Ahmed Medhat Zayed*, Glynis Frans and Nicolas Delvaux

Evaluating large language models as clinical laboratory test recommenders in primary and emergency care: a crucial step in clinical decision making

https://doi.org/10.1515/cclm-2025-0647
Received May 28, 2025; accepted July 29, 2025;
published online August 14, 2025

Abstract

Objectives: Large language models (LLMs), such as OpenAI's GPT-4o, have demonstrated considerable promise in transforming clinical decision support systems. In this study, we focused on a single but crucial task of clinical decision-making: laboratory test ordering.

Methods: We evaluated the self-consistency and performance of GPT-4o as a laboratory test recommender for 15 simulated clinical cases of different complexities across primary and emergency care settings. Through two prompting strategies – zero-shot and chain-of-thought – the model's recommendations were evaluated against expert consensus-derived gold-standard laboratory test orders categorized into essential and conditional test orders.

Results: We found that GPT-4o exhibited high self-consistency across repeated prompts, surpassing the consistency observed among individual expert orders in the earliest round of consensus. Precision was moderate to high for both prompting strategies (68–82 %), although relatively lower recall (41–51 %) highlighted a risk of underutilization. A detailed analysis of false negatives (FNs) and false positives (FPs) could explain some gaps in recall and precision. Notably, variability in recommendations centered primarily on conditional tests, reflecting the broader diagnostic uncertainty that can arise in diverse clinical contexts. Our analysis revealed that neither prompting strategy, case complexity, nor clinical context significantly affected GPT-4o's performance.

Conclusions: This work underscores the promise of LLMs in optimizing laboratory test ordering while identifying gaps for enhancing their alignment with clinical practice. Future research should focus on real-world implementation, integrating clinician feedback, and ensuring alignment with local test menus and guidelines to improve both performance and trust in LLM-driven clinical decision support.

Keywords: large language models; artificial intelligence; laboratory test utilization; clinical decision support

Introduction

Healthcare delivery resembles an intricate puzzle, where each medical test represents a crucial piece, and the completed image reflects a patient's comprehensive care pathway. Clinical laboratory testing is the most frequently performed medical activity in healthcare with 80 % of clinical practice guidelines recommending the use of laboratory tests as the standard of care for establishing a diagnosis, screening for or managing diseases [1–3]. However, there are substantial concerns regarding laboratory test misutilization, with underutilization, although less studied, occurring more frequently than overutilization, at 45 % compared to 21 % [4].

The root causes of laboratory test misutilization are multifaceted, including the rapid evolution of medical knowledge and the expansion of laboratory test menus, which necessitate ongoing education for healthcare providers [5, 6]. This complexity is highlighted by surveys showing primary care physicians' uncertainty regarding 15 % of their laboratory test orders and 8 % of their interpretations [5]. Improving laboratory test utilization is essential for improving the quality of clinical care, reducing

***Corresponding author: Ahmed Medhat Zayed**, Department of Public Health and Primary Care, Faculty of Medicine, Academic Center for General Practice, KU Leuven, Kapucijnenvoer 7 blok h – box 7001, 3000, Leuven, Belgium; and Laboratory Medicine Department, Menoufia University National Liver Institute, Shebin El-Kom, Egypt,
E-mail: ahmed.zayed@kuleuven.be. https://orcid.org/0000-0001-7797-1655
Glynis Frans, Department of Laboratory Medicine, University Hospitals Leuven, Leuven, Belgium; and Department of Microbiology, Immunology and Transplantation, Clinical and Diagnostic Immunology, KU Leuven, Leuven, Belgium. https://orcid.org/0000-0001-8528-5719
Nicolas Delvaux, Department of Public Health and Primary Care, Faculty of Medicine, KU Leuven, Leuven, Belgium

diagnostic errors, and ensuring the cost-effectiveness of the care delivered [4].

In response to these challenges, Clinical Decision Support (CDS) systems have emerged as vital tools for optimizing laboratory test utilization [7, 8]. Recent advancements in Artificial Intelligence (AI), big data analytics, and Machine Learning (ML) have further enhanced the capabilities of CDS systems, paving the way for more effective interventions [9]. At the forefront of this technological revolution are large language models (LLMs) such as OpenAI's GPT-4 or Google's Gemini models, which have demonstrated human-level performance across various professional and academic evaluations [10, 11]. These models have showcased remarkable abilities in understanding, recall, and applying medical knowledge, as evidenced by their proven capabilities in passing medical licensing exams, interpreting lab results, and diagnostic reasoning [12–17]. Such capabilities indicate the potential of LLMs to mimic the clinical reasoning processes, ensuring ordering the right laboratory test for the right patient at the right time. This advancement promises to enhance the healthcare landscape with Next-Generation CDS systems, fostering personalized, evidence-based patient care.

In this study, we sought to investigate OpenAI's GPT-4o model's utility as a clinical decision support tool to recommend laboratory tests. By simulating the intricate process of clinical reasoning that combines patient data analysis with clinical guidelines across a variety of simulated clinical case scenarios in primary and emergency care, our evaluation provides insight into the model's potential. These laboratory test orders recommended by GPT-4o were evaluated against gold-standard laboratory test orders, established through expert consensus for each case.

Additionally, as advised by Perlis and Fihn, we examined two distinct aspects of the model's performance: first, the effectiveness of various prompt engineering strategies, including zero-shot prompting and chain-of-thought (CoT) reasoning, in optimizing the model's recommendations; and second, the variability in model outputs when the same prompt was used repeatedly for the same clinical case [18]. zero-shot prompting uses no examples or reasoning steps, while CoT prompting includes step-by-step reasoning, which has been shown to improve LLM performance on complex tasks [19, 20]. By investigating both the impact of different prompting strategies and the consistency of the model's outputs, this study aims to establish best practices for interfacing with LLMs in clinical settings to maximize the utility and reliability of LLMs as clinical laboratory test recommenders.

Materials and methods

Simulated clinical cases

Fifteen simulated clinical cases were generated using GPT-4o, including 10 cases from General Practice (GP) visits and five cases from Emergency Room (ER) admissions. These cases represented diverse clinical contexts that end up with a challenging lab test order. Each case included comprehensive patient information: demographics (age, gender), past medical history, presenting symptoms, physical examination findings, and any relevant prior diagnostic data. The complexity of cases was classified as simple, moderate, or complex based on clinical reasoning requirements. The prompt used to generate the cases and the full list of resulting simulated cases are available in Supplementary Appendix 1. To minimize potential bias, the clinical case generation was conducted using GPT-4o on the web rather than through APIs to ensure being completely separated from the test recommendation phase. Since this study did not involve real-world patient data, no Ethical Approval was required.

Establishment of gold-standard lab test orders

To define the gold standard for necessary laboratory tests, three expert panels were formed, each comprising three clinicians (GPs or ER specialists) and one laboratory medicine specialist. At least one clinician in each panel was at a mid-senior or senior level with a minimum of eight years of clinical experience.

The Delphi method was employed to achieve expert consensus on categorizing laboratory tests for each clinical case as essential, conditional, or unnecessary [21]. Essential tests were defined as those critical for diagnosis or management, conditional tests as those that provide useful but non-critical information, and unnecessary tests as those that add no value to patient care, as detailed in Supplementary Table 1.

The consensus process included three rounds of independent evaluations and feedback, followed by a final consensus meeting to resolve any remaining disagreements. The consensus was defined as at least 75 % agreement among the panellists, with three out of four experts agreeing on the categorization of a test. Test recommendations were guided by evidence-based clinical guidelines, clinical judgment,

cost-effectiveness, and patient safety considerations. The laboratory test menu available for ordering was based on the Leuven University Hospitals' (UZ Leuven) laboratory test menu [22].

Prompting the model for lab test recommendations

With the public release of ChatGPT in November 2022, numerous LLMs have been introduced. For this study, we selected GPT-4o due to its advanced ranking on the Public LLMs leaderboard, as assessed by Vectara's Hughes Hallucination Evaluation Model [23, 24]. At the time of model utilization, GPT-4o was identified as the model with the lowest hallucination rate, making it well-suited for clinical decision-support tasks.

We accessed GPT-4o model using OpenAI's Assistant Application Programming Interface (API), which enables interactions with models, tools, and files to respond to user queries [25]. Through OpenAI's enterprise API platform, we created two assistants that utilized the same tools and model configurations but differed in their instructions to represent the two prompting strategies tested; zero-shot and CoT. zero-shot prompting means that the prompt used to interact with the model does not contain examples, demonstrations, or reasoning steps [19]. We directly instructed the zero-shot assistants to perform the task of recommending laboratory test with clarifying the input and the required output with the predefined criteria of different laboratory test order categories.

On the other hand, CoT prompting includes stepwise intermediate reasoning steps, which showed to improve the performance of LLMs on complex reasoning tasks [20]. We instructed the CoT assistant to break down the task into smaller, logically ordered reasoning steps, explicitly outlining the rationale for each step in the decision-making process. The steps included systematically processing the case information, using step-by-step hypothesis deduction to establish a differential diagnosis, and applying logical and sequential decision-making to differentiate between essential and conditional tests while minimizing unnecessary tests. In the CoT instructions, we incorporated approaches and examples for establishing a diagnosis based on laboratory tests as described by Wians [3]. The prompts used to instruct both zero-shot and CoT assistants are detailed in Supplementary Appendix 2.

To ensure consistent and reliable outputs, we configured a key model hyperparameter, sampling temperature, which controls the determinism of the model's responses and ranges between 0 and 2 [26]. Lower temperature values favor deterministic outputs by prioritizing the highest probability tokens, while higher values increase randomness and diversity in responses. Given the context of our task as clinical decision support, we set the sampling temperature to 0.1 to maximize the generation of exact and reliable answers.

For each clinical case, the assistants were repeatedly prompted across 10 distinct threads to evaluate consistency. A thread represents a conversation session with the assistant, retaining context for seamless interaction [25]. Both assistants were instructed to provide the list of recommended tests in JavaScript Object Notation (JSON) format to be stored as a database. Each recommended test included fields for the case ID, thread ID, test type (essential or conditional), test name, and test code from the test menu. Additional fields included the order category (screening, monitoring, or diagnosis), and clinical reasoning, which provided an explanation for ordering the test in the specific case. The output also included an evidence-based rule citing guidelines from reputable medical associations and references pointing to relevant clinical guidelines or publications. This structured output facilitated the systematic evaluation of the model's recommendations.

Preprocessing of model recommendations

To ensure alignment with the local laboratory test menu, the UZ Leuven Test Menu, we utilized OpenAI's File Search tool to attach the same menu that was used by experts. However, during initial trials, the model's outputs did not fully adhere to the attached menu. To address this limitation, we developed a third assistant designed to automatically correct and standardize test codes in the recommendation outputs of the two recommender assistants.

Following this step, the recommended tests generated by the zero-shot and CoT assistants, post-correction by the third assistant, underwent further preprocessing. This included both automated and manual validation to ensure complete alignment with the test menu. As a result, 91.6 % of the recommendations corresponded to valid tests listed in the UZ Leuven Test Menu. Instances of recommendations with missing or incorrect codes (3.3 and 2.5 %, respectively) were either codified or corrected for inclusion in the evaluation. Conversely, recommendations for non-laboratory tests (e.g., Chest X-rays) or tests not part of the menu (0.7 and 1.9 %, respectively) were excluded from the evaluation. This meticulous preprocessing step ensured that the final dataset was robust and strictly adhered to the defined laboratory test local menu.

Evaluation of laboratory test recommendations

We employed statistical methods to quantitatively evaluate the concordance between GPT-4o's recommendations and the gold-standard laboratory test orders. The assessment was conducted across individual test categories (essential and conditional) as well as for the combined set of all recommended tests, regardless of category. Figure 6 in the result illustrates how numbers of true positive (TP), true negative (TN), false negative (FN) and false positive (FP) tests were calculated.

Given the extensive test menu (1,259 tests), a disproportionately large number of unnecessary tests (TNs) were identified, resulting in a highly imbalanced distribution of class labels. This imbalance disproportionately inflated metrics dependent on TNs, such as specificity and overall accuracy, which primarily reflected the abundance of correctly excluded tests rather than the model's actual discriminative capabilities. To address this, we prioritized the use of precision and recall as evaluation metrics.

Precision measured the correctness of the recommended tests, crucial for avoiding overutilization. On the other hand, recall measured the proportion of gold-standard tests identified by the model, ensuring critical tests are not missed, minimizing the risk of underutilization. The F1 score combines precision and recall to provide a balanced evaluation of the model's performance in clinical decision-making.

Statistical analysis and implementation tools

The interaction with OpenAI APIs was executed using Python (version 3.12.6) to operate the zero-shot and CoT assistants, leveraging the OpenAI library. The generated recommended tests was evaluated using statistical analyses conducted in R (version 4.4.2).

Statistical Methods included calculating Jaccard similarity to assess the consistency of repeated model outputs, followed by pairwise comparisons using the Wilcoxon rank-sum test with Holm's method for p-value adjustment to account for multiple comparisons. These comparisons evaluated differences in mean Jaccard similarity between recommenders within each test category and between test categories within each recommender. Metrics such as precision, recall, and the F1 score were computed to quantify the concordance between the model's recommendations and the gold-standard orders. Additionally, a linear mixed-effects (LME) model was fitted using restricted maximum likelihood (REML) to evaluate the influence of fixed and random effects on F1 scores, while accounting for the hierarchical structure of the data.

Results

Figure 1 summarizes the study design and key findings.

Model consistency in repeat prompting

Before evaluating the model's responses against the gold standard, we first assessed the consistency of model's outputs across repeated prompts (Figure 2). The analysis of mean Jaccard similarity revealed overall high agreement in the model's combined list of recommended tests using both recommenders: CoT (66.5 %, SD: 16 %) and zero-shot (62 %, SD: 13 %). For comparison, we calculated the same metric to evaluate the consistency of human experts' individual responses during the first Delphi round of the consensus process to establish the gold-standard orders, which showed substantially lower consistency (41.4 %, SD: 11 %).

Both recommenders demonstrated significantly higher consistency than the individual experts' responses across both the combined list and individual test categories ($p<0.005$), while there were no statistically significant differences in consistency between CoT and zero-shot. The conditional test category showed greater variability across both recommenders and human experts compared to other categories, with significantly lower consistency than both the essential test category and the combined list ($p<0.001$). This highlights the intrinsic variability of recommendations in conditional tests, even among human experts.

Modelling evaluation outcomes: fixed and random effects

The evaluation of GPT-4o's laboratory test recommendations against gold-standard orders used the F1 score as a key metric, revealing variability across different clinical cases and prompting conditions (Figure 3). This variability was observed at both intra-case and inter-case levels. Intra-case variability referred to fluctuations in F1 scores within a single case across different threads generated by the same recommender, while inter-case variability highlighted differences in average F1 scores between cases.

To quantify these influences while appropriately modelling correlated observations within and across cases, we employed an LME model. This model included fixed effects (predictors of interest) for recommender, context,

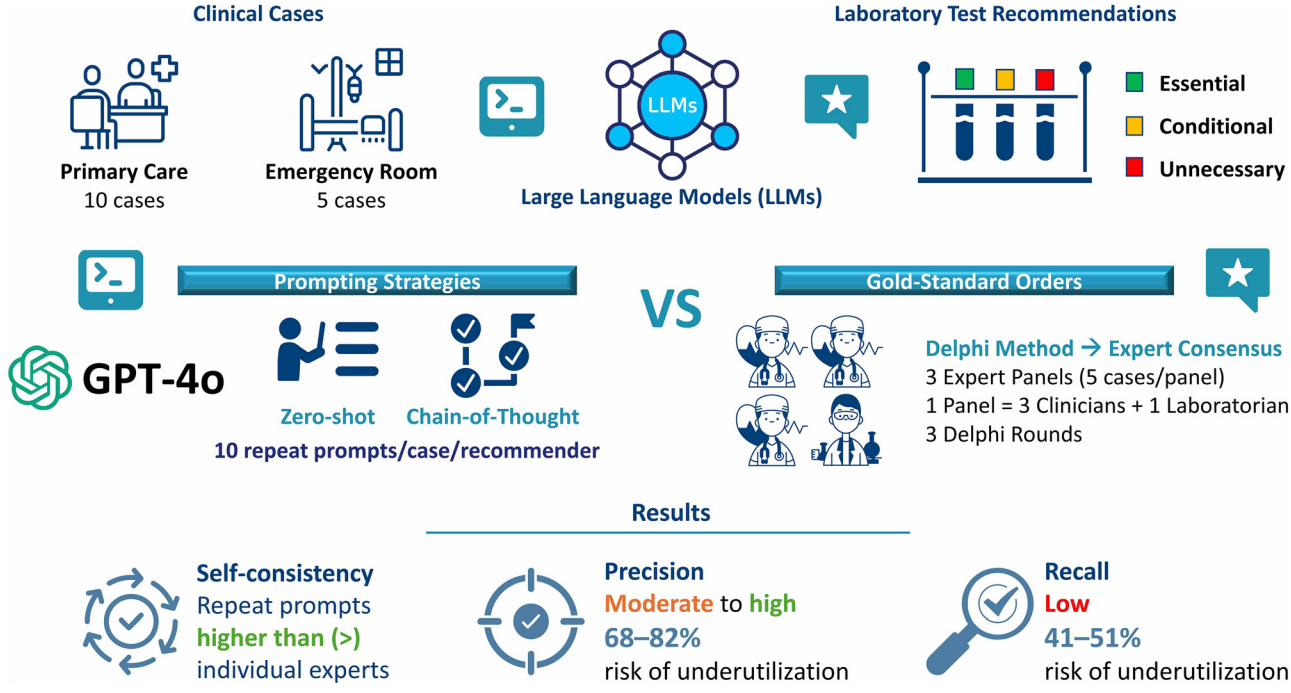

Figure 1: Overview of the study design and key findings evaluating GPT-4o, a large language model (LLM), for clinical laboratory test recommendations. The study simulated 15 clinical cases — 10 from primary care and 5 from emergency room settings — and used two prompting strategies (zero-shot and chain-of-thought) to assess GPT-4o's test recommendation performance. Model outputs were compared against expert consensus-derived gold-standard test orders, developed through the Delphi method (3 expert panels, 3 rounds each). The evaluation focused on self-consistency, precision, and recall. GPT-4o demonstrated higher self-consistency than individual experts, with moderate-to-high precision (68–82 %) and lower recall (41–51 %), indicating a potential risk of underutilization. Test recommendations were categorized as essential, conditional, or unnecessary.

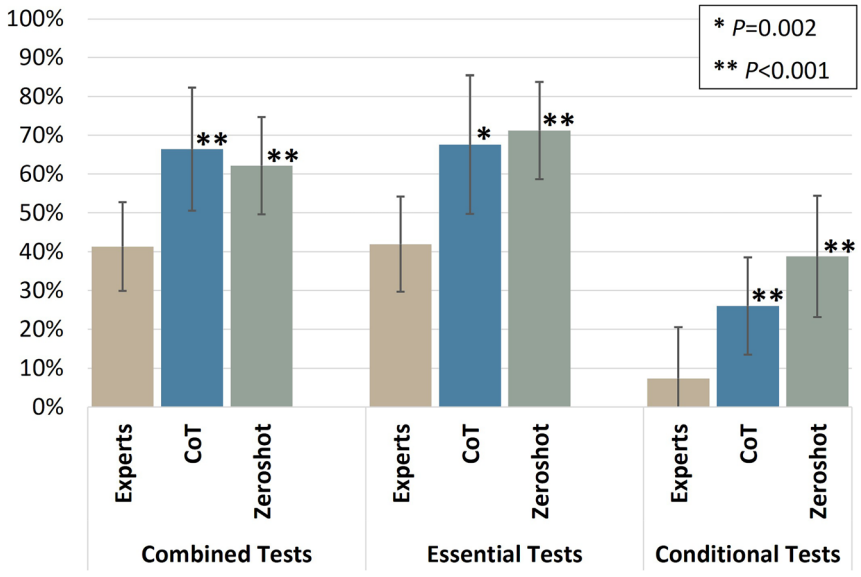

Figure 2: Consistency of model recommended tests in repeat prompting. The mean Jaccard similarity between lists of recommended lists in each category is shown for both prompting strategies (zero-shot and CoT) compared to individual responses from human experts. Both prompting strategies showed significantly higher consistency than individual experts across the combined list and both test categories ($p<0.005$), with no statistically significant difference between CoT and zero-shot. The conditional test category had greater variability than both the essential category and the combined list, highlighting the inherent variability in conditional recommendations, even among human experts.

complexity, and test category as well as random effects to accommodate variability at three hierarchical levels: case ID, thread ID, and repeated prompting ID; a specific variable that groups all repeated threads belonging to the same conditions (i.e., specific combinations of case ID, recommender, and test category).

Fixed effects analysis of the LME model showed that the prompting strategy (recommender) exerted a minor but

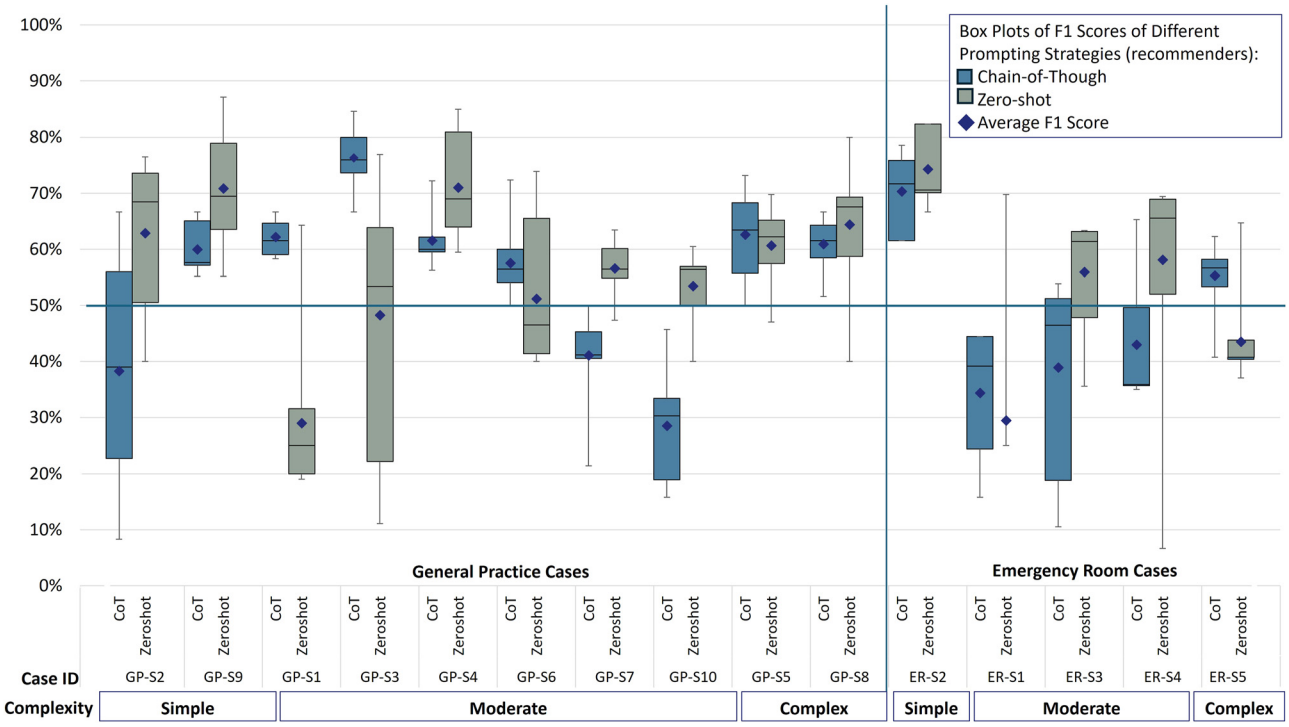

Figure 3: F1 scores of recommender GPT in all cases against gold-standard (combined list of tests). Box plots show the distribution of F1 scores across the threads (10 repeated prompts) of all cases of different clinical contexts (GP or ER) and different complexities (simple, moderate, or complex) using both prompting strategies (CoT and zero-shot recommenders). The average F1 score values for each case for each recommender are also shown.

marginally significant effect, with CoT underperforming zero-shot by ~0.05 ($p<0.05$) (Figure 4). In contrast, the test category showed the strongest impact. Conditional tests underperformed significantly, with F1 scores ~0.39 lower than the combined list of tests ($p<0.001$), while essential tests performed slightly better (~0.04 higher, though not significant). Case complexity and clinical context did not significantly alter performance, once variability across cases and threads was accounted for. Random effects analysis was detailed in Supplementary Figures 1 and 2.

Aggregated evaluation of GPT-4o recommenders

Building on the variability analysis and mixed-effects modelling of GPT-4o's laboratory test recommendations, we aggregated performance metrics to assess the recommenders at a broader level. Specifically, we analysed both the combined list of tests and individual test categories provided by both recommenders, using precision, recall, and accuracy in addition to the F1 score for a more nuanced evaluation (Figure 5).

The aggregated analysis revealed consistently high accuracy (>98 %) across all test categories for both recommenders, highlighting the model's ability to exclude unnecessary tests effectively. However, as noted in Methods, precision and recall more informative indicators.

Both recommenders achieved moderate to high precision rates for the combined list of tests: approximately 68 % for zero-shot and 82 % for CoT. Recall rates were lower, ranging from low to moderate: 41 % for CoT and 51 % for zero-shot. Precision consistently outperformed recall across categories, reflecting the recommenders' strong focus on minimizing the risk of overutilization but suggesting a potential trade-off with recall. Additionally, the aggregated precision and recall values showed variability across test categories, with conditional tests exhibiting significantly lower values than essential tests, aligning with patterns observed in the fixed-effects modelling.

False negatives and false positives: insights for performance optimization

The variability in the aggregated metrics underscored the need to move beyond and explore the underlying lists of FN and FP tests in detail. Understanding the nature of these errors provided actionable insights into the clinical relevance of the model's recommendations and inform targeted

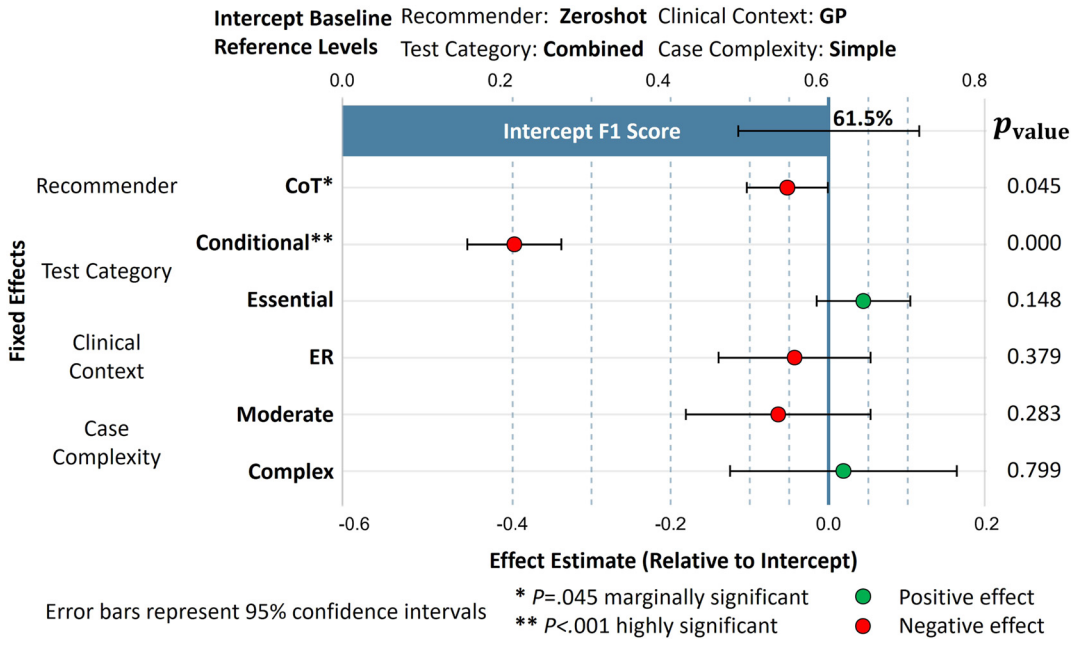

Figure 4: Fixed coefficient plot of fixed effects on recommender evaluation (F1 score). The intercept, representing the baseline F1 score for the reference levels (recommender=zero-shot, test category=combined, context=GP, complexity=simple), was estimated at 0.615. The prompting strategy (recommender) exerted a minor but marginally significant effect, with CoT underperforming zero-shot by ~0.05 ($p<0.05$). In contrast, the test category showed the strongest impact. Conditional tests underperformed significantly, with F1 scores ~0.39 lower than the combined list of tests ($p<0.001$), while essential tests performed slightly better (~0.04 higher, though not significant). Case complexity and clinical context did not significantly alter performance.

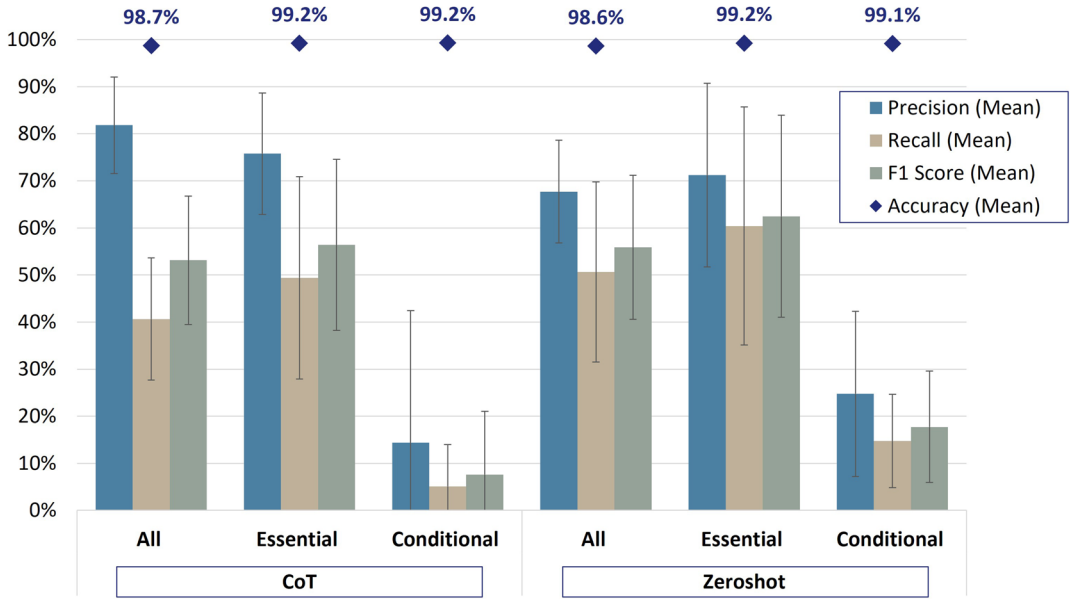

Figure 5: Aggregated evaluation metrics of GPT-4o Recommenders across different test categories.

strategies to address the most impactful gaps to optimize the balance between precision and recall and subsequent risks of overutilization and underutilization (Figure 6). Table 1 also provides an illustrative example of a simulated complex GP clinical case (GP-S8) with laboratory test recommendations from GPT-4o and expert consensus.

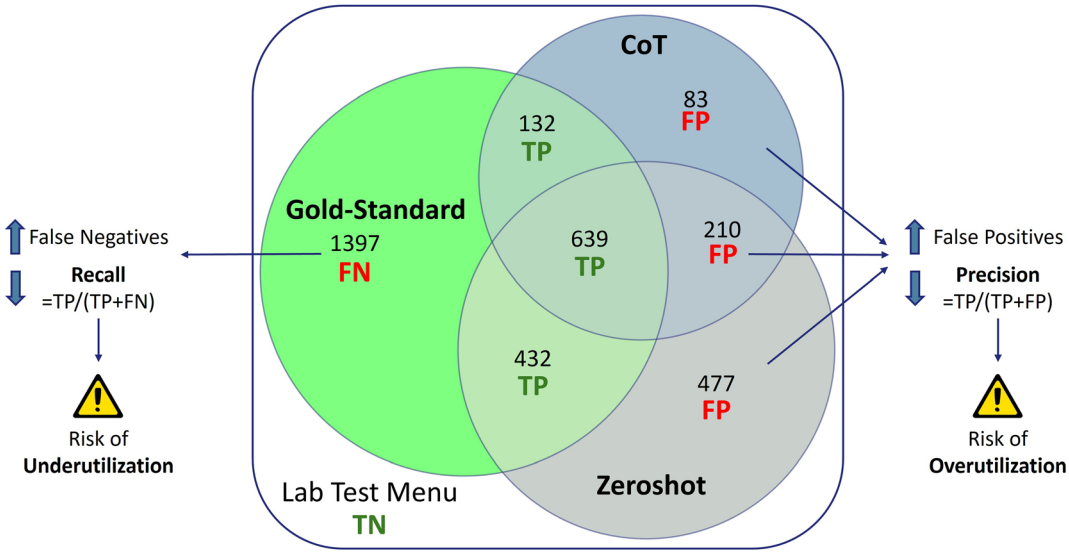

Figure 6: Proportional venn diagram illustrating the overlap between gold-standard orders, zero-shot, and chain-of-thought (CoT) recommendations. Each circle represents the set of test recommendations by each group. True positives (TP) indicate correct test recommendations that overlap between the respective recommenders and the gold standard. False positives (FP) highlight incorrect tests recommended by the respective recommenders (zero-shot or CoT) that do not align with the gold-standard. False negatives (FN) represent necessary tests identified by the gold-standard but missed by both recommenders. The bottom-left true negatives (TN) region outside the circles represents all other tests in the lab test menu that were correctly excluded by all groups.

False negatives: understanding underutilization

To evaluate the risk of underutilization by the GPT-4o recommenders, we analyzed the false negative (FN) tests, focusing on those included in the gold-standard orders for at least three cases but frequently missed by the model. Figure 7 highlights the frequency of these tests and their missing rates as aggregated across various cases and threads. A notable finding was the prevalence of tests from the Comprehensive Metabolic Panel (CMP) among FNs, with 11 tests contributing substantially to the overall FN count and the reduction in recall (sensitivity). This pattern suggests the model's limitation in recommending individual tests when comprehensive panels are needed, likely due to the absence of such panels in the local test menu. This discrepancy underscores the importance of aligning models with local test menus to fit clinical needs in different healthcare settings.

False positives: understanding overutilization

We examined the potential for overutilization by identifying false positive (FP) tests generated by the GPT-4o recommenders. Cross-checking these recommendations against the first Delphi round of expert consensus revealed two distinct groups. The first group included 66 tests recommended in 433 threads (44 % of all FP occurrences) that were also recommended by at least one human expert. In contrast, the second group included 148 tests recommended in 547 threads (56 % of all FP occurrences) that did not appear on any expert's list. Notably, the two groups differed in their average recommendation frequency (about seven threads per test vs. 4 threads per test, respectively) suggesting that the model more repeatedly (and possibly more confidently) recommended those tests that had some expert support.

Discussion

In this study, we evaluated the performance and consistency of GPT-4o in recommending laboratory tests for diverse simulated clinical scenarios in primary and emergency care settings. Unlike previous studies on LLMs has often focused on diagnostic reasoning capabilities in "one-shot" formats [14, 16], this study focused on a critical yet distinct component of clinical decision-making: laboratory test ordering. The real-world clinical environment is a dynamic process where decisions are refined over multiple encounters, making this aspect of model evaluation particularly relevant.

Table 1: Illustrative example of a simulated complex GP clinical case (GP-S8) with laboratory test recommendations from GPT-4o and expert consensus.

GP-S8: Progressive fatigue, pallor, and dark stools (complex case)			
Clinical context details	A 60-year-old man presents with progressive fatigue, pallor, and shortness of breath on exertion for the past two months. He has also noticed dark stools and occasional dizziness.		
Past medical history	– Chronic diseases: hypertension, gastroesophageal reflux disease (GERD) – Previous investigations: colonoscopy two years ago, which was normal – Medications: lisinopril 20 mg daily – omeprazole 20 mg daily		– Allergies: none – Lifestyle factors: former smoker (quit 15 years ago), drinks alcohol occasionally, works as an accountant, sedentary lifestyle
Presenting symptoms	– Progressive fatigue – Pallor	– Shortness of breath on exertion – Dark stools	– Occasional dizziness – No significant weight loss or fever
Physical examination findings	– Vital signs: – Temperature: 36.9 °C – Blood pressure: 110/70 mmHg – Heart rate: 95 bpm – Respiratory rate: 18/min – General appearance: appears pale and fatigued		– Skin: pallor, no rashes or bruising – Cardiovascular: regular rhythm, no murmurs – Respiratory: clear breath sounds – Abdominal: soft, non-tender, mild epigastric tenderness, no palpable masses – Rectal examination: dark, tarry stool

Gold standard tests (with TP rates)				**Other test recommendations (with FP rates)**			
	Test	**Zero-shot**	**CoT**	**Test**	**Zero-shot**	**CoT**	**Experts, n**
Essential	CRP (blood)	10	10	Occult blood in stool	10	10	0
	Complete blood count, incl. WBC differential count	9	10	Folate (blood)	4	9	1
	Iron and transferrin (incl. Transferrin saturation) (blood)	7	7	Urea (blood)	2	2	1
	Creatinine (blood)	2	1	Sodium (blood)	1	2	2
	Gamma GT (blood)	2	0	Potassium (blood)	1	1	2
	Ferritin (blood)	1	0	Chloride (blood)	1	0	2
Conditional	NT-proBNP (blood)	9	4	Bicarbonate (blood)	1	0	2
	Reticulocyte count (blood)	5	4	Prothrombin time (PT) (blood)	0	2	1
	Bilirubin total and direct (blood)	3	0	aPTT (blood)	0	2	1
	ALT (GPT) (blood)	3	0	AST (GOT) (blood)	0	3	0
	Alk. Phosphatase (blood)	2	0	Haptoglobin (blood)	0	2	0
	Vitamin B12 (blood)	0	0	TSH, free T4, free T3 (blood)	0	1	0

F1 score (median)	Zero-shot	CoT	**Precision (median)**	Zero-shot	CoT	**Recall (median)**	Zero-shot	CoT
	67.5 %	61.5 %		77.4 %	81.8 %		58.8 %	47.1 %

This Table provides the clinical case details including patient context, presenting symptoms, examination findings, and gold-standard laboratory tests derived through expert consensus. The table displays how frequently each essential or conditional test was correctly recommended (true positives, TP) by GPT-4o under zero-shot and chain-of-thought (CoT) prompting strategies across 10 threads. It also lists other tests recommended by the model that were not part of the gold standard (false positives, FP), along with the number of experts (n) who independently recommended them during the first Delphi round. Summary performance metrics (median F1 score, precision, and recall) are presented at the bottom to reflect the model's performance on this case.

Strengths

A distinguishing feature of our study was its benchmarking approach. Rather than relying on expert evaluation of the model's recommendations, we evaluated these against independently established gold standard test orders. This design choice minimized bias and ensured a more objective reference point.

Notably, while some studies have leveraged published case scenarios with known "correct" answers [16, 27, 28], we instead simulated fictional clinical cases tailored to our task and derived consensus-based gold standards specific to

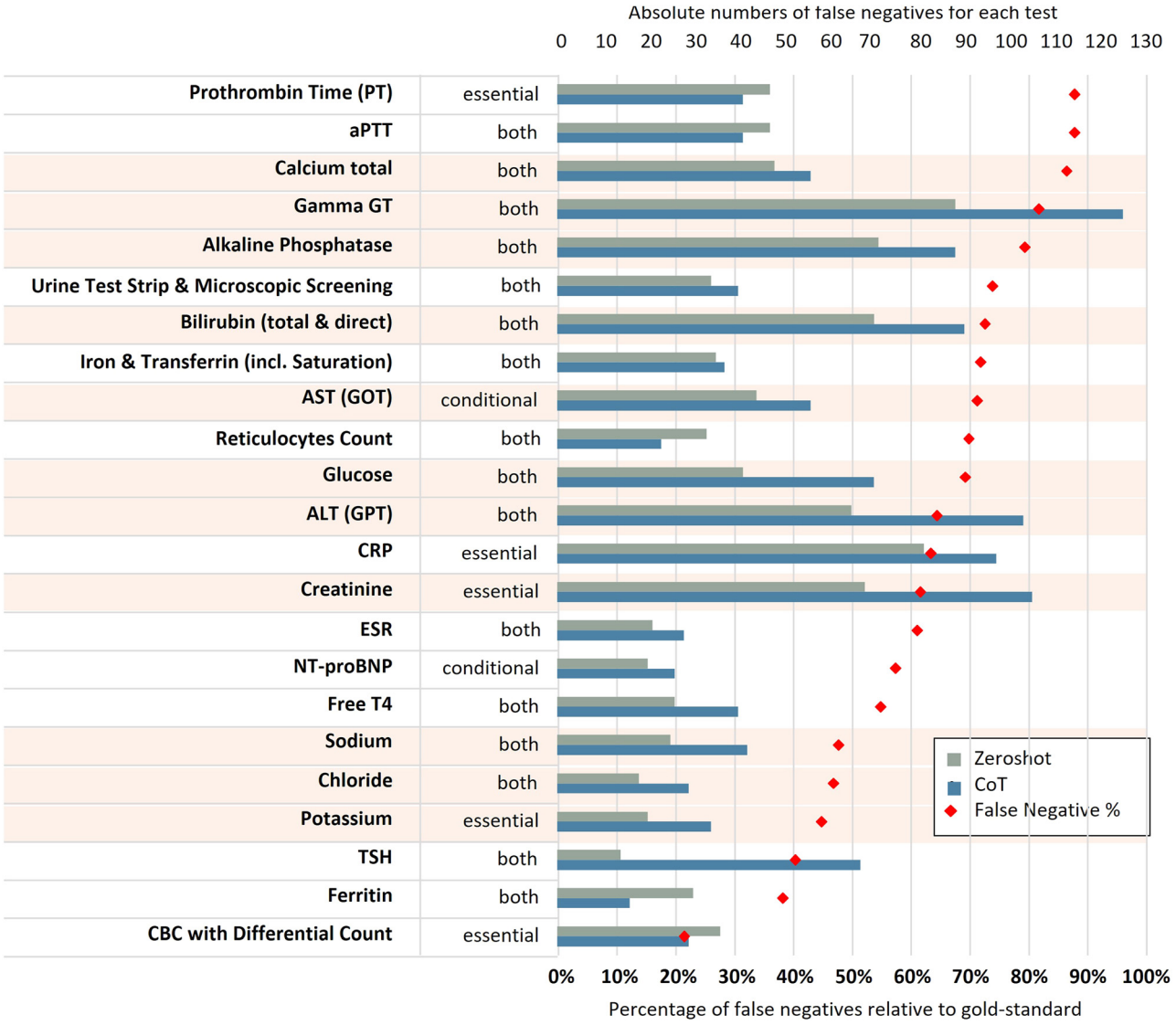

Figure 7: List of aggregated false negative tests. Bars represent the number of threads in which each test was missed, grouped by recommender (zero-shot and CoT). Red diamonds indicate the FN percentage relative to the total number of recommendations for each test according to the gold standard. Tests are categorized as essential, conditional, or included in both test categories. Shaded rows highlight tests from the comprehensive metabolic panel (CMP), which account for a substantial portion of the FN count.

laboratory test ordering. This approach also reduces the likelihood of overlap with the model's training data, strengthening the validity of our results.

Our findings underscore the importance of evaluating repeat prompting. GPT-4o displayed higher self-consistency across multiple threads than was observed among individual human experts during their first Delphi round, confirming results from other studies [29]. Interestingly, despite prior research suggesting that different prompting strategies can significantly affect model performance in medical tasks [16, 30], no significant differences between zero-shot and CoT were observed in this study. This contrast raises the possibility that more advanced models like GPT-4o may be less sensitive to traditional prompt engineering strategies [31].

Beyond prompting strategies, we used an LME model to disentangle various fixed and random effects on F1 scores. Our analysis revealed that case complexity and clinical context did not significantly alter performance, indicating the broad applicability of general-purpose LLMs across diverse medical scenarios.

Moreover, distinguishing between essential and conditional test order categories was instrumental in localizing a substantial amount of variability to the conditional test category, as confirmed by self-consistency results and the

LME model's findings. The intrinsic variability of this category reflects the wide diagnostic landscape of clinical cases. Additionally, distinguishing conditional tests may aid in designing and recommending automated reflex-testing laboratory orders. However, for some aggregated performance metrics, we chose to highlight the combined list of recommended tests as it considers the overall model recommendation capability and accounts for necessary tests recommended (TPs) but assigned to a different category than the gold-standard test orders.

Limitations

Despite these strengths, certain limitations underscore the need for cautious interpretation and future work. First, the clinical scenarios employed here, while crafted to be realistic, remain simulated rather than based on real-world patient data, a choice largely driven by ethical and regulatory constraints preventing clinical data sharing with third-party APIs. Consequently, whether the model would handle the complexities, data gaps, and nuances of authentic clinical environments remains an open question. Additionally, we did not evaluate the model's evidence-based rationale – an aspect crucial for clinicians' trust and safe implementation. Finally, while moderate-to-high precision was observed, lower recall rates may lead to underutilization of necessary tests.

Balancing underutilization vs. overutilization

While both models demonstrated moderate to high precision (68 % for zero-shot and 82 % for CoT), their lower recall rates (41 and 51 %, respectively) suggest a cautious approach that limits excessive testing but risks omitting necessary ones.

The FN analysis identified missing tests commonly included in the CMP panel as a substantial contributor to underutilization, highlighting the model's challenges in recommending individual tests that are typically ordered as part of a larger panel. This underscores the need to align recommendations with local test menus and standardized panels. However, since these tests are routine in clinical practice, real-world omissions are unlikely.

On the FP side, recommendations were categorized into two groups: those overlapping with expert-provided individual responses (44 %) and those unsupported by any expert (56 %). This distinction helps differentiate between reasonable clinical reasoning and potential overutilization. Further evaluation of the model's references and rationale is needed to confirm the appropriateness of these recommendations.

Conclusions

The findings suggest targeted strategies for improvement, which include finetuning model outputs to align recommendations with local test menus and guidelines and verifying model's justifications. A key next step involves evaluating LLMs in real-world clinical settings to assess their performance, considering patient data variability, guidelines, and user inputs. These studies should include clinician feedback, examine medicolegal implications, and measure patient outcomes to ensure LLM-driven clinical decision support is effective and safe [32].

Acknowledgments: The authors would like to extend their sincere gratitude to the experts and clinicians who participated in the consensus process, providing invaluable insights and recommendations that significantly enhanced the quality and applicability of this research. Their expertise and dedication were instrumental in shaping the outcomes of this study.
Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: A.Z., N.D., and G.F. conceived the study. A.Z. implemented the LLM experiments, performed data curation, formal analysis, and software development, and drafted the manuscript. N.D. led resource acquisition and supervision. G.F. led validation and, along with N.D., recruited experts for the consensus process. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: ChatGPT (developed by OpenAI) was used as a writing assistant to review and improve the language and clarity of the manuscript. No content was generated by the tool without author oversight, and all intellectual and scientific contributions are those of the authors.
Conflict of interest: The authors state no conflict of interest.
Research funding: The authors disclose and acknowledge the support provided through OpenAI's Researcher Access Program, which granted access to OpenAI APIs.
Data availability: The datasets generated and/or analyzed during the current study are available in the Github repository, [https://github.com/Ahmed-Medhat-Zayed/lab_test_recommender_gpt].

References

1. Freedman DB. Towards better test utilization – strategies to improve physician ordering and their impact on patient outcomes. EJIFCC 2015;26: 15–30.
2. The Lewin Group, Inc. The value of diagnostics innovation, adoption and diffusion into health care 2005. [Internet] [cited 2023 Apr 3]; Available from: https://www.lewin.com/content/dam/Lewin/Resources/Site_Sections/Publications/ValueofDiagnostics.pdf.
3. Wians FH. Clinical laboratory tests: which, why, and what do the results mean? Lab Med 2009;40:105–13.
4. Zhi M, Ding EL, Theisen-Toupal J, Whelan J, Arnaout R. The landscape of inappropriate laboratory testing: a 15-year meta-analysis. PLoS One 2013;8:e78962.
5. Hickner J, Thompson PJ, Wilkinson T, Epner P, Shaheen M, Pollock AM, et al. Primary care physicians' challenges in ordering clinical laboratory tests and interpreting results. J Am Board Fam Med 2014;27:268–74.
6. Lam JH, Pickles K, Stanaway FF, Bell KJL. Why Clinicians overtest: development of a thematic framework. BMC Health Serv Res 2020;20:1011.
7. Delvaux N, Van Thienen K, Heselmans A, de Velde SV, Ramaekers D, Aertgeerts B. The effects of computerized clinical decision support systems on laboratory test ordering: a systematic review. Arch Pathol Lab Med 2017;141:585–95.
8. Delvaux N, Piessens V, Burghgraeve TD, Mamouris P, Vaes B, Stichele RV, et al. Clinical decision support improves the appropriateness of laboratory test ordering in primary care without increasing diagnostic error: the ELMO cluster randomized trial. Implement Sci 2020;15:100.
9. Shortliffe EH, Sepúlveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018;320:2199–200.
10. Open AI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report 2023. [Internet] [cited 2024 Jan 30]; Available from: http://arxiv.org/abs/2303.08774.
11. Gemini Team Google, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, et al. Gemini: a family of highly capable multimodal models 2024. [Internet] [cited 2024 Dec 24]; Available from: http://arxiv.org/abs/2312.11805.
12. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620:1–9.
13. Cadamuro J, Cabitza F, Debeljak Z, Bruyne SD, Frans G, Perez SM, et al. Potentials and pitfalls of ChatGPT and natural-language artificial intelligence models for the understanding of laboratory medicine test results. An assessment by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) Working Group on Artificial Intelligence (WG-AI). Clin Chem Lab Med 2023;61:1158–66.
14. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7:e2440969.
15. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models 2023. [Internet] [cited 2024 Dec 24]; Available from: http://arxiv.org/abs/2305.09617.
16. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med 2024;7:1–7.
17. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of gemini models in medicine [internet] 2024. [cited 2024 Apr 30]; Available from: http://arxiv.org/abs/2404.18416.
18. Perlis RH, Fihn SD. Evaluating the application of large language models in clinical research contexts. JAMA Netw Open 2023;6: e2335924.
19. Wei J, Bosma M, Zhao V, Guu K, Yu AW, Lester B, et al. Finetuned language models are zero-shot learners 2022. [Internet] [cited 2024 Dec 25]; Available from: https://openreview.net/forum?id=gEZrGCozdqR.
20. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought prompting elicits reasoning in large language models 2023. [Internet] [cited 2024 Apr 18]; Available from: http://arxiv.org/abs/2201.11903.
21. Dalkey N, Helmer O. An experimental application of the Delphi method to the use of experts. Manag Sci 1963;9:458–67.
22. Laboboeken – UZ leuven [Internet]. [cited 2025 Feb 13]; Available from: https://laboboeken.nexuzhealth.com/toc/pboek/internet/GHB
23. Vectara. Leaderboard comparing LLM performance at producing hallucinations when summarizing short documents. GitHub. [Internet] [cited 2024 Dec 24];Available from: https://github.com/vectara/hallucination-leaderboard.
24. GPT-4o system card. [Internet]. [cited 2025 Feb 18];Available from: https://openai.com/index/gpt-4o-system-card/.
25. OpenAI platform - assistants API overview. [Internet]. [cited 2024 Dec 24]; Available from: https://platform.openai.com/docs/assistants/overview.
26. Prompt Engineering Guide. DAIR.AI. Prompt engineering guide – LLM settings 2024. [Internet] [cited 2024 Dec 25]; Available from: https://www.promptingguide.ai/introduction/settings.
27. Savage T, Wang J, Gallo R, Boukil A, Patel V, Safavi-Naini SAA, et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. J Am Med Inf Assoc 2025;32:139–49.
28. Arvidsson R, Gunnarsson R, Entezarjou A, Sundemo D, Wikberg C. ChatGPT (GPT-4) versus doctors on complex cases of the Swedish family medicine specialist examination: an observational comparative study. BMJ Open 2024;14:e086148.
29. Gallo RJ, Baiocchi M, Savage TR, Chen JH. Establishing best practices in large language model research: an application to repeat prompting. J Am Med Inf Assoc 2024;31:ocae294.
30. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med 2024;7:1–9.
31. Wang G, Sun Z, Gong Z, Ye S, Chen Y, Zhao Y, et al. Do advanced language models eliminate the need for prompt engineering in software engineering? 2024. [Internet] [cited 2025 Jan 22]; Available from: http://arxiv.org/abs/2411.02093.
32. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. J Am Med Assoc 2025;333:319–28.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cclm-2025-0647).