Evaluating Large Language and Large Reasoning Models as Decision Support Tools in Emergency Internal Medicine

Josip Vrdoljak MD PhD, University of Split, School of Medicine, Department of Pathophysiology, Split, Croatia
Zvonimir Boban Msc Phys. PhD, University of Split, School of Medicine, Department of Medical Physics, Split, Croatia
Ivan Males MD, University Hospital Split, Department of Surgery, Split, Croatia
Roko Skrabic MD, University Hospital Split, Department of Nephrology, Split, Croatia
Marko Kumric MD PhD, University Hospital Split, Department of Cardiology, Split, Croatia
Anna Ottosen, University of Split, School of Medicine, Split, Croatia
Alexander Clemencau, University of Split, School of Medicine, Split, Croatia
Sebastian Völker* Msc. PhD, IU International University of Applied Sciences, Department of Health, Erfurt, Germany
Josko Bozic MD PhD, University of Split, School of Medicine, Department of Pathophysiology, Split, Croatia

***Correspondence:**

Sebastian Völker * Msc PhD, International University of Applied Sciences, Department of Health, 99084 Erfurt, Germany; email: sebastian.voelker@iu.org, phone: +49 231 945362 117

Word count: 5019

Abstract

Background
Large Language Models (LLMs) hold promise for clinical decision support, but their real-world performance varies. We compared three leading models (OpenAI's "o1" Large Reasoning Model (LRM), Anthropic's Claude-3.5-Sonnet, and Meta's Llama-3.2-70B) to human experts in an emergency internal medicine setting.

Methods
We conducted a prospective comparative study on 73 anonymized patient cases from the Emergency Internal Medicine ward of the University Hospital Split, Croatia (June–September 2024). Two independent internal medicine specialists, blinded to model identity, graded the LLM-generated reports in two steps: (1) they evaluated the relevance of recommended diagnostic tests based on the patient's signs, symptoms, and medical history; (2) after reviewing the actual diagnostic test results, they assessed each model's final diagnosis, therapy plan, and follow-up recommendations. The same evaluative framework was applied to human-authored reports. Likert scales (1–4 or 1–3) were used, and statistical comparisons included the Friedman and Wilcoxon signed-rank tests.

Results
The o1 model achieved a mean final rating (3.63) statistically indistinguishable from human physicians (3.67; $p=0.62$). Claude-3.5-Sonnet (3.38) and Llama-3.2-70B (3.23) scored significantly lower ($p<0.01$ vs. o1), largely due to errors in therapy planning and non-medication recommendations. Despite this gap, all three models demonstrated ≥90% accuracy in final diagnoses and patient admission decisions. The o1 model correctly classified all abnormal lab values (100%), while Claude-3.5-Sonnet and Llama-3.2-70B showed minor errors (99.5% and 99% accuracy, respectively).

Conclusions
When evaluated on real-world emergency cases, an advanced LLM with enhanced reasoning (o1) can match expert-level clinical performance, underscoring its potential utility as a decision-support tool.

Keywords: Artificial Intelligence; Natural Language Processing; Decision Support Systems, Clinical; Emergency Medicine; Internal Medicine

Introduction

Large Language Models (LLMs), and multimodal generative artificial intelligence (AI) systems hold the potential to transform healthcare, particularly in areas like medical education, clinical decision support, and healthcare administration [1]. State-of-the-art (SOTA) LLMs are trained on large amounts of high-quality data parsed from the internet, textbooks, video transcripts, and others, with the goal of predicting the next token (word in a sequence) [2].

The pretraining of large language models through self-supervised learning results in remarkable performance across a diverse range of downstream tasks and benchmarks [3]. For instance, leading models such as GPT-4 and Claude 3.5 Sonnet attain exceptionally high scores on the Massive Multi-task Language Understanding (MMLU) benchmark, with scores of 86.4% and 88.3%, respectively [4-6]. The MMLU benchmark is specifically designed to evaluate a model's understanding and problem-solving abilities across various topics and domains, including mathematics, computer science, law, and medicine. For comparison, an expert-level human in a specific subject typically achieves an average score of 89.8% [5].

One of the most interesting areas of potential LLM usage in healthcare is in clinical decision support. LLMs demonstrate significant potential in clinical decision support by facilitating the selection of appropriate diagnostic protocols, optimizing treatment recommendations, and serving as verification systems to reduce medical errors and oversights [1,7].

Current research shows somewhat mixed results in LLM accuracy in clinical decision-making. For example, Gumilar et al. report how Gemini Advanced achieves 81.87 % accuracy in clinical decision-making when tested on questions based on real-world cases [8]. Similarly, in a study evaluating ChatGPT's therapeutic recommendations for head and neck cancers, Marchi et al. analyzed 68 hypothetical cases and 204 clinical scenarios against NCCN Guidelines [9]. Results showed that ChatGPT demonstrated high sensitivity and overall accuracy in addressing NCCN-related queries, although some inaccuracies were noted, particularly in primary treatment scenarios [9]. Moreover, one other study evaluated the appropriateness and usefulness of LLM recommendations in neuroimaging clinical decision support. The models were tested on 24 neuroradiology clinical scenarios, and GPT-4 delivered optimal recommendations in 23/24 cases [10].

One of the potential use cases for LLMs in decision support, is to provide guidance to untrained doctors, as demonstrated by a recent comparative study on emergency department triage by Masanneck et al.. They evaluated multiple LLMs (including GPT-4) against both untrained doctors and professionally trained emergency department staff, using the Manchester Triage System [11]. While the best LLM-based tools achieved triage performances comparable to untrained clinicians, they still fell short of matching the gold-standard results achieved by professional triage personnel. Of note, the LLMs tended to over-triage, underscoring the continuing need for model refinement and domain-specific tuning in high-stakes environments like the emergency department [11].

While the aforementioned studies exhibited good LLM performance in clinical decision support, there are also other reports where LLMs failed to match human performance or achieve sufficient accuracy in diagnostic and therapeutic recommendations [7,12]. For example, in a study by Hager et al., researchers have evaluated open-source LLMs on 2,400 real patient cases (across four common abdominal pathologies), and have shown that existing open-source LLMs (specifically Llama 2 Chat (70B), Open Assistant (70B), WizardLM (70B), Camel (70B) and Meditron (70B)) struggled to adhere to diagnostic and treatment guidelines, and encountered difficulties with fundamental tasks such as laboratory result interpretation. Likewise, in a study by Benary et al., researchers evaluated four LLMs (ChatGPT-3.5, Galactica, Perplexity, and BioMedLM) as support tools for precision oncology and have reported how tested LLMs underperformed human experts with lower precision and recall in generating treatment options for 10 fictional advanced cancer cases [12]. However, it is important to note that both this study and the one by Benary et al. may not reflect the capabilities of the most recent open-source models, as the most recent open-source models, including Llama-3-70b and 405b, have shown performance levels matching GPT-4, suggesting these earlier studies do not capture the current state of open-source AI capabilities. To further contextualize the current landscape of AI vs. human clinical decision-making, we note the findings from Goh et al. in their randomized clinical trial published in JAMA Network Open[13]. This study investigated how large language models influence diagnostic reasoning and found that while LLMs outperformed physicians in standalone diagnostic accuracy, physicians who used LLMs did not show improved diagnostic performance compared to those using standard resources. This highlights a critical gap between AI capability and effective clinical integration, suggesting that simply providing access to advanced AI tools may not enhance diagnostic reasoning without proper implementation strategies.

What is more, a recent systematic review on LLMs in health care applications, concluded how out of 519 examined studies, only 5% of them used real patient care data for LLM evaluation, indicating a lack of LLM research on real-world patient data [14]. There is also new paradigm in the progress of LLMs, with the advancement to Large Reasoning Models (LRMs) like the o1 model from OpenAI, where incorporating specific reinforcement learning techniques during post-training leads to significant performance gains on reasoning tasks [15]. Such models excel on all reasoning benchmarks, but to our knowledge, have still not been tested on real-world medical tasks.

Given the mixed findings on the utility of LLMs in real-world clinical decision support, and given the lack of evidence concerning the performance of LRMs, this study aims to comprehensively evaluate the performance of state-of-the-art LLMs (Claude-3.5 Sonnet, Llama-3.2-70b, and LRM the o1 model) in supporting clinical decision-making within an internal medicine emergency department. The evaluation is conducted using real-world patient cases to assess their practical applicability and accuracy in high-stakes medical scenarios. Our investigation will focus on two critical aspects of clinical decision-making: first, the

appropriateness of diagnostic test recommendations based on initial patient presentation, medical history, and current symptoms, and second, the accuracy of diagnostic conclusions and therapeutic recommendations when provided with actual diagnostic test results. This real-world evaluation will provide insights into the practical utility of current LLMs as clinical decision-support tools and their potential role in enhancing emergency medical care.

Methodology

Study Design and Participants

This prospective comparative study evaluates LLMs' performance against human experts in emergency internal medicine. Data were prospectively collected during a two-month period, from the 10th of June to the 30th of September, 2024., in the Emergency Internal Medicine ward of the University Hospital of Split, Croatia. Patient selection criteria required that patient reports be complete, with information about the reason for visit, signs and symptoms, used diagnostic tests and procedures, as well as final diagnosis and treatment with recommendations. Each day, two or more complete patient reports (with documented reason for visit, symptoms, diagnostic tests and procedures, and final diagnosis/treatment) were chosen through random sampling. On certain days, data collection was not conducted due to the unavailability of the research team. This random selection strategy was used to minimize selection bias and capture a diverse range of clinical presentations; thus, the resulting sample distribution aligns with the real-world patient population encountered in the ward. From a total of 80 collected reports, 7 were removed due to incomplete findings and missing structure (missing conclusion or diagnostic test results), leaving a total of 73 complete patient reports. Prior to LLM prompting, all patient data was anonymized by removing direct identifiers. The evaluation process consisted of two distinct phases. In Phase 1, LLMs were presented with initial patient data (presenting symptoms, medical history, and current illness description) and asked to recommend appropriate diagnostic tests. In Phase 2, after being provided with the actual diagnostic test results, the LLMs were tasked with generating diagnoses, treatment recommendations, and follow-up plans. The models were also evaluated for the understanding of laboratory value reference ranges. All LLM assessments were based on standard clinical inputs from real-world emergency department cases, without access to the actual clinical decisions or outcomes. We open-source the de-identified patient reports, as well as the LLM-generated reports. All the data is available at the following repository: https://github.com/vrda23/Evaluating-Large-Language-and-Large-Reasoning-Models-as-Decision-Support-Tools-in-Medicine/tree/main.

The study protocol was approved by the Ethics Board of the University Hospital of Split (approval number: 2181-147/01-06/LJ.Z.-24-02). Furthermore, the study strictly adhered to data protection regulations and securely storing datasets on encrypted servers in compliance with GDPR standards.

LLM Patient Report Generation

We prompted three different LLMs, Claude-3.5-Sonnet from Anthropic AI, open-sourced Llama-3.2-70B from Meta AI, and the o1 model from OpenAI. We picked Claude-3.5-Sonnet because it is currently the best standard LLM on most benchmarks. Similarly, we picked Llama-3.2-70B as currently the best open-source LLM, while we picked the o1 model because it is presently the best large reasoning model, and the best model overall (on all the major benchmarks).

Claude-3.5-Sonnet and Llama-3.2-70b were accessed via their respective APIs (provided by Anthropic and Together.ai), while the o1 model was accessed via OpenAI's web user interface. Two specific prompts (available at supplementary material) were used. First, the LLM is prompted to identify and recommend the specific diagnostic tests and procedures it considers necessary to arrive at an accurate diagnosis for a given patient case. Second, after receiving complete patient information, including the results of the recommended diagnostic tests, the LLM is prompted to generate a final diagnosis, treatment plan, and corresponding recommendations. Additionally, we have extracted 400 real-world laboratory values from the reports to test the understanding of laboratory reference ranges.

To access the APIs we used the Python programming language (version 3.10.), along with libraries "together" and "anthropic". The "temperature" parameter for all models was set at zero.

Patient Report Evaluation

LLM-generated reports were evaluated in multiple steps. The evaluation system was developed based on recommendations from the QUEST evaluation framework, which recommends the use of Likert scale for evaluating accuracy and the relevance of LLM responses [16].

Two independent expert internal medicine physicians (with 5 and 7 years of work experience) first graded the real-world patient reports (human-generated), on a 1-4 Likert scale. Subsequently, they graded the reports generated by the LLMs (GPT-o1, Claude-3.5.-Sonnet, and Llama-3-70b), without knowing

which specific model generated each report, ensuring they were blinded during the grading process. Prior to grading the patient reports, the two expert internal medicine physicians discussed and agreed upon the grading criteria to ensure consistency in their evaluations. There was no statistically significant difference between the observers (P = 0.89, Wilcoxon signed rank test), indicating a high level of consistency and reliability in the grading process. Moreover, the Cohen's kappa between the raters for the o1 model was 0.87, while for the Claude-3.5-Sonnet it was 0.85, and for Llama-3-70b 0.86, indicating substantial agreement between the raters. In order to approach the use of LLMs conservatively and retain the interpretability of the grades, only the lower of the two grades was used in all further analyses.

Firstly, the evaluators graded the appropriateness of recommended diagnostic tests and procedures given the initial input patient data (signs and symptoms, medical history, and current reason for arrival). This first part was evaluated and graded on a Likert scale (1-3), where the 3 corresponded to the LLM, giving all the correct diagnostic recommendations that were also in line with what was ordered in the actual real-world patient report. Grade 2 was given when the LLM missed some of the required tests, while grade 1 was given when the LLM missed most of the required tests or missed a critical test for a particular patient case.

Subsequently, in the next part of the evaluation (where the LLM is given the entire report data but without the final diagnosis and treatment recommendations), they first evaluated the correctness of the final diagnosis. This was performed in a binary manner (0-incorrect, 1-correct), where the final LLM-generated diagnosis was compared to the diagnosis from the doctor's report. Then, they evaluated both the validity of the therapy prescribed and the recommendations given. The therapeutic interventions and clinical recommendations were evaluated using a standardized three-point Likert scale assessment framework. A score of 3 denoted optimal prescribing behavior, characterized by comprehensive medication selection (single or multiple therapeutic agents) with accurate dosing parameters and administration schedules. A score of 2 indicated partial adherence to clinical guidelines, through either incomplete medication coverage or minor deviations in dosing regimens or administration frequencies. A score of 1 signified substantial therapeutic inadequacies, where the prescribed intervention demonstrated significant clinical inappropriateness for the presented patient case. Similarly, for recommendations, a score of 3 denoted comprehensive and clinically appropriate guidance, encompassing all necessary lifestyle modifications, monitoring parameters, and follow-up protocols. A score of 2 indicated partially complete recommendations with minor omissions or imprecise specifications, while a score of 1 represented recommendations that were either clinically inappropriate or substantially incomplete for the patient's presenting condition. For both therapy and recommendations, the graders were instructed to take into account both European and American guidelines, so the grades were not decreased based on, for example, the model using a different medication, but from the same medication group, when compared to the real-world finding.

The model's understanding of laboratory value reference ranges was tested in a sequential prompt. For each model, a set of laboratory values was provided, with a prompt to highlight the values that were either decreased or increased.

Furthermore, also in a binary manner (0-incorrect, 1-correct), we have evaluated the LLM recommendations for patient admission, where the correct class was given when the LLM recommendations adhered to the admission status from the real-world report.

Finally, the quality of the entire LLM-generated report was graded on a Likert scale (1-4). As mentioned, the same was also performed for the actual real-world finding (human doctor-generated report). This grading system encompasses the previous diagnosis, therapy, and recommendations grades. The highest grade (4) was given when the report implemented adequate diagnostic procedures, the correct differential diagnosis, and correct treatment recommendations that adhered to the latest medical guidelines. The "mostly correct" grade (3) was given when the report had minor deficiencies, for example, if some minor treatment options were missed in the treatment recommendation. The "partially correct" grade (2) was given when the report missed some important information (wrong diagnostic procedure, partially wrong differential diagnosis, or wrong treatment options). Finally, the "completely incorrect" grade (1) was given when the differential diagnosis or treatment was completely missed and did not reflect the specific case correctly.

Statistical Analysis

The grading results of reports generated by LLMs and humans were compared by first evaluating the normality of the data distribution using the Shapiro-Wilk test. To assess overall differences between the grading distributions of the three LLMs (o1, Claude-3.5-Sonnet, and Llama3-70b) and human grades, we employed the Friedman test (due to non-normal data distribution). Likewise, post-hoc analyses were conducted using pairwise Wilcoxon signed-rank tests with Bonferroni correction for multiple comparisons to control the familywise error rate ($\alpha = 0.05$/number of comparisons). For binary diagnosis classifications (0/1), we calculated the percentage of positive diagnoses (1's) for each LLM with 95% confidence intervals computed using the binomial exact method (Clopper-Pearson interval). The Chi^2 test was used to determine the differences in the final diagnosis between the models. The effect size was approximated with Cohen's d. Inter-rater agreement was determined with Wilcoxon signed-rank test, as well as with Cohen's kappa. Statistical relevance was set at $P= 0.05$.

Results were expressed as the mean (with a 95% confidence interval), median (alongside interquartile range, IQR), and count (percentage %). All statistical analyses were performed using Python

(version 3.10) with the "numpy," "pandas," and "scipy.stats" libraries. Data visualization was performed using the R programming language (version 4.2.1), using the "ggplot2" and "likert" libraries.

Results

Patient Report Characteristics

The reports represent real-world cases from the Emergency Department of Internal Medicine. A total of 73 cases were analyzed, with most of the cases stemming from heart and lung pathologies, as well as abdominal diseases. Some of the most common case pathologies were: J18.9 Pneumonia, I48 Atrial fibrillation and flutter, N39.0 Urinary tract infection, I50.0 Congestive heart failure, and J44.9 Chronic Obstructive Pulmonary Disease, among others. More specifically, a total of 22 cases (30.1%) were from cardiovascular diseases, 13 cases (17.8%) from respiratory diseases, 16 cases (21.9%) from abdominal diseases, 5 (6.9%) from infectious diseases. In comparison, the rest of the 17 cases (23.3%) were grouped as "other diagnoses" (Table 1.). A detailed complete list of diagnoses is provided in Supplementary data.

Table 1. Disease frequency by group

Group	**Count**	**Percentage**
Cardiovascular Diseases	22	30.1 %
Respiratory Diseases	13	17.8 %
Abdominal and Gastrointestinal Diseases	16	21.9 %
Infectious Diseases	5	6.9 %
Other Diagnoses	17	23.3 %
Total	73	100 %

Patient Report Grading

For the total score (Likert scale 1-4), the mean final rating for human physicians was 3.67 (95% CI: 3.54-3.81) (Figure 1. (A)). Subsequently, the mean final rating for the o1 model was 3.63 (95% CI: 3.49-3.77), for Claude-3.5-Sonnet 3.38 (95% CI: 3.21-3.56), and for Llama-3-70B 3.23 (95% CI: 3.05-3.41) (Figure 1. (A)). The median scores for human physicians, o1 model, and Claude-3.5-Sonnet were 4 [IQR: 3-4] in all three cases, while the median score for Llama-3-70B was 3 [IQR: 3-4]. Human physicians were graded with the grade 4 in 53 cases (72.6%), the o1 model was graded with 4 in 50 cases (68.5%), Claude-3.5-Sonnet in 38 cases (52.1%), and Llama-3-70B in 30 cases (41.1%). Human physicians and the o1 model were never graded with the worst grade (grade 1), while Llama-3-70B was graded with 1 in two instances (2.7%), and Claude-3.5-Sonnet in one instance (1.4%). A more detailed representation of grade frequency is presented in Figure 1. (B).

The Friedman test indicated that there was a statistically significant difference in final grades between the LLMs and human doctors, as well as between the LLMs themselves (P= 0.002)

There was no statistically significant difference in scores between human physicians and the o1 model (P= 0.6174, Wilcoxon Signed-Rank Test, Bonferroni corrected, Cohen's d=-0.057). On the other hand, there was a significant difference between human ratings and Llama-3-70B (P= 0.0002, Wilcoxon Signed-Rank Test, Bonferroni corrected, Cohen's d=0.333) and between human physicians and Claude-3.5-Sonnet (P= 0.0077, Wilcoxon Signed-Rank Test, Bonferroni corrected, Cohen's ≈ -0.364). Moreover, the o1 model significantly outperformed both Claude-3.5.-Sonnet (P= 0.008, Wilcoxon Signed-Rank Test, Bonferroni corrected, Cohen's d≈ 0.333), and Llama-3-70B (P= 0.0002, Wilcoxon Signed-Rank Test, Bonferroni corrected, Cohen's d≈ 0.480).

Finally, to better illustrate the differences in the quality of LLM generated responses, we provide a concrete example were the human graders graded the o1's output with 4, Claude-3.5-Sonnet's output with 2, and Llama-3-70B's output with 1 (Check Supplementary Materials). In this example, both Claude-3.5-Sonnet and Llama-3-70B failed to recognize the need to administer antibiotic treatment (along with failing to recognize pneumonia as the final diagnosis), while the o1 model correctly recognized pneumonia and recommended appropriate antibiotic treatment (when compared to the real-world report) (Supplementary Materials).

Figure 1: Comparison of final grades for human and large language model (LLM)-generated reports in emergency internal medicine. (A) Bar chart showing the mean final grades (on a 1–4 Likert scale) for human physicians and three LLMs (o1 model, Claude 3.5 Sonnet, and Llama 3.1 70B), with error bars representing 95% confidence intervals. Grades reflect the overall quality of diagnostic and treatment recommendations, where 1 = completely incorrect, 2 = partially correct, 3 = mostly correct, and 4 = fully correct according to clinical guidelines. (B) Stacked bar chart displaying the distribution of grades (1–4) for each group, with percentages indicating the proportion of 73 real-world patient cases receiving each grade. The numbers within each grade segment denote the percentage of total graded reports for that category.

For the scoring of the correctness of the recommended diagnostic tests (Likert scale 1-3), the o1 model scored the highest, with a mean score of 2.92 (95% CI: 2.85 - 2.98) and a median of 3.00 (IQR: 3.00 - 3.00) (Figure 2. (A)). Followed by Claude-3.5-Sonnet with the mean score of 2.82 (95% CI: 2.72 - 2.93), and the median of 3.00 (IQR: 3.00 - 3.00) (Figure 2. (A)). Meanwhile, the Llama-3-70B had a mean score of = 2.78 (95% CI: 2.66 - 2.91) and a median of 3.00 (IQR: 3.00 - 3.00) (Figure 2. (A)). The o1 model had the highest frequency of grade 3 for the relevance of diagnostic tests (67/73), followed by Llama-3-70B (62/73), and Claude-3.5-Sonnet (61/73) (Figure 2. (B)). Moreover, the o1 model had 0 (0%) grade 1 scores, Claude-3.5-Sonnet had 2 (2.7%), and Llama-3-70B had 4 (5.5%) grade 1 scores, which indicate cases with major flaws in recommended diagnostic tests (Figure 2. (B)).

There was no statistically significant difference in the relevance of recommended diagnostic tests between the models (P= 0.5611, Friedmann test).

Figure 2: Comparison of grades for diagnostic tests recommended by human physicians and large language models (LLMs) in emergency internal medicine. (A) Bar chart showing the mean grades (on a 1–3 Likert scale) for the o1 model, Claude 3.5 Sonnet, and Llama 3.1 70B, with error bars representing 95% confidence intervals. Grades reflect the appropriateness of recommended diagnostic tests based on patient signs, symptoms, and medical history, where 1 = missed most critical tests, 2 = missed some required tests, and 3 = recommended all correct tests in line with actual clinical orders. (B) Stacked bar chart displaying the distribution of grades (1–3) for each LLM, with percentages indicating the proportion of 73 real-world patient cases receiving each grade. The numbers within each grade segment denote the percentage of total graded reports for that category.

For the scoring of the relevance of prescribed therapy (Likert 1-3), the o1 model again scored the highest, with a mean score of 2.77 (95% CI: 2.65 - 2.88), and a median of 3.00 (IQR: 3.00 - 3.00) (Figure 3. (A)). The second highest scoring model in therapy recommendation was Claude-3.5-Sonnet, with a mean score of 2.58 (95% CI: 2.44 - 2.72) and a median of 3.00 (IQR: 2.00 - 3.00) (Figure 3. (A)). Finally, Llama-3-70B scored the mean score of 2.51 (95% CI: 2.37 - 2.65), and the median of 3.00 (IQR: 2.00 - 3.00) (Figure 3. (A)). The o1 model was graded with the highest grade for therapy prescriptions in 58 cases (79.5%). Followed by Claude-3.5-Sonnet that was graded with 3 in 46 cases (63.0%), and Llama-3-70B in 41 cases (56.2%). The o1 model was found to have minor errors in prescribed therapy (grade 2) in 13 cases (17.8%), while more major inconsistencies were found in just two cases (2.7%) (Figure 3. (B)). On the other hand, both Claude-3.5.-Sonnet and Llama-3-70B had major errors in prescribed therapy in 4 cases (5.5%) (Figure 3. (B)).

There was a significant difference in model ratings concerning therapy recommendations (P= 0.002, Friedmann test). Specifically, the significant difference was between the o1 model and Llama-3-70B (P= 0.001, Wilcoxon Signed-rank Test, Bonferroni corrected, Cohen's d ≈ 0.403), as well as between the o1 model and Claude-3.5-Sonnet (P= 0.01, Wilcoxon Signed-rank Test, Bonferroni corrected, Cohen's d ≈

0.311). On the other hand, there was no significant difference between Claude-3.5-Sonnet and Llama-3-70B (P= 0.35, Wilcoxon Signed-rank Test, Bonferroni corrected, Cohen's d ≈ 0.109).

Figure 3: Comparison of grades for prescribed therapies recommended by human physicians and large language models (LLMs) in emergency internal medicine. (A) Bar chart showing the mean grades (on a 1–3 Likert scale) for the o1 model, Claude 3.5 Sonnet, and Llama 3.1 70B, with error bars representing 95% confidence intervals. Grades reflect the validity of therapy recommendations, where 1 = substantial therapeutic inadequacies, 2 = partial adherence with minor deviations in dosing or administration, and 3 = optimal prescribing behavior with accurate medication selection and dosing. (B) Stacked bar chart displaying the distribution of grades (1–3) for each LLM, with percentages indicating the proportion of 73 real-world patient cases receiving each grade. The numbers within each grade segment denote the percentage of total graded reports for that category.

Next, concerning the correctness of the given recommendations (non-medication), the o1 model again scored the highest, with a mean score of 2.90 (95% CI: 2.83 - 2.97) and a median of 3.00 (IQR: 3.00 - 3.00) (Figure 4.). Claude-3.5-Sonnet scored the second best in giving recommendations, with a mean score of 2.84 (95% CI: 2.75 - 2.92) and a median of 3.00 (IQR: 3.00 - 3.00). Llama-3-70B scored the mean score of 2.68 (95% CI: 2.55 - 2.82), and the median of 3.00 (IQR: 2.00 - 3.00). The o1 model outperformed other LLMs in the number of highest grades for giving recommendations, with grade 3 scored in 66 cases (90.4%). Followed by Claude-3.5-Sonnet that was graded with 3 in 61 cases (83.6%), and Llama-3-70B in 54 (74.0%) (Figure 4.). Notably, both the o1 model and Claude-3.5-Sonnet had no cases in which they failed to give the

proper recommendations in a major way, while Llama-3-70B had major errors in recommendations in 4 cases (5.5%) (Figure 4.).

There was a statistically significant difference between model performance in offering recommendations (P= 0.005, Friedmann test), with the o1 model significantly outperforming Llama-3-70B (P= 0.002, Wilcoxon Signed-rank Test, Bonferroni corrected, Cohen's d ≈ 0.376). There was no significant difference in giving recommendations between the o1 model and Claude-3.5-Sonnet, as well as between Claude-3.5-Sonnet and Llama-3-70B (P = 0.07, Wilcoxon Signed-rank Test, Bonferroni corrected, Cohen's d ≈ 0.254).

Figure 4: Comparison of grades for non-medication recommendations provided by human physicians and large language models (LLMs) in emergency internal medicine. (A) Bar chart showing the mean grades (on a 1–3 Likert scale) for the o1 model, Claude 3.5 Sonnet, and Llama 3.1 70B, with error bars representing 95% confidence intervals. Grades reflect the appropriateness of recommendations (e.g., lifestyle modifications, follow-up protocols), where 1 = clinically inappropriate or substantially incomplete, 2 = partially complete with minor omissions, and 3 = comprehensive and clinically appropriate guidance. (B) Stacked bar chart displaying the distribution of grades (1–3) for each LLM, with percentages indicating the proportion of 73 real-world patient cases receiving each grade. The numbers within each grade segment denote the percentage of total graded reports for that category.

Upon testing the understanding of laboratory reference ranges, the o1 model has correctly classified all 400 values, with an accuracy of 100% (95%CI: 0.990 - 1.000), while the Claude-3.5-Sonnet has made

two (accuracy of 99.5% (95%CI: 0.982 - 0.999)) and Llama-3-70b four mistakes (accuracy of 99% (95%CI: 0.975 - 0.996)), respectively.

All three models performed well in determining the patient admission status. The o1 model and Claude-3.5-Sonnet achieved the same accuracy of 91.8% (95% CI: 0.832 - 0.962), followed by Llama-3-70B with an accuracy of 90% (95% CI: 0.815 - 0.953). There was no statistically significant difference in patient admission accuracy between the tested models ($P > 0.05$, Chi^2 test).

Concerning the accuracy in defining the final diagnosis (when given all patient input and diagnostic data from the real-world report), the o1 model performed the best with an accuracy of 0.973 (95% CI: 0.905-0.997). The other models have also shown good performance in defining the final diagnosis, with Claude-3.5-Sonnet scoring accuracy of 0.932 (95% CI: 0.847-0.977) and Llama-3-70B achieving the same score of 0.932 (95% CI: 0.847-0.977). There was no statistically significant difference in final diagnosis accuracy between the tested models ($P > 0.05$, Chi^2 test).

Discussion

Our study demonstrates that a state-of-the-art (SOTA) LLM with reasoning capabilities (the o1 model) performed comparably to human experts when diagnosing and treating real-world cases in emergency internal medicine. These results align with the findings from previous studies that have exhibited LLMs matching expert-level performance on various medical tasks [17,18]. However, our observations also show that other SOTA LLMs (Claude-3.5-Sonnet and Llama-3-70B) did not attain the same level of accuracy or comprehensiveness as the o1 model or human physicians, indicating that LLM capabilities can vary greatly depending on model size, and, pre-/-post training approaches. Specifically, the o1 model significantly outperformed Claude-3.5-Sonnet and Llama-3-70B in recommending/prescribing the correct therapy, as well as significantly outperformed Llama-3-70B in giving non-medication recommendations (e.g., lifestyle changes, patient follow-up, etc.). However, it is also important to note that failing to find a statistically significant difference does not constitute proof of equivalence between the models. Our results indicate that we were unable to detect a significant disparity, rather than confirming that the LLMs and human physicians are definitively performing at identical levels.

Our results further showcase how the LLMs, without additional reinforcement learning post-training (that was applied in training the o1 model), are still not ready for complex medical tasks like recommending diagnostic tests and treatment. The o1 model belongs to a new family of LLM models, sometimes referred to as Large Reasoning Models (LRMs), due to specific differences from vanilla LLMs (like post-training reinforcement learning) [15]. These new LRMs spend much more time “thinking” or

"reasoning" through a problem by employing longer (and more problem-related) chains of thought, effectively allowing the models to spend more compute resources on a given problem [15]. Real-world medical cases, especially the ones concerning internal medicine, are often complex and contain many details and intricacies vital for providing correct diagnosis and treatment recommendations. Hence, spending more compute on a given case, allows the models to capture better all the important details that are necessary for solving the patient case correctly.

Interestingly, all three models performed similarly well in choosing the initial diagnostic tests. The models did show a tendency to "over-diagnose", ie. to order more tests than were needed to solve a particular patient case, but this could also be explained given the specific prompt used, where we did not explicitly state that the LLM should choose only the most important tests (check Supplementary Materials). Moreover, our results highlight how Claude-3.5-Sonnet and Llama-3-70B mostly underperformed in prescribing the appropriate medication and giving the appropriate recommendation. This underperformance in therapy prescription and non-medication recommendations could be mitigated by incorporating strategies like retrieval-augmented generation (RAG) and model fine-tuning [19]. By incorporating a RAG system connected to medical knowledge bases like UpToDate or StatsPearls, relevant treatment data, and guidelines could be fetched based on semantic similarity and pushed to the model's context [20]. While we did not observe classical "AI hallucinations" (i.e., completely fabricated facts or references) in our dataset, some inaccuracies arose due to the models omitting crucial therapeutic details or failing to recommend certain key interventions, what we term 'partial misses.' These oversights underscore the importance of robust clinical oversight and illustrate that even advanced LLMs may require targeted safeguards.

Another notable finding from our study, in contrast to previous research reporting poor performance of LLMs on medical tasks (e.g., Hager et al.), is that all three LLMs demonstrated a strong understanding of laboratory values and their corresponding reference ranges [7]. In the study by Hager et al., open-source models based on Llama-2-70B exhibited a complete inability to comprehend or interpret basic laboratory values [7]. This discrepancy highlights the significant advancements achieved within the same model family (Llama by Meta AI) through a single iteration of model development.

The differences in model performance complement the heterogeneous findings reported in recent studies. For instance, a large randomized clinical trial by Goh et al. examined whether an LLM (GPT-4) improves diagnostic reasoning in practicing physicians, finding that direct LLM access did not significantly enhance clinicians' diagnostic performance compared with conventional resources alone [13]. Intriguingly, GPT-4 alone (without human input) demonstrated higher performance than the physicians in some diagnostic tasks, emphasizing the nuanced and sometimes paradoxical nature of human-AI collaboration [13]. In a parallel line of investigation, Williams et al. evaluated GPT-3.5 and GPT-4 using a high-volume sample of 1000 emergency department visits, highlighting both models' tendency to overestimate risk (high

sensitivity, lower specificity) when making key clinical recommendations [21]. Their findings underscore that existing LLMs require additional calibration and domain-specific constraints to avoid overly conservative management strategies. We have noticed similar behavior when testing the LLM performance in determining patient admission. While all the tested LLMs performed well overall (accuracy > 90%), in nearly all the cases, the discrepancy in admission status, when compared to the real-world finding, mainly occurred in cases where the human physician did not recommend admission.

Our findings also highlight critical considerations for integrating LLMs into internal medicine. While these models demonstrate promise as decision-support tools, they should not (and are not ready to) replace clinical expertise. Instead, they can serve as valuable complements to clinicians by synthesizing complex data, suggesting possible diagnoses and treatments, and flagging errors. This model of AI-assisted healthcare emphasizes augmentation rather than replacement of human judgment. Additionally, given that our data show lower performance by certain models (Claude-3.5-Sonnet and Llama-3.2-70b) in therapy planning and non-medication recommendations, there is a tangible risk of patient harm if these models were utilized unsupervised in high-acuity environments such as the Emergency Department. This underscores the importance of incorporating strict human oversight, and model-specific guardrails before any real-world deployment, thereby minimizing the possibility of incorrect or incomplete therapeutic guidance. On the other hand, we can also envision a future where human overreliance on these increasingly capable models, along with a corresponding decline in critical thinking, might foster "automation bias," allowing subtle errors to go unchallenged and potentially compromising patient safety[22]. Moreover, there is also the issue of potential adversarial attacks, as demonstrated by the recent work of Khaleel et al. (2024), which demonstrates that medical LLMs remain vulnerable to adversarial attacks manipulating <1.2% of model parameters to generate harmful outputs, necessitating defense strategies like adversarial training and cross-domain monitoring for clinical deployment[23]. Finally, there is also the question of AI legal liability, as highlighted by Mensah et al. (2025), Ghana's legal case study reveals current frameworks struggle with AI liability apportionment, advocating for updated laws with proportional responsibility standards and mandated transparency in medical AI system approvals[24,25].

After solving the previously mentioned obstacles to deployment, LLMs could help standardize care in the emergency department by delivering consistent, evidence-based recommendations (especially within a RAG framework) [19]. They may also reduce diagnostic errors and improve patient outcomes, particularly in resource-limited environments or facilities without ready access to specialists, potentially mitigating critical gaps in care.

Several limitations must be considered in interpreting our findings. First, although our study draws on real-world patient data (from one institution), a larger multi-institutional sample would further enhance generalizability. Second, while we covered a broad spectrum of internal medicine conditions, a more

extensive study with rare disease cases would discern a more complete status of LLM performance in this use case. Lastly, the cost-effectiveness and workflow integration of implementing LLM-driven decision support remain unresolved, especially given the large compute costs of running the o1 model [15].

Given the expert-level performance of the o1 model and its attention to detail, one of the interesting future research areas would be to test the model's ability to discern mistakes in medical reports. We believe that the first modality of LLM integration into real-world clinical workflows will be as error-checkers, where the errors would be flagged and later reviewed by a human physician, potentially leading to a drastic reduction in costs attributed to medical mistakes. Future investigations could also incorporate randomized trials in clinical settings and study optimal ways of fostering human-LLM collaboration.

Conclusions

The o1 model performs on par with human experts in recommending diagnostic tests and treatment for real-world emergency internal medicine cases, indicating a potential for clinical integration. In the near term, this could include triage support, second-opinion recommendations for complex cases, and error-flagging during chart review. In contrast, other SOTA LLMs tested did not meet this benchmark. These findings underscore the potential of advanced language models for clinical decision support; however, further multi-center validation research trials are needed before they can be widely implemented in clinical settings.

References:

1. **Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M. & Božić, J. in Preprints (Preprints, 2024).**
2. **Yin, S. et al. A Survey on Multimodal Large Language Models. arXiv preprint (2024).** doi:10.48550/arXiv.2306.13549 (2024).
3. **Park, Y.-J. et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Medical Informatics and Decision Making 24, 72, doi:10.1186/s12911-024-02459-6 (2024).**
4. **Wang, Y., Ma, X. & Chen, W. Augmenting black-box llms with medical textbooks for clinical question answering. arXiv preprint arXiv:2309.02233 (2023).**
5. **Hendrycks, D. et al. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020).**

6. Anthropic. Claude-3.5-Sonnet; Available at: https://www.anthropic.com/news/claude-3-5-sonnet. (2024).

7. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine 30, 2613-2622, doi:10.1038/s41591-024-03097-1 (2024).

8. Gumilar, K. E. et al. Assessment of Large Language Models (LLMs) in decision-making support for gynecologic oncology. Computational and Structural Biotechnology Journal 23, 4019-4026, doi:https://doi.org/10.1016/j.csbj.2024.10.050 (2024).

9. Marchi, F., Bellini, E., Iandelli, A., Sampieri, C. & Peretti, G. Exploring the landscape of AI-assisted decision-making in head and neck cancer treatment: a comparative analysis of NCCN guidelines and ChatGPT responses. Eur Arch Otorhinolaryngol 281, 2123-2136 (2024).

10. Miller, L. et al. A Comparative Evaluation of Large Language Model Utility in Neuroimaging Clinical Decision Support. J Imaging Inform Med 7, 024-01161 (2024).

11. Masanneck, L. et al. Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. J Med Internet Res 26, e53297, doi:10.2196/53297 (2024).

12. Benary, M. et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Network Open 6, e2343689-e2343689, doi:10.1001/jamanetworkopen.2023.43689 (2023).

13. Goh, E. et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. JAMA network open 7, e2440969-e2440969, doi:10.1001/jamanetworkopen.2024.40969 (2024).

14. Bedi, S. et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA 15, 21700 (2024).

15. Jaech, A. et al. Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024).

16. Tam, T. Y. C. et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine 7, 258, doi:10.1038/s41746-024-01258-7 (2024).

17. Sandmann, S., Riepenhausen, S., Plagwitz, L. & Varghese, J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 15, 024-46411 (2024).

18. Gumilar, K. E. et al. Assessment of Large Language Models (LLMs) in decision-making support for gynecologic oncology. Comput Struct Biotechnol J 23, 4019-4026 (2024).

19. Bora, A. & Cuayáhuitl, H. Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. Machine Learning and Knowledge Extraction 6, 2355-2374 (2024).

20. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. arXiv preprint arXiv:2402.13178 (2024).

21. Williams, C. Y. K., Miao, B. Y., Kornblith, A. E. & Butte, A. J. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. Nature Communications 15, 8236, doi:10.1038/s41467-024-52415-1 (2024).

22. Goddard, K., Roudsari, A. & Wyatt, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19, 121-127 (2012).

23. Khaleel, Y. L. Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches. Applied Data Science and Analysis 2024, 121-147, doi:10.58496/adsa/2024/011 (2024).

24. Bottomley, D. & Thaldar, D. Liability for harm caused by AI in healthcare: an overview of the core legal concepts. Front Pharmacol 14 (2023).

25. Mensah, G. B., Mijwil, M. M., Abotaleb, M., Ali, G. & Dutta, P. K. High Performance Medicine: Involving Artificial Intelligence Models in Enhancing Medical Laws and Medical Negligence Matters A Case Study of Act, 2009 (Act 792) in Ghana. SHIFAA 2025, 1-6, doi:10.70470/shifaa/2025/001 (2025).

Supplementary Materials

Table 1. Total diagnosis frequency

Diagnosis (ICD-10)	Count	Percentage
J18.9 Pneumonia, unspecified	5	6.8 %
I48 Atrial fibrillation and flutter	4	5.5 %
N39.0 Urinary tract infection	4	5.5 %

I50.0 Congestive heart failure	4	5.5 %
J44.9 Chronic Obstructive Pulmonary Disease, Unspecified	4	5.5 %
I.10 Essential Hypertension	3	4.1 %
I21 Acute myocardial infarction	3	4.1 %
K29. 70 Gastritis, unspecified	2	2.7 %
R07 Pain in throat and chest	2	2.7 %
K64. 9 Unspecified hemorrhoids	2	2.7 %
K92. 1 Melena	2	2.7 %
D50 Iron deficiency anemia	2	2.7 %
R10 Abdominal and pelvic pain	2	2.7 %
I26 Pulmonary embolism	2	2.7 %
K85.0 Acute inflammation of the pancreas (acute pancreatitis)	2	2.7 %
K81.0 Acute cholecystitis	2	2.7 %
J20 Acute bronchitis	2	2.7 %
R21 Rash, unspecified	1	1.4 %
R01.1 Heart murmur, unspecified	1	1.4 %
N13.30 Unspecified hydronephrosis	1	1.4 %
E11.10 Type 2 diabetes mellitus with ketoacidosis without coma	1	1.4 %
U07.1 COVID-19	1	1.4 %
E87.1 Hypo-osmolality and hyponatraemia	1	1.4 %
A09.0 Diarrhea and gastroenteritis of presumed infectious origin	1	1.4 %
I47.1 Supraventricular tachycardia	1	1.4 %
K92.0 Hematemesis	1	1.4 %
K52.9 Noninfective gastroenteritis and colitis	1	1.4 %
I63.81 Lacunar ischemic stroke	1	1.4 %

F41.0 Panic disorder [episodic paroxysmal anxiety]	1	1.4 %
R60 Edema, unspecified	1	1.4 %
K56 Unspecified ileus	1	1.4 %
I80.0 Acute thrombophlebitis of right lower leg	1	1.4 %
M02 Postreactive arthritis	1	1.4 %
K83.0 Inflammation of Bile Duct (Cholangitis)	1	1.4 %
C34 Malignant neoplasm of the bronchus and lung	1	1.4 %
I87.2 Venous insufficiency (chronic)	1	1.4 %
D63.0 Anemia in neoplastic diseases	1	1.4 %
J22.0 Unspecified acute lower respiratory tract illness	1	1.4 %
K51 Suspected Ulcerative Colitis	1	1.4 %
R10.4 Other and unspecified abdominal pain	1	1.4 %
K92 Other diseases of the digestive system	1	1.4 %
M11.20 Other chondrocalcinosis, unspecified site	1	1.4 %
Note: Some patients had multiple diagnoses		

Prompt #1:

“

You are an experienced emergency internal medicine physician. Given the following patient information, provide:

1. A list of potential preliminary diagnoses (differential diagnosis)
2. A detailed list of diagnostic tests you would order, including specific lab tests and imaging studies

Patient Information:

[Insert patient's signs, symptoms, medical history, and reason for arrival here, excluding any actual diagnosis or conclusion]

Format your response as follows:
Preliminary Differential Diagnosis:
1. [Diagnosis 1]
2. [Diagnosis 2]
3. [Diagnosis 3]
...

Recommended Diagnostic Tests:
Laboratory Tests:
1. [Test 1]
2. [Test 2]
...

Imaging Studies:
1. [Study 1]
2. [Study 2]
...

Other Diagnostic Procedures:
1. [Procedure 1]
2. [Procedure 2]
...

Provide a brief explanation for each recommended test or study.
“

Prompt #2:
“
You are an experienced emergency medicine physician. Earlier, you ordered diagnostic tests for a patient based on their initial presentation. Now, you have received the results of these tests. Given this new information:
1. Provide a final diagnosis (or diagnoses if multiple conditions are present)
2. Develop a detailed treatment plan and follow-up recommendations

Patient Information:
[Same as before]

Diagnostic Test Results:
[Insert all diagnostic test results here]

Format your response as follows:
Final Diagnosis:
1. [Primary Diagnosis]
2. [Secondary Diagnosis (if applicable)]
...

Treatment Plan (also provide specific treatment that will be administered while the patient is in the ER, if applicable):
1. Medications:
   - [Medication 1]: [Dosage and frequency]
   - [Medication 2]: [Dosage and frequency]
   ...

2. Procedures (if necessary):
   - [Procedure 1]
   - [Procedure 2]
   ...

3. Lifestyle Recommendations:
   - [Recommendation 1]
   - [Recommendation 2]
   ...

4. Follow-up Care:
   - [Follow-up appointment 1]
   - [Follow-up test 1]
   ...

5. Patient Education:
   - [Education point 1]
   - [Education point 2]
   ...

6. Other Recommendations (if necessary):
   -
7. Hospital Admission (should the patient be admitted or not)

Provide a brief explanation for each component of the treatment plan, including the rationale behind medication choices and follow-up recommendations.
“

Differences in LLM responses (examples):

Real-world report (human):

Date of Birth: 1938

Triage Category: 4
Diagnoses:
   • J18.9 Pneumonia, unspecified
   • I10 Essential (primary) hypertension
   • J90.0

• Z00 General examination and investigation of persons without complaints or diagnosis (systematic examination)

Triage Data:

- Triage Category: 4
- Blood Pressure: 120/80 mmHg
- Pulse: 62/min
- SpO2: 96%
- Temperature: 36.1°C
- VAS: 4 - Moderate pain
- GCS: 0

Reason for Visit:

- Pain in the left shoulder and left side of the chest

Medical History**:**

Two years ago, the patient experienced thromboembolism of the right lower leg (17.09.2022 Thromboembolectomy AP dex. sec Fogarty) and has regular follow-ups with a vascular surgeon. He is treated for arterial hypertension and gout and is aware of lumboischialgia. Hospitalized in Gastroenterology from 23.4. to 30.4.2024 due to abdominal pain. Diagnostic evaluation confirmed a splenic infarction. Abdominal surgery consultation found no indication for urgent surgical treatment. On 29.6.2024, he had a private gastroenterological examination including colonoscopy at Polyclinic Dr. Petrović: duplex colonic polyp, colonic diverticulosis, internal hemorrhoids. Polypectomy was not performed due to ongoing anticoagulant therapy. Advised to undergo polypectomy in the rectal area and the oral part of the sigmoid colon for one polyp with a short stalk, approximately 1.5 cm in size. He has not been on anticoagulant therapy since.

Current Illness:

For the past month, the patient has felt pain in the left shoulder and left side of the chest, which intensified yesterday. No cough or sputum. Afebrile. Denies dyspneic symptoms.

Function:

- Appetite: Reduced
- Weight Loss: Yes, but does not know how much; denies weight fluctuation
- Stool: Regular, last two days ago, without pathological admixtures
- Urination: Normal, without dysuria

Habits:

- Non-smoker
- Does not drink alcohol

Therapy:

- Pantoprazole 40 mg 1x1
- Ormidol 50 mg 1x1
- Tomid 10 mg 1x1
- Allopurinol 100 mg 1x1
- Colospa 200 mg 1x1

• Xarelto 2.5 mg 2x1 - not taken for 7 days!
• Buscol tablets as needed
• Betaserc 24 mg 1x1
• Detralex as needed
• Zaldiar tablets as needed

Allergies: Denies medication allergies.

Status:

• Musculoskeletal build: Robust, obese, conscious, cooperative, oriented, mobile with a cane, non-icteric, afebrile, well-hydrated, skin and visible mucous membranes well-perfused, tongue moist, not coated, midline, without peripheral lymphadenopathy, neck veins well filled, no vascular murmurs over the carotids, lumbar percussion bilaterally negative.
• Chest: Normal shape, symmetrical, respirations evenly movable.
• Lungs: On auscultation, reduced to inaudible breath sounds on the left. SpO2 97%.
• Heart: Regular rhythm, muffled tones, no murmurs heard, BP 120/80 mmHg, pulse 62/min.
• Abdomen: Higher than the level of the chest, soft, elastic, non-tender on palpation, no organomegaly, audible peristalsis.
• Limbs: Symmetrical, trophic changes in both calves, no edema, peripheral pulses palpable.

Diagnostic and Consultative Examinations:

• EKG: Sinus rhythm, rate 62/min, physiological electrical axis, single VES, no ST segment changes
• Blood tests: WBC 7.1 x10^9/L, RBC 5.23 x10^12/L, HGB 159 g/L, HCT 0.468 L/L, MCV 89.5 fL, MCH 30.4 pg, MCHC 340 g/L, RDW-CV 14.6%, Platelets 126 x10^9/L, MPV 8.8 fL, NEU% 60.7%, LYM% 26.8%, MONO% 10.1%, EOS% 1.8%, BASO% 0.6%, NEU 4.28 x10^9/L, LYM 1.89 x10^9/L, MONO 0.71 x10^9/L, EOS 0.13 x10^9/L, BASO 0.04 x10^9/L, NRBC/100 0/100 WBC, Urea 4.6 mmol/L, Creatinine 91 µmol/L, eGFR 66.0 mL/min/1.73m², AST / U/L, ALT 30 U/L, GGT 55 U/L, LDH / U/L, CK 55 U/L, Protein 80 g/L, Albumin 38.5 g/L, NT-proBNP 465 pg/mL, hs-TNT 21.7 ng/L, Na 136 mmol/L, K / mmol/L
• Chest X-ray (PA and Lateral): Homogeneous shadowing of the lower and middle lung fields on the left, indicating pleural effusion reaching the 5th rib with consolidated lung parenchyma parahilar region. Vascular hila with pronounced terminal branching. Normal lung parenchyma transparency on the right. Right hemidiaphragm normal. Heart shadow not adequate for analysis. Calcified aortic arch.
• Abdominal X-ray (standing): No signs of free air under the diaphragm, intestinal distension, or formed air-fluid levels. Pleural effusion on the left.
• Lab values: PT-INR 1.01, APTT 22.3 s, APTT Ratio 0.83, Glucose Serum 6.0 mmol/L, Urea 4.6 mmol/L, Creatinine 93 µmol/L, eGFR 64.3 mL/min/1.73m², AST 54 U/L, ALT 29 U/L, GGT 58 U/L, LDH 577 U/L, Albumin 39.2 g/L, Protein 81 g/L, CRP 19.7 mg/L, NT-proBNP 458 pg/mL, hs-TNT 21.8 ng/L, Na 137 mmol/L, K 3.8 mmol/L, CK 46 U/L
• Lung Ultrasound: Thoracocentesis left - moderate to large anechoic pleural effusion. 1000 ml of slightly hemorrhagic clear fluid was evacuated. No complications.
• Pleural effusion: Specific gravity 1.015 kg/L, Glucose 6.3 mmol/L, Total Protein 50.6 g/L, Albumin 30.2 g/L, LD 745 U/L – EXUDATE

Therapy:

• Tramal 50 mg SC 1x

Conclusion:
Senior service consulted! Clinical examination, lab tests, cardiac enzyme monitoring, and EKG ruled out acute coronary event. During observation, the patient was in good general condition, without anginal symptoms or dyspnea, cardiopulmonary compensated, and arrhythmologically stable. Thoracentesis performed with evacuation of 1000 ml of slightly hemorrhagic effusion of exudative biochemical characteristics. Advised to perform MSCT of the chest and follow-up with a pulmonologist with findings. Advised to take Klavocin 2x1 for 10 days with Sumamed 500 mg 1x1 for 3 days and Xarelto 15 mg 1x1 (patient does not prefer Fragmin SC, do not exclude from therapy, please refer to discharge summary and anamnesis). For pain, take Zaldiar 2-3x1 or Paracetamol 1000 mg 3-4x1. Gastroenterologist follow-up on an outpatient basis. Control of tumor markers - CYFRA 21.1, NSE, SCC, CEA, PSA, fPSA, AFP.
In case of worsening, report to the appropriate emergency service!

The o1 model report:

Response to prompt #1:

Preliminary Differential Diagnosis:

1. Left-sided pleural effusion or pulmonary infiltrate (e.g., pneumonia or lung mass)
2. Musculoskeletal shoulder pathology (e.g., rotator cuff tendinopathy, arthritis, referred pain from cervical spine)
3. Pancoast tumor (apical lung tumor) causing shoulder and chest pain
4. Cardiac ischemia or atypical angina
5. Recurrent pulmonary embolism or pulmonary infarction (given past thromboembolic history)
6. Subdiaphragmatic pathology (splenic issues, subdiaphragmatic abscess) with referred shoulder (Kehr's sign) pain

---

Recommended Diagnostic Tests:
Laboratory Tests:

1. Complete Blood Count (CBC):
    - To assess for signs of infection (elevated WBC count), anemia, or other hematological abnormalities.
2. C-Reactive Protein (CRP) and Erythrocyte Sedimentation Rate (ESR):
    - To detect inflammation or infection. Elevated markers may indicate pneumonia, pleural effusion, or another inflammatory process.
3. Cardiac Enzymes (Troponin, CK-MB):
    - To rule out cardiac ischemia or ongoing myocardial injury as a cause of chest pain.
4. D-Dimer:
    - Given the history of thromboembolism, a D-dimer test can help rule out acute venous thromboembolism (e.g., pulmonary embolism) if clinical suspicion persists.

5. Basic Metabolic Panel (Electrolytes, Renal Function, Glucose):
    - To assess overall metabolic and renal status, ensuring safe use of contrast if imaging with CT is needed and to detect any electrolyte imbalances.
6. Liver Function Tests:
    - To obtain a baseline since the patient has a complex history including GI issues and splenic infarction.

Imaging Studies:

1. Chest X-ray (Posteroanterior and Lateral Views):
    - First-line imaging to evaluate reduced/inaudible breath sounds on the left.
    - Can identify pleural effusion, consolidation, lung mass, or other structural abnormalities.
2. Chest Ultrasound:
    - If a pleural effusion is suspected on the chest X-ray, an ultrasound can better characterize fluid presence and guide possible thoracentesis.
3. Electrocardiogram (ECG):
    - Although not an "imaging" study per se, it is critical to assess for cardiac ischemia, arrhythmias, or other cardiac causes of chest pain.
4. Computed Tomography (CT) Scan of the Chest (if indicated by initial findings):
    - For further evaluation of suspicious lesions on the X-ray, suspected tumors (e.g., Pancoast tumor), or subtle pulmonary infarcts not clearly visible on standard imaging.
    - Can also evaluate mediastinal structures and rule out aortic pathologies if considered.

Other Diagnostic Procedures:

1. Echocardiogram:
    - To evaluate cardiac function, rule out pericardial effusion, or structural cardiac abnormalities, especially if muffled heart sounds are a concern or if chest X-ray/ECG suggests cardiac involvement.
2. Orthopedic or Physiotherapy Assessment (if indicated**)**:
    - If initial cardiopulmonary investigations are non-diagnostic, an assessment of the shoulder girdle and cervical spine can identify musculoskeletal causes of pain.
3. Possible Gastroenterological Follow-Up:
    - If chest imaging and labs are unremarkable, consider repeating abdominal imaging (e.g., abdominal ultrasound) to evaluate for any subdiaphragmatic process that could refer pain to the shoulder (such as a splenic issue or subphrenic abscess).

Explanation for Each Recommended Test/Study:

- CBC, CRP, ESR: Identify signs of infection or inflammation that could point toward pneumonia, pleuritis, or another inflammatory condition.
- Cardiac Enzymes & ECG: Evaluate for ischemic heart disease, as cardiac pain can radiate and mimic musculoskeletal or pleural pain.
- D-Dimer: Given past thromboembolic events, this helps assess the risk of new pulmonary embolism.
- Chest X-ray: The most crucial initial imaging to identify structural lung or pleural abnormalities.
- Chest Ultrasound: Provides a more detailed look at potential pleural effusions.
- CT Chest: Offers detailed imaging if the X-ray is inconclusive or if suspicious findings (like a mass or subtle infiltrates) arise.
- Echocardiogram: Rules out cardiac causes like pericardial effusion or valvular issues contributing to chest discomfort.
- Orthopedic/Physiotherapy Assessment: If no cardiopulmonary cause is found, musculoskeletal issues must be considered.
- Abdominal Imaging: If suspicion arises that abdominal pathology (e.g., splenic complications) is causing referred chest/shoulder pain.

Response to prompt #2:

Final Diagnosis:

1. Left-sided parapneumonic pleural effusion secondary to pneumonia
2. Underlying stable comorbidities (arterial hypertension, gout)

Treatment Plan:

1. Medications:
    - IV Antibiotics (e.g., Ceftriaxone 2 g IV once daily + Azithromycin 500 mg IV once daily):
      Rationale: Given the presence of a pleural effusion with consolidation on imaging, an exudative effusion, and mild elevated inflammatory markers, a parapneumonic effusion is likely. Broad-spectrum empirical antibiotics targeting typical and atypical organisms involved in community-acquired pneumonia are indicated.
    - Analgesics (e.g., Paracetamol 1 g IV/PO every 6–8 hours as needed):
      Rationale: To manage chest and shoulder pain, ensuring patient comfort and facilitating deep breathing to prevent atelectasis.
    - Diuretics (if indicated, e.g., Furosemide 20 mg IV once):
      Rationale: May be considered if there is any suspicion of fluid overload or borderline cardiac function. However, the primary issue is exudative, not transudative, so the mainstay is antibiotic therapy, not diuresis.

- Continue Current Medications for Comorbidities:
  Rationale: Ensure stable management of hypertension, gout, and gastric protection.

2. Procedures (if necessary):
   - Thoracocentesis (already performed**)**:
     Rationale: Diagnostic and therapeutic thoracentesis has already been done, draining 1000 mL of effusion. This improves respiratory mechanics and helps alleviate symptoms.
   - Chest Tube Placement (if recurrent effusion or complicated effusion):
     Rationale: If the fluid re-accumulates rapidly or if there are signs of a complicated (loculated) effusion, a chest tube placement may be required for adequate drainage.
3. Lifestyle Recommendations:
   - Encourage Adequate Fluid Intake and Balanced Nutrition:
     Rationale: Supports overall health and recovery, aids in preventing constipation, and supports immune function.
   - Smoking Cessation Advice (if applicable):
     Rationale: Patient is currently a non-smoker, so continue to avoid tobacco. This is essential for lung health.
4. Follow-up Care:
   - Pulmonology Follow-up in 1–2 Weeks:
     Rationale: To reassess lung status, ensure resolution of effusion and pneumonia, and evaluate if further drainage or investigations are required.
   - Repeat Chest Imaging (Chest X-ray) in 2–3 Weeks:
     Rationale: To monitor resolution of the pleural effusion and confirm improvement of pneumonia.
   - Pleural Fluid Analysis (if indicated):
     Rationale: If suspicion of complicated effusion or malignancy arises, additional fluid analysis, cytology, or a repeat sampling may be needed.
5. Patient Education:
   - Explain the Condition (Parapneumonic Effusion):
     Rationale: Ensuring the patient understands that the chest pain and effusion are related to an underlying lung infection that requires antibiotic treatment.
   - Medication Adherence and Warning Signs:
     Rationale: Instruct on taking antibiotics as prescribed, recognizing warning signs such as increased shortness of breath, fever, worsening pain, or inability to eat/drink adequately that would prompt earlier reevaluation.
6. Other Recommendations (if necessary):
   - Physical Therapy / Deep Breathing Exercises:
     Rationale: Encouraging the patient to perform incentive spirometry or deep breathing exercises to improve lung expansion and prevent atelectasis.

7. Hospital Admission:
    - Given the patient's age, complexity of presentation, presence of a significant parapneumonic effusion, and need for IV antibiotics, admission for at least short-term inpatient treatment is recommended.
      Rationale: Close monitoring for clinical improvement, response to antibiotics, and potential recurrence of the effusion is warranted. The patient's advanced age and comorbidities justify careful in-hospital observation.

Explanation:

- The final diagnosis of a parapneumonic pleural effusion is supported by the presence of a left-sided pleural effusion with consolidation on imaging, exudative fluid analysis (high protein, high LDH), and the clinical picture of chest pain and mild inflammatory markers. Treatment includes appropriate IV antibiotics to address the underlying pneumonia, analgesia for symptom control, and hospitalization for close monitoring and supportive care. Follow-up imaging and specialist evaluation ensure resolution of the effusion and prevent complications.

Claude-3.5-Sonnet report:

Response to prompt #1:

Preliminary Differential Diagnosis:

Pleural Effusion (Left-sided) - Given reduced/inaudible breath sounds on left side
Pulmonary Embolism - History of thromboembolism and recent cessation of anticoagulation
Acute Coronary Syndrome - Age and risk factors, though atypical presentation
Musculoskeletal Pain - Could be referred from shoulder pathology
Pneumothorax - Given reduced breath sounds, though vital signs stable
Pericardial Effusion - Given muffled heart tones
Thoracic Aortic Pathology - Age and hypertension history
Pneumonia - Though afebrile and denies respiratory symptoms

Recommended Diagnostic Tests: Laboratory Tests:

Complete Blood Count (CBC) - Evaluate for infection, anemia, thrombocytosis
Basic Metabolic Panel - Assess kidney function and electrolytes
Cardiac Enzymes (Troponin, CK-MB) - Rule out cardiac injury
D-dimer - Screen for thromboembolism
BNP (Brain Natriuretic Peptide) - Evaluate for heart failure
Arterial Blood Gas - Assess oxygenation and acid-base status
CRP and ESR - Evaluate inflammatory response
PT/INR, aPTT - Coagulation status given anticoagulation history

Imaging Studies:

Chest X-ray (PA and Lateral) - First-line imaging to evaluate pleural effusion, pneumothorax, infiltrates
CT Pulmonary Angiogram - Gold standard for PE diagnosis, also evaluates other thoracic pathology
Bedside Echocardiogram - Evaluate cardiac function, pericardial effusion
CT Aorta with contrast - If concerned for aortic pathology
Ultrasound of left shoulder - If musculoskeletal cause suspected after ruling out serious pathology

Other Diagnostic Procedures:

12-lead ECG - Evaluate for cardiac ischemia or other cardiac abnormalities
Pulse oximetry monitoring - Continuous monitoring of oxygenation
Thoracentesis - If significant pleural effusion identified, for diagnostic and potentially therapeutic purposes
Shoulder examination - Detailed physical exam to evaluate for musculoskeletal cause

Key considerations:

Patient's recent cessation of anticoagulation (Xarelto) is particularly concerning given history of thromboembolism
Reduced/inaudible breath sounds on left side strongly suggests pleural pathology
Muffled heart tones warrant cardiac evaluation
Age and risk factors increase concern for serious cardiopulmonary pathology

Response to prompt#2:

Final Diagnosis:

Exudative Pleural Effusion (Primary) - confirmed by diagnostic thoracentesis and fluid analysis
Suspected Pulmonary Embolism - given history and recent anticoagulation cessation
Chronic Arterial Hypertension (Controlled)
Chronic Gout

Treatment Plan:

Medications:
  Xarelto (Rivaroxaban): Resume at 2.5mg BID immediately
  Furosemide: 40mg PO daily for 7 days
  Continue current medications:
    Pantoprazole 40mg daily
    Allopurinol 100mg daily
    Other prescribed medications as before
Procedures:
  Completed therapeutic thoracentesis with removal of 1000ml fluid
  Schedule CT Pulmonary Angiogram within 24 hours to evaluate for PE
  Consider repeat thoracentesis if symptoms recur
Lifestyle Recommendations:
  Daily weight monitoring

- Moderate sodium restriction (<2g/day)
- Regular physical activity as tolerated
- Report any increased shortness of breath or chest pain immediately

Follow-up Care:

Immediate (within 24-48 hours):
- CT Pulmonary Angiogram
- Follow-up chest X-ray

Within 1 week:
- Primary care physician visit
- Vascular surgery consultation

Within 2 weeks:
- Pulmonology consultation for pleural effusion management
- Repeat chest X-ray

Patient Education:
- Importance of medication compliance, especially anticoagulation
- Warning signs requiring immediate medical attention:
  - Increased shortness of breath
  - Chest pain
  - Fever
  - Leg swelling or pain
- Proper technique for daily weight monitoring
- Dietary sodium restrictions

Other Recommendations:
- Consider outpatient pulmonology referral for long-term management
- Monitor renal function and electrolytes while on diuretics
- Consider screening for underlying malignancy given exudative effusion
- Document baseline functional status for comparison

Rationale:

- Exudative effusion confirmed by fluid analysis (protein and LDH criteria met)
- Anticoagulation crucial given history of thromboembolism and recent cessation
- Diuretic therapy to prevent fluid reaccumulation
- Close follow-up needed due to:
  - Age and comorbidities
  - Need to identify underlying cause of effusion
  - Risk of recurrence
  - Need for anticoagulation monitoring
- CT Pulmonary Angiogram essential to rule out PE given risk factors
- Multiple specialist involvement needed for comprehensive care

The treatment plan addresses both immediate symptom management and long-term prevention of recurrence while investigating the underlying cause of the exudative effusion.

Llama-3-70B report:

Response to prompt#1:

Preliminary Differential Diagnosis:
1.Pulmonary Embolism
2.Pneumothorax
3.Pleural Effusion
4.Acute Coronary Syndrome
5.Pericarditis
6.Musculoskeletal Pain (e.g., costochondritis)
7.Splenic Infarction (recurrence or extension)
8.Pneumonia
9.Aortic Dissection
Recommended Diagnostic Tests:
Laboratory Tests:
1.Complete Blood Count (CBC): To assess for infection, anemia, or elevated white blood cell count.
2.Basic Metabolic Panel (BMP): To evaluate electrolyte imbalances and kidney function.
3.Cardiac enzymes (Troponin, CK-MB): To rule out myocardial infarction.
4.D-dimer: To help rule out pulmonary embolism.
5.NT-proBNP: To assess for heart failure.
6.Arterial Blood Gas (ABG): To evaluate oxygenation and acid-base status.
7.C-reactive protein (CRP) and Erythrocyte Sedimentation Rate (ESR): To assess for inflammation.
8.Liver Function Tests (LFTs): To evaluate liver function and potential splenic involvement.

Imaging Studies:
1.Chest X-ray (PA and lateral views): To evaluate for pneumothorax, pleural effusion, pneumonia, or cardiac enlargement.
2.CT Pulmonary Angiography (CTPA): To rule out pulmonary embolism and assess for other thoracic pathologies.
3.Echocardiogram: To evaluate cardiac function, pericardial effusion, and potential valvular issues.
4.CT Abdomen with contrast: To reassess the splenic infarction and evaluate for other abdominal pathologies.

Other Diagnostic Procedures:
1.12-lead ECG: To evaluate for acute coronary syndrome, pericarditis, or other cardiac abnormalities.
2.Pulse oximetry: Continuous monitoring of oxygen saturation.
3.Thoracentesis (if pleural effusion is identified): For diagnostic and potentially therapeutic purposes.
Each test and study is recommended based on the patient's presenting symptoms, medical history, and physical examination findings. The combination of these diagnostic tools will help differentiate between the potential diagnoses and guide appropriate treatment. Given the patient's history of thromboembolism and recent cessation of anticoagulant therapy, particular attention should be paid to ruling out pulmonary embolism. The reduced breath sounds on the left side also warrant thorough investigation of potential pleural or pulmonary pathologies.

Response to prompt#2:

Final Diagnosis:
1.Exudative Pleural Effusion (Left Side)

2.Possible Pulmonary Embolism (given history of thromboembolism and recent cessation of anticoagulation)
3.Hypertension (pre-existing condition)
4.Gout (pre-existing condition)

Treatment Plan:
1.Medications:
•Enoxaparin: 1 mg/kg subcutaneously twice daily
•Acetaminophen: 1000 mg orally every 6 hours as needed for pain
•Resume previous medications:
•Pantoprazole 40 mg orally once daily
•Tomid (bisoprolol) 10 mg orally once daily
•Allopurinol 100 mg orally once daily
2.Procedures:
•CT Pulmonary Angiography (CTPA) to rule out pulmonary embolism
•Repeat thoracentesis if effusion reaccumulates

3.Lifestyle Recommendations:
•Gradual increase in physical activity as tolerated
•Maintain proper hydration
•Follow a low-sodium diet to help manage hypertension and prevent fluid retention

4.Follow-up Care:
•Outpatient follow-up with pulmonologist in 1 week
•Repeat chest X-ray in 1 week
•Follow-up with vascular surgeon within 2 weeks
•Continue regular follow-ups with gastroenterologist for polyp management

5.Patient Education:
•Explain the nature of pleural effusion and its potential causes
•Stress the importance of anticoagulation therapy and medication adherence
•Discuss signs and symptoms that should prompt immediate medical attention (e.g., increased shortness of breath, chest pain, fever)

6.Other Recommendations:
•Consider hematology consultation for further evaluation of hypercoagulable state

Explanation:
The primary diagnosis is an exudative pleural effusion based on the chest X-ray findings and the results of the thoracentesis. The fluid analysis meets Light's criteria for an exudate (protein > 30 g/L and LDH > 200 U/L), suggesting an inflammatory process. Given the patient's history of thromboembolism and recent cessation of anticoagulation, pulmonary embolism remains a possibility and needs to be ruled out with CTPA. Enoxaparin is prescribed as a bridge anticoagulation therapy until the CTPA results are available and a long-term anticoagulation plan can be established. Acetaminophen is prescribed for pain management, as it is safe and effective for moderate pain. The patient's previous medications for hypertension and gout are resumed to manage these chronic conditions. The follow-up plan includes monitoring the pleural effusion with a repeat chest X-ray and pulmonologist evaluation. The vascular

surgeon follow-up is crucial for managing the patient's thromboembolism risk. The gastroenterologist follow-up is maintained for polyp management. Patient education focuses on understanding the condition, recognizing warning signs, and the importance of medication adherence, particularly anticoagulation therapy. The hematology consultation is recommended to evaluate for any underlying hypercoagulable state that may be contributing to the patient's recurrent thrombotic events. This comprehensive approach addresses the acute pleural effusion, investigates potential underlying causes, manages chronic conditions, and provides a framework for ongoing care and monitoring.

Author Contributions

J.V. was heavily involved for project planning and the initial idea, conducted model prompting, and drafted the final manuscript.
J.B. served as the co-lead, contributed to project planning, and reviewed the final manuscript.
S.V. served as co-lead, aided in project planning and methodological design, and reviewed the final manuscript.
Z.B. performed the statistical analyses, contributed to data visualization, and supported the writing of the manuscript.
R.S. contributed to data collection and model prompting, and was also involved in writing and reviewing the manuscript draft.
I.M. contributed to data collection, assisted in prompting the models, and helped with literature searching.
A.O. assisted in data collection, model prompting, and contributed to literature searching.
A.C. also assisted in data collection, model prompting, and contributed to literature searching.
M.K. aided in data collection, offered input during the manuscript drafting phase, and supported the literature search.
All authors read and approved the final manuscript.

Competing interests

All authors declare no financial or non-financial competing interests.

Data availability

De-identified patient data, and all LLM generated reports are available at the following repository: https://github.com/vrda23/Evaluating-Large-Language-and-Large-Reasoning-Models-as-Decision-Support-Tools-in-Medicine/tree/main.