# Does Biomedical Training Lead to Better Medical Performance?

**Amin Dada**[1] **Osman Alperen Koraş** [1] **Marie Bauer**[1] **Jean-Philippe Corbeil**[2]
**Amanda Butler Contreras**[3] **Constantin Marc Seibold**[1] **Kaleb E Smith**[3]
**Julian Friedrich**[1] **Jens Kleesiek**[1*]
[1]Institute for AI in Medicine, University Hospital Essen, Germany
[2] Microsoft Healthcare & Life Sciences [3] NVIDIA

## Abstract

Large Language Models (LLMs) hold significant potential for improving healthcare applications, with biomedically adapted models promising enhanced performance on medical tasks. However, the effectiveness of biomedical domain adaptation for clinical tasks remains uncertain. In this study, we conduct a direct comparison of 12 biomedically adapted models and their general-domain base counterparts across six clinical tasks. Our results reveal that 11 out of 12 biomedical models exhibit performance declines, challenging prior findings that reported positive effects of biomedical adaptation. Notably, previous positive results primarily relied on multiple-choice evaluations, which may not reflect performance in real-world clinical applications. To promote reproducibility and further research, we open-source our evaluation pipeline, providing a resource for the development of models with practical benefits in healthcare settings.

## 1 Introduction

Large Language Models (LLMs) have the potential to transform healthcare by enhancing patient care quality and efficiency (Moor et al., 2023). Open-source biomedical LLMs, designed for medical applications, promise improved performance with fewer parameters than general models (Luo et al., 2023; Chen et al., 2023; Labrak et al., 2024). However, recent research questions the effectiveness of biomedical domain adaptation (Jeong et al., 2024; Ceballos-Arroyo et al., 2024; Dada et al., 2025).

In this study we perform a direct comparison of 12 biomedically adapted models with their general-domain base models on six clinical tasks. Our results reveal performance declines in 11 of 12 biomedical models. This is in contrast to previous findings that reported positive effects of biomedical training (Chen et al., 2023; Gururajan et al., 2024; Christophe et al., 2024). However, these studies primarily relied on multiple-choice evaluations that did not incorporate real-world clinical documents. This suggests that the observed benefits of biomedical adaptation may not translate effectively to practical healthcare settings.

To facilitate reproducibility and enable future development of models with practical benefits in healthcare settings, we open-source our evaluation pipeline. By providing a standardized framework for assessing biomedical LLMs on real-world clinical tasks, we aim to bridge the gap between benchmark performance and real-world applicability.

## 2 Related Work

The need for specialized healthcare tools has recently accelerated biomedical LLM development, yielding commercial models like Med-PaLM (Singhal et al., 2023) and MedGemini (Saab et al., 2024), and open-source alternatives such as Meditron (Chen et al., 2023), Biomistral (Labrak et al., 2024), Internist.ai (Griot et al., 2024), and Med42 (Christophe et al., 2024).

Although biomedical LLMs initially outperformed general-domain models on tasks like multiple-choice question-answering (MCQA) exams, recent studies (Jeong et al., 2024; Ceballos-Arroyo et al., 2024; Dada et al., 2025) challenge this view. Jeong et al. (2024) found no clear advantage for biomedical LLMs with model-specific prompt tuning, and Ceballos-Arroyo et al. (2024) suggest domain adaptation might impair instruction-following.

## 3 Evaluation Tasks

We introduce the clinical language understanding evaluation (CLUE) consisting of six tasks on clinical notes, consumer health questions, electronic

[*]Other affiliations: Cancer Research Center Cologne Essen (CCCE), German Cancer Consortium (DKTK, Partner site Essen) and Department of Physics of TU Dortmund (Dortmund, Germany).



*Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025)*, pages 46–59
July 31 - August 1, 2025 

| Dataset | Samples | Words | Documents | Focus |
|---|---|---|---|---|
| **Level 1** | | | | |
| MedNLI | 1425 | 21 | Clinical Notes | Clinical reasoning |
| MeQSum | 1000 | 61 | Consumer Health Questions | Summarization |
| Problem Summary | 237 | 124 | Clinical Notes | Information extraction |
| **Level 2** | | | | |
| LongHealth | 400 | 5537 | EHR | Information extraction |
| MeDiSumQA | 453 | 1452 | Discharge Summary | Simplification/Clinical reasoning |
| MeDiSumCode | 500 | 1515 | Discharge Summary | Information extraction / Coding |

Table 1: An overview of the characteristics of the tasks. We split the tasks into the difficulties level 1 and level 2.

health records (EHR) and discharge summaries, encompassing information extraction, summarization, clinical reasoning, simplification, and coding. Table 1 summarizes the characteristics of these tasks. We divide the tasks into two levels. Level 1 includes simpler tasks with short inputs, while Level 2 has complex tasks with long inputs. We provide prompt examples for each task in Figures 2, 3, 4, 5, 6, and 7 in Appendix B.3.

**MedNLI** (Romanov and Shivade, 2018) is based on clinical notes from MIMIC-III (Johnson et al., 2016). It evaluates models on predicting the logical relationship—contradiction, neutrality, or entailment—between a premise and hypotheses, testing clinical reasoning with short input lengths.

**MeQSum** (Ben Abacha and Demner-Fushman, 2019) contains 1,000 consumer health inquiries summarized by medical experts. This task evaluates whether models can understand lay language, extract key information, and reformulate patient queries into concise, medically sound questions.

**Problem Summary** Derived from SOAP-structured clinical notes, this task was first described by Gao et al. (2022) and utilizes the Subjective and Assessment sections for predicting a patient's health problems (Weed, 1964). Like MedNLI, its short input length tests basic information extraction abilities.

**LongHealth** (Adams et al., 2024) consists of 20 fictional patient records designed to challenge LLMs on long input comprehension. Evaluation involves answering questions on multiple long documents, handling added irrelevant information, and recognizing when data is unavailable. This task assesses comprehension, long-input retention, and hallucination tendencies.

**MeDiSumQA** (Dada et al., 2025) requires models to comprehend MIMIC-IV (Johnson et al., 2021) discharge summaries, extract key information, answer patient-related queries, and simplify medical information. Additionally, models must apply medical knowledge to provide appropriate follow-up advice.

Using MIMIC-IV, we create **MeDiSumCode**, an ICD-10 prediction dataset by linking discharge summaries with annotated ICD-10 codes via hospital admission IDs. This dataset provides discharge summaries as inputs and ICD-10 codes as labels for model evaluation.

**MeDiSumCode** involves assigning ICD-10 codes to diagnoses and procedures in discharge summaries, a critical task for patient records, billing, and healthcare analysis (Organization, 2004). This challenge requires models to extract diagnoses from complex clinical text, comprehend over 70,000 ICD-10 codes, and accurately match diagnoses to the correct codes.

## 4 Experimental setup

We evaluated 24 language models, including biomedically trained models, their base models, and additional general-domain models as reference. Our evaluation aims to (1) measure the effects of continuous biomedical training, (2) assess whether biomedical models or general-domain models are more suitable for specific medical scenarios, and (3) rank current openly available models. Appendix A.1 describes the metrics we applied to each task. For each task, we report the average over all metrics.

### 4.1 Models

We evaluate the following biomedical LLMs: Meditron-7B and 70B (Chen et al., 2023), Internist.ai (Griot et al., 2024), BioMistral (Labrak et al., 2024), Llama3-Aloe-8B-Alpha (Gururajan et al., 2024), Llama3-OpenBioLLM-8B and 70B (Ankit Pal, 2024), Med42-Llama3-8B and 70B (Christophe et al., 2024), and Meditron3-8B and 70B (OpenMeditron, 2024). More details are in Table 5 in Appendix B.2. We did not evaluate Llama2-based models on Level 2 tasks due to their

| Model | Level 1 | | | Level 2 | | |
|---|---|---|---|---|---|---|
| | MedNLI | Prob. Sum. | MeQSum | LongHealth | MeDiSumQA | MeDiSumCode |
| Llama-2-7B | 29.5 | 16.8 | 14.0 | - | - | - |
| - Meditron-7B | 2.4 (-27.1) | 21.6 (+4.8) | 15.1 (+1.1) | - | - | - |
| Llama-2-70B | 76.3 | 18.6 | 10.6 | - | - | - |
| - Meditron-70B | 63.5 (-12.7) | 18.7 (+0.1) | 9.6 (-1.1) | - | - | - |
| Mistral-7B-Instruct-v0.1 | 64.8 | 25.0 | 31.1 | 30.0 | 25.5 | 13.9 |
| - BioMistral-7B | 62.8 (-2.0) | 25.1 (+0.1) | 33.9 (+2.8) | 26.7 (-3.3) | 22.8 (-2.7) | 22.0 (+8.2) |
| - BioMistral-7B-DARE | 66.8 (+2.0) | 28.4 (+3.4) | 34.5 (+3.4) | 30.5 (+0.5) | 25.7 (+0.2) | 21.3 (+7.4) |
| - Internist.ai 7b | 76.3 (+11.5) | 23.1 (-1.9) | 15.2 (-15.9) | 44.2 (+14.2) | 19.8 (-5.6) | 21.9 (+8.0) |
| Zephyr 7B | 68.5 | 25.5 | 34.2 | 33.3 | 22.7 | 28.5 |
| Meta-Llama-3-8B-Instruct | 74.1 | 31.6 | 39.5 | 58.8 | 30.3 | 27.8 |
| - OpenBioLLM-8B | 44.9 (-29.1) | 21.7 (-9.9) | 33.0 (-6.4) | 26.9 (-31.9) | 30.4 (+0.1) | 18.9 (–8.9) |
| - Med42-8B | 77.5 (+3.4) | 32.4 (+0.8) | 42.8 (+3.3) | 57.8 (-1.0) | 29.7 (-0.6) | 25.2 (-2.6) |
| - Aloe-8B-Alpha | 73.9 (-0.1) | 21.3 (-10.3) | 32.3 (-7.2) | 49.7 (-9.1) | 21.4 (-8.9) | 19.8 (-8.0) |
| Meta-Llama-3-70B-Instruct | 79.4 | 34.7 | 43.0 | 83.8 | 33.3 | 50.9 |
| - OpenBioLLM-70B | 80.8 (+1.5) | 23.7 (-11.0) | 38.1 (-4.8) | 72.9 (-10.8) | 30.0 (-3.3) | 33.8 (-17.2) |
| - Med42-70B | 76.1 (-3.2) | 24.3 (-10.4) | 33.9 (-9.0) | 56.4 (-27.4) | 24.2 (-9.1) | 42.0 (-9.0) |
| Meta-Llama-3.1-8B-Instruct | 79.1 | 29.8 | 42.1 | 70.5 | 32.9 | 33.4 |
| - Meditron3-8B | 74.0 (-5.1) | 27.9 (-1.9) | 40.8 (-1.3) | 50.5 (-20.0) | 31.1 (-1.8) | 10.1 (-23.3) |
| Meta-Llama-3.1-70B-Instruct | 84.9 | 34.5 | 43.7 | 87.7 | 32.6 | 52.8 |
| - Meditron3-70B | 82.6 (-2.3) | 31.8 (-2.7) | 42.1 (-1.6) | 67.7 (-20.0) | 32.1 (-0.5) | 47.7 (-5.0) |
| Mistral-7B-Instruct-v0.2 | 69.9 | 29.2 | 40.3 | 57.4 | 29.4 | 30.0 |
| Phi-3-mini-instruct | 66.6 | 28.4 | 36.7 | 45.9 | 25.8 | 41.1 |
| Mixtral-8x7B-Instruct-v0.1 | 80.1 | 18.4 | 13.8 | 58.1 | 28.8 | 40.8 |
| Mixtral-8x22B-Instruct-v0.1 | 76.5 | 27.3 | 39.6 | 79.7 | 30.0 | 43.9 |

Table 2: The aggregated average scores over the individual metrics for each task of our evaluation on CLUE. For biomedical models we include performance gains and losses compared to their respective base model.

limited context size of 4k tokens.

We also evaluate the base models of the biomedical LLMs and the following additional models: Zephyr-7B-Beta (Tunstall et al., 2023), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Phi-3-Mini-128k-Instruct (Abdin et al., 2024), Mixtral-8x7B, and Mixtral-8x22B (Jiang et al., 2024).

## 5 Results

Table 2 presents average results for each task, while Table 3 summarizes the relative performance differences between biomedical models and their base models compared to previous MCQA evaluations. Only BioMistral-7B-DARE shows a consistent performance advantage across all six tasks. In contrast, 11 models show performance losses in at least one task, and four biomedical models exhibit declines on all tasks, indicating that domain-specific fine-tuning can harm general task performance.

Most performance gains are observed in models based on Llama-2 and Mistral-7B-v0.1, while models derived from more recent LLMs frequently underperform after adaptation. Additionally, improvements are more common in models with up to 8B parameters, whereas larger models tend to lose performance after biomedical training. Figure 1 shows a comparison between the best-performing biomedical models and their base models. We find slight performance gains for Mistral-7B-v0.1 but clear performance losses for models based on better-performing general-domain LLMs.

| Model | MCQA | Level 1 | Level 2 |
|---|---|---|---|
| MEDITRON-7B | +6.07 | -7.08 | - |
| MEDITRON-70B | +3.63 | -4.59 | - |
| BioMistral-7B | +4.13 | +0.26 | +0.71 |
| BioMistral-7B-DARE | +4.57 | +2.93 | +2.7 |
| Internist.ai 7b | - | -2.07 | +5.52 |
| OpenBioLLM-8B | -0.63 | -15.17 | -13.54 |
| OpenBioLLM-70B | +1.46 | -4.78 | -10.45 |
| Med42-8B | +0.47 | +2.51 | -1.4 |
| Med42-70B | +2.8 | -7.57 | -15.14 |
| Aloe-8B-Alpha | +2.21 | -5.87 | -8.67 |
| Meditron3-8B | - | -2.76 | –15.04 |
| Meditron3-70B | - | -2.18 | -8.51 |

Table 3: A direct comparison between biomedical models and their respective base models Llama-2-(7B/70B), Mistral-7B-v0.1, Meta-Llama-3-(8B/70B) and Meta-Llama-3.1-(8B/70B). The scores show the difference between each model before and after domain adaptation. MCQA shows the reported performance difference averaged over (MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019)) while Level 1 and 2 show the differences on CLUE.

Task complexity also plays a key role: gains are mainly seen in Level 1 tasks, while performance on more complex Level 2 tasks often declines. This suggests biomedical models may struggle with tasks requiring language understanding and reasoning.

Unlike previous reports of biomedical LLM improvements on MCQA evaluations, only two models show slight average gains on both Level 1 and Level 2 tasks on CLUE (see Table 3).

Overall, general-domain LLMs remain strongest, with Llama3.1-70B emerging as the top performer. Although Llama3-Med42-8B slightly outperforms its base model on simple tasks (+0.56%), it shows a large drop on Level 2 tasks (-8.03%).

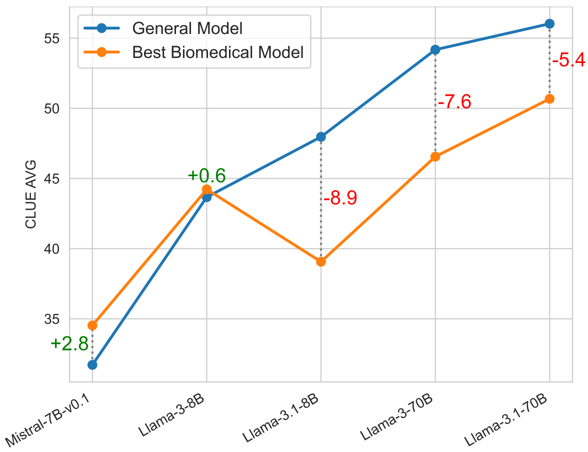


Figure 1: Comparison of average scores between general-domain models and highest scoring biomedical models.

### 5.1 Error Analysis

Primary contributors to biomedical model performance drops are LongHealth task 3 and MeDiSumCode valid code scores (Table 4). Biomedical Mistral-7B-based models improve, whereas Llama3-based models show performance decreases of up to 79.15%.

LongHealth task 3 measures how often a model correctly returns no answer when information is absent, reflecting hallucination rates. Similarly, MeDiSumCode's valid code scores reveal ICD-10 code fabrication, with low-scoring models incrementing numbers instead of predicting valid codes (see Appendix B.4). Notably, Meta-Llama-3-8B-Instruct scored 56.25 on LongHealth task 3, whereas Llama3-OpenBioLLM-8B dropped to 1.55. Llama3-OpenBioLLM-70B also underperforms compared to Meta-Llama-3-70B-Instruct.

Beyond hallucinations, biomedical models often fall into repetition loops, generating the same tokens repeatedly and producing incoherent outputs. Additionally, models struggle with instruction adherence, particularly in long-input tasks like

| | LH Task3 | Valid Codes |
|---|---|---|
| BioMistral-7B | +4.15 | +17.26 |
| BioMistral-7B-DARE | +0.95 | +18.79 |
| Internist.ai 7b | +45.55 | +16.32 |
| OpenBioLLM-8B | -40.05 | -10.77 |
| Med42-8B | -12.7 | -6.8 |
| Aloe-8B-Alpha | -22.55 | -17.09 |
| OpenBioLLM-70B | -28.80 | -20.29 |
| Med42-70B | -79.15 | -15.39 |
| Meditron3-8B | -52.15 | -49.19 |
| Meditron3-70B | -54.6 | -4.76 |

Table 4: Mistral-7B-v0.1 , Meta-Llama-3-(8B/70B) and Meta-Llama-3.1-(8B/70B) based models on LongHealth task 3 and percentage of valid ICD-10 codes in MeDiSumCode

LongHealth. This supports previous similar observations (Ceballos-Arroyo et al., 2024).

## 6 Discussion

Performance declines are observed across various training methods, except for BioMistral-DARE, which uses weight merging, indicating a potential mitigation strategy. However, the superior performance of Mistral-7B-Instruct-v0.2 (Table 2) suggests that improved general-domain training has a more significant impact than biomedical training.

Many SFT models used generated data, suggesting data quality affects performance. Internist.ai 7b, trained on high-quality data, performed best on Level 2 tasks, reinforcing this hypothesis.

Improvements were almost exclusive to the lower-performing Mistral-7B-Instruct-v0.1 models, suggesting that recent general models like Llama-3 and Mistral-7B-v0.2 already address these gaps. Tables 4 and 3 further support this.

## 7 Conclusion

Our study suggests that biomedical LLMs are not competing effectively with general-domain models on clinical tasks. While some biomedical models have shown improvements, more recent and larger models are underperforming. Fine-tuning these models with domain-specific data often leads to reduced performance, introducing hallucinations and decreased model stability. This stands in contrast to traditional MCQA evaluations, where biomedical models have previously demonstrated superior performance. Our evaluation provides a more practical assessment of LLM capabilities in real-world healthcare settings. To support further progress in this field, we open-source our evaluation scripts, allowing for broader validation and replication of our results.

## Limitations

Our study has several limitations that should be considered. Due to the significant computational resources required to run LLMs with up to 141 billion parameters, we did not explore the impact of various model configurations, such as temperature settings, or advanced techniques like chain-of-thought prompting on model performance. Future research should investigate these aspects to gain a more comprehensive understanding of their effects. Additionally, the datasets we use are publicly available resources. As such, we cannot completely prevent data contamination. This limitation underscores the need for future research into robust methods for mitigating data contamination, which is crucial for ensuring the validity of any public LLM benchmark. While we presented novel insights in this paper, their application to clinical data requires further investigation. Future work should refine these methods to enhance their applicability and reliability in clinical settings. Furthermore, our evaluation primarily focused on tasks involving clinical documents and their relevance, but it was not conducted in a realistic clinical setting. Therefore, extensive evaluation through prospective clinical trials is necessary to meet the required safety levels before applying these models to clinical environments.

## References


Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2024. LongHealth: A Question Answering Benchmark with Long Clinical Documents. *Preprint*, arxiv:2401.14490.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28th - August 2*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical) LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *Preprint*, arxiv:2311.16079.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco AF Pimentel, and Shadab Khan. 2024. Med42 – evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches.

Amin Dada, Osman Alperen Koras, Marie Bauer, Amanda Butler, Kaleb E Smith, Jens Kleesiek, and Julian Friedrich. 2025. Medisumqa: Patient-oriented question-answer generation from discharge letters. *arXiv preprint arXiv:2502.03298*.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew M. M. Churpek, and Majid Afshar. 2022. Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2024. Impact of high-quality, mixed-domain data on the performance of medical language models. *Journal of the American Medical Informatics Association*, page ocae120.

Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, Sergio Alvarez-Napagao, Eduard Ayguadé-Parra, and Ulises Cortés Dario Garcia-Gasulla. 2024. Aloe: A family of fine-tuned open healthcare llms. *Preprint*, arXiv:2405.01886.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2021. Mimic-iv.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *Preprint*, arxiv:2402.10373.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. *arXiv preprint*. ArXiv:2308.09442 [cs] version: 2.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

OpenMeditron. 2024. Meditron3 model card.

World Health Organization. 2004. Icd-10 : international statistical classification of diseases and related health problems : tenth revision.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *CoRR*, abs/1808.06752.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

| Model Name | Base Model | Type of Training |
|---|---|---|
| Meditron-7B | Llama2-7B | Continued pretraining |
| Internist.ai 7B | Mistral-7B-v0.1 | Continued pretraining + SFT |
| BioMistral-7B | Mistral-7B-Instruct-v0.1 | Continued pretraining |
| BioMistral-7B-DARE | Mistral-7B-Instruct-v0.1 | Continued pretraining +DARE |
| Llama3-OpenBioLLM-8B | Meta-Llama-3-8B-Instruct | SFT + DPO |
| Llama3-Med42-8B | Meta-Llama-3-8B-Instruct | SFT + DPO |
| Llama3-Aloe-8B-Alpha | Meta-Llama-3-8B-Instruct | SFT + DPO |
| Meditron3-8B | Meta-Llama-3.1-8B-Instruct | - |
| Meditron-70B | Llama-2-70B | Continued pretraining |
| Llama3-OpenBioLLM-70B | Meta-Llama-3-70B-Instruct | SFT + DPO |
| Llama3-Med42-70B | Meta-Llama-3-8B-Instruct | SFT + DPO |
| Meditron3-70B | Meta-Llama-3.1-70B-Instruct | - |

Table 5: Evaluated Biomedical Models

Lawrence L. Weed. 1964. Medical records, patient care, and medical education. *Irish Journal of Medical Science*, 39(6):271–282.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

# A Task Details

## A.1 Metrics

For open-ended tasks, we report the F1-score between the model predictions and ground truth unigrams (ROUGE-1), bigram (ROUGE-2), and the longest common subsequence (ROUGE-L)[1] (Lin, 2004). We compute the BERTScore (Zhang et al., 2019) on clinical documents to measure semantic similarity using an encoder trained on MIMIC III[2] (Alsentzer et al., 2019). We first tuned the score rescaling baselines for MIMIC IV discharge summaries. For Problem Summaries and MeDiSumQA, we also extract the Unified Medical Language System (UMLS) (Bodenreider, 2004) entities with scispacy (Neumann et al., 2019) and compute their F1-score to consider medical abbreviations and synonyms. When evaluating MedDiSumCode, we calculate the ratio of valid ICD-10 codes. We use the python package icd10-cm[3] to probe the validity of ICD-10 codes. We distinguish between exact match (EM) and the match of the first three characters of the codes, which is an approximate match (AP) based on the hierarchical structure of ICD-10 codes.

[1] https://huggingface.co/spaces/evaluate-metric/rouge
[2] emilyalsentzer/Bio_ClinicalBERT
[3] https://pypi.org/project/icd10-cm/

# B Experimental setup

## B.1 Computational Resources

All experiments were conducted on an NVIDIA DGX A100 640GB node with 8x NVIDIA A100 80GB Tensor Core GPUs within three days, resulting in approximately 1536 GPU hours.

## B.2 Models

Table 5 lists all biomedical models we evaluated.

## B.3 Prompting

We apply few-shot prompting and use the instruction template on Hugging Face for the instruction-tuned models. For the other models, we concatenate the system prompt, few-shot examples, and user prompt into one string separated by double newlines. For the level one evaluation, we performed 3-shot prompting. For level two, we provide one shot with the exception of LongHealth, where we provide no examples due to the content length.

Figures 2, 3, 4, 5, 6, and 7 are showing the prompt formats we are using for the different benchmark tasks. If the input length allowed this, we also included sample texts from the datasets.

## B.4 Error Analysis

Figure 8 shows some examples of the described type of error with regard to counting.

**System Prompt**

You are a highly skilled assistant, specifically trained to assist patients. Your primary responsibility will be to summarize patient inquiries as concise question. You will be given such a patient inquiry. You will be expected to summarize and rewrite the inquiry as a concise question. Only write out the question. Do not add any other text.

**3-shot examples**

**User Prompt**

--------------PATIENT INQUIRY--------------
SUBJECT: hearing loss
MESSAGE: have you experience in hearing loss due to autoimmune disorder called Cogan syndrome? If yes I will contact you for my 18year old son.
--------------END PATIENT INQUIRY--------------

**Assistant Response**

Question: Can Cogan syndrome cause hearing loss?

⋮

**User Prompt**

--------------PATIENT INQUIRY--------------
[...]
--------------END PATIENT INQUIRY--------------

Figure 2: MeQSum prompt format with example.

System Prompt

You are a highly skilled and detail-oriented assistant, specifically trained to assist medical professionals in interpreting and extracting key information from medical documents. Your primary responsibility will be to analyze discharge letters from hospitals. You will receive an excerpt of such a discharge letter. Your task is to summarize the diagnoses and problems that led to the patient's hospitalization.

User Prompt

**3-shot examples**

--------------BEGIN DISCHARGE LETTER--------------
Chief Complaint: 24 Hour Events: -post cath check okay -epistaxis resolved - RISB 68, converted to PSV 15/5. -Tele with frequent PVCs Allergies: No Known Drug Allergies

78 year-old man with history as above who was referred from OSH for a
cardiac catheterization secondary to persistent shortness of breath. Pt
had [**Year (4 digits) **] placed in distal RCA for 90% lesion. Procedure was complicated
for sigificant epistaxis following NGT placement. Integrelin and
heparin held. Pt did received plavix load and aspirin.
--------------END DISCHARGE LETTER--------------
Now respond with the list of diagnoses and patient problems. Do not generate anything else.

Assistant Response

Diagnoses/Patient problems: Coronary Artery Disease, Chronic systolic heart failure

•
•
•

User Prompt

--------------BEGIN DISCHARGE LETTER--------------
[...]
--------------END DISCHARGE LETTER--------------

Figure 3: Problem Summary prompt format with example.

**System Prompt**

You are a highly skilled assistant, specifically trained to assist medical professionals. You will receive two sentences labeled 'SENTENCE_1' and 'SENTENCE_2', respectively. Your task is to determine the logical relation between the two sentences. Valid answers are: ENTAILMENT, NEUTRAL or CONTRADICTION.

---

**3-shot examples**

**User Prompt**

SENTENCE_1: In the ED, initial VS revealed T 98.9, HR 73, BP 121/90, RR 15, O2 sat 98% on RA.
SENTENCE_2: The patient is hemodynamically stable

**Assistant Response**

entailment

---

**User Prompt**

SENTENCE_1: [ ...]
SENTENCE_2: [ ...]

Figure 4: MedNLI prompt format with example.

System Prompt

You are a highly skilled and detail-oriented assistant, specifically trained to assist medical professionals in interpreting and extracting key information from medical documents. Your primary responsibility will be to analyze discharge letters from hospitals. When you receive one or more of these letters, you will be expected to carefully review the contents and accurately answer multiple-choice questions related to these documents.

Your answers should be:
1. Accurate: Make sure your answers are based on the information provided in the letters.
2. Concise: Provide brief and direct answers without unnecessary elaboration.
3. Contextual: Consider the context and specifics of each question to provide the most relevant information.

Remember, your job is to streamline the physician's decision-making process by providing them with accurate and relevant information from discharge summaries. Efficiency and reliability are key.

User Prompt

--------------BEGIN DOCUMENTS--------------

{documents}

--------------END DOCUMENTS--------------

{question_text}
{options}

Please answer using the following format:
1. Begin your answer with the phrase "The correct answer is".
2. State the letter of the correct option (e.g., A, B, C, D, E).
3. Follow the letter with a colon and the exact text of the option you chose.
4. Make sure your answer is a single, concise sentence.

For example, if the correct answer to a question is option C, and the text for C is 'Acute Bronchitis', your answer should be:
'The correct answer is C: Acute bronchitis.'

Figure 5: LongHealth prompt format.

System Prompt

You are a highly skilled assistant, specifically trained to assist patients. Your primary responsibility will be to work with discharge letters from hospitals. You should carefully review the contents and accurately answer questions related to the described case. Keep you answer as short as possible only focussing on the most relevant infromation. Simplify the information in a patient-friendly way and avoid extensive details or expert terminology. If the requested information is not given in the document, try to deduce it on the basis of the information provided.

Here are some examples for good answers:
--------------BEGIN EXAMPLES--------------
Question: What type of medication was prescribed for my high blood pressure?
Answer: We prescribed a beta-blocker called metoprolol to help manage your high blood pressure.

Question: How was my condition diagnosed?
Answer: We performed a chest X-ray and a CT scan, which revealed that you had fluid in your lungs.

Question: What was the reason for my persistent cough, and what was the treatment?
Answer: Your persistent cough was due to an upper respiratory infection, and we treated it with a course of antibiotics to address the infection and a cough suppressant to relieve symptoms.

Question: What kind of test was performed to check my thyroid function?
Answer: We performed a blood test called a thyroid function test to measure your hormone levels.

Question: What type of vaccine did I receive today?
Answer: You received the influenza vaccine to help protect you against the flu this season.
--------------END EXAMPLES--------------
Use a similar choice of words and level of detail as in the examples.

**1-shot example**

User Prompt

--------------BEGIN DISCHARGE LETTER--------------
{discharge_summary}
--------------END DISCHARGE LETTER--------------
Question: What was the outcome of my virtual colonoscopy?

Assistant Response

Answer: We did not find any polyps, masses, or signs of inflammatory disease in your examination.

User Prompt

--------------BEGIN DISCHARGE LETTER--------------
{discharge_summary}
--------------END DISCHARGE LETTER--------------
What side effect did I experience from taking Clozapine, and how was it managed?

Figure 6: MeDiSumQA prompt format.

**System Prompt**

You are a highly skilled and detail-oriented assistant, specifically trained to assist medical professionals in interpreting and extracting key information from medical documents. Your primary responsibility will be to analyze discharge letters from hospitals. You will be given such a discharge letter. Your task is to identify all primary and secondary diagnoses from the report and list their respective ICD-10 codes.

**1-shot example**

**User Prompt**

--------------BEGIN DISCHARGE LETTER--------------
{discharge_summary}
--------------END DISCHARGE LETTER--------------
Now return the list of diagnoses ICD-10 codes you found. Only list the ICD-10 codes. Do not generate anything else.

**Assistant Response**

ICD-10 Codes: F321, F1010, R45851

**User Prompt**

--------------BEGIN DISCHARGE LETTER--------------
{discharge_summary}
--------------END DISCHARGE LETTER--------------
Now return the list of diagnoses ICD-10 codes you found. Only list the ICD-10 codes. Do not generate anything else.

Figure 7: MeDiSumCode prompt format.

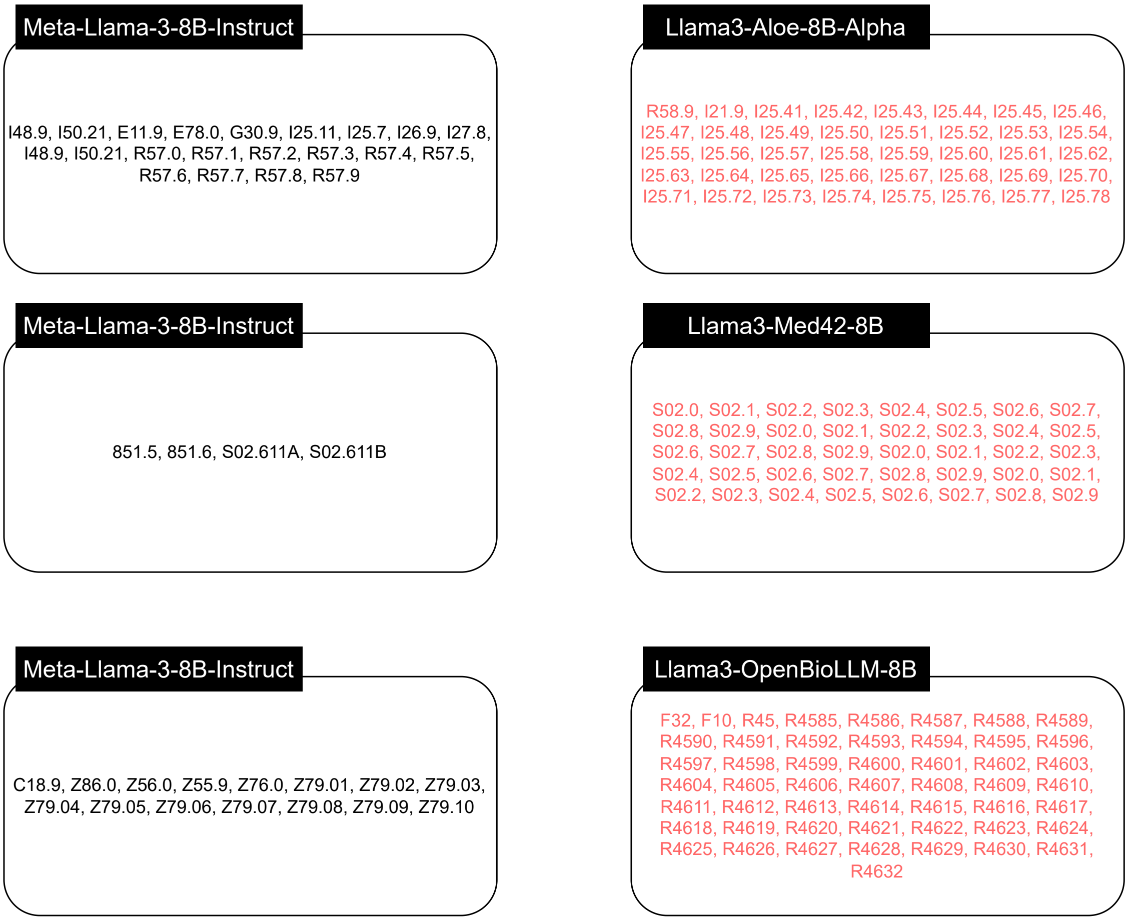


Figure 8: Biomedical models that show the described counting behavior compared to their base model.