AI Medicine
https://www.sciltp.com/journals/aim

Review

Large Language Models for Differential Diagnosis: A Survey of Performance, Collaboration, and Technical Strategies

Yunjia Wu †, Qi Yan † and Dingcheng Tian *

College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110016, China
* Correspondence: 2310520@stu.neu.edu.cn
† These authors contributed equally to this work and share first authorship.

How To Cite: Wu, Y.; Yan, Q.; Tian, D. Large Language Models for Differential Diagnosis: A Survey of Performance, Collaboration, and Technical Strategies. *AI Medicine* **2026**, *3*(2), 7. https://doi.org/10.53941/aim.2026.100007

Received: 25 January 2026
Revised: 17 August 2026
Accepted: 21 August 2026
Published: 25 August 2026

Abstract: Errors in differential diagnosis often arise while clinicians are generating and comparing candidate explanations. This review examines the use of large language models (LLMs) for this part of diagnostic reasoning. Internal medicine and pediatrics are the main focus; evidence from radiology, surgical subspecialties, infectious disease, and mental health is used to examine how findings change across specialties. Reported performance depends on the clinical setting, the quality of the input, the prompt, model adaptation, and the evaluation design. Some studies place LLMs near trainees and find that they produce wider, better-organized differentials. Experienced clinicians, however, remain more reliable overall. Domain adaptation, external knowledge, and interactive workflows have improved performance in specific evaluations, but hallucinations and automation bias remain, alongside unresolved questions of governance. Current evidence therefore supports clinician-supervised use of artificial intelligence (AI) systems rather than autonomous diagnosis, pending prospective and specialty-specific evaluation.

Keywords: large language models; differential diagnosis; clinical decision support; diagnostic reasoning; human–AI collaboration

1. Introduction

Clinical decisions depend on which diagnoses enter the differential and how diagnostic confidence is revised as evidence accumulates. The task is to compare plausible explanations for the same presentation and decide which one best accounts for the available findings. In primary care, emergency medicine, pediatrics, and complex internal medicine, common errors include premature closure, anchoring, and failure to generate a sufficiently broad set of hypotheses. Computational systems are being investigated because they may help clinicians retain or organize alternatives that would otherwise be overlooked.

Clinical use of large language models (LLMs) draws on their ability to process unstructured text, summarize clinical information, and produce context-sensitive reasoning. Reviews cover applications in medical education, summarization, triage, and information retrieval [1]. Differential diagnosis places a different demand on the model: it must propose several plausible hypotheses, weigh competing evidence, and communicate uncertainty, with clinically consequential errors possible at each step.

Early evaluations based on standardized vignettes or other structured diagnostic tasks reported strong results for advanced LLMs, including GPT-4 [2]. The later evidence is less uniform because it includes real and simulated cases from internal medicine [3–5], pediatrics [6,7], radiology [8–10], surgical subspecialties [11,12], infectious disease [13], and mental health [14–17]. Meta-analyses of these heterogeneous studies place trained physicians above LLMs on average; model performance is closer to that of trainees or non-experts [18,19].

Most reviews of medical LLMs group differential diagnosis with broader diagnostic or clinical applications, although the generation of competing hypotheses is a distinct task. Consequently, the rapidly growing literature has received little analysis focused specifically on differential generation. This review treats that step separately. Structured alternatives and explicit reasoning may help clinicians notice overlooked possibilities, but the same outputs can contain hallucinations or encourage automation bias. Usefulness and risk therefore need to be assessed together.

We examine the evidence from three linked perspectives: clinical specialty, comparison with clinicians, and the technical choices that shape diagnostic accuracy. The review asks (1) where LLMs most reliably generate differential lists, (2) how their outputs compare with clinician performance under different study designs, and (3) which technical strategies appear feasible for clinical use. Rather than assigning a single capability level to LLMs, the analysis considers domain context, data distribution, and interaction design as sources of variation.

2. Background and Methods

2.1. Background and Conceptual Framework

Differential diagnosis is an iterative process in which competing diagnostic hypotheses are generated and revised as clinical information changes. In this review, a 'differential diagnosis list' is the set of candidate diagnoses produced by a model or clinician. 'Diagnostic reasoning' denotes the weighing of evidence, explanation of alternatives, and updating of those candidates. Unlike single-label diagnostic classification, differential diagnosis must account for uncertainty, alternative explanations, and relationships among signs, symptoms, and disease mechanisms. It therefore examines capacities that extend beyond pattern recognition. Reviews of clinical LLM use describe applications in documentation, decision support, and patient communication, alongside unresolved technical and ethical problems [1,2]. Diagnostic applications warrant particular caution because an error can directly affect patient safety.

Most LLM-based diagnostic systems can be described in four stages. Clinical information is taken from vignettes, electronic health records (EHRs), radiology reports, or interview transcripts. Preprocessing or knowledge integration may then be applied. An LLM or an ensemble generates diagnostic hypotheses and associated reasoning, after which task-specific metrics are used for evaluation. Figure 1 maps this pipeline and the principal design choices at each stage.

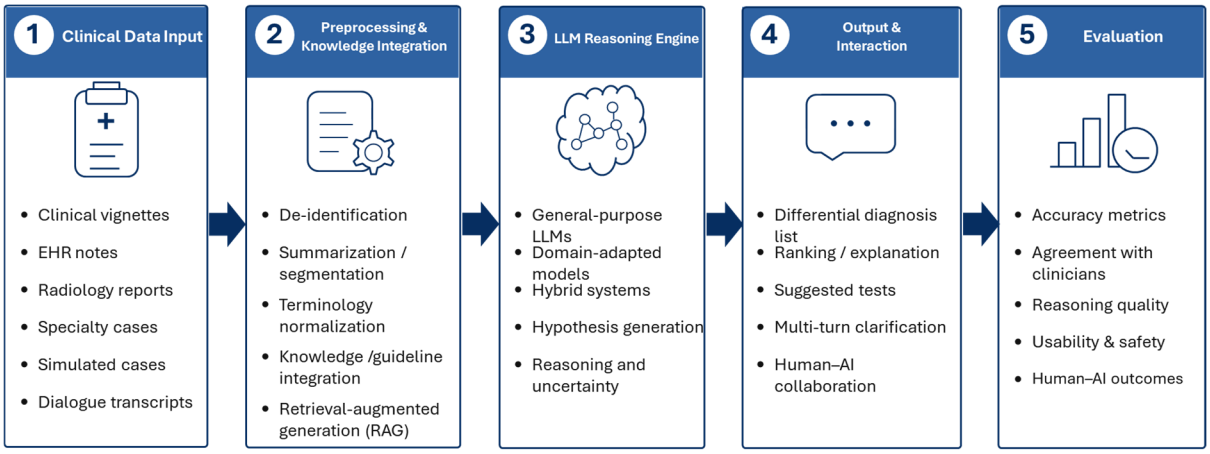

Figure 1. Pipeline of Large Language Model–Based Differential Diagnosis. Abbreviations: AI, artificial intelligence; EHR, electronic health record; LLM, large language model; RAG, retrieval-augmented generation.

Studies do not begin from comparable inputs. Standardized or textbook-derived vignettes present curated information in a single narrative [3,4,20], whereas real EHR notes are noisy, incomplete, and heterogeneous [5–7]. Radiology evaluations usually use transcribed imaging reports [8–10]. Surgical subspecialty and ophthalmology studies more often rely on focused case descriptions [11,12], and mental health work may use multi-turn interview transcripts or synthetic dialogues [14]. These differences in input structure are one reason reported performance varies across specialties.

General-purpose LLMs such as GPT-3.5 and GPT-4 are often evaluated with zero-shot or few-shot prompts [3,8,12,20]. Other studies fine-tune models on institutional EHRs, pediatric or pediatric intensive care unit (PICU) data, or mental health corpora to match local language and documentation patterns [4–7,14]. One way to constrain hallucinations and make outputs more explainable is to connect the model to a knowledge graph or dual-inference framework [21]. A different line of work changes who participates in the decision. MedSyn

combines clinicians and models in simulated physician–LLM dialogues, but its effectiveness with practicing physicians is not yet known [22]. Human–AI collectives aggregate clinician and model judgments; in vignette-based evaluations, this combination was more accurate than either component alone [23].

Prompt format determines both what an LLM returns and how that output is presented to clinicians. A direct request usually produces a differential list; diagnostic-reasoning or chain-of-thought (CoT) prompts can yield more structured explanations [20,24]. Other systems request explicit uncertainty estimates or use self-consistency sampling. The Articulate Medical Intelligence Explorer (AMIE) instead conducts a multi-turn dialogue that more closely resembles a clinical encounter [4]. These inference-time choices affect measured accuracy as well as clinicians' interpretation and use of the output.

Studies use several non-equivalent endpoints. Common measures include top-k accuracy, inclusion of the correct diagnosis anywhere in a differential list, and overlap with expert-generated differentials [3,8,11–13]. Structured clinical reasoning rubrics are used to compare the quality of LLM and physician reasoning [25,26], while randomized vignette trials test whether LLM suggestions change clinicians' diagnostic performance [9,10,26,27]. Meta-analyses across these designs find lower overall performance for LLMs than for physicians, although LLM estimates are often close to those of non-experts or trainees [18,19]. Performance can also fall when a prompt is derived from structured EHR data rather than a well-formed case narrative [28].

2.2. Literature Search and Narrative Synthesis

Literature selection was conducted as a targeted narrative search rather than a systematic review or formal meta-analysis. PubMed, Scopus, Google Scholar, and arXiv were searched for English-language publications published between 2019 and 2025. Search terms included model-related terms such as 'large language model', 'LLM', and 'GPT'; diagnostic terms such as 'differential diagnosis', 'diagnostic reasoning', and 'clinical decision support'; and specialty-related terms including 'pediatrics', 'internal medicine', 'radiology', and 'mental health'. The search strategy and eligibility criteria are summarized in Table 1. The original targeted search covered publications through 30 December 2025. Thirty potentially relevant publications were initially retained; after removal of three duplicate or overlapping records, 27 publications were included. A focused supplementary search of the mental-health literature was conducted on 11 August 2026. Three additional eligible mental-health studies were incorporated into the narrative synthesis [15–17], resulting in 30 publications in the final narrative synthesis, and the bibliographic record for the previously included WiseMind study was updated to its peer-reviewed version [14]. Targeted governance sources concerning liability, bias, and patient trust were also consulted to strengthen the discussion of real-world implementation [29–32]; these contextual sources were not counted as diagnostic studies.

Table 1. Search strategy summary.

Items	Specification
Original targeted search cutoff	30 December 2025
Focused supplementary search	11 August 2026
Supplementary-search scope	Mental-health evidence gaps; three additional eligible studies were incorporated [15–17]. The WiseMind citation was updated to its peer-reviewed version [14]. Governance sources [29–32] were used for contextual discussion and were not counted as diagnostic studies.
Databases and other sources searched	PubMed, Scopus, Google Scholar, and arXiv
Publication timeframe	Original search: 2019–2025; focused update: through 11 August 2026
Search terms used	'large language model', 'LLM', 'GPT', 'differential diagnosis', 'diagnostic reasoning', 'clinical decision support', 'pediatrics', 'internal medicine', 'radiology', and 'mental health'
Inclusion criteria	Empirical studies, comparative evaluations, randomized vignette trials, human-AI collaboration studies, and relevant systematic reviews or meta-analyses that directly assessed differential diagnosis generation, diagnostic reasoning quality, clinician-LLM comparison, clinician use of LLM-generated diagnostic outputs, or technical strategies for improving diagnostic reliability.

Table 1. *Cont.*

Items	Specification
Exclusion criteria	Studies limited to administrative summarization, patient education, image segmentation, coding, general medical question answering without a diagnostic task, or non-diagnostic clinical natural language processing tasks.
Selection process	Literature identification and selection were conducted by one author.
Study selection	Original search: 30 potentially relevant publications retained; 3 duplicate or overlapping records removed; 27 publications included. Focused supplementary search: 3 additional mental-health studies included. Final narrative synthesis: 30 publications.
Data extracted	Clinical specialty, data source, model configuration, adaptation strategy, comparator type, diagnostic task, evaluation metrics, and major findings.
Synthesis approach	Narrative comparative synthesis because the included studies differed substantially in clinical tasks, datasets, comparator groups, model configurations, and outcome metrics.

For each included study, we recorded the clinical specialty, data source (for example, vignette, EHR, radiology report, subspecialty case, or interview transcript), model configuration and adaptation strategy, comparator type, diagnostic task, evaluation metrics, and major findings. Studies were retained as direct evidence when they examined differential diagnosis generation, diagnostic reasoning quality, or clinician-model collaboration; background articles were used only for clinical and ethical context. The clinical tasks, datasets, comparator groups, model configurations, and outcome metrics differed enough that quantitative pooling was not appropriate. We instead used a narrative comparative synthesis with three axes: performance across specialties, comparisons with clinicians and human-AI collectives, and technical strategies affecting diagnostic performance and interpretability. Recent systematic reviews and meta-analyses informed this comparison [18,19]. No formal risk-of-bias assessment was performed because the review used a targeted narrative design.

Table 2 summarizes the key characteristics of the studies included in this survey, including specialty, data source, model type, and high-level outcomes.

Table 2. Overview of Representative Studies on LLM-based Differential Diagnosis.

Specialty	Data Source	LLMs	Task Type	Key Finding
Radiology [8]	Transcribed imaging reports	GPT-3.5, GPT-4	Differential generation	GPT-4 more accurate and complete than earlier models.
Radiology [10]	Magnetic resonance imaging (MRI) reports	LLM assistant	Interactive differential diagnosis	LLM-assisted workflow improved diagnostic completeness.
Pediatrics [6]	Outpatient EHRs	Fine-tuned GPT-3	Differential diagnosis prediction	Performance comparable to pediatricians.
PICU [7]	Critical-care notes	Domain-specific LLMs	EHR-based differential diagnosis	Specialized models outperformed general LLMs.
Internal Medicine [4]	Clinical vignettes	AMIE, GPT-4	Multi-turn reasoning	Broader, more structured differentials than unaided clinicians.
Internal Medicine [3]	Complex oncology cases	GPT-3.5, GPT-4	Differential diagnosis generation	GPT-4 improved accuracy but retained hallucination risks.
Infectious Disease [13]	Simulated infectious disease cases	GPT-4, Llama, Gemini	Differential diagnosis lists	Low concordance with expert differentials.
Surgery [11]	Hand/nerve injury cases	GPT-4, Isabel	Differential diagnosis ranking	GPT-4 outperformed rule-based tools.
Ophthalmology [12]	Complex eye-disease cases	Multiple LLMs	Differential diagnosis + management	Improved differential lists but unstable management.
Mental Health [14]	Multi-turn interviews	WiseMind	Conversational diagnosis	Multi-agent framework improved consistency.

3. LLM Performance Across Medical Specialties

Cross-specialty estimates of LLM diagnostic performance are not directly comparable. Radiology and some surgical subspecialties use relatively standardized inputs and constrained hypothesis spaces, and report more stable performance [8–12]. Internal medicine, infectious disease, and mental health use less uniform data and show wider ranges, lower expert concordance, and greater sensitivity to input format [3–5,13,14,28]. Table 3 summarizes the typical data sources, modeling strategies, and approximate study-level performance ranges reported for the major specialties. These ranges are not head-to-head estimates because datasets and comparator groups differ. Figure 2 is limited to the four specialties with approximate ranges for both LLMs and clinicians. Mental health is not plotted because its studies combine vignette, transcript, and risk-assessment outcomes without a consistent head-to-head clinician baseline [14–17].

Table 3. Summary of LLM Diagnostic Performance by Medical Specialty.

Specialty	Data Source	Best LLM Strategy	LLM Performance	Clinician Baseline	Notes
Radiology [8–10]	Imaging findings	CoT prompting; interactive user interface	55–75%	70–90%	Explanations amplify usefulness.
Pediatrics [6,7]	EHR encounters	Fine-tuning	60–80%	70–85%	Data alignment > model size.
Internal Medicine [3–5,25,26,33]	Vignettes, real EHR	Domain adaptation; dialogue	40–70%	70–90%	Highly variable.
Surgical Subspecialties [11,12]	Case descriptions	GPT-4 > rule-based	50–85%	60–95%	Strong in differential generation.
Mental Health [14–17]	Vignettes, interviews, and crisis-text transcripts	Multi-agent dialogue	Heterogeneous; not directly comparable	No consistent clinician baseline	Evidence limited to vignettes, transcript-based assessment, and proxy outcomes.

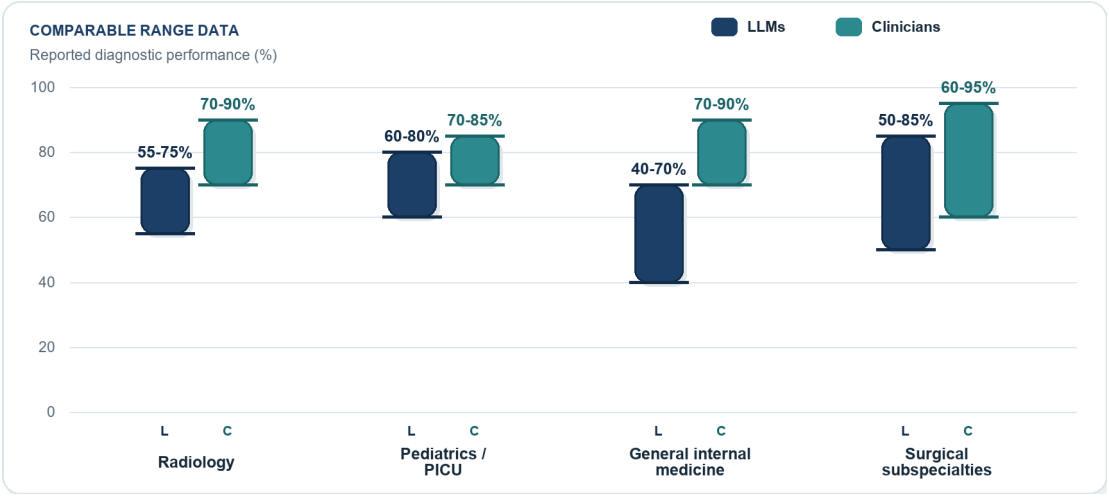

Figure 2. Reported Diagnostic Performance in Specialties with Comparable Study-Level Data. Note: Figure 2 presents only radiology, pediatrics, internal medicine, and surgical subspecialties, for which approximate study-level ranges are available for both LLMs and clinicians. The ranges are descriptive summaries rather than pooled head-to-head estimates; mental health is excluded because published studies use heterogeneous tasks and outcome measures without a consistent clinician baseline [14–17]. The displayed ranges synthesize radiology studies [8–10], pediatric studies [6,7], internal-medicine studies [3–5,25,26,33], and surgical-subspecialty studies [11,12].

3.1. Radiology

Radiology provides relatively structured inputs for LLM-based differential diagnosis. Textbook-style cases based on transcribed imaging findings provide one example: GPT-4 generated more accurate and complete differentials than GPT-3.5 and was more likely to include and highly rank the correct diagnosis [8]. When contextual information was incomplete, however, it also produced hallucinated statements [8,28]. Output format mattered in a randomized study of radiologists. Structured CoT explanations were associated with higher diagnostic accuracy than either final diagnoses or unstructured lists, suggesting that the reasoning text directed

attention to relevant imaging features [9]. In a separate brain MRI usability study, residents using an LLM-assisted workflow produced more complete differentials without taking longer to interpret the case [10]. Together, these findings locate the benefit in the combination of structured input and explanation design, rather than in model output alone.

3.2. Pediatrics and Pediatric Critical Care

Diagnostic accuracy varies more in pediatrics because EHR encounters differ in completeness, note style, developmental stage, and presenting symptoms. A fine-tuned GPT-3 model performed comparably to pediatricians on real-world outpatient EHR encounters for both accuracy and clinically acceptable differential lists [6]. In the PICU, models trained on local critical-care notes outperformed general-purpose and biomedical LLMs. Pediatric intensivists rated their differentials as more appropriate and complete [7]. The comparison indicates that alignment with local data and documentation can matter more than model size in this setting. Meta-analyses still report variable pediatric performance that is generally below that of experienced clinicians [18,19], supporting an assistive rather than autonomous role.

3.3. Surgical Subspecialties and Ophthalmology

For hand and peripheral nerve injuries, GPT-4 included the correct diagnosis more often than the rule-based Isabel differential diagnosis generator and ranked it higher [11]. Hand surgeons also judged the GPT-4 suggestions to be more clinically plausible and useful. In challenging cases from JAMA Ophthalmology, subscription-based LLMs generated better differentials than free versions, although their absolute accuracy and management recommendations were not adequate for standalone use [12]. In these structured subspecialty tasks, LLMs may serve as an additional source of diagnostic candidates, while clinicians remain responsible for treatment decisions.

3.4. General Internal Medicine and Infectious Disease

Complex internal medicine and infectious disease place different demands on diagnostic models. For internal medicine and oncology cases, GPT-4 included the correct diagnosis more often than GPT-3.5 and produced lists closer to the hypotheses of expert discussants [3]. In a challenging case-report setting, standalone AMIE achieved higher top-10 diagnostic accuracy and produced more appropriate and comprehensive differential-diagnosis lists than unaided clinicians [4]. A multispecialty study of real hospital cases reported the highest diagnostic scores for the domain-adapted Med-Go model, including in internal medicine scenarios [5]. Results in infectious disease were less favorable. On simulated cases, the differentials produced by GPT-4, Llama, and Gemini had only modest overlap with expert lists and weak agreement in ranking [13]. Temporal and epidemiologic reasoning thus remain important sources of error.

3.5. Mental Health

Mental health studies do not share a single differential-diagnosis benchmark, even though diagnosis in this field depends heavily on conversation and context. Results from 20 Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) clinical cases differed across models and disorders; GPT-3.5 and GPT-4 generally performed better than the comparator models [15]. A cross-sectional comparison of four LLMs against norms from mental health professionals also found that performance depended on the condition. Results were stronger for depression and post-traumatic stress disorder than for complex presentations such as early schizophrenia [16]. These vignette studies assess performance under constrained conditions and cannot be treated as comparisons with comprehensive psychiatric assessment.

Several mental health evaluations address proxy tasks rather than diagnosis. One study used 100 youth crisis text-line transcripts rated by four human experts and found that results changed with prompting and temperature. Its best configurations reached balanced accuracy of approximately 0.54–0.71, although validity was poor for several clinically important items [17]. WiseMind and related conversational or multi-agent systems address a different task: organizing interviews, symptom summaries, and diagnostic reasoning [14]. Outcomes such as agreement on suicide-risk ratings, symptom classification, or conversational consistency therefore should not be interpreted as top-k diagnostic accuracy or as direct comparisons with practicing psychiatrists.

Current mental health studies do not provide a clinician baseline that can be recovered from the reported data. Most use small or curated vignettes, synthetic dialogue, text-only transcripts, or reference labels. They consequently omit parts of psychiatric assessment that depend on longitudinal course, behavioral observation, collateral history, cultural context, and therapeutic interaction. Prospective multisite studies should compare

assisted and unassisted clinicians on the same cases, report subgroup performance and safety outcomes, and determine whether suggestions improve calibration without increasing automation bias [1,2,14–19,28]. Table 3 therefore records the gap, and Figure 2 excludes mental health from the quantitative comparison.

3.6. Cross-Specialty Patterns

Performance differences across specialties are closely related to study context. Relatively structured radiology reports and focused subspecialty cases support more stable estimates, whereas internal medicine, infectious disease, and mental health expose weaknesses when longitudinal information is heterogeneous [3–8,11–17,28]. Fine-tuning and local data alignment reduce the gap with clinicians in pediatrics, PICU, and multispecialty hospital decision support [5–7]. Even with these adaptations, meta-analyses and comparative studies find that physicians outperform LLMs on average; LLM performance is more often close to trainee-level accuracy and may add complementary hypotheses [18,19,25–27,33]. Studies of human–AI interaction and collective decision-making support a co-diagnostic rather than replacement role [9,10,22,23,26,27].

4. Comparative Analysis: LLMs vs. Human Clinicians

Comparisons with clinicians place the specialty-specific results in a clinical context. Across studies, LLMs usually perform below experienced physicians but near trainees or non-experts. Systematic reviews report that the difference is most apparent in complex or high-stakes cases, even when an LLM produces a coherent differential [18,19]. Some individual studies also identify reasoning that complements clinicians' judgments, which makes interaction design relevant to the comparison [4,9,10,22,23].

Table 4 summarizes comparative studies in radiology, pediatrics, internal medicine, infectious diseases, global health, and collective intelligence. It records relative diagnostic results together with explanation format, domain-specific fine-tuning, and human-AI interaction design. The comparator groups differ across studies, ranging from pediatricians and residents to infectious disease specialists and human–AI collectives. These baselines do not yield a single cross-specialty performance gap. The table therefore reports the direction and size of each difference in relation to its study design.

Table 4. Summary of Comparative Findings: LLMs vs. Human Clinicians.

Study Context	Comparator	Outcome	Takeaway
Radiology [9]	Radiologists: no LLM support vs. GPT-4 advice	CoT reasoning improved human accuracy	Explanation format shapes utility
Pediatrics [6]	Fine-tuned GPT-3 vs. pediatricians	Comparable performance	Domain-specific tuning narrows gap
Internal Medicine [4]	Standalone AMIE vs. unaided clinicians	Broader differential lists	LLM excels in hypothesis breadth
Infectious Disease [13]	GPT-4, Llama 3, and Gemini 1.5 vs. expert-consensus differential lists	Low concordance	High-risk domain remains fragile
Clinical Vignettes [26]	Clinicians ± GPT-4	Mixed results	Raw suggestions insufficient without structured reasoning
Low-Resource Randomized Controlled Trial [27]	AI-trained physicians: LLM plus conventional resources vs. conventional resources alone	LLM access improved diagnostic reasoning scores	AI-literacy training may support effective clinician–LLM collaboration
Collective Decision Models [23]	Multi-model + multi-clinician	Highest accuracy	Hybrid ensembles outperform individuals

Randomized vignette trials directly test whether LLM output changes clinician performance. In a large trial of GPT-4 and diagnostic reasoning, access to model output did not reliably improve clinicians' results, particularly when the suggestions lacked structured reasoning [26]. Trials using CoT explanations reported better diagnostic completeness and calibration. Within those study settings, structured reasoning was associated with more effective assistance, but the result does not establish a general causal mechanism [9,24]. In a resource-limited setting, another randomized trial found significantly higher diagnostic reasoning scores among AI-trained clinicians with LLM access than among AI-trained clinicians using conventional resources alone [27].

Real and semi-structured cases reveal a different set of comparisons from vignette trials. On structured reasoning tasks, LLMs have matched or exceeded residents on rubric-based assessments, including the Revised-

IDEA scoring rubric (R-IDEA; Interpretive summary, Differential diagnosis, Explanation of reasoning, and Alternative diagnoses explained) [25,26]. The gap between physicians and LLMs is larger when cases require epidemiologic reasoning, rare disease recognition, or management of complex multimorbidity, as reported in infectious disease and internal medicine evaluations [3,5,13].

Aggregating human and model judgments produced the highest overall accuracy in a multi-reader, multi-model study, exceeding either clinicians or LLMs alone [23]. Studies of structured explanations and a user-tested MRI assistant also indicate that the design of the interaction can support clinician performance while preserving human oversight [9,10]. MedSyn addresses the same broad problem through simulated physician–LLM interactions, but it has not been validated with practicing physicians [22].

The comparative evidence supports a circumscribed role for LLMs. An LLM may broaden the hypothesis set and expose alternative reasoning paths. It can also supply explanations for clinician review, but these functions do not replace clinical judgment. This role depends on workflow integration, interfaces that support scrutiny instead of automatic acceptance, and training that addresses both the strengths and limitations of model-generated reasoning.

5. Technical Strategies Shaping Diagnostic Performance

Prompt design, model adaptation, knowledge integration, interactive workflows, and evaluation strategy each affect LLM diagnostic performance. Their effects extend beyond accuracy to the structure of reasoning, clinical safety, and workflow use. The implementation burden is not uniform. Prompting can be piloted with limited infrastructure; fine-tuning, retrieval systems, and multi-agent tools require additional data governance, computing resources, maintenance, and clinical oversight. Figure 3 organizes these approaches according to the technical component used.

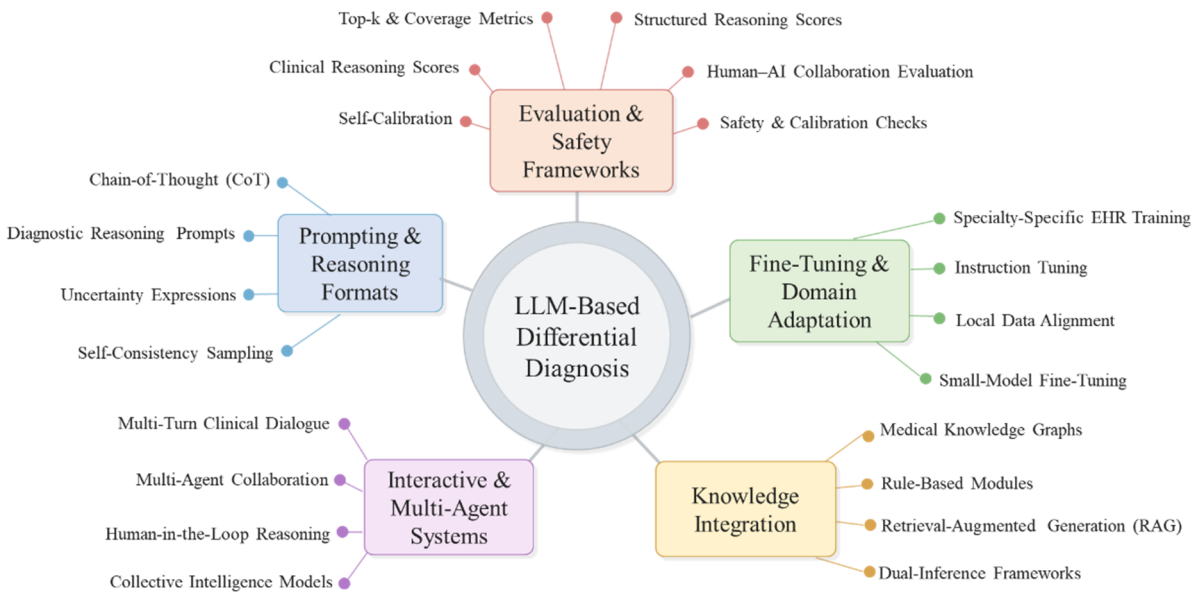

Figure 3. Taxonomy of Technical Strategies for LLM-Based Differential Diagnosis. Abbreviations: CoT, chain-of-thought; EHR, electronic health record; LLM, large language model; RAG, retrieval-augmented generation.

5.1. Prompting and Reasoning Formats

Prompt format affects diagnostic output. CoT and diagnostic-reasoning prompts have produced more structured explanations than direct-answer prompts and have improved hypothesis completeness in the evaluated tasks [20,24]. Some systems also use self-consistency sampling or ask the model to express uncertainty. These choices do not remove sensitivity to wording. Privacy-preserving or highly structured prompts can reduce accuracy when they omit contextual cues [28]. A clinical prompting protocol therefore requires local validation and version control, even though prompting has a lower implementation cost than model adaptation. Clinicians also need to know when a template may distort or exclude relevant context.

5.2. Fine-Tuning and Domain Adaptation

Fine-tuning on domain data has yielded consistent gains in several diagnostic studies. Local pediatric and PICU EHR models outperformed larger general-purpose LLMs [6,7]. The hospital-adapted Med-Go model

likewise outperformed GPT-4 on internal medicine, emergency, and surgical tasks [5]. In these studies, matching the training data to the target setting mattered more than model size. That advantage comes with operational risks: overfitting, privacy loss, and distribution drift. Deployment therefore depends on de-identified data, secure computing, drift monitoring, and retraining when documentation practices or patient populations change.

5.3. Knowledge Integration and Hybrid Architectures

Unsupported model output can be constrained by tying generation to structured medical knowledge. Knowledge graphs, retrieval-augmented generation (RAG), rule-based modules, and dual-inference systems implement this link in different ways. Dual-Inference LLMs, for example, compare generated hypotheses against structured representations of medical evidence, improving consistency and reducing unsupported differential suggestions [21]. The language model remains flexible, but the evidence source sits outside it. Reliability therefore depends on keeping ontologies, guidelines, and retrieval corpora current and on retaining a traceable source for each diagnostic suggestion.

5.4. Interactive and Multi-Agent Systems

Interactive systems can obtain additional information before a differential is finalized, but they do so through different forms of coordination. AMIE gathers information through iterative questioning and improved accuracy in internal medicine scenarios [4]. Multi-agent systems distribute work among roles such as interviewer, summarizer, and reasoner; this organization has been used to limit model drift and improve consistency in mental health assessment [14]. MedSyn divides tasks between clinicians and models, although evaluation has so far been limited to simulated physician–LLM interactions rather than practicing physicians [22]. Human–AI collectives aggregate judgments, and in vignette evaluations they outperformed humans or models used alone [23]. Each approach requires more than prompt design: the clinical interface must support audit trails, escalation, and a clear assignment of responsibility for the final decision.

5.5. Evaluation Frameworks and Diagnostic Metrics

Evaluation endpoints answer different questions. Top-k accuracy and inclusion rates indicate whether the target diagnosis appears, not whether the reasoning is sound. R-IDEA addresses this distinction by scoring hypothesis generation, evidence integration, and justification separately [25,26]. Clinician-centered evaluations measure a different outcome: changes in decision-making, anchoring, or calibration. For comparisons across model versions, specialties, and data sources, meta-analyses have called for common benchmarks [18,19]. A prospective randomized study protocol comparing an LLM with a human coach proposes evaluating interaction design and user engagement alongside diagnostic outcomes [34].

6. Discussion

Variation across studies is closely tied to the structure of the clinical input. In radiology and some surgical subspecialties, GPT-4 has produced plausible differential lists from comparatively structured case descriptions, sometimes performing near trainees [8–12]. Internal medicine and infectious disease require more multimorbidity, temporal, or epidemiologic reasoning; accuracy and specialist concordance are correspondingly more variable [3–5,13]. Mental health adds subjective narratives and conversational context, while the available studies use no common diagnostic endpoint [14–17].

Task structure also affects comparisons with clinicians. On some structured reasoning tasks, LLMs match or exceed medical trainees; in real clinical scenarios, attending physicians retain an advantage. Meta-analyses report higher physician performance overall, particularly in high-acuity or ambiguous cases [18,19]. The wider and more systematic organization of some LLM differentials may be useful for expanding hypotheses, but presentation quality is not the same as factual grounding. Rubric-based studies found that grounding remained variable even when the diagnostic rationale was well organized [25,26].

Clinician performance does not improve simply because an LLM output is available. In randomized vignette trials, raw suggestions produced no reliable gain and could introduce automation bias when they were wrong [26]. CoT explanations and interactive systems such as AMIE or multi-turn MRI assistants were associated with more complete or better calibrated differentials [4,9,10]. MedSyn provides a simulation-based account of this interaction and still requires evaluation with practicing physicians [22]. Human–AI collectives achieved the highest accuracy reported in their study by aggregating clinician and model judgments [23].

Technical strategies differ in both effect and feasibility. Diagnostic-reasoning prompts and structured explanations can make an output easier for clinicians to interpret [20,24]. Models adapted to local EHR or pediatric data have outperformed larger general-purpose models in the corresponding settings [6,7], while dual-inference and retrieval-augmented architectures use external knowledge to reduce unsupported output [21]. The resources required are different for each approach. Prompting is relatively easy to pilot; fine-tuning requires institutional data and governance; RAG requires maintained knowledge sources; and multi-agent systems require changes to clinical workflow. Selection should therefore reflect the resources and safety requirements of the intended setting.

Important limitations remain across these approaches. Current systems have difficulty with temporal ordering of symptoms, prevalence calibration, rare disease recognition, and dynamic clinical data. Hallucinated diagnoses, overconfidence, and sensitivity to small prompt changes are continuing safety concerns [28]. Common benchmarks rarely assess whether a system communicates uncertainty, provides sound reasoning, or changes clinician decisions. Prospective evaluations embedded in clinical workflows are needed to measure those outcomes before routine deployment.

Governance, Accountability, Bias, and Patient Trust

Diagnostic accuracy alone does not resolve the governance questions raised by LLM-based differential diagnosis. Outputs are probabilistic, and behavior can change between model versions. Responsibility for an erroneous, delayed, or overconfident diagnosis may therefore be shared among developers, deploying institutions, and clinicians, while liability standards remain jurisdiction-dependent and unsettled [29,32]. A deployment policy should identify the final decision-maker, define triggers for escalation or specialist review, and record model versions, prompts, retrieved sources, clinician actions, and adverse events. General disclaimers do not substitute for institutional review, incident reporting, and post-deployment monitoring.

Training data, clinical documentation, and retrieval sources may underrepresent groups or encode stereotypes. An LLM can reproduce these patterns through omitted diagnoses or inappropriate prevalence assumptions for racial, ethnic, sex, age, language, disability, or socioeconomic groups [1,2,28,29]. Aggregate accuracy may not reveal the resulting errors. Local validation should report subgroup results, audit fairness and errors, assess language and accessibility barriers, and monitor distribution drift. Institutions should also identify populations and settings that were absent from development and validation data.

Disclosure and integration into the clinical relationship affect patient trust. In a vignette study, AI-based clinical decision support reduced perceived physician competence or integrity in some high-risk settings and subgroups, so technical performance alone did not determine trust [31]. Patients should be told when an LLM materially contributes to diagnostic reasoning, what data it processes, what its limitations are, and how a clinician reviews the output. They should be able to ask questions and, where feasible, decline AI-supported workflows. World Health Organization guidance also identifies stakeholder participation, transparent communication, independent audit, autonomy, and privacy as conditions for trustworthy deployment [29].

Current evidence supports using LLMs to assist, rather than conduct, differential diagnosis. Their clearest contribution is to bring overlooked possibilities into consideration and organize the rationale that a clinician must review. Safeguards, interpretable output, and human oversight remain necessary because the evidence does not support autonomous diagnosis.

7. Conclusions

LLMs can produce organized differential diagnosis lists, but the evidence across specialties does not support a single estimate of their performance. Results are strongest when inputs are structured, as in radiology and surgical subspecialties. The more heterogeneous information required in internal medicine, infectious disease, and mental health remains harder for these models to handle. Across comparative studies and meta-analyses, performance is generally closer to that of skilled trainees than experienced clinicians.

Human–AI interaction can alter both the model output and how clinicians use it. In the evaluated settings, structured explanations, multi-turn dialogue, and external knowledge were associated with more complete or better calibrated clinician differentials. Dual-inference models change how evidence is incorporated, whereas multi-agent frameworks and human–AI collectives change how tasks or judgments are distributed. These approaches remain dependent on study design and do not remove the need for clinician review.

Prospective studies now need to test whether LLM support improves reliability, reduces cognitive load, or expands access to expertise. Such evaluations require domain-specific adaptation, high-quality clinical data, safety controls, and standardized diagnostic endpoints. The intended use remains diagnostic support rather than clinician

replacement. Before adoption, systems also need explicit accountability, subgroup fairness audits, and transparent disclosure [29,32], together with continued attention to patient trust [30,31].

Author Contributions

Y.W.: conceptualization, methodology, initial literature search, evidence synthesis, data curation, visualization, and writing—original draft preparation. Q.Y.: supplementary literature search, evidence screening and verification, methodology revision, visualization, writing—original draft preparation, and writing—reviewing and editing. D.T.: conceptual and methodological supervision, interpretation of the evidence, project administration, and writing—reviewing and editing. Y.W. and Q.Y. contributed equally to this work and share first authorship. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This article is a narrative review of previously published literature and did not involve human participants, animals, or the collection of new clinical data.

Informed Consent Statement

Not applicable. This article is a narrative review and did not involve the enrollment of human participants or the use of identifiable patient information.

Data Availability Statement

All data and information discussed in this narrative review are available in the cited publications. No new datasets were generated for this study.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT (OpenAI) for language polishing and assistance in improving the organization of the manuscript. After using this tool, the authors critically reviewed and edited all AI-assisted content as needed and take full responsibility for the content of the published article.

References

1. Jung, K.H. Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations. *Healthc. Inform. Res.* **2025**, *31*, 114–124. https://doi.org/10.4258/hir.2025.31.2.114.
2. Yang, X.; Li, T.; Su, Q.; et al. Application of Large Language Models in Disease Diagnosis and Treatment. *Chin. Med. J.* **2025**, *138*, 130–142. https://doi.org/10.1097/cm9.0000000000003456.
3. Ríos-Hoyo, A.; Shan, N.L.; Li, A.; et al. Evaluation of Large Language Models as a Diagnostic Aid for Complex Medical Cases. *Front. Med.* **2024**, *11*, 1380148. https://doi.org/10.3389/fmed.2024.1380148.
4. McDuff, D.; Schaekermann, M.; Tu, T.; et al. Towards Accurate Differential Diagnosis with Large Language Models. *Nature* **2025**, *642*, 451–457. https://doi.org/10.1038/s41586-025-08869-4.
5. Wang, X.; Ye, H.; Zhang, S.; et al. Evaluation of the Performance of Three Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases. *J. Med. Syst.* **2025**, *49*, 23. https://doi.org/10.1007/s10916-025-02152-9.
6. Mansoor, M.A.; Ibrahim, A.F.; Grindem, D.J.; et al. Evaluating the Accuracy and Reliability of Large Language Models in Assisting with Pediatric Differential Diagnoses: A Multicenter Diagnostic Study. *medRxiv* **2024**. https://doi.org/10.1101/2024.08.09.24311777.
7. Akhondi-Asl, A.; Yang, Y.; Luchette, M.; et al. Comparing the Quality of Domain-Specific versus General Language Models for Artificial Intelligence-Generated Differential Diagnoses in PICU Patients. *Pediatr. Crit. Care Med.* **2024**, *25*, e273–e282. https://doi.org/10.1097/pcc.0000000000003468.

8. Sun, S.H.; Huynh, K.; Cortes, G.; et al. Testing the Ability and Limitations of ChatGPT to Generate Differential Diagnoses from Transcribed Radiologic Findings. *Radiology* **2024**, *313*, e232346. https://doi.org/10.1148/radiol.232346.
9. Spitzer, P.; Hendriks, D.; Rudolph, J.; et al. The Effect of Medical Explanations from Large Language Models on Diagnostic Accuracy in Radiology. *npj Digit. Med.* **2026**, *9*, 333. https://doi.org/10.1038/s41746-026-02619-0.
10. Kim, S.H.; Wihl, J.; Schramm, S.; et al. Human-AI Collaboration in Large Language Model-Assisted Brain MRI Differential Diagnosis: A Usability Study. *Eur. Radiol.* **2025**, *35*, 5252–5263. https://doi.org/10.1007/s00330-025-11484-6.
11. AlShenaiber, A.; Datta, S.; Mosa, A.J.; et al. Large Language Models in the Diagnosis of Hand and Peripheral Nerve Injuries: An Evaluation of ChatGPT and the Isabel Differential Diagnosis Generator. *J. Hand Surg. Glob. Online* **2024**, *6*, 847–854. https://doi.org/10.1016/j.jhsg.2024.07.011.
12. Shanmugam, S.K.; Browning, D. Comparison of Large Language Models in Diagnosis and Management of Challenging Clinical Cases. *Clin. Ophthalmol.* **2024**, *18*, 3239–3247. https://doi.org/10.2147/opth.s488232.
13. Mondal, A.; Karad, R.; Bhattacharjee, B.; et al. Evaluating the Performance of Artificial Intelligence in Generating Differential Diagnoses for Infectious Diseases Cases: A Comparative Study of Large Language Models. *medRxiv* **2024**. https://doi.org/10.1101/2024.06.28.24309694.
14. Wu, Y.; Wan, G.; Li, J.; et al. WiseMind: A Knowledge-Guided Multi-Agent Framework for Accurate and Empathetic Psychiatric Diagnosis. *npj Digit. Med.* **2026**, *9*, 575. https://doi.org/10.1038/s41746-026-02559-9.
15. Gargari, O.K.; Fatehi, F.; Mohammadi, I.; et al. Diagnostic Accuracy of Large Language Models in Psychiatry. *Asian J. Psychiatry* **2024**, *100*, 104168. https://doi.org/10.1016/j.ajp.2024.104168.
16. Levkovich, I. Evaluating Diagnostic Accuracy and Treatment Efficacy in Mental Health: A Comparative Analysis of Large Language Model Tools and Mental Health Professionals. *Eur. J. Investig. Health Psychol. Educ.* **2025**, *15*, 9. https://doi.org/10.3390/ejihpe15010009.
17. Thomas, J.; Elyoseph, Z.; Kuchinke, L.; et al. Large Language Model Performance versus Human Expert Ratings in Automated Suicide Risk Assessment. *Sci. Rep.* **2025**, *15*, 39231. https://doi.org/10.1038/s41598-025-22402-7.
18. Takita, H.; Kabata, D.; Walston, S.L.; et al. A Systematic Review and Meta-Analysis of Diagnostic Performance Comparison between Generative AI and Physicians. *npj Digit. Med.* **2025**, *8*, 175. https://doi.org/10.1038/s41746-025-01543-z.
19. Shan, G.; Chen, X.; Wang, C.; et al. Comparing Diagnostic Accuracy of Clinical Professionals and Large Language Models: Systematic Review and Meta-Analysis. *JMIR Med. Inform.* **2025**, *13*, e64963. https://doi.org/10.2196/64963.
20. Wu, C.K.; Chen, W.L.; Chen, H.H. Large Language Models Perform Diagnostic Reasoning. *arXiv* **2023**, arXiv:2307.08922.
21. Zhou, S.; Lin, M.; Ding, S.; et al. Explainable Differential Diagnosis with Dual-Inference Large Language Models. *npj Health Syst.* **2025**, *2*, 12. https://doi.org/10.1038/s44401-025-00015-6.
22. Sayin, B.; Schlicht, I.B.; Hong, N.V.; et al. MedSyn: Enhancing Diagnostics with Human-AI Collaboration. In Proceedings of the Hybrid Human-Artificial Intelligence (HHAI) 2025 Workshops, Pisa, Italy, 9–10 June 2025; pp. 494–505.
23. Zöller, N.; Berger, J.; Lin, I.; et al. Human–AI Collectives Most Accurately Diagnose Clinical Vignettes. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2426153122. https://doi.org/10.1073/pnas.2426153122.
24. Savage, T.; Nayak, A.; Gallo, R.; et al. Diagnostic Reasoning Prompts Reveal the Potential for Large Language Model Interpretability in Medicine. *npj Digit. Med.* **2024**, *7*, 20. https://doi.org/10.1038/s41746-024-01010-1.
25. Cabral, S.; Restrepo, D.; Kanjee, Z.; et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared with Physicians. *JAMA Intern. Med.* **2024**, *184*, 581–583. https://doi.org/10.1001/jamainternmed.2024.0295.
26. Goh, E.; Gallo, R.; Hom, J.; et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw. Open* **2024**, *7*, e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.
27. Qazi, I.A.; Ali, A.; Khawaja, A.U.; et al. Large Language Model Diagnostic Assistance for Physicians in a Lower-Middle-Income Country: A Randomized Controlled Trial. *Nat. Health* **2026**, *1*, 198–205. https://doi.org/10.1038/s44360-025-00007-8.
28. Reese, J.T.; Danis, D.; Caufield, J.H.; et al. On the Limitations of Large Language Models in Clinical Diagnosis. *medRxiv* **2024**. https://doi.org/10.1101/2023.07.13.23292613.
29. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models*; World Health Organization: Geneva, Switzerland, 2024.
30. Choudhury, A.; Chaudhry, Z. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. *J. Med. Internet Res.* **2024**, *26*, e56764. https://doi.org/10.2196/56764.
31. Zondag, A.G.M.; Rozestraten, R.; Grimmelikhuijsen, S.G.; et al. The Effect of Artificial Intelligence on Patient-Physician Trust: Cross-Sectional Vignette Study. *J. Med. Internet Res.* **2024**, *26*, e50853. https://doi.org/10.2196/50853.
32. Mello, M.M.; Guha, N. Understanding Liability Risk from Using Health Care Artificial Intelligence Tools. *N. Engl. J. Med.* **2024**, *390*, 271–278. https://doi.org/10.1056/nejmhle2308901.

33. Hirosawa, T.; Harada, Y.; Mizuta, K.; et al. Evaluating ChatGPT-4's Accuracy in Identifying Final Diagnoses within Differential Diagnoses Compared with Those of Physicians: Experimental Study for Diagnostic Cases. *JMIR Form. Res.* **2024**, *8*, e59267. https://doi.org/10.2196/59267.
34. Kämmer, J.E.; Hautz, W.E.; Krummrey, G.; et al. Effects of Interacting with a Large Language Model Compared with a Human Coach on the Clinical Diagnostic Process and Outcomes among Fourth-Year Medical Students: Study Protocol for a Prospective, Randomised Experiment Using Patient Vignettes. *BMJ Open* **2024**, *14*, e087469. https://doi.org/10.1136/bmjopen-2024-087469.