

BRIDGE: Benchmarking Large Language Models for Understanding Real-world Clinical Practice Text

Jiageng Wu^{1,*}, Bowen Gu^{1,*}, Ren Zhou², Kevin Xie³, Doug Snyder^{4,5}, Yixing Jiang⁶,
Valentina Carducci⁴, Richard Wyss¹, Rishi J Desai¹, Emily Alsentzer⁶, Leo Anthony Celi^{5,7,8},
Adam Rodman⁹, Sebastian Schneeweiss¹, Jonathan H. Chen^{10,11,12}, Santiago Romero-Brufau^{4,5},
Kueiyu Joshua Lin^{1, #}, and Jie Yang^{1, 13, 14, 15, #}

¹ Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

² Siebel School of Computing and Data Science, The Grainger College of Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA

³ Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

⁴ Department of Otorhinolaryngology – Head & Neck Surgery, Mayo Clinic, Rochester, MN, USA

⁵ Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

⁶ Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA

⁷ Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA

⁸ Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

⁹ Division of General Internal Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

¹⁰ Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA

¹¹ Division of Hospital Medicine, Stanford University, Stanford, CA, USA

¹² Stanford Clinical Excellence Research Center, Stanford University, Stanford, CA, USA

¹³ Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA

¹⁴ Broad Institute of MIT and Harvard, Cambridge, MA, USA

¹⁵ Harvard Data Science Initiative, Harvard University, Cambridge, MA, USA

***Contribute equally to this paper**

Co-senior authorship.

Correspondence:

Jie Yang, PhD (jyang66@bwh.harvard.edu) and Kueiyu Joshua Lin, MD, ScD (jklin@bwh.harvard.edu), Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital & Harvard Medical School, 75 Francis St, Boston MA 02115, USA

Abstract

Large language models (LLMs) hold great promise for medical applications and are evolving rapidly, with new models being released at an accelerated pace. However, current evaluations of LLMs in clinical contexts remain limited. Most existing benchmarks rely on medical exam-style questions or PubMed-derived text, failing to capture the complexity of real-world electronic health record (EHR) data. Others focus narrowly on specific application scenarios, limiting their generalizability across broader clinical use. To address this gap, we present BRIDGE, a comprehensive multilingual benchmark comprising 87 tasks sourced from real-world clinical data sources across nine languages. We systematically evaluated 52 state-of-the-art LLMs (including DeepSeek-R1, GPT-4o, Gemini, and Llama 4) under various inference strategies. With a total of 13,572 experiments, our results reveal substantial performance variation across model sizes, languages, natural language processing tasks, and clinical specialties. Notably, we demonstrate that open-source LLMs can achieve performance comparable to proprietary models, while medically fine-tuned LLMs based on older architectures often underperform versus updated general-purpose models. The BRIDGE and its corresponding leaderboard serve as a foundational resource and a unique reference for the development and evaluation of new LLMs in real-world clinical text understanding.

Keywords: Large Language Models, Electronic Healthcare Records, Multilingual, Real-world clinical tasks, Benchmark.

Introduction

Recent advances in large language models (LLMs) have demonstrated a transformative potential in improving healthcare delivery and clinical research.¹ By combining extensive pretraining on vast corpora with supervised instruction tuning across diverse tasks,^{2,3} LLMs exhibit exceptional capabilities in textual understanding, generation, and reasoning,⁴⁻⁶ and show promise in medical applications. The prompt-based instruction provides an intuitive and easy-to-use approach to interact with LLMs on diverse tasks. For clinicians, LLMs can support the drafting of clinical documentation^{7,8} and assist in clinical decision-making⁹, enhancing efficiency and reducing workload burdens.¹⁰ LLMs

also offer the promise to benefit patients by providing simple interpretations of complex medical information¹¹ and personalized preventive advice,¹² promoting patient engagement, treatment adherence, and overall disease management.^{13,14} Consequently, LLMs hold significant promise for improving the quality, cost and accessibility of healthcare services worldwide.¹⁵ However, concerns persist regarding the reliability and clinical validity of LLM-generated outputs,¹⁶ particularly given the high diversity of clinical tasks, specialties, and languages.¹⁷ Moreover, LLMs are rapidly evolving, with new models released almost every week, and there is significant diversity among them—ranging from proprietary to open-source, medical-specific to general-purpose, and from small to large models. The general-purpose nature of LLM usage complicates institutional model selection and comparison. While clinical trials are necessary for the most high-risk cases, they are slow, expensive, and cannot possibly investigate every single use case. Even quality improvement evaluation paradigms, which largely utilize evaluation data that is already collected, cannot possibly keep up with the pace of model development. Clinical benchmarks – automated, timely, and systematic evaluations of model performance – remain essential for clinicians, patients, health systems, and regulators, providing both understanding of the usability and trustworthiness of LLMs across diverse clinical scenarios.

The most commonly-used benchmarks of LLMs in medicine focus on medical questions sourced from medical examinations or derived from academic literature,¹⁸ exemplified by the United States Medical Licensing Examination (USMLE)¹⁹ or PubMed-based datasets.^{20,21} These standardized question sets offer a basic and rapid assessment, such as the GPT-4²² and Med-PaLM 2²³, achieving expert-level scores on the USMLE. However, such datasets were not sourced from real-world clinical practice and failed to fully characterize the complexities of clinical environments.²⁴ The simplified nature of examination-style questions overlooks the multifaceted, context-rich scenarios routinely encountered by clinicians.²⁵ Furthermore, most of these knowledge-retrieval benchmarks have become effectively saturated. Text from medical exams and PubMed is well formatted and grammatically correct, in stark contrast to electronic health records (EHRs) from real-world clinical systems,²⁶ which often feature abbreviations, acronyms, varied structures, and non-standard expressions.²⁷ Additionally, many critical clinical tasks, such as phenotype extraction, have so far received insufficient attention in large-scale performance evaluations despite their importance and necessity in practical settings.²⁸

While some studies have evaluated LLMs in real clinical settings, they typically focused on specific use cases, making it difficult to generalize the findings to other clinical applications.^{29–31} In addition, limited research on multilingual medical benchmarks impedes the broader applicability of LLMs in global healthcare, raising concerns about the potential bias arising from underrepresented languages and regions.^{32–35}

The rapid evolution of LLMs underscores a high demand for comprehensive and continuously updated medical leaderboards.³⁶ With advanced technology and models emerging every few weeks, the landscape of LLMs is dynamic and ever-changing. Notably, the recent introduction of medical LLMs such as Med-PaLM1/2^{23,37}, MeLLaMA,³⁸ and Med-Found³⁹ also highlights the growing focus on improving performance and clinical relevance in healthcare. Leaderboards – objective displays of model performance across a wide variety of tasks, are essential for providing fair comparisons of LLM capabilities and tracking performance variations, thereby offering valuable guidance for subsequent model development and clinical implementation. Such leaderboards have already been widely implemented in non-medical fields⁴⁰ but have not yet been adopted in clinical domains. As LLMs are progressively integrated and deployed into clinical practice, robust benchmarking is vital for proactively identifying and mitigating potential risks and biases before they impact patient care.⁴¹ Therefore, establishing a unified, realistic, and multilingual clinical benchmark is crucial for bridging the gap between the theoretical capabilities of LLMs and their practical implementation in specific use cases and care settings.^{28,42}

To address the above challenges, we developed a large-scale and comprehensive benchmark that evaluates LLM performance on multilingual, real-world clinical text across diverse tasks. Building upon our systematic review of global clinical text resources,⁴³ this study proposes BRIDGE, a multilingual LLM benchmark that comprises 87 real-world clinical text tasks spanning nine languages and more than one million samples. Reference standards for benchmark evaluation are sourced from the original data releases, including various forms of manual review and labels derived from structured data linked to the source EHRs.⁴³ To our knowledge, BRIDGE is the largest benchmark for LLM in medicine to date. We evaluated 52 advanced LLMs, including DeepSeek-R1,⁴⁴ Llama 4,⁴⁵ GPT-4o,⁴⁶ and Google Gemini,⁴⁷ under three different commonly employed inference strategies (zero-shot, few-

shot,⁴⁸ and chain-of-thought⁴⁹). By integrating a systematic task taxonomy, we established a comprehensive leaderboard that not only provides a holistic perspective on LLM performance but also investigates their capabilities across various clinical settings, including inference strategies, languages, task types, and clinical specialties. This study provides critical insights and resources for integrating LLM into clinical practice, bridging the gap between LLM development and clinical applications.

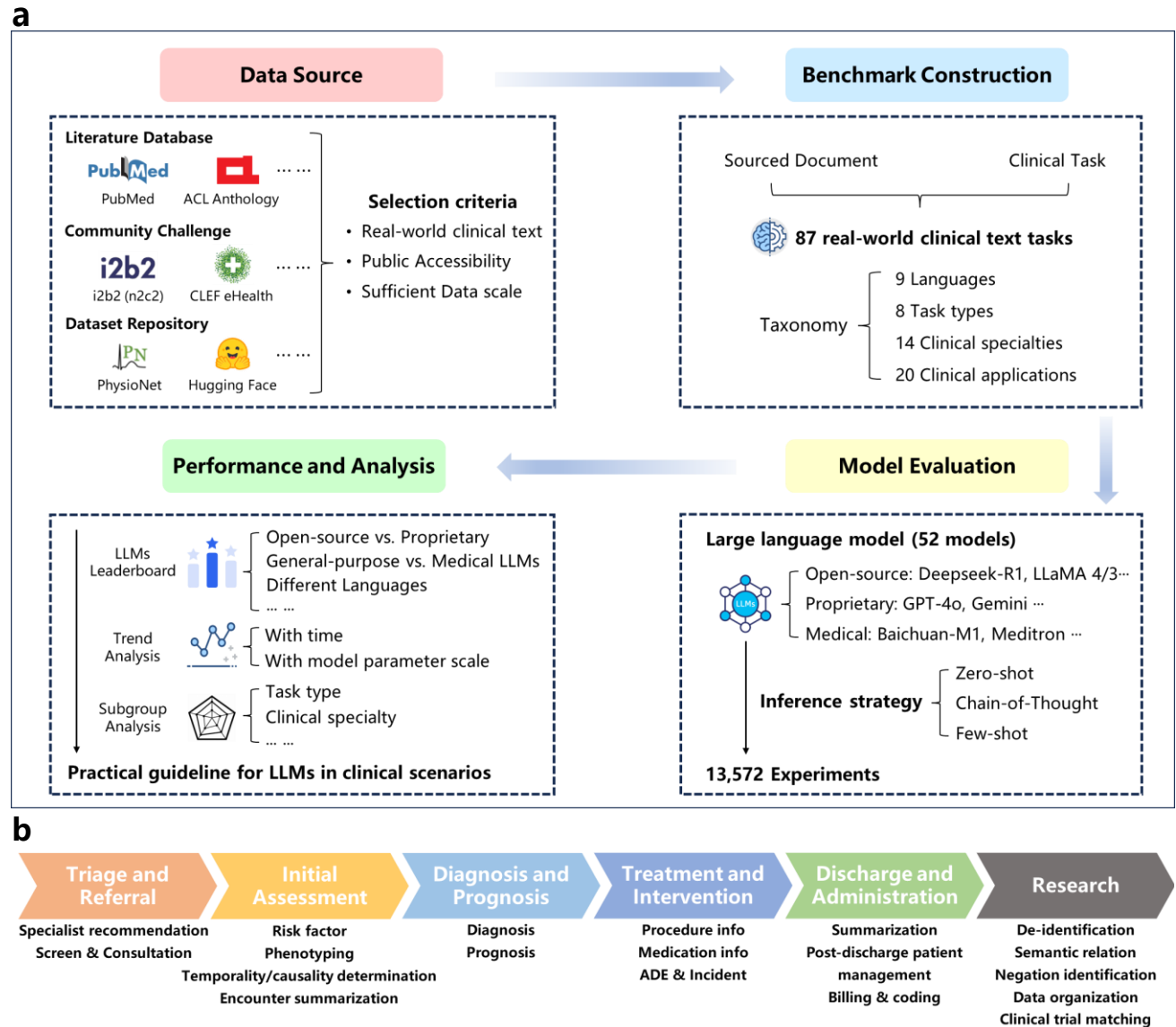


Figure 1. Overview of benchmarking large language model in clinical text understanding. (a) Workflow of benchmark construction, model evaluation, and performance analysis; (b) Clinical applications supported by the benchmark across different stages of patient care.

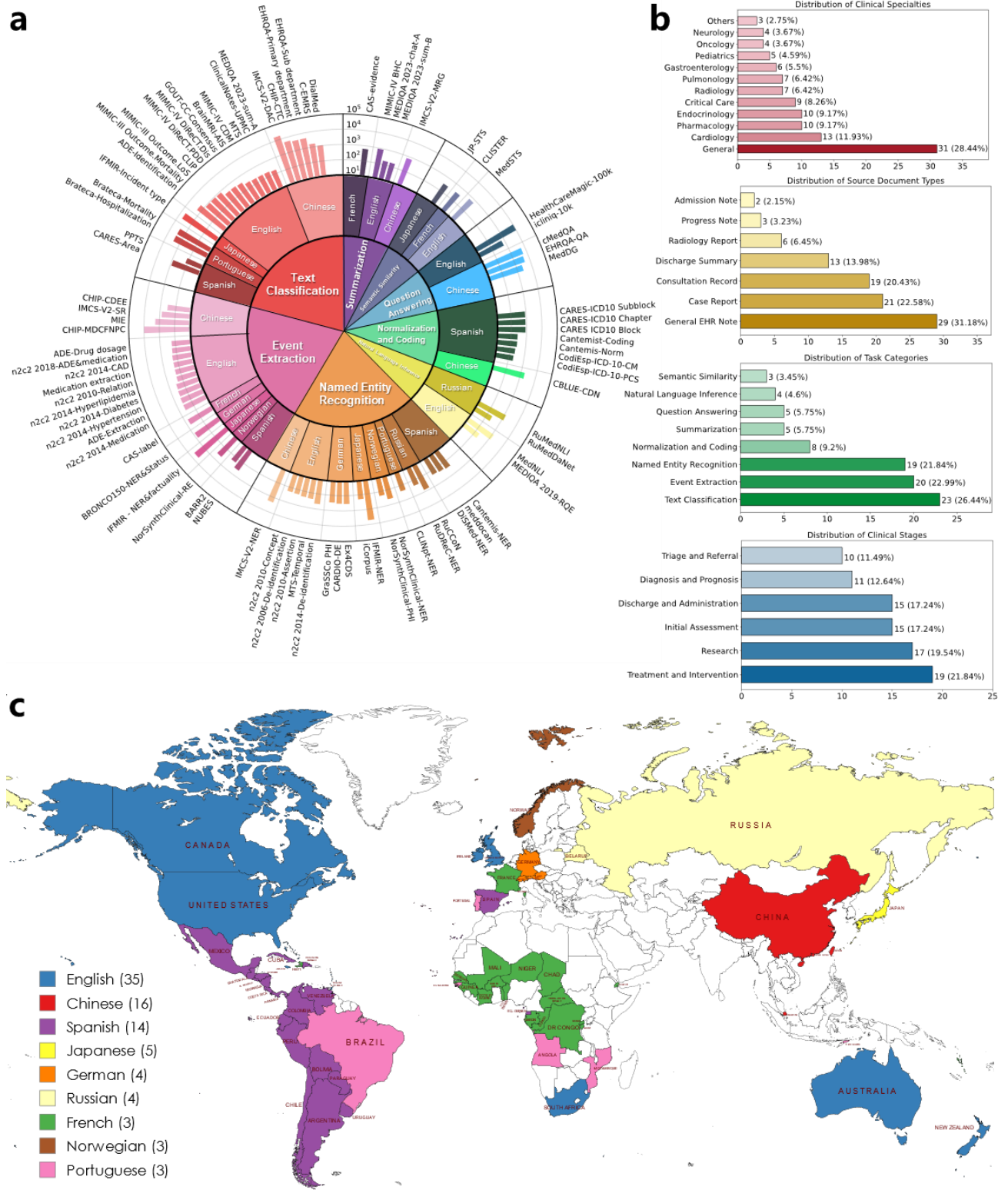


Figure 2. Overview of Benchmark Characteristics and Task Distribution. (a) Distribution of task types and associated languages, (b) Statistics on the distribution of clinical specialties, source document types, and task categories., and (c) Geographic distribution of countries where our benchmark covers the official languages.

Results

Benchmark Overview

The overall workflow of this study is illustrated in Figure 1. In total, our benchmark encompasses 87 tasks spanning nine languages and 1,418,042 samples, with 138,472 samples reserved for testing. Among them, 68 tasks (78.2%) are sourced from real-world EHR notes or clinical case reports, and 19 tasks (21.8%) are derived from real-world online patient-doctor consultation records. Figures 2a and 2b visualize the distribution of these tasks, covering 8 types, such as named entity recognition (e.g., phenotyping), classification (e.g., disease prediction), question answering, EHR summarization, and others. Figure 2c highlights this benchmark’s coverage of nine languages, distributed as follows: English (35 tasks, 40.2%), Chinese (16, 18.4%), Spanish (14, 16.1%), Japanese (5, 5.8%), German (4, 4.6%), Russian (4, 4.6%), French (3, 3.5%), Norwegian (3, 3.5%), and Portuguese (3, 3.5%). Detailed information about all tasks can be found in the Section Methods.

As shown in Figure 3a, we evaluated 52 state-of-the-art (SOTA) LLMs, covering proprietary, open-source models (1B to 671B parameters), and medically specialized models. Each LLM was assessed under three distinct inference strategies: Zero-shot, Chain-of-Thought (CoT), and Few-shot. Comprehensive descriptions of the models, alongside their technical specifications, are available in Supplementary Table S1. Additionally, we investigated potential data contamination of the benchmark and found that most tasks did not appear to have been included in the training corpora of these models. Detailed analyses are presented in Supplementary Figure S1. The constructed benchmark and leaderboard are publicly available and will be regularly updated.

Overall Performance

We assessed LLMs using an overall score, defined as the average of their primary-metric scores across all tasks, to reflect their comparative performance across the entire benchmark, with the score ranging from 0 to 100 (Supplementary Table S2 for details). Figure 3b demonstrates the zero-shot performance of LLMs, while Figure 4 highlights the leading models under each inference strategy. Under the zero-shot setting, the three best-performing LLMs were DeepSeek-R1⁴⁴ (44.2 [43.5, 45.0], 95% CI), GPT-4o⁴⁶ (44.2 [43.4, 45.0]), and Gemini-1.5-Pro⁴⁷ (43.8 [43.1, 44.6]). With CoT prompting, the overall

performance did not improve in general, with DeepSeek-R1 maintaining the top position (42.1 [41.3, 42.9]), followed by Gemini-2.0-Flash (42.0 [41.2, 42.8]) and GPT-4o (40.7 [39.9, 41.4]). In contrast, few-shot prompting led to substantial performance gains, with Gemini-1.5-Pro (55.5 [54.7, 56.3]) leading, followed by Gemini-2.0-Flash (53.3 [52.5, 54.2]) and GPT-4o (52.6 [51.8, 53.4]). Among the medical LLMs, the Baichuan-M1-14B-Instruct⁵⁰ emerged as the best-performing model, achieving the highest overall scores of 36.08 ([35.26, 36.89]), 34.4 ([33.6, 35.2]), and 48.3 ([47.4, 49.2]) for zero-shot, CoT, and few-shot, respectively. Overall, few-shot learning emerged as the most effective inference strategy for these clinical text tasks. Although LLM performance varied across different tasks, our leaderboard reveals the substantial ability gap of current LLMs for comprehensive clinical text understanding across diverse tasks, with the highest overall score on the whole BRIDGE only 55.5 under few-shot setting, indicating considerable space for further improvements.

LLMs Comparison

Figure 3b reveals a generally upward trajectory in overall performance under zero-shot setting across the evolving landscape of LLMs, illustrating both the rapid development and substantial potential of LLMs. While proprietary models, represented by GPT-4o and the Google Gemini series, maintain a performance lead, open-source LLMs have been rapidly advancing and narrowing the gap. Notably, the newly released DeepSeek-R1 (671B) outperforms all proprietary LLMs in both zero-shot and CoT settings. Other popular open-source models also achieved comparable performance, including Mistral-Large-Instruct (score of 42.28),⁵¹ Qwen2.5-72B-Instruct (41.62),⁵² Gemma-3-27B-it (39.90),⁵³ and Llama-3.3-70B-Instruct (39.86)⁵⁴ with the derived variant Athene-V2-Chat (72B) (41.69),⁵⁵ under zero-shot setting. Surprisingly, Llama-4-Scout-17B-16E-Instruct⁴⁵ showed a notable decline in performance compared to its predecessor, Llama-3.3-70B-Instruct, despite having a larger number of parameters. Despite the emergence of specialized medical LLMs, they did not outperform their general-purpose counterparts. For example, MeLLaMA-70B-chat (32.26) and Llama-3-70B-UltraMedical (33.40) performed worse than the related Llama-3.1-70B-Instruct (39.09), while some medical variants even lagged behind their foundation models (e.g., Llama-3.1-8B-UltraMedical [20.16] vs. Llama-3.1-8B-Instruct [28.98]).

Figure 3b illustrates the performance gains associated with increasing model size, with larger models generally outperforming smaller ones. As shown in Figure 3c, comparisons between DeepSeek-R1 and its variants with different model sizes demonstrate a consistent improvement in performance as the size of model parameters increases, aligning with the trend observed in the Llama, Qwen, and MeLLaMA model families. Together, these results highlight the effectiveness of scaling laws in enhancing LLM performance for clinical applications.² Models with around 70B parameters represent the most common category of open-source LLMs and typically achieve robust performance, led by Athene-V2-Chat (72B) (41.69), Qwen2.5-72B (41.62), and Llama-3.3-70B (39.86). Among LLMs with around 30B parameters, Gemma-3-27b-it (39.90) and DeepSeek-R1-Distill-Qwen-32B (39.75) stand out. Notably, smaller yet high-performing models such as Phi-4 (14B) (36.13) and Baichuan-M1-14B-Instruct (36.08) closely approach the performance of certain 70B models; the latter, being a recently developed medically specialized LLM, demonstrates the effectiveness of domain adaptation strategies.

Inference Strategy Performance

Figure 4a compares the performance of representative LLMs under three inference strategies: zero-shot, CoT, and five-shot. Compared to zero-shot, almost all models (51/52, 98.1%) achieved better performance using few-shot, with 38 models (73.1%) achieving gains exceeding 20%. This widespread performance boost highlights the effectiveness of few-shot prompting, even with a small number of randomly selected examples. Furthermore, few-shot enhancements benefited both top-tier and lower-ranked LLMs. Among the leading models under zero-shot, DeepSeek-R1 improved from 44.2 to 51.4 (+7.2, +16.3%), while Gemini-1.5-Pro rose from 43.8 to 55.5 (+11.7, +26.7%), indicating few-shot further augments even the strong LLMs. Moreover, models initially underperforming in zero-shot mode also exhibited significant improvements with few-shot prompting (see Supplementary Table S2), such as the smaller LLMs (e.g., Llama-3.2-1B-Instruct from 12.7 to 24.4, +92.1%). In particular, several medical LLMs benefited the most from the few-shot strategy. The Llama3-OpenBioLLM-8B showed the largest improvement, growing from 14.0 to 33.1 (+19.1, 136.4%), followed closely by Meditron-70B (15.7 to 32.1, +16.4, +104.5%). Other medical models, such as MMed-Llama-3, Llama-3.1-UltraMedical (8B and 70B), and Baichuan-M1-14B-Instruct, also exhibited improvements ranging

from 32.9% to 67.2%. In contrast, explicitly applying step-by-step reasoning through CoT did not yield the expected performance gains for most LLMs, with only two models showing slight improvement.

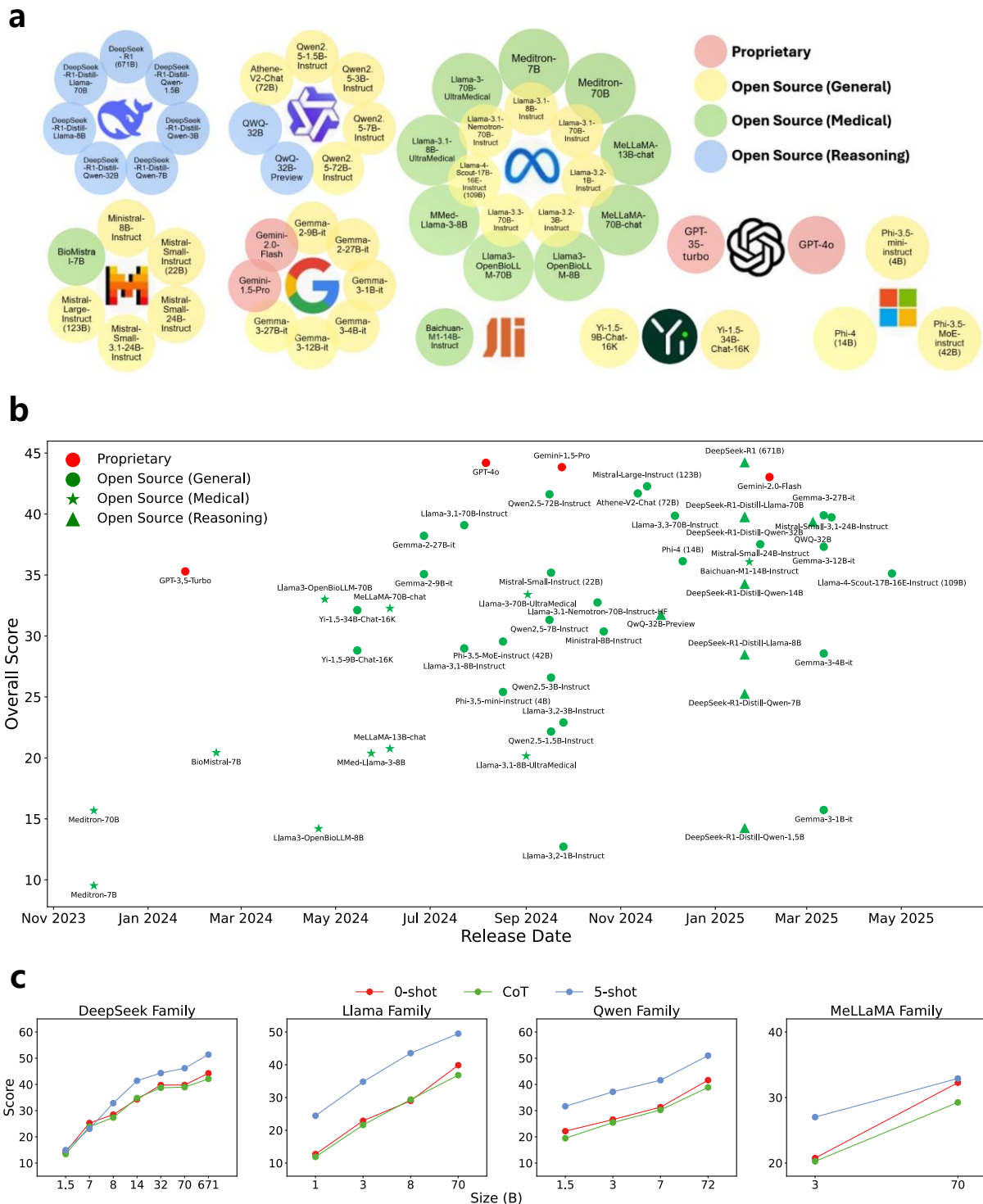
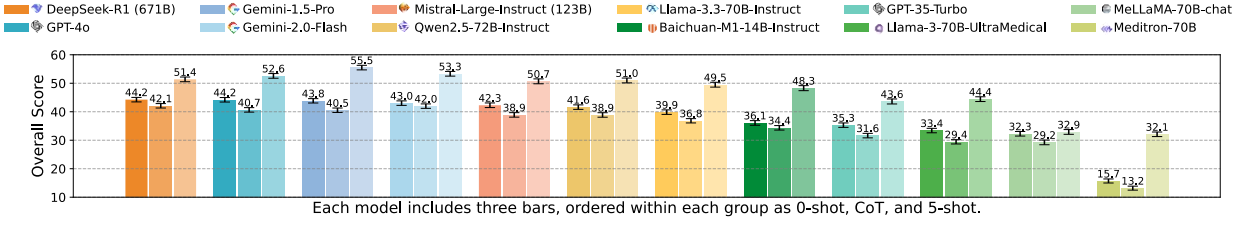


Figure 3. Overview of evaluated LLMs and their performance. (a) Categorization and information of evaluated LLMs. (b) Benchmark performance (zero-shot) of LLMs with their release dates. (c) Comparative performance analysis of LLMs of varying sizes within the same model family.

a



b

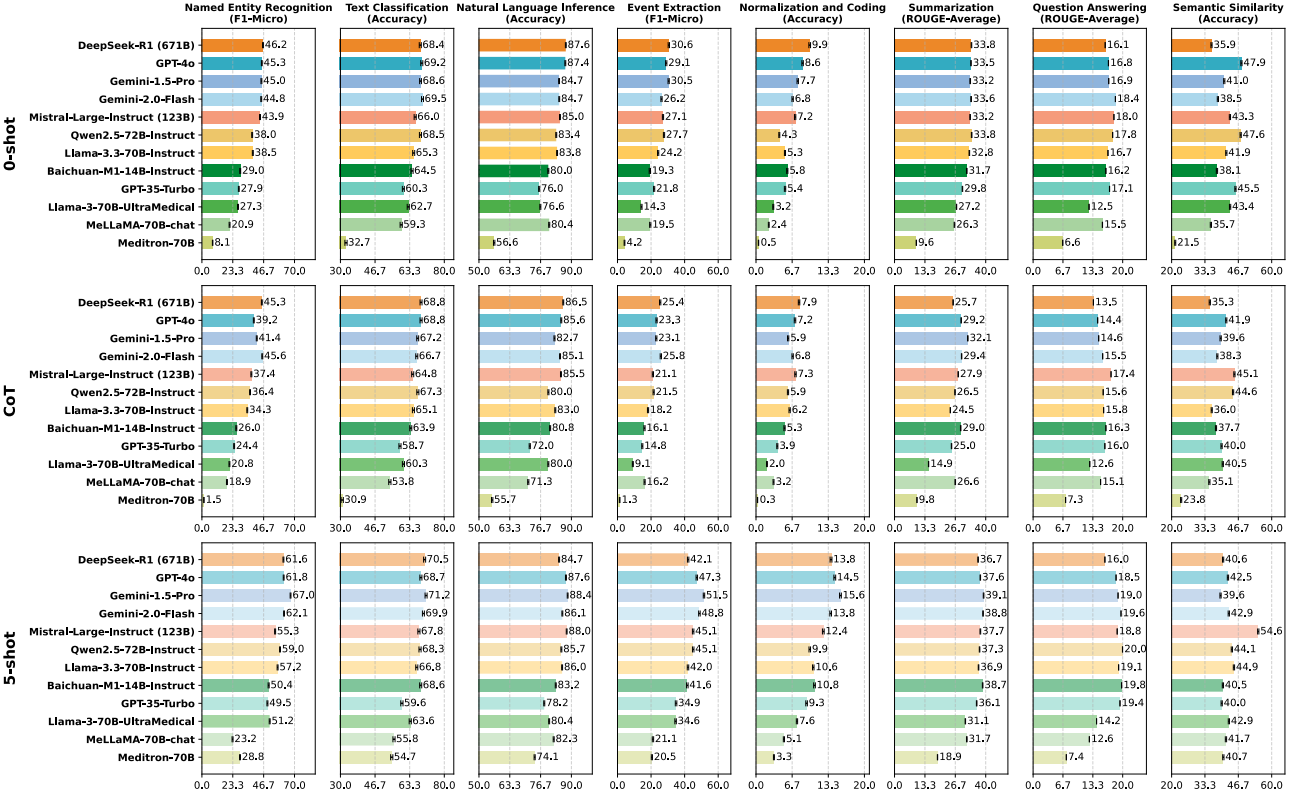


Figure 4. Comparative performance of 12 leading and commonly used LLMs under different inference strategies. (a) Overall score of LLMs evaluated across three inference strategies: zero-shot, CoT, and 5-shot prompting. For each model, three bars are displayed in the order of zero-shot, CoT, and 5-shot. (b) Performance of LLMs across different task categories under the three inference strategies. Error bars represent 95% CI.

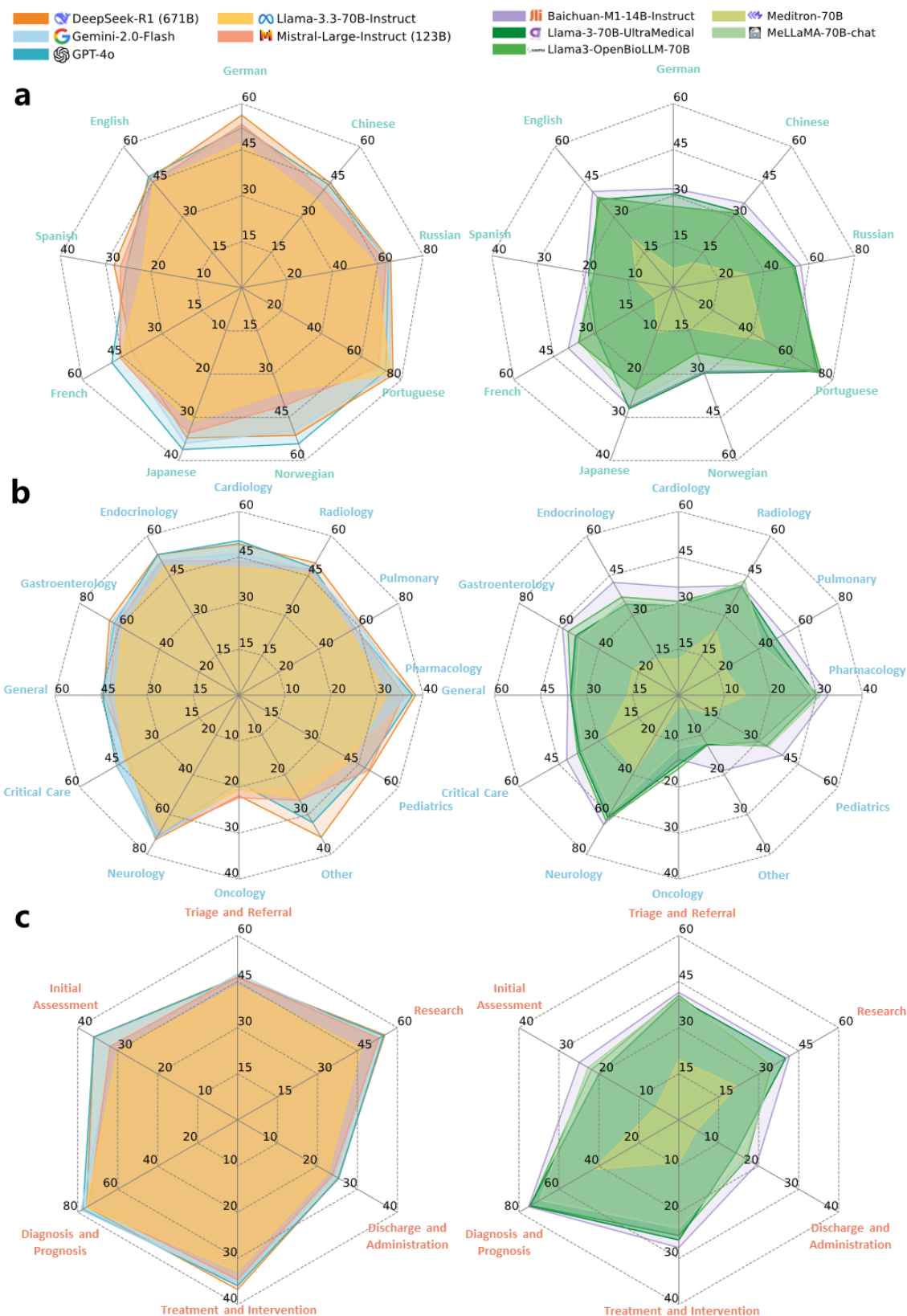


Figure 5. Zero-shot performance of 5 leading and commonly used LLMs in both the general and medical domains across different BRIDGE subgroups. (a) languages, (b) clinical specialties, and (c) clinical stages.

Performance analysis for different task types

This benchmark encompasses a broad range of clinical tasks, and Figure 4b provides a broad abilities assessment across different task types for both general and medical LLMs (Supplementary Table S3 for full details). In zero-shot setting, DeepSeek-R1 achieved the highest scores in four types: 46.2 at Named Entity Recognition (NER), 30.6 at Event Extraction, 87.6 at Natural Language Inference (NLI), and 9.9 at Normalization and Coding, while Gemini-2.0-Flash excelled in Text Classification (69.5) and Question-Answering (QA) (18.4). Athene-V2-Chat led in Semantic Similarity with a score of 48.1, and Gemma-2-27B-it achieved the best score (34.7) in Summarization. In contrast, most medically specialized LLMs encountered difficulties in adapting to multiple tasks, with average ranks consistently below 20 out of 52 LLMs. Figure 4b also highlights the performance variations of LLMs across different inference strategies. We observed that LLMs typically performed best on text classification and NLI, both of which offer well-defined, discrete label sets and thus present fewer ambiguities in the output. In contrast, information extraction tasks, such as NER and event extraction, benefited substantially from few-shot prompting, suggesting that more complex and context-dependent clinical tasks require examples to clarify detailed definitions and criteria. Meanwhile, normalization and coding tasks, which demand alignment with standardized medical coding systems (e.g., ICD-10), remained particularly challenging, as many LLMs lacked built-in mappings to these codes.⁵⁶ Although few-shot prompts brought modest improvements, performance in these coding tasks remained relatively low (around 15%). Text generation tasks, including QA and summarization, show an average performance of around 20%, indicating that LLMs face significant challenges in clinical text generation.

Performance analysis for different languages

Figure 5a demonstrates the performance of LLMs across different languages, with additional details available in Supplementary Table S4. The results reveal that many advanced LLMs exhibit robust cross-linguistic adaptability, consistently delivering fine performance for different languages. For instance, DeepSeek-R1 achieved first place in four languages: Chinese, Spanish, German, and Russian, while GPT-4o excelled in Japanese and Norwegian, Gemini-1.5-Pro led in English. Phi-3.5-MoE-instruct outperformed other models in Portuguese, and Qwen2.5-72B-Instruct took the top rank in

French. Notably, LLMs built on the Qwen base model, such as Athene-V2-Chat (72B), Qwen2.5-72B-Instruct, and DeepSeek-R1-Distill-Qwen-32B, achieved high scores across all languages of 46.3, 46.1, and 44.7. Especially, DeepSeek-R1-Distill-Qwen-32B outperformed several larger 70B variants (e.g., Llama-3.3-70B-Instruct and DeepSeek-R1-Distill-Llama-70B, scoring 44.1), this underscores the strong multilingual potential of well-optimized foundation models. Among specialized medical LLMs, Baichuan-M1-14B-Instruct exhibited the best versatility by achieving top scores in six languages (German, English, Spanish, French, Russian, and Chinese), surpassing all other 70B medical models. In contrast, English-centric medical LLMs (e.g., Meditron, MeLLaMA, and BioMistral) perform comparatively lower when applied to other languages. These results highlight the necessity for more diverse multilingual corpora and language-specific tuning to ensure effective global deployment of LLM-based solutions.

Performance analysis for different clinical specialties and clinical stages

We further examined model performance within various clinical specialties, reflecting the diverse specialties from which datasets originated or the specific clinical challenges they addressed (see Supplementary Table S5). As Figure 2b shows, this benchmark comprises 14 specialties; “General” denotes datasets that span more than five specialties or that are not explicitly indicated, while “Others” includes nephrology, dermatology, and psychology (one dataset each). Figure 5b highlights that DeepSeek-R1 delivered the highest scores in radiology (49.8), pulmonology (61.4), neurology (72.5), endocrinology (53.0), gastroenterology (64.9), and the “Others” category (35.7). In contrast, Gemini-1.5-Pro led in pharmacology, critical care (46.6), pediatrics (50.4), and oncology (22.5), whereas GPT-4o excelled in cardiology and “General” tasks. Despite domain-focused pretraining and supervised fine-tuning, the specialized medical LLMs did not show superiority over their general-purpose counterparts. Among these medical models, Baichuan-M1-14B-Instruct demonstrated the broadest versatility by leading in 10 clinical specialties. Llama3-OpenBioLLM-70B outperformed other medical LLMs in oncology-specific tasks. MeLLaMA-70B-chat performed best in radiology, likely aided by substantial training on radiology-focused datasets such as MIMIC-CXR.⁵⁷ Figure 5c further presents LLM performance across different clinical stages, and more details are provided in Supplementary Table S6. Gemini-1.5-Pro achieved the best results in Initial Assessment, Diagnosis

and Prognosis, and Treatment and Intervention, while Gemini-2.0-Flash led in Triage and Referral. GPT-4o performed best in Discharge and Administration, and DeepSeek-R1 outperformed others in Research-related tasks. Among medical LLMs, Baichuan-M1-14B-Instruct consistently delivered the highest performance across all stages. Overall, LLMs demonstrated stronger performance in the Diagnosis and Prognosis stage, likely due to the prevalence of well-structured classification tasks. In contrast, other stages often involve more complex tasks such as information extraction (e.g., phenotyping, temporal or causal relations) and text generation (e.g., summarization), which may pose greater challenges.

Discussion

This study represented the largest benchmark to date for LLMs evaluations on multilingual, real-world clinical text, encompassing 87 tasks in nine languages. We developed a systematic framework and leaderboard for categorizing tasks and defining corresponding evaluation methods, enabling a thorough assessment of 52 state-of-the-art LLMs with 13,572 experiments. Beyond providing a holistic view of current LLM capabilities, our analyses examined performance variations across different perspectives, including inference strategies, languages, task types, and clinical specialties. With comprehensive analyses and a continuously updated leaderboard, clinicians can employ this benchmark to determine the candidate LLMs that best fit specific clinical or research contexts and deployment environments, while AI developers in healthcare can leverage it as a robust reference for further model fine-tuning and system integration. For patients, the benchmark serves as a preliminary assessment of the reliability of LLM outputs, thereby promoting better transparency and confidence in AI-assisted healthcare services.

Unlike scientific literature or licensing exam questions, the information in our benchmark is drawn from actual patient care – the EHR and online doctor-patient interactions.^{58,59} Differing from the simplified and standardized multiple-choice evaluations,¹⁸ BRIDGE includes tasks specific to the administration and provision of health care, better reflecting the multifaceted capabilities of LLMs. Although certain LLMs achieved scores above 80 on standardized exams^{23,60}—for example, Deepseek-R1 reached a score of 92 on the USMLE dataset,⁶⁰ it attained an overall score of only 44.2 out of 100

on our benchmark. While this was the highest among the models evaluated, there remains substantial room for improvement. This stark discrepancy is sobering and highlights limitations of the current LLMs in clinical applications and the necessity of more clinically oriented evaluation before integrating LLMs into clinical practice.⁶¹ Furthermore, by supporting nine distinct languages, our benchmark facilitates more equitable and globally applicable advancements in medical AI, extending the promise of LLMs across diverse healthcare systems worldwide.⁶²

Given the rapid and transformative developments of LLMs,⁶³ especially fueled by open-source initiatives such as LLaMA, Qwen, and DeepSeek,^{52,54} our leaderboard of 52 cutting-edge LLMs provides valuable guidance on integrating LLMs into clinical settings. We document the remarkable progress of open-source models, exemplified by Deepseek-R1 surpassing the proprietary models.⁶⁴ Although these open-source foundation models also support the development of medical LLMs, current medical-specific variants generally underperform their general-purpose counterparts. This gap partially stems from the outdated base models these medical LLMs were built on, and the limited clinical relevance of their training data,⁶⁵ which is primarily drawn from medical textbooks or literature (e.g., PubMed) rather than EHR data.^{66,67} Additionally, most medical LLMs are fine-tuned on a limited set of medical tasks (e.g., Meditron was only trained and evaluated on multiple-choice questions).⁶⁸ The lack of task diversity may lead to overfitting and reduce the LLM's generalizability across tasks.⁶⁹ This limitation is evident in their weaker performance under zero-shot settings and the substantial gains from few-shot prompts, which effectively introduce task-specific information.⁷⁰ Despite the multilingual capabilities derived from extensive web-data pretraining,^{71,72} most medical LLMs are primarily optimized for the English context, leaving genuinely multilingual clinical foundation models relatively underexplored.⁷³ Meanwhile, open-source solutions facilitate on-site deployments within hospitals, enabling more localized data and model governance, thereby reducing potential privacy and security risks.⁷⁴ But our experiments also confirm that scaling laws² persist within clinical tasks – the larger models significantly and consistently outperform the small ones, highlighting a continuing need for efficient deployment strategies to enhance LLM usability, particularly for resource-limited regions.

Inference strategy plays a pivotal role in the practical deployment of LLMs, particularly since most clinical applications will likely not involve LLM fine-tuning but rather rely on task-specific

prompting.⁷⁵ Our findings indicate that few-shot prompting proves highly effective for clinical text tasks, significantly enhancing task-specific comprehension and contextualization with only five randomly selected examples. To enhance interpretability, CoT explicitly instructs LLMs to generate step-by-step reasoning before arriving at a final answer.⁴⁹ Contrary to observations in other domains,⁷⁶ CoT did not yield consistent performance gains in our benchmark and mostly impaired results.⁷⁷ This discrepancy appears to stem from the lack of sufficiently grounded medical knowledge to support accurate multi-step reasoning and the heightened risk of hallucinations in such a knowledge-intensive setting.^{78,79} Meanwhile, the newly developed reasoning LLMs (e.g., DeepSeek-R1 and QWQ), which employ reinforcement learning to strengthen reasoning capabilities, achieved superior results on our benchmark.^{44,80} These models demonstrated a promising direction for developing interpretable LLMs that more closely mirror human decision-making processes. Beyond improved reasoning, the integration of external domain knowledge, such as via retrieval-augmented generation,⁸¹ can further enhance LLM performance and reliability, leveraging their in-context learning abilities while mitigating the risk of misinformation in clinical applications.⁸¹

Because of the tremendous potential of LLMs in healthcare scenarios as well as the considerable risks to patient safety, robust benchmarking will be essential in order to adopt LLMs into clinical practice.^{82,83} Future benchmarks should encourage closer collaboration among clinicians, patients, and LLMs to better simulate real-world interactions, enhance overall clinical impact, and provide a validate human baseline.⁸⁴ Additionally, LLMs have been observed to exhibit overoptimism in their inferences,⁸⁵ which is critical to address in medical applications. Our benchmark provides a foundation for evaluating the trustworthiness of such predictions. Given the complexity of clinical practice, the scores in our benchmark do not fully equate to LLM performance on specific clinical applications, which require further rigorous assessment. However, BRIDGE provides timely and comprehensive comparisons across diverse tasks and models, serving as a valuable starting point for model selection and filtering, guiding decisions before committing to resource-intensive evaluations or further development of selected base models.

This study has certain limitations. First, given the large scale of our dataset, we do not provide a human baseline for comparison. When considering the safety of an LLM for any specific task, a human

baseline – and ideally a distribution of human baselines is essential. Second, given the breadth of this project, we could not validate the robustness of the reference labels provided by each data source. However, the focus of this work is on the relative performance between LLMs, which is only marginally affected by such shortcomings. Third, the overall score was calculated by averaging the primary metrics across different task types, which introduces inconsistency due to varying metrics. However, we provide fine-grained evaluation in the leaderboard across task types under the same metric to ensure fair comparisons. Finally, while we investigated 52 advanced models, several newly released LLMs (e.g., OpenAI o1, Gemini-Pro-2.5, and Med-PaLM2) were not evaluated due to the constraints of model access and resources within healthcare systems. This leaderboard is periodically updated to maintain its relevance and currency, offering a dynamic resource for tracking and advancing LLM performance in clinical text understanding.

In conclusion, this study established a comprehensive benchmark and systematically evaluated LLMs on real-world clinical text understanding. By centering on real-world EHR-based tasks and capturing the complexity of clinical text, our findings highlight the gap between current LLM capabilities and the demands of clinical practice, while also revealing substantial performance variability across models, languages, and clinical scenarios. These insights provide critical guidance for optimizing LLMs in healthcare and inform future efforts to align model development with the practical needs of clinical applications.

Methods

Clinical Text Dataset Collection

To comprehensively evaluate the LLMs performance on real-world clinical text data, we systematically identified and curated a diverse collection of clinical text datasets representative of authentic clinical scenarios.

This process was initially guided by our prior systematic review of clinical text datasets,⁴³ which conducted a global survey of publicly available resources. Building upon this foundation, we expanded our search scope and employed a standardized protocol to ensure that the included datasets fully satisfied the benchmarking criteria of clinical relevance, diversity, and suitability.

Specifically, we targeted three primary sources:

1. **Literature Databases:** Widely recognized biomedical literature databases, including PubMed and MEDLINE, and computational linguistic repositories, notably the ACL Anthology, a leading digital archive of Natural Language Processing (NLP)-focused journal articles and conference proceedings.
2. **Community Challenges:** Commonly used and actively maintained clinical NLP challenges and benchmarks, such as the National NLP Clinical Challenges⁸⁶ (n2c2, formerly i2b2) and CLEF eHealth.⁸⁷
3. **Dataset Repositories:** Biomedical dataset platforms (PhysioNet⁸⁸) and NLP-focused dataset hubs (Hugging Face⁸⁹), which store extensive collections of biomedical datasets and are frequently updated with new resources.

Detailed search strategies, including specific Medical Subject Headings (MeSH) terms and keywords, can be found in our prior review.⁴³

Criteria for Dataset Selection

Datasets identified through these sources underwent screening based on the following predefined inclusion and exclusion criteria:

1. **Real-World Clinical Text Data:** Eligible datasets were required to consist of authentic clinical texts derived directly from real-world medical settings, such as electronic health

records (EHRs), clinical case reports, or healthcare consultations. Non-clinical sources (e.g., textbooks, social media) or datasets relying primarily on non-textual (e.g., genomic or protein sequences) or multimodal inputs were excluded.

2. **Public Accessibility and Availability:** Datasets included in this investigation are publicly available or accessible through standardized request procedures to ensure reproducibility and transparency.
3. **Sufficient Data Scale:** Only datasets containing at least 200 samples were included to ensure reliable evaluations and robust statistical analyses.

Finally, the curated datasets provided a diverse corpus representative of authentic clinical scenarios.

Benchmark Construction

Based on the included datasets, we constructed a set of clinical text tasks tailored for assessing LLMs. These tasks simulate diverse clinical scenarios characterized by complexity, contextual variability, and multi-source information requirements. Unlike traditional NLP methods, which typically rely on supervised training with task-specific model architectures, LLMs perform tasks by interpreting textual prompts without dedicated training. Therefore, precise task design and standardization are critical for fair and objective evaluations. Detailed information about all tasks can be found in Supplementary Section 4 and Section 5, and the metadata of tasks are in the Supplementary Table S8.

To ensure task suitability and consistency, we transformed and standardized datasets through the following structured process:

Task Definition and Categorization

Task objectives and evaluation criteria were primarily derived from original dataset descriptions or primary publications (hereafter collectively referred to as dataset source). Tasks were categorized into different types:

1. Text classification: Determine or predict categorical labels (e.g., diagnosis, risk stratification) based on the provided clinical texts.
2. Semantic similarity: Assessing the similarity of two sentences or clinical notes.
3. Natural Language Inference (NLI): Evaluating the logical relationships (e.g., entailment, contradiction, neutrality) between paired texts.

4. Normalization and coding: Map the whole clinical note or the extracted entities to standardized clinical code systems (e.g., ICD, SNOMED)
5. Named Entity Recognition (NER): Identify the medical entities and label them with appropriate types (e.g., symptom, disease, examination).
6. Event extraction: Identify the medical entities and capture additional attributes or relations beyond simple entity types (e.g., temporal status, severity).
7. Question-Answering (QA): Generating accurate responses to healthcare inquiries.
8. Summarization: Condense clinical notes into concise summaries by retaining essential information, with extraction or generation methods.

Input Text Preparation and Standardization

Relevant textual information for each task was systematically extracted from the original datasets and integrated using standardized templates. For instance, the required text fields were distilled from the whole EHR database and then condensed into structured inputs with templates (e.g., "Chief Complaint: ..., Examination: ..."). Additionally, for tasks introducing structured metadata (e.g., demographic information and examination results), we transformed these structured features into explicitly labeled textual forms, integrating them seamlessly with clinical notes. The template for input and output can be found in Supplementary Section 4.1.

Output Standardization and Formatting

Given the heterogeneity of original task outputs, including classification logits, BIO-style entity tags, or structured annotations, all outputs were instructed to be standardized into clear, structured textual responses for automatic result processing and analysis on evaluation. Specifically:

1. Tasks for Text classification, Semantic similarity, NLI, and Normalization and Coding (document level): Outputs were standardized into explicit textual labels indicating predicted categories.
2. Tasks for NER, Event extraction, and Normalization and Coding (entity level): Outputs were formatted into structured textual annotations clearly indicating subject spans and other required attributes (e.g., "Entity: ..., Type: ..., Status: ...").

3. Tasks for QA and Summarization: Outputs were directly formulated as concise, structured free-text outputs.

Outputs that failed to follow the required format were regarded as invalid responses during the evaluation phase. This uniform output format enables automated extraction and quantitative evaluation of LLM-generated results, ensuring efficient and objective performance assessment.

Task Instruction (Prompt) Definition

Due to the complexity of clinical text tasks, which often rely on professional definitions and domain-specific terminology, we prioritized the use of task instructions from authoritative dataset sources, including original dataset papers, annotation guidelines, and supplementary data descriptions. To minimize variability and potential biases introduced by differing prompts, we adopted straightforward and concise instruction templates for all tasks.

Given the multilingual nature of our benchmark, we preferentially retained the task description aligned with their original language if available, preserving context-specific semantics critical for accurate interpretation. Meanwhile, the instructions for other tasks and the base templates (both input and output may involve) were uniformly provided in English, as the existing LLMs all support English and yield fine performance in English.

Dataset Splitting

Dataset partitions followed the official splits defined by dataset sources whenever available, facilitating direct comparability with prior studies. For datasets without predefined splits, we applied the following selection strategy: for datasets with over 2000 samples, 10% were randomly selected as the test set; for datasets with 1000 to 2000 samples, 20% were selected; and for datasets with fewer than 1000 samples, all samples were used for testing except for 20 cases reserved as a pool for selecting few-shot examples. For each dataset, five samples outside the test set were randomly selected as few-shot examples. All the benchmark experiments were conducted on the split test partitions.

Task Taxonomy and Characteristics

To systematically investigate the abilities of LLMs across different clinical scenarios, we extracted key features for each task and mapped them into standardized taxonomies. These include

Language, Sourced Clinical Document, Clinical Specialty, and Clinical Stage and Application, which can refer to Supplementary Section 4.2 Task Taxonomy and Characteristics.

Model Implementation

We included a diverse range of state-of-the-art LLMs, covering both proprietary and open-source LLMs. Detailed information for all models can be found in Supplementary Table S1. All experiments, including data selection and model inference, used a fixed random seed (42) across all tasks and models to ensure reproducibility. For decoding configuration, the greedy decoding strategy was employed for all models, with specific parameters (temperature = 0, top_p = None, top_k = None), to eliminate randomness and produce deterministic outputs.

The model inference was conducted using the following computational setups:

1. Open-source models: all open-sourced models (except DeepSeek-R1[671B]) were deployed locally on Mass General Brigham (MGB) institutional server with 8 NVIDIA H100 GPUs. The inference process was accelerated using the vLLM framework⁹⁰ to optimize efficiency. DeepSeek-R1 (671B) is deployed and used on Microsoft Azure via infrastructure managed by Stanford University.
2. Proprietary models: Due to privacy and security considerations, proprietary models were evaluated via institutional cloud infrastructure within the compliance of HIPAA:
 - a) OpenAI models (GPT-3.5-Turbo-0125 and GPT-4o-0806): Deployed on Microsoft Azure server at Mass General Brigham.
 - b) Google models (Gemini-2.0-Flash-001 and Gemini-1.5-Pro-002): Deployed on Google Cloud Platform at Mayo Clinic.

All records of inference requests and responses were securely stored in accordance with MGB data governance policies.

Inference Strategy

In this study, we systematically evaluated three distinct inference strategies:

1. Zero-shot: Only the task instructions and input data were provided. The LLM was prompted to directly produce the target outputs without any support.

2. Chain-of-Thought (CoT): Task instructions explicitly directed the LLM to generate a step-by-step explanation of its reasoning process before providing the final output, which can significantly improve the model's interpretability.
3. Few-shot: Five reserved independent samples serve as examples, which leverage the LLM's capability of in-context learning to guide the model to conduct tasks. For models supporting conversational interactions, examples were presented sequentially to simulate realistic user-system dialogues; otherwise, input-output pairs were directly appended to the instruction.

Details about the prompt for different inference strategies can be found in Supplementary Section 4.1.

Evaluation Framework

To ensure consistent, objective, and automated evaluations, we established standardized evaluation schemes covering reference standard, result extraction, task-specific metrics, and statistical analysis.

Reference Standard

For each task in our benchmark, the reference standard is sourced from the labels released with the original source datasets. These labels were generated through different mechanisms, including expert manual annotation and derivation from structured EHR systems, and undergone the quality-control procedures defined by dataset creators. To maintain consistency with prior work and preserve the integrity of each dataset, we adopted these original labels without additional modification.

Result Extraction

We develop an automated script for each task separately to extract results from the standardized LLM outputs described previously. For outputs failing to meet the required formatting standards, we regarded them as invalid responses. We calculated the valid rate for each experiment setting and presented the results in Supplementary Table S7. For tasks under the types of text classification, semantic similarity, NLI, and normalization/coding (with explicitly defined labels), invalid model outputs were replaced with randomly assigned labels from the valid label set. For the remaining task types, invalid outputs were retained as empty responses.

Evaluation Metrics

Representative metrics were carefully selected for each task category, with a designated primary metric facilitating overall benchmarking comparisons:

1. Text classification, Semantic similarity, NLI, and Normalization and Coding (document level): We evaluate these tasks with Accuracy (primary metric), micro F1-score, and macro F1-score. Accuracy directly reflects overall classification performance by measuring the proportion of correct predictions. The F1 scores provide complementary insights by considering precision and recall across classes. The micro F1-score emphasizes performance in common classes, while the macro F1-score equally weights all classes, highlighting performance in less frequent categories.
2. NER, Event extraction, and Normalization and Coding (entity level): We evaluate these tasks with subject-level F1-score and event-level F1-score (both calculated by micro-scoring). The subject-level F1-score only evaluates the model's ability to identify the correct subjects without considering their attributes, providing preliminary performance insights. Event-level F1-score, the primary metric, comprehensively evaluates model accuracy by measuring extraction precision and recall across entities and their attributes.
3. QA and Summarization: We evaluate these tasks with BLEU-4, ROUGE-average (primary metric), and BERTScore. ROUGE-average⁹¹ is the average score of ROUGE-1, ROUGE-2, and ROUGE-L, thus capturing the recall of unigrams, bigrams, and longest common subsequences between the candidate and reference texts. BLEU-4⁹² combines the precision of 1-, 2-, 3-, and 4-gram matches between generated outputs and references. BERTScore⁹³ evaluates semantic alignment between generated and reference texts by leveraging contextual embeddings from BERT. All these metrics generally show a consistent trend across tasks, while the ROUGE-average exhibits more distinctions among models in our experiments. Therefore, we adopt ROUGE-average as the primary metric.

These metrics were computed using standard libraries to ensure reproducibility: *Scikit-learn*⁹⁴ for classification and extraction tasks, *nlTK*⁹⁵ for BLEU-4, *rouge_scorer*⁹⁶ for ROUGE-average, and *bert_score*⁹⁷ for BERTScore.

Performance Calculation

To enable quantitative comparisons, we compute an overall score for each LLM by averaging its primary-metric values across all tasks. This aggregate measure reflects a model’s relative performance

on the benchmark as a whole. For subgroup analyses, such as task type, language, clinical specialty and clinical stages, the same averaging procedure is applied to the subset of tasks that meet the specified criteria, yielding a focused performance estimate within that domain.

Statistical Analysis

To robustly estimate model performance and assess variability, we performed bootstrapping (1,000 resamples) with replacement to calculate mean scores and corresponding 95% confidence intervals (CIs). The resulting CIs provided statistically rigorous estimates reflecting the uncertainty and reliability of model performance.

Data Contamination Analysis

The advanced LLMs typically undergo extensive training on vast data, raising the possibility of unintended data exposure of benchmark. To assess this potential data contamination, we employed a text completion-based approach to detect possible leakage of benchmark data into the evaluated models' training corpora.⁹⁸ Specifically, we tokenized each test sample using model-specific tokenizers, truncating sequences at predetermined positions (tokens 10, 15, 20, 25, and 30) and prompting the LLM to predict the subsequent five tokens. The accuracy of predicted tokens compared to actual tokens (5-gram accuracy) was measured. A test sample was classified as potentially leaked if predictions exactly matched actual tokens at three or more truncation positions. To balance sufficient contextual information while avoiding overly simplistic long-context completions, truncation points began from token position 10, incrementing by intervals of 5 tokens.

Acknowledgments

We thank Xiaocong Liu, Wanxin Li, Qingcheng Zeng, Zichang Su, and Xiaoyue Wang for the initial collection and process of datasets. This study was partially funded by PCORI ME-2022C1-25646, Goldberg Scholarship and Brigham Research Institute. L.A.C. is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, the National Science Foundation through ITEST #2148451, and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00403047). S.S. is funded in part by FDA research contracts (HHSF223201710186C and HHSF223201710146C), the FDA Sentinel Innovation Center (75F40119F19002), the NIH (NHLBI R01-HL141505, NIAMS R01-AR080194), the Burroughs Wellcome Fund, and PCORI. J.H.C. was supported by NIH/National Institute of Allergy and Infectious Diseases (1R01AI17812101), NIH-NCATS-Clinical & Translational Science Award (UM1TR004921), NIH/National Institute on Drug Abuse Clinical Trials Network (UG1DA015815 - CTN-0136), Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (IIP) [R12] [JHC], NIH/Center for Undiagnosed Diseases at Stanford (U01 NS134358), Stanford Institute for Human-Centered Artificial Intelligence (HAI), and Gordon and Betty Moore Foundation (Grant #12409).

Author contribution

J.Y. designed this study. J.W. and B.G. constructed the benchmark and conducted the experiments. J.W., B.G., and J.Y. analyzed the results and drafted the initial manuscript. R.Z. and K.X. contributed to the benchmark construction, J.W., B.G., and R.Z. contributed to the data contamination experiments, and K.X. and J.W. contributed to building the leaderboard. D.S., V.C., S.R-B, and Y.J. contributed the inferences of Gemini-series and Deepseek-R1. R.W., R.J.D., E.A., L.A.C., A.R., S.S., J.H.C., S.R-B., K.J.L., and J.Y. contributed to manuscript refinement. J.Y. and K.J.L. supervised this study. All authors revised, read, and approved the manuscript.

Competing Interest

K.J.L. has received research grants from Takeda, AbbVie, and UCB for projects unrelated to this study. S.S. is participating in investigator-initiated grants to the Brigham and Women’s Hospital from Boehringer Ingelheim, Takeda, and UCB unrelated to the topic of this study. He is an advisor to and owns equity in Aetion Inc., a software manufacturer. S.S. is an advisor to Temedica GmbH, a patient-oriented data generation company and his interests were declared, reviewed, and approved by the Brigham and Women’s Hospital in accordance with their institutional compliance policies. J.H.C. reports cofounding Reaction Explorer, that develops and licenses organic chemistry education software, and receive medical expert witness fees from Sutton Pierce, Younker Hyde MacFarlane, Sykes McAllister, Elite Expert, consulting fees from ISHI Health, and honoraria or travel expenses for invited presentations by insitro, General Reinsurance Corporation, Cozeva, and other industry conferences, academic institutions, and health systems.

Data availability

All fully open-access datasets in BRIDGE are shared at <https://huggingface.co/datasets/YLab-Open/BRIDGE-Open> under their respective data use agreements. All dataset sources and corresponding links for additional data access are listed in Supplementary Section 4. The leaderboard is available at <https://huggingface.co/spaces/YLab-Open/BRIDGE-Medical-Leaderboard>, where we also provide contact information for submitting new models for evaluation without requiring direct data access. Additional metadata supporting this study are available from the corresponding author upon reasonable request.

Code availability

The corresponding benchmark evaluation code can be found at <https://github.com/YLab-Open/BRIDGE>.

References

1. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at <http://arxiv.org/abs/2001.08361> (2020).
3. Zhang, S. *et al.* Instruction tuning for large language models: A survey. (2024).
4. Min, B. *et al.* Recent advances in natural language processing via large pre-trained language models: A survey. *Acm Comput. Surv.* **56**, (2023).
5. Qin, C. *et al.* Is ChatGPT a general-purpose natural language processing task solver? in *Proceedings of the 2023 conference on empirical methods in natural language processing* 1339–1384 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.emnlp-main.85.
6. Huang, J. & Chang, K. C.-C. Towards reasoning in large language models: a survey. in *Findings of the association for computational linguistics: ACL 2023* 1049–1065 (Association for Computational Linguistics, Toronto, Canada, 2023). doi:10.18653/v1/2023.findings-acl.67.
7. Tang, L. *et al.* Evaluating large language models on medical evidence summarization. *Npj Digit. Med.* **6**, 1–8 (2023).
8. Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
9. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
10. Goh, E. *et al.* GPT-4 assistance for improvement of physician performance on patient care tasks: A randomized controlled trial. *Nat. Med.* 1–6 (2025) doi:10.1038/s41591-024-03456-y.
11. Jeblick, K. *et al.* ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur. Radiol.* (2023) doi:10.1007/s00330-023-10213-1.
12. Goodman, R. S. *et al.* Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw. Open* **6**, e2336483 (2023).
13. Sheng, B. *et al.* Large language models for diabetes care: Potentials and prospects. *Sci. Bull.* **69**, 583–588 (2024).
14. Zaretsky, J. *et al.* Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Netw. Open* **7**, e240357 (2024).
15. Wang, X. *et al.* ChatGPT: promise and challenges for deployment in low- and middle-income countries. *Lancet Reg. Health – West. Pac.* **41**, (2023).
16. Ferryman, K., Mackintosh, M. & Ghassemi, M. Considering Biased Data as Informative Artifacts in AI-Assisted Health Care. *N. Engl. J. Med.* **389**, 833–838 (2023).
17. McCoy, L. G., Manrai, A. K. & Rodman, A. Large language models and the degradation of the medical record. *N. Engl. J. Med.* **391**, 1561–1564 (2024).
18. Raji, I. D., Daneshjou, R. & Alsentzer, E. It’s time to bench the medical exam benchmark. *NEJM AI* **2**, AIe2401235 (2025).
19. Jin, D. *et al.* What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, (2021).

20. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. 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)* 2567–2577 (Association for Computational Linguistics, Hong Kong, China, 2019). doi:10.18653/v1/D19-1259.
21. Chen, Q. *et al.* Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat. Commun.* **16**, 3280 (2025).
22. OpenAI. GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2023).
23. Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nat. Med.* 1–8 (2025).
24. Mahmood, F. A benchmarking crisis in biomedical machine learning. *Nat. Med.* 1–1 (2025) doi:10.1038/s41591-025-03637-3.
25. Harris, E. Large Language Models Answer Medical Questions Accurately, but Can't Match Clinicians' Knowledge. *JAMA* **330**, 792–794 (2023).
26. Peng, C. *et al.* A study of generative large language model for medical research and healthcare. *Npj Digit. Med.* **6**, 1–10 (2023).
27. Kreimeyer, K. *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **73**, 14–29 (2017).
28. Bedi, S. *et al.* Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* (2024) doi:10.1001/jama.2024.21700.
29. Perlis, R. H. & Fihn, S. D. Evaluating the Application of Large Language Models in Clinical Research Contexts. *JAMA Netw. Open* **6**, e2335924 (2023).
30. Goh, E. *et al.* Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Netw. Open* **7**, e2440969 (2024).
31. Nayak, A. *et al.* Comparison of History of Present Illness Summaries Generated by a Chatbot and Senior Internal Medicine Residents. *JAMA Intern. Med.* **183**, 1026–1027 (2023).
32. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).
33. Liu, X. *et al.* Uncovering Language Disparity of ChatGPT on Retinal Vascular Disease Classification: Cross-Sectional Study. *J. Med. Internet Res.* **26**, e51926 (2024).
34. Kim, J., Cai, Z. R., Chen, M. L., Simard, J. F. & Linos, E. Assessing Biases in Medical Decisions via Clinician and AI Chatbot Responses to Patient Vignettes. *JAMA Netw. Open* **6**, e2338050 (2023).
35. Zeng, Q. *et al.* GreenPLM: Cross-Lingual Transfer of Monolingual Pre-Trained Language Models at Almost No Cost. in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* 6290–6298 (International Joint Conferences on Artificial Intelligence Organization, Macau, SAR China, 2023). doi:10.24963/ijcai.2023/698.
36. Chang, Y. *et al.* A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* (2024) doi:10.1145/3641289.
37. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
38. Xie, Q. *et al.* Medical foundation large language models for comprehensive text analysis and beyond. *Npj Digit. Med.* **8**, 1–10 (2025).

39. Liu, X. *et al.* A generalist medical language model for disease diagnosis assistance. *Nat. Med.* 1–11 (2025) doi:10.1038/s41591-024-03416-6.
40. Chiang, W.-L. *et al.* Chatbot arena: An open platform for evaluating LLMs by human preference. in *Proceedings of the 41st international conference on machine learning* (JMLR.org, Vienna, Austria, 2024).
41. Minssen, T., Vayena, E. & Cohen, I. G. The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. *JAMA* **330**, 315–316 (2023).
42. Tang, Y.-D., Dong, E.-D. & Gao, W. LLMs in medicine: The need for advanced evaluation systems for disruptive technologies. *The Innovation* **5**, (2024).
43. Wu, J. *et al.* Clinical Text Datasets for Medical Artificial Intelligence and Large Language Models — A Systematic Review. *NEJM AI* (2024) doi:10.1056/AIra2400012.
44. DeepSeek-AI *et al.* DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. Preprint at <https://doi.org/10.48550/arXiv.2501.12948> (2025).
45. The llama 4 herd: The beginning of a new era of natively multimodal AI innovation. *Meta AI* <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
46. OpenAI *et al.* GPT-4o system card. Preprint at <https://doi.org/10.48550/arXiv.2410.21276> (2024).
47. Team, G. *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint at <https://doi.org/10.48550/arXiv.2403.05530> (2024).
48. Brown, T. *et al.* Language models are few-shot learners. in *Advances in neural information processing systems* (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H.) vol. 33 1877–1901 (Curran Associates, Inc., 2020).
49. Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. in *Proceedings of the 36th International Conference on Neural Information Processing Systems* 24824–24837 (Curran Associates Inc., Red Hook, NY, USA, 2022).
50. Wang, B. *et al.* Baichuan-M1: Pushing the medical capability of large language models. Preprint at <https://doi.org/10.48550/arXiv.2502.12671> (2025).
51. Jiang, A. Q. *et al.* Mistral 7B. Preprint at <https://doi.org/10.48550/arXiv.2310.06825> (2023).
52. Qwen *et al.* Qwen2.5 technical report. Preprint at <https://doi.org/10.48550/arXiv.2412.15115> (2025).
53. Team, G. *et al.* Gemma 3 technical report. Preprint at <https://doi.org/10.48550/arXiv.2503.19786> (2025).
54. Grattafiori, A. *et al.* The llama 3 herd of models. Preprint at <https://doi.org/10.48550/arXiv.2407.21783> (2024).
55. Nexusflow/athene-V2-chat · hugging face. <https://huggingface.co/Nexusflow/Athene-V2-Chat> (2024).
56. Soroush, A. *et al.* Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI* **1**, AIdbp2300040 (2024).
57. Johnson, A. E. W. *et al.* MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
58. Percha, B. Modern Clinical Text Mining: A Guide and Review. *Annu. Rev. Biomed. Data Sci.* **4**, 165–187 (2021).

59. Huang, K. *et al.* A foundation model for clinician-centered drug repurposing. *Nat. Med.* 1–13 (2024) doi:10.1038/s41591-024-03233-x.
60. Tordjman, M. *et al.* Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning. *Nat. Med.* 1–1 (2025) doi:10.1038/s41591-025-03726-3.
61. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *Npj Digit. Med.* **7**, 1–7 (2024).
62. Pfohl, S. R. *et al.* A toolbox for surfacing health equity harms and biases in large language models. *Nat. Med.* 1–11 (2024) doi:10.1038/s41591-024-03258-2.
63. DeepSeek-AI *et al.* DeepSeek LLM: Scaling open-source language models with longtermism. Preprint at <https://doi.org/10.48550/arXiv.2401.02954> (2024).
64. Buckley, T. A., Crowe, B., Abdulnour, R.-E. E., Rodman, A. & Manrai, A. K. Comparison of frontier open-source and proprietary large language models for complex diagnoses. *JAMA Health Forum* **6**, e250040 (2025).
65. Yang, X. *et al.* A large language model for electronic health records. *Npj Digit. Med.* **5**, 1–9 (2022).
66. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
67. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *Npj Digit. Med.* **6**, 1–10 (2023).
68. Dorfner, F. J. *et al.* Biomedical large languages models seem not to be superior to generalist models on unseen medical data. Preprint at <https://doi.org/10.48550/arXiv.2408.13833> (2024).
69. Meshkin, H. *et al.* Harnessing large language models’ zero-shot and few-shot learning capabilities for regulatory research. *Brief. Bioinform.* **25**, bbae354 (2024).
70. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing 1998–2022* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
71. Qiu, P. *et al.* Towards building multilingual language model for medicine. *Nat. Commun.* **15**, 8384 (2024).
72. Qin, L. *et al.* A survey of multilingual large language models. *Patterns* **6**, 101118 (2025).
73. Liu, F. *et al.* A multimodal multidomain multilingual medical foundation model for zero shot clinical diagnosis. *Npj Digit. Med.* **8**, 1–12 (2025).
74. Das, B. C., Amini, M. H. & Wu, Y. Security and privacy challenges of large language models: A survey. *ACM Comput Surv* **57**, 152:1-152:39 (2025).
75. Lehman, E. *et al.* Do we still need clinical language models? in *Proceedings of the Conference on Health, Inference, and Learning* (2023).
76. Wang, X. *et al.* Self-consistency improves chain of thought reasoning in language models. Preprint at <https://doi.org/10.48550/arXiv.2203.11171> (2023).
77. Liévin, V., Hother, C. E., Motzfeldt, A. G. & Winther, O. Can large language models reason about medical questions? *Patterns* **5**, (2024).
78. Wu, J. *et al.* Large language models leverage external knowledge to extend clinical insight beyond language boundaries. *J. Am. Med. Inform. Assoc.* ocae079 (2024) doi:10.1093/jamia/ocae079.

79. Wu, J., Wu, X. & Yang, J. Guiding clinical reasoning with large language models via knowledge seeds. in vol. 8 7491–7499 (2024).
80. Team, Q. QwQ-32B: Embracing the power of reinforcement learning. *Qwen* <https://qwenlm.github.io/blog/qwq-32b/> (2025).
81. Zakka, C. *et al.* Almanac — Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* **1**, AIoa2300068 (2024).
82. Shah, N. H., Entwistle, D. & Pfeffer, M. A. Creation and Adoption of Large Language Models in Medicine. *JAMA* (2023) doi:10.1001/jama.2023.14217.
83. Karargyris, A. *et al.* Federated benchmarking of medical artificial intelligence with MedPerf. *Nat. Mach. Intell.* **5**, 799–810 (2023).
84. Rodman, A., Zwaan, L., Olson, A. & Manrai, A. K. When it comes to benchmarks, humans are the only way. *NEJM AI* **2**, AIe2500143 (2025).
85. Gu, B., Desai, R. J., Lin, K. J. & Yang, J. Probabilistic medical predictions of large language models. *Npj Digit. Med.* **7**, 1–9 (2024).
86. National NLP clinical challenges (n2c2). <https://n2c2.dbmi.hms.harvard.edu/home>.
87. CLEF eHealth lab series. <https://clefehealth.imag.fr/clefehealth.imag.fr/index.html>.
88. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* **101**, e215–e220 (2000).
89. Hugging face – the AI community building the future. <https://huggingface.co/> (2025).
90. Kwon, W. *et al.* Efficient memory management for large language model serving with PagedAttention. in *Proceedings of the 29th Symposium on Operating Systems Principles* 611–626 (Association for Computing Machinery, New York, NY, USA, 2023). doi:10.1145/3600006.3613165.
91. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. in *Text Summarization Branches Out* 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004).
92. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* 311 (Association for Computational Linguistics, Philadelphia, Pennsylvania, 2001). doi:10.3115/1073083.1073135.
93. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating text generation with BERT. in *International conference on learning representations* (2020).
94. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
95. Bird, S. NLTK: The natural language toolkit. in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions* (ed. Curran, J.) 69–72 (Association for Computational Linguistics, Sydney, Australia, 2006). doi:10.3115/1225403.1225421.
96. google-research/rouge at master · google-research/google-research. *GitHub* <https://github.com/google-research/google-research/tree/master/rouge>.
97. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating text generation with BERT. Preprint at <https://doi.org/10.48550/arXiv.1904.09675> (2020).
98. Xu, R., Wang, Z., Fan, R.-Z. & Liu, P. Benchmarking benchmark leakage in large language models. Preprint at <https://doi.org/10.48550/arXiv.2404.18824> (2024).