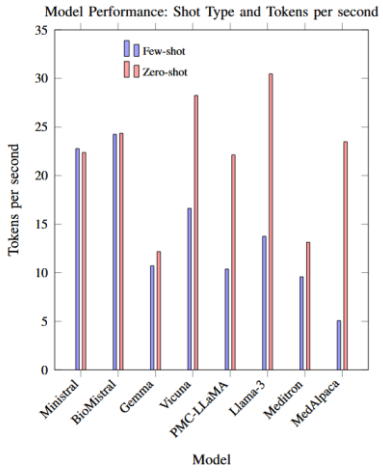

Figure 3. Tokens per second performance comparison between few-shot and zero-shot modes across all evaluated models.

5. DISCUSSION

5.1. Implications of General Model Superiority

The consistent outperformance of general- purpose models challenges the core assumption that domain-specific training is required for high-quality performance in medical applications. The stronger scores from models like Ministral-8B-Instruct suggest that advances in general architectures and broad training data may offer more benefits than narrow, domain- specific pretraining.

This finding aligns with observations by Hong et al. [6] regarding the diminishing returns of domain-specific training as model scale increases. The strong performance of general models can likely be attributed to their exposure to highly diverse training data—which almost certainly includes substantial medical content—allowing for more robust generalization. Furthermore, it is plausible that recent architectural improvements in models like Llama 3 and Ministral are a more significant factor in their success than the fine-tuning of older architectures with specialized data.

5.2. Zero-Shot Learning Capabilities

The strong zero-shot performance seen in our study lends support to the idea that modern language models have inherent capabilities for domain adaptation, even without task-specific examples. This effect might be partially explained by the concept of "task contamination" proposed by Li and Flanigan [11], where models could have encountered similar medical content during pretraining.

However, the consistent pattern across multiple models and datasets suggests that architectural improvements and scale are also major contributors. The minimal performance gap between few-shot and zero-shot settings indicates that modern instruction-tuned models can effectively leverage their extensive pretraining to handle specialized tasks without needing domain-specific examples.

5.3. Efficiency Considerations

Our efficiency analysis reveals important trade-offs for practical deployment. While a medical model like BioMistral-7B showed competitive processing speeds (24.245 tokens/second), its lower performance scores raise questions about the cost-benefit analysis of deploying specialized models. The significant variation in processing speeds, particularly the high throughput of Llama-3-8B-Instruct in zero-shot mode (30.462

tokens/second), suggests that model architecture is a key driver of computational efficiency. The importance of computational efficiency in AI applications extends to various domains where resource optimization is critical [30], making our efficiency findings particularly relevant for practical deployment.

5.4. Performance Considerations

For healthcare applications where both accuracy and response time are critical, these findings suggest that newer general- purpose models may offer a better overall value. Their ability to deliver stronger performance with comparable or even superior processing speeds makes them an attractive option for real-world deployment.

5.5. Theoretical Implications

Our results add to a growing body of evidence suggesting the traditional model of domain-specific fine-tuning may need to be reconsidered. The success of general models indicates that broad, high-quality pretraining combined with effective instruction tuning can be more valuable than narrow domain specialization. This is consistent with findings from Rosa et al. [9] regarding the importance of model scale for domain generalization.

The strong zero-shot performance also suggests the boundary between "general" and "specialized" knowledge is more fluid than previously believed. As models continue to scale and training methodologies advance, the need for explicit domain specialization could diminish, with implications that extend beyond healthcare into other specialized fields.

5.6. Practical Recommendations

The results strongly suggest that teams should prioritize adopting modern, general- purpose models, as they consistently demonstrate superior performance and efficiency. This strategic shift implies that focusing resources on newer model architectures is likely more effective than creating domain-specific versions of older ones. In addition, the strength of these models in zero-shot settings indicates that the extensive effort of creating few-shot prompts may be unnecessary for achieving high performance in many applications. However, while our automated metrics show clear patterns, they are not a substitute for clinical validation. Any system must be thoroughly vetted by domain experts before deployment in a healthcare setting.

5.7. Limitations and Future Directions

Despite a detailed evaluation, this study has several limitations. First, our evaluation relied on automated metrics and did not include validation by human experts, which limits our ability to assess the clinical accuracy and safety of model responses. Second, our model selection was constrained to the 7B-13B parameter range; these trends may differ for larger or smaller models. Third, the three datasets used, while diverse, do not capture the full complexity of medical language tasks in real-world clinical practice.

Future work should address these limitations by incorporating expert evaluations, exploring a wider range of model scales, and assessing performance on more complex tasks like clinical dialogue and decision support. Furthermore, investigating the specific mechanisms responsible for the superior performance of general models could inform future development strategies.

6. CONCLUSION

Our comparative evaluation of medical and general-purpose language models presents two key findings: general-purpose models consistently outperform their medical- specific counterparts on specialized medical tasks, and zero-shot performance often matches or surpasses few-shot results. These findings directly challenge long-held assumptions about the need for domain- specific pretraining in specialized applications.

The implications of this work suggest a potential shift in strategy for developing specialized AI. Instead of concentrating resources on creating domain-specific model variants, the field may see greater benefit from continued improvements in general model architectures and training methodologies. As language models evolve, the distinction between general and specialized capabilities will likely become increasingly blurred, pointing toward a future where powerful, universal models can handle a diverse range of specialized tasks effectively without requiring explicit, domain-specific training.

ETHICS STATEMENT

This study was conducted using publicly available datasets (PubMedQA, BioASQ, and WikiDoc) and did not involve human subjects, patient data, or clinical trials. As a computational study analyzing the performance of language models on existing question-answer pairs, no institutional review board (IRB) approval was required. The datasets used in this research are openly accessible and have been previously published for research purposes. No personally identifiable information or sensitive medical data was collected or processed during this study. Both authors contributed equally to this study, and both approved the final version of the paper.

AUTHORS' CONTRIBUTIONS

Daniel Quillan ROXAS: Conducted the experiments, analyzed the results, and wrote the manuscript.

Hakan EMEKCİ: Supervised the research, revised the manuscript, and wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Nori H., King N., McKinney S. M., Carignan D. and Horvitz E., "Capabilities of GPT-4 on medical challenge problems," arXiv preprint arXiv:2303.13375, (2023).

[2] Jin Q., Dhingra B., Liu Z., Cohen W. and Lu X., "PubMedQA: A Dataset for Biomedical Research Question Answering," ***Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing***, 2567-2577, (2019).

[3] Krithara A., Nentidis A., Bougiatiotis K. and Paliouras G., "BioASQ-QA: A manually curated corpus for Biomedical Question Answering," ***Scientific Data***, 10: 170, (2023).

[4] "WikiDoc Medical Encyclopedia," [Online], Available: https://www.wikidoc.org/index.php/Main_Page.

[5] Singhal K. et al., "Towards Expert-Level Medical Question Answering with Large Language Models," arXiv preprint arXiv:2305.09617, (2023).

[6] Hong Z., Ajith A., Pauloski J., Duede E., Chard K. and Foster I., "The diminishing returns of masked language models to science," ***Findings of the Association for Computational Linguistics: ACL 2023***, 1270-1283, (2023).

[7] Mustafa M. A., Erdem O. A. and Söğüt E., "Use of Chest X-ray Images and Artificial Intelligence Methods for Early Diagnosis of COVID-19," ***Politeknik Dergisi***, (2025).

[8] Eriç A., Özgür E. G., Asker Ö. F. and Bekiroğlu N., "ChatGPT ve Sağlık Bilimlerinde Kullanımı," ***Celal Bayar Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi***, 11: 176-182, (2024).

[9] Rosa G. M. et al., "No parameter left behind: How distillation and model size affect Zero-Shot retrieval," arXiv preprint arXiv:2206.02873, (2022).

[10] Özden Gürcan G., Gokdas H. and Turan Kızıldoğan E., "Artificial Intelligence in Healthcare: Fall Risk Assessment in Older Adults Using Machine Learning Techniques," ***Politeknik Dergisi***, (2025).

[11] Li C. and Flanigan J., "Task contamination: Language models may not be few-shot anymore," arXiv preprint arXiv:2312.16337, (2023).

[12] Herbold S., "Semantic similarity prediction is better than other semantic similarity measures," arXiv preprint arXiv:2309.12697, (2023).

[13] Koroteev M. V., "BERT: A Review of Applications in Natural Language Processing and Understanding," arXiv preprint arXiv:2103.11943, (2021).

[14] Glushkova T., Zerva C. and Martins A. F. T., "BLEU Meets COMET: Combining lexical and neural metrics towards robust machine translation evaluation," arXiv preprint arXiv:2305.19144, (2023).

[15] Lin C.-Y., "ROUGE: A package for automatic evaluation of summaries," ***Text Summarization Branches Out***, 74-81, (2004).

[16] Papineni K., Roukos S., Ward T. and Zhu W.-J., "BLEU: A method for automatic evaluation of machine translation," ***Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics***, 311-318, (2002).

[17] Zhang T., Kishore V., Wu F., Weinberger K. Q. and Artzi Y., "BERTScore: Evaluating text generation with BERT," arXiv preprint arXiv:1904.09675, (2019).

[18] Reimers N. and Gurevych I., "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," arXiv preprint arXiv:1908.10084, (2019).

[19] "Hugging Face - The AI community building the future," [Online], Available: https://huggingface.co/.

[20] EPFL LLM Team, "Meditron-7B," Hugging Face, [Online], (2024).

[21] BioMistral Team, "BioMistral-7B," Hugging Face, [Online], (2024).

[22] MedAlpaca Team, "MedAlpaca-13B," Hugging Face, [Online], (2024).

[23] Axiong X., "PMC-LLaMA-13B," Hugging Face, [Online], (2024).

[24] Mistral AI, "Ministral-8B-Instruct-2410," Hugging Face, [Online], (2024).

[25] Google Research, "Gemma-2-9B-it," Hugging Face, [Online], (2024).

[26] LMSYS Org, "Vicuna-13B-v1.5," Hugging Face, [Online], (2024).

[27] Meta AI, "Llama-3-8B-Instruct," Hugging Face, [Online], (2024).

[28] Gao T., Yao X. and Chen D., "SimCSE: Simple contrastive learning of sentence embeddings," arXiv preprint arXiv:2104.08821, (2021).

[29] Guo Z. and Hua Y., "Continuous training and fine-tuning for domain-specific language models in medical question answering," arXiv preprint arXiv:2311.00204, (2023).

[30] Ersöz O. Ö. et al., "Makine Öğrenmesi ile Kestirimci Bakım ve Yedek Parça Yönetimi," ***Politeknik Dergisi***, (2025).