Paper Title	Year	Architecture	LLM Used	Evaluations	Summary
ClinicalGPT	2023	Based on a large and diverse medical dataset including cMedQA2, cMedQA-KG, MD-EHR, MEDQA-MCMLE, and MedDialog.	BLOOM-7B	Evaluated on medical conversation, medical examination, diagnosis suggestion, and medical question answering. ClinicalGPT outperformed other LLMs on diagnosis tasks, achieving an average accuracy of 80.9% across all disease groups.	ClinicalGPT is a large language model that is fine-tuned on a diverse range of medical data. The authors find that ClinicalGPT demonstrates superior capabilities in understanding and generating medical and clinical-related responses. It performs particularly well in diagnosis tasks, demonstrating its potential for real-world clinical applications.
BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains	2024	The BioMistral model is constructed using Mistral 7B Instruct v0.1 as its foundation. The model is further pre-trained on PubMed Central.	Mistral 7B Instruct v0.1.	Evaluated on a benchmark of 10 medical question answering (QA) tasks in English. The benchmark was also automatically translated into 7 other languages to assess multilingual generalisation.	BioMistral 7B is an open-source large language model that is specifically tailored for the biomedical domain. It leverages Mistral 7B Instruct v0.1 as its foundation and is further pre-trained on PubMed Central. The study demonstrates BioMistral's superior performance compared to other open-source medical models and its competitive edge against proprietary models in both English and multilingual medical QA tasks.
Med-Gemini	2024	Med-Gemini leverages the Gemini family of models, specifically highlighting the use of Med-Gemini-L 1.0 and Med-Gemini-M 1.5 in different contexts.	The source does not explicitly state the base Gemini model architectures. It mentions that Med-Gemini builds upon the existing capabilities of the Gemini family.	Evaluated on multiple medical reasoning benchmarks, including MedQA (USMLE), NEJM CPC, and GeneTuring. Also evaluated for multimodal capabilities on tasks like visual question answering and long-context processing in clinical and research scenarios.	Med-Gemini is a family of large language models specifically trained for medical applications. It combines the conversational capabilities of the original Gemini model with fine-tuned medical knowledge, allowing it to engage in medical dialogues, perform diagnoses, and generate summaries from medical records and research papers.
Aloe	2024	Used SFT(Supervised Fine tuning)which included synthetic medical data, High quality general datasets and curated public medical datasets, models are adjusted with human preferences using DPO.	The source focuses on evaluating and comparing various LLMs in the medical domain, including Aloe, without detailing their specific architectures.	Evaluated using the Medprompt strategy, which incorporates nearest neighbour examples into prompts. Also assessed on bias and toxicity benchmarks, demonstrating competitive performance while highlighting areas for further improvement.	The study evaluates Aloe and other LLMs in the medical domain using a novel prompting strategy called Medprompt. Aloe shows promising performance on medical question answering and bias and toxicity benchmarks, indicating its potential for healthcare applications. However, the study also highlights the need for continued research and development to address ethical considerations and potential risks associated with using LLMs in healthcare.
MedTrinity-25M	2024	Leverages a multimodal approach using image-ROI-description triplets. It incorporates a medical knowledge database built using PubMed, StatPearls, and medical textbooks.	The study explores the use of various multimodal foundation models, including LLaVA-Med, Med-Flamingo, and Med-PalM, without specifying their underlying LLM architectures.	Evaluated by fine-tuning various MLLMs on the dataset for medical VQA and captioning tasks. LLaVA-Med++, an enhanced version of LLaVA-Med, achieved state-of-the-art results on the VQA-RAD and SLAKE benchmarks.	MedTrinity-25M is a large-scale medical dataset composed of image-ROI-description triplets. It includes multi-granular textual descriptions for each image and incorporates external medical knowledge retrieved from sources like PubMed and StatPearls. The study finds that MLLMs trained on MedTrinity-25M show substantial improvements in medical VQA and captioning tasks, demonstrating the dataset's value in advancing medical AI.
Med42-v2	2024	The study focuses on fine-tuning strategies for LLMs in the clinical domain using Med42-v2, which is built upon the Llama3 architecture.	The study employs various Llama3 models, including Med42-Llama3-8B, Med42-Llama3-70B, Med42-Llama3.1-8B, and Med42-Llama3.1-70B.	Evaluated using Eleuther AI's evaluation harness framework for zero-shot performance on medical benchmarks like MMLU, MedMCQA, MedQA, USMLE, PubMedQA, and ToxiGen.	Med42-v2 is a suite of clinical LLMs that are fine-tuned from Llama3 models. The study evaluates these models on a range of medical benchmarks, showcasing their strong zero-shot performance and potential for clinical use.