

Artificial Intelligence and the Accuracy of Disease Diagnosis

Taruna Agarwal, John Leddo, Dev Agarwal, Ashna Melanathuru, Rishaan Prabakaran, Mihika Mamidi, Srishti Kallu, Shree Garg, Dhairya Garg, Heidi Vineet, Yousef Malik, Tatva Duggineni, Anika Dubey, Maahika Challagulla, Vihaan Penumatsa and Mahathi Girla

MyEdMaster, LLC

Virginia, US

DOI: 10.46609/IJSSER.2026.v11i09.027 URL: <https://doi.org/10.46609/IJSSER.2026.v11i09.027>

Received: 19 August 2026 / Accepted: 15 September 2026 / Published: 26 September 2026

ABSTRACT

Artificial intelligence (AI) and large language models (LLMs) have demonstrated strong diagnostic performance in narrowly defined medical tasks, but less is known about their accuracy when diagnosis is open ended and spans diverse disease categories. The present study evaluated GPT-4o as a general diagnostic system using 1,000 real-patient cases derived from the MIMIC-IV clinical database. Clinical information from each case was parsed and assembled into a structured summary, and GPT-4o generated a ranked list of five possible diagnoses. Three prompting strategies were compared: zero-shot, few-shot, and chain-of-thought prompting. Performance was evaluated at both the disease level and the broader medical-category level using Top-1, Top-3, and Top-5 accuracy. Zero-shot prompting produced the strongest disease-level performance, with Top-1 accuracy of 55.7%, Top-3 accuracy of 69.5%, and Top-5 accuracy of 73.0%. Few-shot prompting yielded corresponding accuracies of 52.1%, 65.9%, and 69.6%, while chain-of-thought prompting produced substantially lower accuracies of 34.2%, 46.7%, and 50.3%. At the category level, zero-shot Top-1 accuracy was 72.3%, compared with 69.4% for few-shot and 56.2% for chain-of-thought prompting. Zero-shot performance varied considerably across disease categories, with particularly high Top-5 disease-level accuracy for musculoskeletal (98%) and endocrine/metabolic (91%) cases and lower accuracy for cardiovascular (66%) and other (63%) cases. Accuracy also appeared to stabilize after approximately 300–400 evaluation samples. These findings suggest that GPT-4o can identify the correct diagnosis within a short differential diagnosis for a substantial proportion of heterogeneous real-patient cases, but performance depends on prompting strategy and disease category. Further research should address possible information leakage from discharge-summary framing, automated diagnosis-matching limitations, unequal category sample sizes,

and prospective evaluation under conditions that more closely approximate real-time clinical diagnosis.

Introduction

Accurate diagnosis is a foundational requirement of effective medical care. Diagnostic errors can delay appropriate treatment, expose patients to unnecessary testing or therapy, and contribute to preventable morbidity and mortality. Artificial intelligence (AI) has therefore attracted considerable attention as a potential means of improving diagnostic performance. Earlier medical AI systems were typically developed for narrowly defined tasks, such as classifying a radiographic abnormality or distinguishing malignant from benign lesions. More recently, large language models (LLMs) have created the possibility of a substantially broader form of diagnostic AI: a system can be given a patient's symptoms, history, examination findings, and test results and asked to determine the most likely diagnosis from an essentially open-ended universe of diseases. These two research paradigms—single-disease or narrowly constrained classification versus open-ended diagnosis—address different questions and should not be treated as equivalent measures of AI's ability to diagnose disease.

The strongest evidence for AI diagnostic performance historically has come from narrow, well-defined tasks, particularly image classification. In these studies, the AI system is usually trained or evaluated for a specific disease or limited group of diseases, and the relevant input is standardized, such as a dermoscopic image, mammogram, pathology slide, or computed tomography (CT) scan. The task may be reduced to determining whether a lesion is malignant or benign, whether a particular abnormality is present, or which of a small number of predefined categories best describes an image. This design has important methodological advantages: the reference diagnosis can often be established by pathology or follow-up, sensitivity and specificity can be calculated directly, and AI performance can be compared with physicians who evaluate the same images.

Research on skin cancer illustrates the maturity of this narrow-task literature. Salinas et al. (2024) conducted a systematic review and meta-analysis comparing AI algorithms with clinicians in skin cancer classification. Across the studies included in their meta-analysis, AI achieved pooled sensitivity of 87.0% and specificity of 77.1%, compared with sensitivity of 79.8% and specificity of 73.6% for clinicians overall. The comparison changed when physician expertise was considered: AI substantially outperformed generalists but performed more similarly to expert dermatologists. Thus, the findings provide evidence that a trained AI classifier can equal or exceed human performance on a particular diagnostic task, but they do not establish that the same AI could determine what disease an undifferentiated patient has when the correct diagnosis is not constrained to skin cancer or even dermatology.

A similar pattern is found in cancer imaging. Wang et al. (2024) systematically reviewed 20 studies comparing standalone deep-learning systems with human experts for lung cancer diagnosis using chest CT. Pooled sensitivity was approximately 82% for deep learning and 81% for human experts, whereas pooled specificity was approximately 75% for deep learning and 69% for experts. The difference in specificity favored AI significantly, while the sensitivity difference was not statistically significant. Importantly, performance varied by imaging modality and task. These results demonstrate both the promise and the limitation of disease-specific AI evaluation: a model may perform at or above specialist level within the distribution of cases and images for which it was designed, while providing little evidence about its ability to reason across unrelated diseases, integrate heterogeneous clinical information, or recognize that an unexpected diagnosis outside its target category is more likely.

The emergence of general-purpose LLMs has made it possible to investigate a different and arguably more demanding question. Rather than asking, for example, whether an image represents lung cancer, an open-ended diagnostic study can present a clinical vignette without specifying the disease category and ask the AI to generate a primary diagnosis or differential diagnosis. This more closely resembles the cognitive problem encountered by a generalist physician: symptoms such as fatigue, chest pain, fever, abdominal pain, dizziness, or weight loss can arise from many organ systems and hundreds of diseases. Success therefore requires not only pattern recognition within a trained disease category but also selection among competing explanations, integration of multiple findings, and prioritization of diagnoses.

Several studies suggest that contemporary LLMs can perform surprisingly well on these broader diagnostic tasks. Ito et al. (2023) compared GPT-4 with three board-certified physicians using 45 clinical vignettes representing a variety of typical health conditions. GPT-4 provided the correct primary diagnosis in 44 of 45 cases (97.8%), compared with 41 of 45 cases (91.1%) for the physician consensus; the difference was not statistically significant. Although the sample was small and vignette-based, the study is important because the model was not being asked merely to identify one predetermined disease from a single type of medical image. It was required to diagnose different health conditions from clinical descriptions.

Hirosawa et al. (2023) evaluated an even more challenging form of open-ended differential diagnosis using 52 complex case-derived clinical vignettes. GPT-4 placed the correct diagnosis first in 60% of cases, within its top five diagnoses in 81%, and within its top 10 in 83%. Physicians placed the correct diagnosis first in 50%, within the top five in 67%, and within the top 10 in 75%; none of the differences between GPT-4 and physicians reached statistical significance. These results also highlight an important issue in defining diagnostic accuracy. An AI system may fail to rank the correct disease first yet still generate it as a clinically useful possibility within a differential diagnosis. Consequently, studies of open-ended AI diagnosis

should distinguish top-1 accuracy from top-k differential-diagnosis accuracy rather than collapsing them into a single measure.

Rutledge (2024) further examined GPT-4 on both common clinical scenarios and previously unpublished challenging internal-medicine cases. On 45 common vignettes, GPT-4 included the correct diagnosis among its top three choices in 100% of cases, compared with 84.3% previously reported for physicians. On 36 challenging cases, GPT-4 included the correct diagnosis among its top six in 61.1% of cases, compared with 43.7% for residents and 49.1% for faculty physicians in the earlier physician evaluations. However, Rutledge emphasized a major methodological concern in LLM research: the common clinical scenarios had previously been published online and may have appeared in GPT-4's training data. Performance on previously published cases can therefore overestimate true generalization if a model has encountered the cases, their answers, or closely related text during training.

The distinction between narrow and open-ended diagnosis has implications for how claims about AI diagnostic accuracy should be interpreted. A single-disease classifier is evaluated within a bounded problem. In many such studies, the model effectively answers a question such as, "Is this lesion malignant?" rather than, "What disease does this patient have?" The latter task has a vastly larger hypothesis space and requires discrimination among diseases that may share symptoms but require very different treatments. High sensitivity and specificity for melanoma, lung cancer, or another single condition therefore cannot simply be generalized to diagnostic medicine as a whole. Conversely, open-ended LLM studies can better approximate differential diagnosis but often rely on written vignettes that omit the uncertainty, incomplete information, physical examination, imaging interpretation, evolving symptoms, and interactive questioning characteristic of actual clinical practice.

A second distinction concerns the relationship between AI and physicians. The clinically relevant question may not ultimately be whether AI can defeat physicians in a head-to-head contest, but whether AI can improve diagnostic performance when incorporated into physician decision making. Evidence on this question is mixed and illustrates the importance of workflow. In a randomized clinical trial, Goh et al. (2024) found that physicians given access to GPT-4 did not achieve a statistically significant improvement in diagnostic reasoning compared with physicians using conventional resources, although GPT-4 alone scored significantly higher than the conventional-resources physician group. This finding suggests that strong standalone AI performance does not automatically translate into improved human-AI team performance; clinicians may discount correct AI suggestions, accept incorrect ones, or fail to use the system optimally.

Taken together, the literature supports a progression in medical AI research from specialized pattern-recognition systems toward systems capable of broad diagnostic reasoning. Disease-specific studies provide compelling evidence that AI can match or exceed physicians on some tightly defined diagnostic tasks, particularly in image-based specialties. Open-ended studies extend this evidence by showing that LLMs can generate plausible and often accurate diagnoses across diverse conditions without being restricted to a single disease category. Nevertheless, the open-ended literature remains less mature and presents additional threats to validity, including small case sets, reliance on artificial vignettes, possible training-data contamination, differences in the amount of information available to AI and physicians, and inconsistent definitions of diagnostic success.

Accordingly, rigorous evaluation of AI as a general diagnostic tool should move beyond demonstrations of performance on individual diseases. A strong test should present AI systems and physicians with the same previously unseen cases spanning multiple medical categories, provide equivalent clinical information to both, establish diagnoses using an independent gold standard, and report both overall and disease-category-specific performance. At minimum, outcomes should include top-1 diagnostic accuracy, the presence of the correct diagnosis within a limited differential diagnosis, sensitivity for serious or time-critical conditions, and the nature of clinically consequential errors. Such a design can address a question that single-disease studies cannot answer: whether an AI system can accurately determine what disease a patient has when the diagnosis is genuinely open ended. The present research is directed toward that broader question and toward determining whether AI diagnostic performance remains competitive with physician performance when the system must discriminate across diverse diseases rather than solve a narrowly predefined classification problem.

Method

Study Design and Data Source

This study evaluated the ability of a general-purpose large language model to generate an open-ended differential diagnosis from real clinical records. The evaluation used 1,000 cases drawn from the MIMIC-IV clinical data resources. The model write-up identified the source records as real hospital discharge summaries linked to MIMIC-IV, with clinical information including the chief complaint, history of present illness, past medical history, physical examination, laboratory results, and discharge diagnosis. The MIMIC-IV-ED resource contains deidentified emergency-department data from Beth Israel Deaconess Medical Center, and MIMIC-IV-Note contains deidentified free-text discharge summaries that can be linked to MIMIC-IV (Johnson et al., 2023a, 2023b).

For the present evaluation, the recorded discharge diagnosis was treated as the reference or ground-truth diagnosis. The objective was not to classify a case into a preselected disease label. Instead, the model was given a structured representation of the patient presentation and asked to generate a ranked list of five possible diagnoses. This design was intended to approximate the open-ended diagnostic task described in the Introduction, in which the correct disease must be selected from a broad clinical hypothesis space.

Clinical-Note Processing and Model Pipeline

Each clinical note was parsed into sections using regular-expression patterns designed to recognize section headers such as Chief Complaint, History of Present Illness, and Past Medical History. Relevant sections were assembled into a structured clinical summary. That summary was inserted into a prompt template and submitted to GPT-4o through the OpenAI API. The model returned five diagnoses ranked from most to least likely. Separately, the reference diagnosis was extracted from the Discharge Diagnosis section using a parser designed to accommodate multiple note formats, including Primary/Secondary headers, bulleted lists, and free text.

The model output was evaluated at two levels. Disease-level matching compared each predicted diagnosis with the reference diagnosis using keyword overlap, expansion of common medical abbreviations (for example, “fx” to “fracture” and “NSTEMI” to “non-ST elevation myocardial infarction”), and substring fallback matching. Category-level matching mapped diagnoses to one of 14 organ-system or clinical categories and evaluated whether the predicted diagnosis belonged to the same category as the reference diagnosis. The categories included Other, Cardiovascular, Gastrointestinal, Trauma, Neurological, Musculoskeletal, Infectious Disease, Hematology/Oncology, Psychiatric/Substance, Respiratory, Renal/Urological, Endocrine/Metabolic, Obstetrics/Gynecology, and Dermatological.

Prompting Conditions

Three prompting strategies were compared. In the zero-shot condition, the model received the clinical presentation and was asked directly for its five most likely diagnoses without examples. In the few-shot condition, two example clinical cases and their correct diagnoses were included before the new case. In the chain-of-thought condition, the prompt instructed the model to identify key findings, generate a differential diagnosis, narrow the possibilities, and then rank the final diagnoses. The underlying data-processing and evaluation pipeline was otherwise unchanged across prompting conditions.

Outcome Measures

The primary outcomes were top-1, top-3, and top-5 disease-level diagnostic accuracy. Top-1 accuracy was the percentage of cases for which the recorded diagnosis was the model's first prediction. Top-3 and top-5 accuracy were the percentages for which the recorded diagnosis appeared anywhere among the first three or first five predictions, respectively. Category-level accuracy was calculated analogously after diagnoses were mapped to the 14 clinical categories. The analysis also examined zero-shot disease-level accuracy by category and the stability of aggregate accuracy as the number of evaluated cases increased from 100 to 1,000.

The available results were aggregate accuracy summaries rather than case-level prediction files. Consequently, the present paper reports descriptive comparisons among prompting strategies and categories; inferential tests of paired differences, confidence intervals, sensitivity, specificity, calibration, and error severity could not be computed from the supplied aggregate results alone.

Results

Overall Disease-Level Accuracy

Zero-shot prompting produced the highest disease-level accuracy at every reported cutoff. Across 1,000 cases, zero-shot top-1 accuracy was 55.7%, top-3 accuracy was 69.5%, and top-5 accuracy was 73.0%. Few-shot prompting produced corresponding accuracies of 52.1%, 65.9%, and 69.6%. Chain-of-thought prompting was substantially lower, with top-1 accuracy of 34.2%, top-3 accuracy of 46.7%, and top-5 accuracy of 50.3%. Thus, relative to zero-shot prompting, few-shot prompting was lower by 3.6 percentage points at top-1 and 3.4 points at top-5, whereas chain-of-thought was lower by 21.5 points at top-1 and 22.7 points at top-5.

Table 1: Disease-Level Diagnostic Accuracy by Prompting Strategy

Strategy	Top-1	Top-2	Top-3	Top-5
Zero-Shot	55.7%	65.3%	69.5%	73.0%
Few-Shot	52.1%	60.9%	65.9%	69.6%
Chain-of-Thought	34.2%	42.3%	46.7%	50.3%

Accuracy as a Function of Evaluation Sample Size

The zero-shot accuracy curves showed modest variation at smaller sample sizes and then became comparatively stable. Top-1 accuracy ranged from approximately 54% to 58% across the displayed sample sizes and converged to 55.7% at 1,000 cases. Top-3 accuracy declined from the

mid-70% range at smaller sample sizes and stabilized near 69% to 70%, ending at 69.5%. Top-5 accuracy similarly declined from the upper-70% range at smaller sample sizes and stabilized near 73%, ending at 73.0%. The pattern supports the descriptive conclusion in the model notes that aggregate estimates were relatively stable after approximately 300 to 400 evaluated cases.

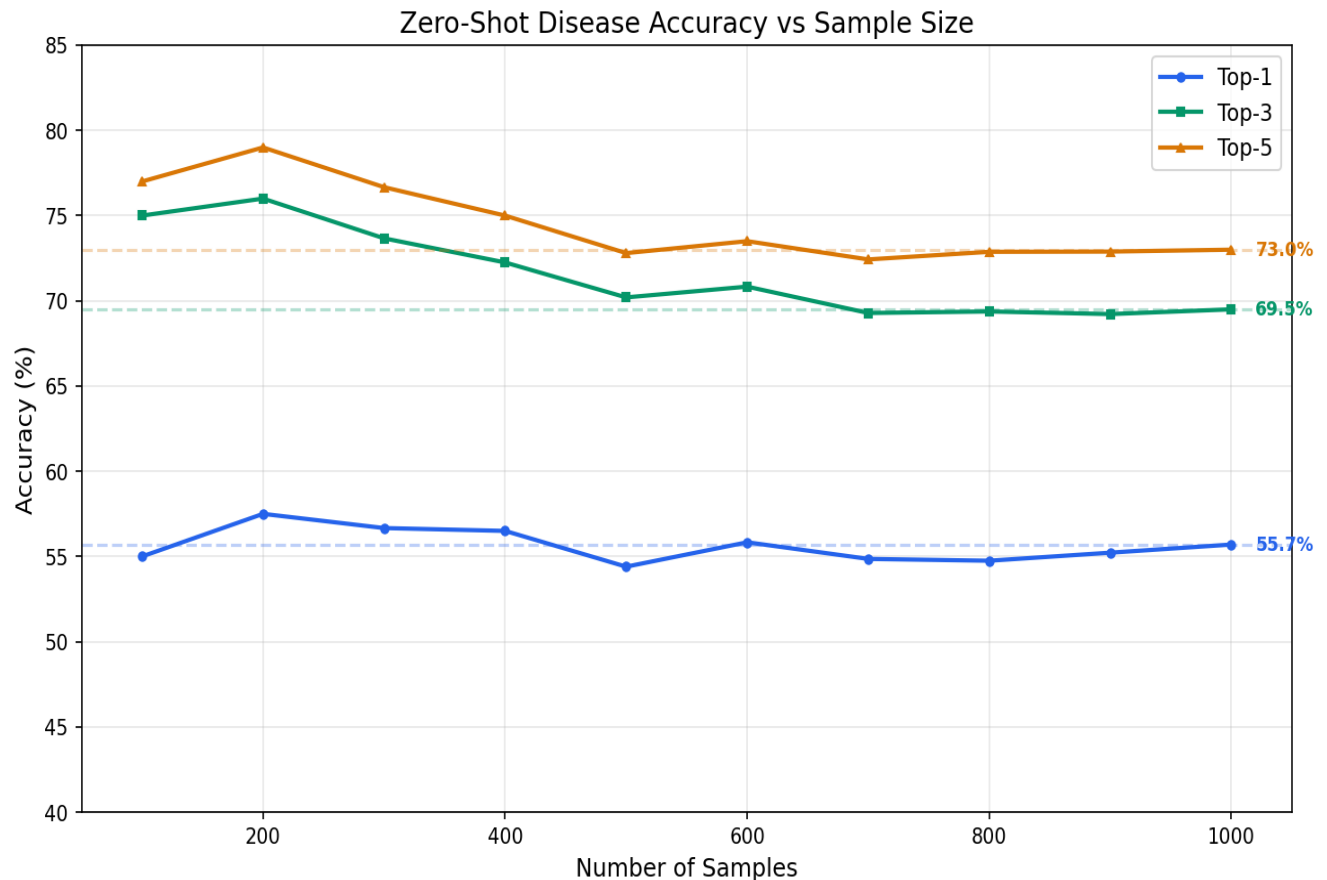


Figure 1. Zero-shot disease-level Top-1, Top-3, and Top-5 accuracy as the cumulative evaluation sample increased from 100 to 1,000 cases. Dashed horizontal lines indicate the final 1,000-case values.

Category-Level Accuracy

When exact diagnoses were collapsed into broader clinical categories, accuracy increased for all three prompting strategies. Zero-shot category-level top-1 accuracy was 72.3%, compared with 55.7% for exact disease-level top-1 accuracy. Few-shot category-level top-1 accuracy was 69.4%, compared with 52.1% at the disease level. Chain-of-thought category-level top-1 accuracy was 56.2%, compared with 34.2% at the disease level. For zero-shot prompting, category-level accuracy reached 83.3% at top-3 and 86.0% at top-5. These results indicate that

the model often identified the correct clinical domain even when it did not rank the exact recorded diagnosis first.

Category-Level Accuracy				
Strategy	Top-1	Top-2	Top-3	Top-5
Zero-Shot	72.3%	80.0%	83.3%	86.0%
Few-Shot	69.4%	77.1%	80.5%	83.1%
Chain-of-Thought	56.2%	67.0%	70.1%	73.1%

Figure 2. Category-level accuracy by prompting strategy. The broader category criterion yields higher accuracy than exact disease-level matching.

Performance by Disease Category

Zero-shot disease-level performance varied substantially across diagnostic categories. At top-5, the highest reported accuracies were Musculoskeletal (98%), Endocrine/Metabolic (91%), Obstetrics/Gynecology (86%), Trauma (82%), and Hematology/Oncology (82%). Intermediate values included Gastrointestinal (79%), Respiratory (79%), Renal/Urological (78%), Psychiatric/Substance (77%), Neurological (74%), and Infectious Disease (72%). Lower values were observed for Cardiovascular (66%) and Other (63%). The Dermatological category had only seven cases and therefore produced an especially imprecise descriptive estimate (71% top-5).

Top-1 performance also differed sharply by category. Musculoskeletal cases had 97% top-1 accuracy, whereas Psychiatric/Substance cases had 40%, Other cases 43%, and Neurological cases 46%. The category counts were highly unequal: Other contained 334 cases and Cardiovascular 140, whereas several categories contained fewer than 30 cases. Accordingly, apparent differences among small categories should be interpreted cautiously because a few cases can materially change their percentages.

Per-Category Disease Accuracy (Zero-Shot)

Category	N	Top-1	Top-3	Top-5
Other	334	43%	57%	63%
Cardiovascular	140	51%	63%	66%
Gastrointestinal	107	62%	75%	79%
Trauma	79	73%	80%	82%
Neurological	70	46%	73%	74%
Musculoskeletal	63	97%	97%	98%
Infectious Disease	36	58%	72%	72%
Hematology/Oncology	33	55%	79%	82%
Psychiatric/Substance	30	40%	77%	77%
Respiratory	29	59%	76%	79%
Renal/Urological	27	70%	78%	78%
Endocrine/Metabolic	23	78%	87%	91%
Obstetrics/Gynecology	22	82%	86%	86%
Dermatological	7	57%	71%	71%

Figure 3. Zero-shot disease-level accuracy by clinical category, including category sample sizes and Top-1, Top-3, and Top-5 accuracy.

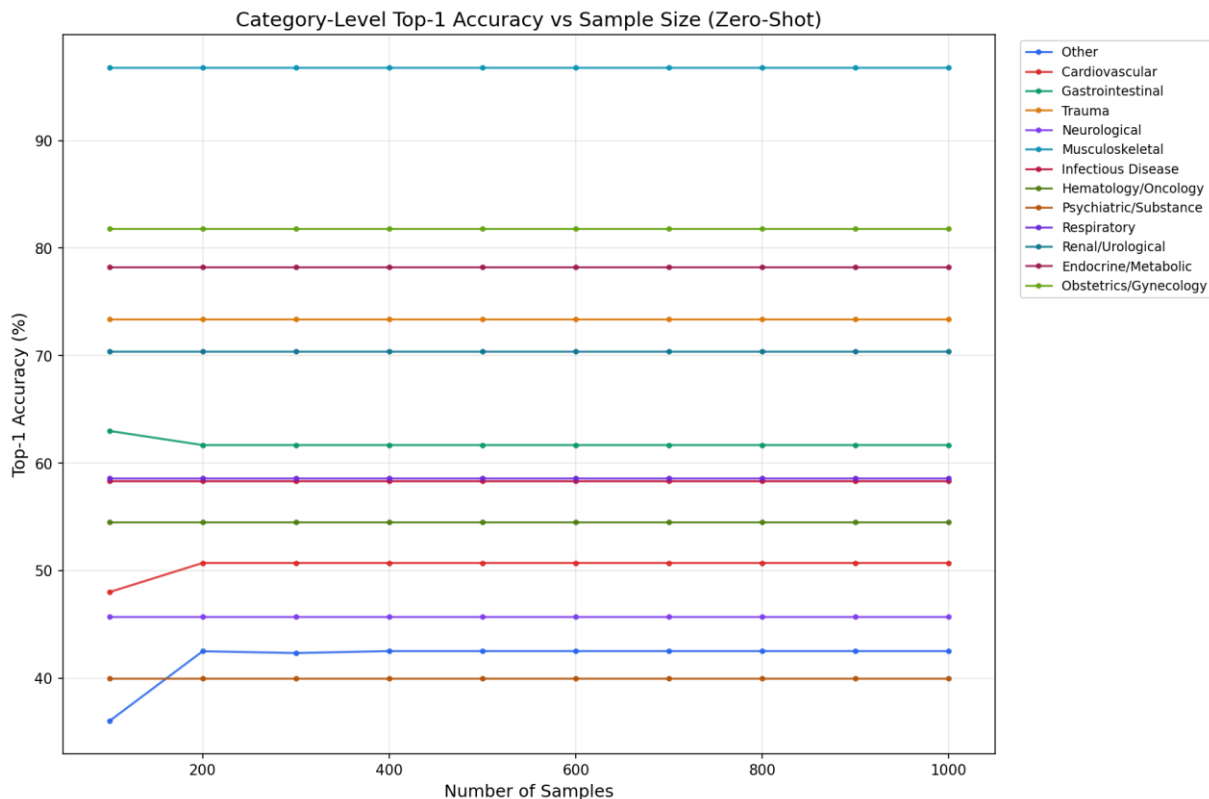


Figure 4. Category-level zero-shot Top-1 accuracy plotted against cumulative sample size. Category estimates should be interpreted in light of the unequal number of cases available in each category.

Prompt Efficiency

The supplied model analysis also reported a computational-efficiency advantage for zero-shot prompting. Approximate token use per API call was 573 tokens for zero-shot, 670 for few-shot, and 1,072 for chain-of-thought. Thus, the prompting strategy with the highest observed diagnostic accuracy was also the least token-intensive of the three conditions in this implementation.

Discussion

The principal finding was that GPT-4o achieved meaningful open-ended diagnostic performance on a heterogeneous set of real clinical records, but performance depended strongly on both the evaluation criterion and the prompting strategy. Under the best-performing zero-shot condition, the model selected the recorded diagnosis first in 55.7% of 1,000 cases and included it within five predictions in 73.0%. When the criterion was broadened from exact disease identity to the

correct clinical category, zero-shot top-1 accuracy increased to 72.3% and top-5 accuracy to 86.0%. This gap is clinically and methodologically important: it suggests that many errors were not failures to recognize the general organ system or clinical domain, but failures to distinguish among diagnoses within that domain.

The zero-shot results are broadly consistent with the Introduction's argument that open-ended diagnosis is a more demanding test than narrow disease classification. Hirose et al. (2023), for example, reported that GPT-4 placed the correct diagnosis first in 60% of complex case-derived vignettes and within the top five in 81%. The present 55.7% top-1 and 73.0% top-5 zero-shot values are somewhat lower, but the current evaluation used real clinical records rather than a small curated vignette set. The results therefore add evidence from a different and potentially more heterogeneous test environment while also underscoring that performance observed on selected vignettes should not automatically be expected on routine clinical records.

A second important result was the ordering of the prompting strategies. Few-shot prompting did not improve performance over zero-shot prompting, and the tested chain-of-thought prompt performed substantially worse. This finding does not establish that examples or structured reasoning are intrinsically harmful. Rather, it shows that the particular prompting implementations tested here were less effective. The fixed few-shot examples may have biased predictions toward diseases represented in those examples, while the chain-of-thought instruction may have encouraged the model to elaborate alternative explanations and move a correct initial hypothesis downward in the ranking. A useful follow-up experiment would compare fixed examples with dynamically selected examples matched to the patient's clinical category and would test an answer-first format in which the ranked differential is produced before a concise explanation.

The marked heterogeneity across disease categories is also noteworthy. Musculoskeletal cases were recognized with very high accuracy, while neurological, cardiovascular, psychiatric/substance, and miscellaneous cases were more difficult. This is plausible because some clinical presentations are relatively specific whereas others share symptoms across many competing diagnoses. The broader diagnostic-error literature similarly shows that diagnostic performance varies by condition and setting; for example, systematic and disease-specific studies have documented diagnostic difficulty in emergency care, depression, urinary tract infection, fractures, and cancer recognition (Mitchell et al., 2009; Newman-Toker et al., 2022; Shallcross et al., 2020; Singh et al., 2009, 2010). The present category analysis should therefore be viewed as a starting point for identifying where an LLM may be more or less useful, not as evidence that the model is uniformly accurate across medicine.

The sample-size analysis provides some reassurance about the stability of the aggregate zero-shot estimates. The curves changed more substantially between the earliest sample sizes but showed comparatively small changes after roughly 300 to 400 cases. Stability, however, should not be confused with validity. A stable estimate can still reflect systematic bias in the case-selection process, reference labels, prompt design, or matching algorithm. In addition, stability of the overall estimate does not imply stability within categories that contain relatively few observations.

The results also have implications for how diagnostic AI should be evaluated. Top-1 accuracy is a stringent measure because it requires the model to rank the reference diagnosis first. Top-3 and top-5 accuracy better reflect a differential-diagnosis use case in which an AI system could alert a clinician to a diagnosis that might otherwise be overlooked. At the same time, top-k accuracy does not measure whether incorrect alternatives are dangerous, whether the ordering is well calibrated, or whether the model would recommend appropriate testing or treatment. Future evaluation should therefore supplement top-k accuracy with clinically weighted error analysis, confidence calibration, and measures of performance on serious or time-critical diagnoses.

Limitations

Several limitations materially affect interpretation. First, the reference diagnosis was derived from the clinical record and was treated as ground truth. MIMIC-IV-ED diagnoses are recorded after the ED encounter and may reflect coding and documentation practices rather than an independently adjudicated diagnostic gold standard. The official MIMIC-IV-ED documentation notes that diagnosis-table labels are coded after discharge and used for billing. In the present pipeline, the reference diagnosis was extracted from the Discharge Diagnosis section of the note, so accuracy should be interpreted as agreement with the recorded diagnosis rather than proof of biological diagnostic truth.

Second, the source material consisted of completed clinical notes/discharge summaries. The model write-up states that relevant sections were assembled into a structured clinical summary and that the discharge diagnosis was separately extracted for scoring; however, the supplied documentation does not provide the exact final prompt text or a field-level audit demonstrating that all direct or indirect diagnostic leakage was excluded from every model input. The development notes themselves identify the “reviewing a hospital discharge” framing as a possible source of answer leakage. A stronger prospective-style evaluation would construct the model input exclusively from information available at the time of initial presentation or at a predefined clinical decision point and would verify programmatically that diagnosis-bearing sections are withheld.

Third, disease-level correctness depended on an automated text-matching procedure using keyword overlap, abbreviation expansion, and substring matching. Such a procedure can create both false matches and false mismatches when diagnoses are synonyms, differ in specificity, or contain overlapping terminology. Manual adjudication of a representative sample, ideally by blinded clinicians, would help quantify matching error.

Fourth, the 1,000-case aggregate was not evenly distributed across disease categories. Several categories had small sample sizes, including Dermatological ($n = 7$), Obstetrics/Gynecology ($n = 22$), Endocrine/Metabolic ($n = 23$), and Renal/Urological ($n = 27$). Category percentages based on these small counts have considerable sampling uncertainty. Fifth, the supplied results did not include case-level predictions, so statistical comparisons among prompting conditions could not be performed and confidence intervals could not be calculated here. If the same cases were evaluated under all three prompts, paired analyses would be preferable to tests that treat the conditions as independent samples.

Finally, MIMIC-IV is a retrospective, deidentified dataset from a single academic medical center. Its patients, documentation practices, disease prevalence, and clinical workflows may differ from those of other institutions and contemporary prospective settings. The findings therefore require external validation before being generalized to other hospitals, outpatient medicine, or real-time clinical use. The study evaluates diagnostic prediction performance and does not establish that the system is safe or effective for autonomous patient care.

Future Research

The present results suggest several concrete next steps. The prompt should be reframed so that the model acts as an emergency-department attending physician evaluating a newly presenting patient rather than reviewing a completed discharge. The ranked diagnoses can be requested before any explanation to test whether this prevents extended reasoning from displacing a strong initial prediction. Few-shot examples can be selected dynamically from clinically similar categories rather than fixed examples. Confidence scores can also be collected for each prediction and evaluated for calibration rather than used simply to discard uncertain cases.

A subsequent study should retain case-level outputs and use a prespecified test set that is untouched during prompt development. Each prompting strategy should be evaluated on the identical cases, with confidence intervals and paired statistical comparisons. A blinded clinician review could classify errors by severity and determine whether a predicted diagnosis is clinically equivalent to, broader than, narrower than, or genuinely inconsistent with the reference diagnosis. External validation on data from other institutions and prospective evaluation using

only information available before diagnosis would provide a substantially stronger test of generalizability and clinical utility.

Conclusion

In this 1,000-case evaluation using real clinical records, GPT-4o showed its strongest performance with the simplest prompting strategy. Zero-shot prompting achieved 55.7% top-1 and 73.0% top-5 exact disease-level accuracy, while category-level accuracy reached 72.3% at top-1 and 86.0% at top-5. Few-shot prompting was modestly less accurate, and the tested chain-of-thought prompt was markedly less accurate. Performance varied widely by clinical category and stabilized descriptively as the overall sample increased. These findings support the feasibility of broad, open-ended AI-assisted differential diagnosis while also demonstrating that accuracy is highly sensitive to prompt design, diagnostic category, scoring method, and the definition of ground truth. Stronger leakage controls, case-level statistical analysis, clinician adjudication, and external prospective validation are necessary before the results can be interpreted as evidence of clinical readiness.

References

- Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>
- Hirosawa, T., Kawamura, R., Harada, Y., Mizuta, K., Tokumasu, K., Kaji, Y., Suzuki, T., & Shimizu, T. (2023). ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: Diagnostic accuracy evaluation. *JMIR Medical Informatics*, 11, e48808. <https://doi.org/10.2196/48808>
- Ito, N., Kadomatsu, S., Fujisawa, M., Fukaguchi, K., Ishizawa, R., Kanda, N., Kasugai, D., Nakajima, M., Goto, T., & Tsugawa, Y. (2023). The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: Evaluation study. *JMIR Medical Education*, 9, e47532. <https://doi.org/10.2196/47532>
- Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., & Horng, S. (2023a). MIMIC-IV-ED (version 2.2). PhysioNet. RRID:SCR_007345. <https://doi.org/10.13026/5ntk-km72>

- Johnson, A., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023b). MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet. RRID:SCR_007345. <https://doi.org/10.13026/1n74-ne17>
- Mitchell, A. J., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: A meta-analysis. *The Lancet*, 374(9690), 609–619. [https://doi.org/10.1016/S0140-6736\(09\)60879-5](https://doi.org/10.1016/S0140-6736(09)60879-5)
- Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., Wilson, L. M., Jia, Y., Omron, R., Tharmarajah, S., Guerin, L., Bastani, P. B., Fracica, E. A., Kotwal, S., & Robinson, K. A. (2022). Diagnostic errors in the emergency department: A systematic review (Comparative Effectiveness Review No. 258; AHRQ Publication No. 22(23)-EHC043). Agency for Healthcare Research and Quality. <https://doi.org/10.23970/AHRQEPCCER258>
- Rutledge, G. W. (2024). Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. *Learning Health Systems*, 8(3), e10438. <https://doi.org/10.1002/lrh2.10438>
- Salinas, M. P., Sepúlveda, J., Hidalgo, L., Peirano, D., Morel, M., Uribe, P., Rotemberg, V., Briones, J., Mery, D., & Navarrete-Dechent, C. (2024). A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *npj Digital Medicine*, 7, Article 125. <https://doi.org/10.1038/s41746-024-01103-x>
- Shallcross, L. J., Rockenschaub, P., McNulty, D., Freemantle, N., Hayward, A., & Gill, M. J. (2020). Diagnostic uncertainty and urinary tract infection in the emergency department: A cohort study from a UK hospital. *BMC Emergency Medicine*, 20, Article 40. <https://doi.org/10.1186/s12873-020-00333-y>
- Singh, H., Daci, K., Petersen, L. A., Collins, C., Petersen, N. J., Shethia, A., & El-Serag, H. B. (2009). Missed opportunities to initiate endoscopic evaluation for colorectal cancer diagnosis. *The American Journal of Gastroenterology*, 104(10), 2543–2554. <https://doi.org/10.1038/ajg.2009.324>
- Singh, H., Hirani, K., Kadiyala, H., Rudomiotov, O., Davis, T., Khan, M. M., & Wahls, T. L. (2010). Characteristics and predictors of missed opportunities in lung cancer diagnosis: An electronic health record–based study. *Journal of Clinical Oncology*, 28(20), 3307–3315. <https://doi.org/10.1200/JCO.2009.25.6636>

Wang, T.-W., Hong, J.-S., Chiu, H.-Y., Chao, H.-S., Chen, Y.-M., & Wu, Y.-T. (2024). Standalone deep learning versus experts for diagnosis lung cancer on chest computed tomography: A systematic review. *European Radiology*, 34, 7397–7407. <https://doi.org/10.1007/s00330-024-10804-6>