International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(6s) 2026 pp. 918-936
https://doi.org/10.70917/ijcisim-2026-2996
Article

Evaluating the Feasibility of Llama 3.2:3b for Automating Bone Scan Report Structuring

Arun Kumar Bashambu[1], Manisha Agarwal[2], Shallu Bashambu[3*]

[1]Banasthali Vidyapith, Jaipur. Email: arunbashambu@rediffmail.com.
[2]Banasthali Vidyapith, Jaipur. Email: manishaagarwal18@yahoo.co.in.
[3*]IT Dept, MAIT, Email: shallubashambu@mait.ac.in

Abstract: Background: Pediatric bone scan reports often contain rich clinical information embedded within lengthy, unstructured narratives, a known challenge in radiology NLP systems [1,2]. Due to inconsistent formatting and variable reporting styles, valuable diagnostic details frequently remain hidden or overlooked, posing major challenges for data integration, research analytics, and automated quality control in nuclear medicine. This growing need for structured, standardized documentation has led to increasing interest in the use of large language models (LLMs) for automated report structuring [3,4]. Objective: This study evaluates the feasibility of locally deployed LLMs for converting unstructured pediatric bone scan reports into structured, machine-readable formats without compromising data privacy, an increasingly important concern in clinical AI applications [5]. Methods: A dataset of 134 de-identified 99mTc-MDP pediatric bone scan reports was analyzed using A dataset of 134 de-identified 99mTc-MDP pediatric bone scan reports was analyzed using the Llama 3.2:3B model under Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT) prompting configurations, reflecting established prompting paradigms in LLM research [6–8]. All experiments were performed through the Ollama framework, ensuring deterministic, on-device inference. Model outputs were benchmarked against expert-annotated ground truth using sensitivity, specificity, PPV, NPV, accuracy, and F1-score, metrics widely used in clinical NLP evaluation [9,10]. Results: The ZS model demonstrated high specificity and NPV (>0.90), indicating strong "rule-out" capability but reduced sensitivity for rare positive findings, consistent with expected behavior under class-imbalanced clinical settings [11]. The FS model achieved balanced improvements in diagnostic extraction (F1 up to 0.91) despite class imbalance. Conversely, CoT prompting amplified conservative bias, increasing false negatives, a known limitation of certain reasoning-style prompts [12]. Conclusion: Local LLM deployment proved technically feasible and clinically valuable, enabling high-throughput, privacy-preserving structuring of pediatric bone scan reports and aligning with current recommendations for privacy-respecting AI in healthcare [5]. These findings establish a reproducible framework for secure, low-cost AI integration into clinical nuclear medicine workflows.

Keywords: Machine Learning; Artificial Intelligence; Bio-Technology

1. INTRODUCTION

Bone scintigraphy is a key tool in pediatric nuclear medicine and helps detect skeletal metastases, infection, trauma, and metabolic bone disorders with high sensitivity [13,14]. The imaging process is standardized, but the associated reports often appear as long, free-text narratives that vary widely in structure and detail across clinicians and institutions. Pediatric reports show even greater variation because of age-specific anatomy, developmental changes, and the need for careful interpretation of tracer uptake patterns. Important findings—such as focal lesions, growth-plate abnormalities, hardware issues, or extra-osseous uptake—may therefore be inconsistently described or hard to locate without manual review.

This unstructured style creates ongoing challenges for clinical workflows, research work, and automated quality assurance. Clinically, the lack of consistent terminology limits interoperability and makes it difficult to compare serial studies in children undergoing follow-up for malignancies or metabolic disorders. For research, variation in free-text reporting slows cohort selection, complicates retrospective audits, and restricts the creation of large datasets needed for epidemiological or outcomes-based studies. Quality control processes also depend on structured metrics, such as

detection rates for key abnormalities, which are difficult to calculate from narrative reports without time-consuming manual extraction. These issues highlight the need for scalable tools that can convert unstructured pediatric nuclear medicine reports into structured, machine-readable formats.

Recent progress in natural language processing (NLP) and large language models (LLMs) offers new opportunities for semantic extraction of diagnostic information from radiology reports [1,2,15]. However, many available solutions rely on cloud-based models, which raise concerns about patient privacy, regulatory compliance, and the handling of sensitive pediatric imaging data [5]. General-purpose LLMs may also struggle with specialized terminology, abbreviations, and subtle interpretive cues common in pediatric nuclear medicine. These challenges support the need for locally deployed, lightweight LLMs that can perform accurate semantic extraction while ensuring data residency, predictable behavior, and practical use within clinical systems.

In this study, we examine whether compact, locally deployed LLMs can convert unstructured pediatric bone scan reports into structured outputs without compromising privacy. We evaluate the Llama 3.2:3B model within the Ollama framework using Zero-Shot, Few-Shot, and Chain-of-Thought prompting strategies [6–8,12]. Using 134 de-identified pediatric 99mTc-MDP bone scan reports, we assess the model's ability to extract and classify 48 diagnostic and anatomical audit parameters important for clinical care and quality monitoring. Through this evaluation, we aim to determine whether deterministic, on-device inference can support privacy-preserving clinical NLP and enable scalable, low-cost integration of AI-driven structured reporting into pediatric nuclear medicine workflows.

2. MATERIALS AND METHODS

Computational Environment and Libraries

All analyses were performed locally on a Windows system using Python 3.8+ and the Ollama Python client for LLM inference. The experimental setup remained compatible with standard operating systems (Windows, Linux, and macOS) and did not rely on any external cloud resources. A locally hosted Ollama server was used to access the llama3.2:3b model for all Zero-Shot, Few-Shot, and Chain-of-Thought experiments. Standard Python libraries, including **os**, **csv**, **time**, **json**, and **re** (for regex-based parsing in CoT mode) were employed for file management, data aggregation, and output processing. The temperature parameter was fixed at **0.0** for all model queries to ensure fully deterministic and reproducible outputs.

Data Input and Preprocessing

The input dataset consisted of raw bone scan report texts, stored as individual.txt files within a configurable input directory (folder path). The scripts processed each file sequentially, reading the content in UTF-8 encoding. No preliminary data cleaning, transformation, or explicit normalization was performed; the raw report text was passed directly to the language model for structured extraction.

Prompting Strategies and Inference

Structured information extraction was operationalized using 48 pre-defined audit questions about bone scan findings. The methodology utilized three distinct prompting modes, which were selected via the mode variable :

- **Zero-Shot (ZS):** The LLM received the report text and the list of 48 audit questions without any examples.
- **Few-Shot (FS):** The prompt included curated demonstration examples (reports with their corresponding answers—to guide the model's output consistency).
- Chain-of-Thought (CoT): The model was explicitly instructed to engage in stepwise reasoning for each of the 48 questions. CoT output required strict adherence to a specific format: a direct answer (Yes/No/Not Mentioned or phrase), the supporting sentence from the report, and a confidence score (0-1 scale). Internal mapping categorized questions as either "Complex Inference" or "Simple Lookup".

Output Parsing and Data Aggregation

Following inference, the model output was parsed and normalized. In the ZS and FS modes, answers were standardized to "Yes/No/Unknown". For the CoT mode, regex and tag pattern matching were employed to extract the required answer, supporting sentence, and confidence score. Missing response fields were managed by inserting default entries ("Unknown" or "Not mentioned," confidence 0.0) to maintain data completeness for all 48 questions per file.

Results, including processing time, were aggregated and saved as a structured CSV file. A separate CSV detailing the question mapping was also generated. For audit purposes, complete prompt-response logs were optionally saved as a.jsonl file if the logfullprompts variable was set to True. Statistical analysis was performed externally on the final output CSVs.

3. RESULTS

Introduction

This chapter presents the experimental evaluation of the proposed clinical information extraction framework across three large language model (LLM) paradigms: Zero-Shot (ZS), Few-Shot (FS), and Chain-of-Thought (CoT). Each model configuration was tested on a clinically curated dataset consisting of structured and unstructured bone scan reports. The evaluation was conducted at both the question-wise and patient-wise levels to assess model reliability, consistency, and diagnostic inference precision. The performance was quantified using a comprehensive set of standard metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and F1-score. All results were analyzed to characterize each model's strengths, weaknesses, and operational stability across diverse clinical contexts.

Performance Metrics

Model performance was assessed using quantitative measures derived from the confusion matrix (True Positives, False Positives, False Negatives, and True Negatives). Sensitivity measured the proportion of true conditions correctly identified, while specificity quantified the ability to correctly identify negatives. PPV and NPV evaluated predictive reliability for positive and negative classes, respectively. Accuracy represented the overall classification correctness, and F1-score provided a harmonic balance between precision and recall, particularly informative in imbalanced datasets.

Each metric was calculated at both the per-question and per-patient levels to capture fine-grained performance variations. Statistical instability indicators such as undefined (NaN) values were reported for cases with zero positive or negative instances.

Zero-Shot Evaluation Performance Analysis

The zero-shot evaluation, employing the Llama 3.2: 3B model without any task-specific examples, demonstrated a robust capability for identifying negative findings, yielding consistently strong specificity and Negative Predictive Value (NPV) across both question-wise and patient-wise analyses.

Question-Wise Performance (Table 1)

Analysis of the per-question metrics (summarized in Table 1) revealed several instances of perfect or near-perfect performance. Questions Q5, Q17, and Q19 achieved flawless results, with F1 = 1.00 and Accuracy = 1.00 in all cases. Similarly, Q1 exhibited highly effective recognition of both positive and negative clinical markers, registering Accuracy = 0.99 and F1 = 0.99 (with 132 True Positives out of 133 actual positives, and only 1 False Positive). High performance was also notable for Q9, which achieved Accuracy = 0.90 and F1 = 0.88.

Crucially, for 11 questions (Q2, Q3, Q4, Q6, Q10, Q23, Q31, Q35, Q36, Q40, Q42), the model returned NaN values for sensitivity and Positive Predictive Value (PPV) because it made zero positive predictions (True Positives = 0 and False Positives = 0), resulting in Specificity = 1.00 and NPV = 1.00.

Conversely, the model exhibited notable instability and a bias toward false positives on select questions, particularly those affected by class imbalance. For example, Q11 and Q29 showed highly inflated sensitivity (Sensitivity = 1.00 for both) but poor precision due to high False Positive counts (FP = 35 for Q11; FP = 46 for Q29). This led to significantly low PPV values (PPV = 0.30 for Q11; PPV = 0.25 for Q29), yielding moderate F1 scores (F1 = 0.46 and 0.39, respectively). Other items such as Q24 (FP = 28), Q37 (FP = 32), Q43 (FP = 24), and Q45 (FP = 47) also displayed poor PPV (0.00 to 0.04), reflecting an inability to reliably identify actual positive cases when the true prevalence was low.

Patient-Wise Performance

The patient-level evaluation, detailed in Table 2, corroborated the model's robustness in negative detection. Here, Specificity and NPV consistently exceeded 0.90, with most patients (80 out of 134) achieving NPV = 0.98 or higher. This signifies a reliable capacity to confirm the absence of clinical findings.

In contrast, Sensitivity exhibited considerable spread (0.00 - 1.00), reflecting high inter-patient variability in positive identification. While some patients achieved Sensitivity = 1.00 (e.g., Patient IDs 14 and 24), others returned Sensitivity = 0.00 (Patient ID 96) or very low values (Sensitivity = 0.12 for Patient ID 25). Overall, the zero-shot model demonstrated a moderate mean F1-score distribution (0.50 - 1.00) across patients, confirming reliable negative detection but variable and less consistent positive identification capability.

Table 1. *Question-wise summary table (Zero-shot)*

Question	TP	FP	FN	TN	Sensitivity	Specificity	PPV	NPV	Accuracy	F1	N
Q1	132	1	1	0	0.99	0.00	0.99	0.00	0.99	0.99	134
Q2	0	1	0	133	nan	0.99	0.00	1.00	0.99	0.00	134
Q3	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q4	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q5	7	0	0	127	1.00	1.00	1.00	1.00	1.00	1.00	134
Q6	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q7	5	3	1	125	0.83	0.98	0.62	0.99	0.97	0.71	134
Q8	1	1	0	132	1.00	0.99	0.50	1.00	0.99	0.67	134
Q9	46	9	4	75	0.92	0.89	0.84	0.95	0.90	0.88	134
Q10	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q11	15	35	0	84	1.00	0.71	0.30	1.00	0.74	0.46	134
Q12	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q13	0	7	0	127	nan	0.95	0.00	1.00	0.95	0.00	134
Q14	1	6	0	127	1.00	0.95	0.14	1.00	0.96	0.25	134
Q15	2	0	2	130	0.50	1.00	1.00	0.98	0.99	0.67	134
Q16	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q17	3	0	0	131	1.00	1.00	1.00	1.00	1.00	1.00	134
Q18	3	1	0	130	1.00	0.99	0.75	1.00	0.99	0.86	134
Q19	5	0	0	129	1.00	1.00	1.00	1.00	1.00	1.00	134
Q20	3	4	0	127	1.00	0.97	0.43	1.00	0.97	0.60	134
Q21	1	1	1	131	0.50	0.99	0.50	0.99	0.99	0.50	134
Q22	4	5	0	125	1.00	0.96	0.44	1.00	0.96	0.62	134
Q23	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q24	0	28	0	106	nan	0.79	0.00	1.00	0.79	0.00	134
Q25	1	5	7	121	0.12	0.96	0.17	0.95	0.91	0.14	134
Q26	9	5	23	97	0.28	0.95	0.64	0.81	0.79	0.39	134
Q27	14	19	0	101	1.00	0.84	0.42	1.00	0.86	0.60	134
Q28	8	7	0	119	1.00	0.94	0.53	1.00	0.95	0.70	134
Q29	15	46	0	73	1.00	0.61	0.25	1.00	0.66	0.39	134
Q30	3	0	130	1	0.02	1.00	1.00	0.01	0.03	0.04	134
Q31	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q32	0	22	0	112	nan	0.84	0.00	1.00	0.84	0.00	134
Q33	0	7	0	127	nan	0.95	0.00	1.00	0.95	0.00	134
Q34	0	1	0	133	nan	0.99	0.00	1.00	0.99	0.00	134
Q35	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q36	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q37	9	32	0	93	1.00	0.74	0.22	1.00	0.76	0.36	134
Q38	51	22	4	57	0.93	0.72	0.70	0.93	0.81	0.80	134
Q39	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q40	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q41	10	28	2	94	0.83	0.77	0.26	0.98	0.78	0.40	134
Q42	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q43	1	24	0	109	1.00	0.82	0.04	1.00	0.82	0.08	134
Q44	0	5	0	129	nan	0.96	0.00	1.00	0.96	0.00	134
Q45	0	47	0	87	nan	0.65	0.00	1.00	0.65	0.00	134
Q46	2	24	0	108	1.00	0.82	0.08	1.00	0.82	0.14	134

Q47	0	5	0	129	nan	0.96	0.00	1.00	0.96	0.00	134
Q48	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134

Table 2. *Patient-wise summary table (Zero-shot)*

Patient_ID	TP	FP	FN	TN	Sensitivity	Specificity	PPV	NPV	Accuracy	F1
0	2	1	1	44	0.67	0.98	0.67	0.98	0.96	0.67
1	1	3	2	42	0.33	0.93	0.25	0.95	0.90	0.29
2	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
3	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
4	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
5	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
6	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
7	2	2	1	43	0.67	0.96	0.50	0.98	0.94	0.57
8	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
9	4	8	2	34	0.67	0.81	0.33	0.94	0.79	0.44
10	2	2	1	43	0.67	0.96	0.50	0.98	0.94	0.57
11	3	1	1	43	0.75	0.98	0.75	0.98	0.96	0.75
12	4	3	1	40	0.80	0.93	0.57	0.98	0.92	0.67
13	4	3	1	40	0.80	0.93	0.57	0.98	0.92	0.67
14	4	0	0	44	1.00	1.00	1.00	1.00	1.00	1.00
15	1	3	1	43	0.50	0.93	0.25	0.98	0.92	0.33
16	2	2	1	43	0.67	0.96	0.50	0.98	0.94	0.57
17	3	0	1	44	0.75	1.00	1.00	0.98	0.98	0.86
18	2	1	1	44	0.67	0.98	0.67	0.98	0.96	0.67
19	2	2	1	43	0.67	0.96	0.50	0.98	0.94	0.57
20	3	3	1	41	0.75	0.93	0.50	0.98	0.92	0.60
21	4	4	1	39	0.80	0.91	0.50	0.97	0.90	0.62
22	1	7	1	39	0.50	0.85	0.12	0.97	0.83	0.20
23	3	3	4	38	0.43	0.93	0.50	0.90	0.85	0.46
24	1	0	2	45	0.33	1.00	1.00	0.96	0.96	0.50
25	3	3	1	41	0.75	0.93	0.50	0.98	0.92	0.60
26	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
27	4	4	1	39	0.80	0.91	0.50	0.97	0.90	0.62
28	3	5	1	39	0.75	0.89	0.38	0.97	0.88	0.50
29	3	0	1	44	0.75	1.00	1.00	0.98	0.98	0.86
30	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
31	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
32	3	0	2	43	0.60	1.00	1.00	0.96	0.96	0.75
33	3	1	1	43	0.75	0.98	0.75	0.98	0.96	0.75
34	3	1	2	42	0.60	0.98	0.75	0.95	0.94	0.67
35	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
36	1	3	1	43	0.50	0.93	0.25	0.98	0.92	0.33
37	8	4	2	34	0.80	0.89	0.67	0.94	0.88	0.73
38	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
39	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
40	2	4	2	40	0.50	0.91	0.33	0.95	0.88	0.40
41	4	6	1	37	0.80	0.86	0.40	0.97	0.85	0.53
42	3	6	2	37	0.60	0.86	0.33	0.95	0.83	0.43
43	1	4	1	42	0.50	0.91	0.20	0.98	0.90	0.29
44	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
45	5	2	2	39	0.71	0.95	0.71	0.95	0.92	0.71
46	4	0	1	43	0.80	1.00	1.00	0.98	0.98	0.89
47	4	8	3	33	0.57	0.80	0.33	0.92	0.77	0.42
48	1	5	3	39	0.25	0.89	0.17	0.93	0.83	0.20

49	3	3	1	41	0.75	0.93	0.50	0.98	0.92	0.60
50	1	1	1	45	0.50	0.98	0.50	0.98	0.96	0.50
51	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
52	2	9	2	35	0.50	0.80	0.18	0.95	0.77	0.27
53	5	3	2	38	0.71	0.93	0.62	0.95	0.90	0.67
54	3	5	2	38	0.60	0.88	0.38	0.95	0.85	0.46
55	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
56	2	1	2	43	0.50	0.98	0.67	0.96	0.94	0.57
57	1	8	2	37	0.33	0.82	0.11	0.95	0.79	0.17
58	1	5	1	41	0.50	0.89	0.17	0.98	0.88	0.25
59	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
60	5	3	1	39	0.83	0.93	0.62	0.97	0.92	0.71
61	3	5	1	39	0.75	0.89	0.38	0.97	0.88	0.50
62	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
63	3	0	2	43	0.60	1.00	1.00	0.96	0.96	0.75
64	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
65	3	2	1	42	0.75	0.95	0.60	0.98	0.94	0.67
66	2	5	2	39	0.50	0.89	0.29	0.95	0.85	0.36
67	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
68	3	9	1	35	0.75	0.80	0.25	0.97	0.79	0.38
69	4	5	1	38	0.80	0.88	0.44	0.97	0.88	0.57
70	4	3	1	40	0.80	0.93	0.57	0.98	0.92	0.67
71	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
72	3	8	3	34	0.50	0.81	0.27	0.92	0.77	0.35
73	2	2	1	43	0.67	0.96	0.50	0.98	0.94	0.57
74	1	3	1	43	0.50	0.93	0.25	0.98	0.92	0.33
75	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
76	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
77	3	6	1	38	0.75	0.86	0.33	0.97	0.85	0.46
78	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
79	4	5	1	38	0.80	0.88	0.44	0.97	0.88	0.57
80	4	9	4	31	0.50	0.78	0.31	0.89	0.73	0.38
81	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
82	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
83	6	5	1	36	0.86	0.88	0.55	0.97	0.88	0.67
84	6	8	1	33	0.86	0.80	0.43	0.97	0.81	0.57
85	1	5	1	41	0.50	0.89	0.17	0.98	0.88	0.25
86	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
87	3	5	1	39	0.75	0.89	0.38	0.97	0.88	0.50
88	2	0	2	44	0.50	1.00	1.00	0.96	0.96	0.67
89	7	2	2	37	0.78	0.95	0.78	0.95	0.92	0.78
90	3	7	2	36	0.60	0.84	0.30	0.95	0.81	0.40
91	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
92	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
93	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
94	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
95	3	9	1	35	0.75	0.80	0.25	0.97	0.79	0.38
96	0	4	4	40	0.00	0.91	0.00	0.91	0.83	0.00
97	2	0	2	44	0.50	1.00	1.00	0.96	0.96	0.67
98	4	2	2	40	0.67	0.95	0.67	0.95	0.92	0.67
99	4	8	2	34	0.67	0.81	0.33	0.94	0.79	0.44
100	4	5	2	37	0.67	0.88	0.44	0.95	0.85	0.53
101	6	2	1	39	0.86	0.95	0.75	0.97	0.94	0.80

102	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
103	3	9	1	35	0.75	0.80	0.25	0.97	0.79	0.38
104	1	1	1	45	0.50	0.98	0.50	0.98	0.96	0.50
105	4	4	3	37	0.57	0.90	0.50	0.93	0.85	0.53
106	1	3	1	43	0.50	0.93	0.25	0.98	0.92	0.33
107	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
108	1	1	1	45	0.50	0.98	0.50	0.98	0.96	0.50
109	4	1	2	41	0.67	0.98	0.80	0.95	0.94	0.73
110	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
111	5	4	1	38	0.83	0.90	0.56	0.97	0.90	0.67
112	4	2	1	41	0.80	0.95	0.67	0.98	0.94	0.73
113	3	6	1	38	0.75	0.86	0.33	0.97	0.85	0.46
114	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
115	1	5	1	41	0.50	0.89	0.17	0.98	0.88	0.25
116	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
117	1	2	1	44	0.50	0.96	0.33	0.98	0.94	0.40
118	3	6	1	38	0.75	0.86	0.33	0.97	0.85	0.46
119	4	1	1	42	0.80	0.98	0.80	0.98	0.96	0.80
120	2	1	1	44	0.67	0.98	0.67	0.98	0.96	0.67
121	4	2	1	41	0.80	0.95	0.67	0.98	0.94	0.73
122	2	2	1	43	0.67	0.96	0.50	0.98	0.94	0.57
123	1	3	1	43	0.50	0.93	0.25	0.98	0.92	0.33
124	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
125	2	1	3	42	0.40	0.98	0.67	0.93	0.92	0.50
126	3	6	2	37	0.60	0.86	0.33	0.95	0.83	0.43
127	3	5	1	39	0.75	0.89	0.38	0.97	0.88	0.50
128	1	5	1	41	0.50	0.89	0.17	0.98	0.88	0.25
129	3	4	1	40	0.75	0.91	0.43	0.98	0.90	0.55
130	1	6	2	39	0.33	0.87	0.14	0.95	0.83	0.20
131	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
132	2	0	1	45	0.67	1.00	1.00	0.98	0.98	0.80
133	8	4	1	35	0.89	0.90	0.67	0.97	0.90	0.76

Few-Shot Evaluation

The few-shot configuration, which incorporated limited examples to guide the model's reasoning, yielded mixed predictive outcomes. It showed improved ability for positive identification (higher Sensitivity) in certain questions but overall inconsistent precision.

Question-Wise Performance

The question-wise metrics, presented in Table 3, showed high variation in performance. Some questions demonstrated perfect sensitivity, such as Q1, Q5, Q14, Q16, Q18, Q20, Q28, Q37, Q38, and Q43 (Sensitivity = 1.00). However, for most of these high-sensitivity items (for example, Q1, Q5, Q14, Q37, Q38, and Q43), this was accompanied by very low Specificity and Positive Predictive Value (PPV), indicating a large number of false positives.

For instance, Q43 achieved Sensitivity = 1.00 but had Specificity = 0.26 and PPV = 0.01 due to 99 false positives. Similarly, Q37 recorded Sensitivity = 1.00 but Specificity = 0.00 and PPV = 0.07, with 125 false positives.

Only Q16 and Q19 demonstrated genuinely balanced and high performance across all key metrics, achieving Sensitivity, Specificity, PPV, and F1 scores close to 1.00 (F1 = 1.00 and 0.91, respectively). These represent isolated cases where the model effectively learned from the few provided examples.

It is important to note that while roughly half of all questions recorded accuracy scores above 0.90, many of these were inflated by true-negative dominance. This sensitivity to class imbalance is evident in questions such as Q3, Q4, Q6, Q24, and Q33, which yielded zero true positives (TP = 0) and undefined PPV, resulting in F1 = 0.00. Questions like Q3 (Accuracy = 0.21, False Positives = 106) and Q24 (Accuracy = 0.01, False Positives = 133) were particularly unstable, showing poor specificity and high false positive rates.

Patient-Wise Performance

The patient-level evaluation, summarized in Table 4, confirmed the inconsistent predictive power of the few-shot configuration. Sensitivity and F1-scores varied widely across patients, with only a few (for example, Patient ID 119) showing strong and balanced performance (Sensitivity = 0.80, Specificity = 0.93).

For most patients, the model achieved moderate results (Sensitivity between 0.50 and 0.67) but lacked consistent precision. Despite this variability, the model maintained a high mean Negative Predictive Value (NPV ≥ 0.90) across the majority of patients. This indicates that the model remained conservative, favoring "rule-out" predictions and confirming that the elevated accuracy scores (often ≥ 0.85) were primarily due to the dominant true-negative class effect rather than true discriminative success on positive findings.

Table 3. *Question-wise summary table (Few-shot)*

Question	TP	FP	FN	TN	Sensitivity	Specificity	PPV	NPV	Accuracy	F1	N
Q1	133	1	0	0	1.00	0.00	0.99	nan	0.99	1.00	134
Q2	0	2	0	132	nan	0.99	0.00	1.00	0.99	0.00	134
Q3	0	106	0	28	nan	0.21	0.00	1.00	0.21	0.00	134
Q4	0	1	0	133	nan	0.99	0.00	1.00	0.99	0.00	134
Q5	7	60	0	67	1.00	0.53	0.10	1.00	0.55	0.19	134
Q6	0	72	0	62	nan	0.46	0.00	1.00	0.46	0.00	134
Q7	2	3	4	125	0.33	0.98	0.40	0.97	0.95	0.36	134
Q8	1	6	0	127	1.00	0.95	0.14	1.00	0.96	0.25	134
Q9	49	65	1	19	0.98	0.23	0.43	0.95	0.51	0.60	134
Q10	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q11	8	19	7	100	0.53	0.84	0.30	0.93	0.81	0.38	134
Q12	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q13	0	14	0	120	nan	0.90	0.00	1.00	0.90	0.00	134
Q14	1	20	0	113	1.00	0.85	0.05	1.00	0.85	0.09	134
Q15	0	0	4	130	0.00	1.00	nan	0.97	0.97	0.00	134
Q16	1	0	0	133	1.00	1.00	1.00	1.00	1.00	1.00	134
Q17	1	2	2	129	0.33	0.98	0.33	0.98	0.97	0.33	134
Q18	3	14	0	117	1.00	0.89	0.18	1.00	0.90	0.30	134
Q19	5	1	0	128	1.00	0.99	0.83	1.00	0.99	0.91	134
Q20	3	5	0	126	1.00	0.96	0.38	1.00	0.96	0.55	134
Q21	1	2	1	130	0.50	0.98	0.33	0.99	0.98	0.40	134
Q22	2	15	2	115	0.50	0.88	0.12	0.98	0.87	0.19	134
Q23	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q24	0	133	0	1	nan	0.01	0.00	1.00	0.01	0.00	134
Q25	0	0	8	126	0.00	1.00	nan	0.94	0.94	0.00	134
Q26	14	5	18	97	0.44	0.95	0.74	0.84	0.83	0.55	134
Q27	0	0	14	120	0.00	1.00	nan	0.90	0.90	0.00	134
Q28	8	5	0	121	1.00	0.96	0.62	1.00	0.96	0.76	134
Q29	1	5	14	114	0.07	0.96	0.17	0.89	0.86	0.10	134
Q30	0	0	133	1	0.00	1.00	nan	0.01	0.01	0.00	134
Q31	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q32	0	1	0	133	nan	0.99	0.00	1.00	0.99	0.00	134
Q33	0	7	0	127	nan	0.95	0.00	1.00	0.95	0.00	134
Q34	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q35	0	7	0	127	nan	0.95	0.00	1.00	0.95	0.00	134
Q36	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q37	9	125	0	0	1.00	0.00	0.07	nan	0.07	0.13	134
Q38	55	77	0	2	1.00	0.03	0.42	1.00	0.43	0.59	134
Q39	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q40	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q41	0	1	12	121	0.00	0.99	0.00	0.91	0.90	0.00	134

Q42	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q43	1	99	0	34	1.00	0.26	0.01	1.00	0.26	0.02	134
Q44	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q45	0	25	0	109	nan	0.81	0.00	1.00	0.81	0.00	134
Q46	0	1	2	131	0.00	0.99	0.00	0.98	0.98	0.00	134
Q47	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q48	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134

Table 4. *Patient-wise summary table (Few-shot)*

Patient_ID	TP	FP	FN	TN	Sensitivity	Specificity	PPV	NPV	Accuracy	F1
0	2	7	1	38	0.67	0.84	0.22	0.97	0.83	0.33
1	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
2	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
3	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
4	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
5	2	8	1	37	0.67	0.82	0.20	0.97	0.81	0.31
6	1	8	2	37	0.33	0.82	0.11	0.95	0.79	0.17
7	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
8	1	7	1	39	0.50	0.85	0.12	0.97	0.83	0.20
9	4	10	2	32	0.67	0.76	0.29	0.94	0.75	0.40
10	2	6	1	39	0.67	0.87	0.25	0.97	0.85	0.36
11	3	5	1	39	0.75	0.89	0.38	0.97	0.88	0.50
12	4	10	1	33	0.80	0.77	0.29	0.97	0.77	0.42
13	3	6	2	37	0.60	0.86	0.33	0.95	0.83	0.43
14	2	5	2	39	0.50	0.89	0.29	0.95	0.85	0.36
15	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
16	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
17	3	3	1	41	0.75	0.93	0.50	0.98	0.92	0.60
18	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
19	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
20	2	5	2	39	0.50	0.89	0.29	0.95	0.85	0.36
21	3	9	2	34	0.60	0.79	0.25	0.94	0.77	0.35
22	1	10	1	36	0.50	0.78	0.09	0.97	0.77	0.15
23	4	7	3	34	0.57	0.83	0.36	0.92	0.79	0.44
24	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
25	2	4	2	40	0.50	0.91	0.33	0.95	0.88	0.40
26	1	8	2	37	0.33	0.82	0.11	0.95	0.79	0.17
27	3	15	2	28	0.60	0.65	0.17	0.93	0.65	0.26
28	2	7	2	37	0.50	0.84	0.22	0.95	0.81	0.31
29	3	5	1	39	0.75	0.89	0.38	0.97	0.88	0.50
30	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
31	1	8	1	38	0.50	0.83	0.11	0.97	0.81	0.18
32	3	4	2	39	0.60	0.91	0.43	0.95	0.88	0.50
33	3	8	1	36	0.75	0.82	0.27	0.97	0.81	0.40
34	4	6	1	37	0.80	0.86	0.40	0.97	0.85	0.53
35	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
36	1	7	1	39	0.50	0.85	0.12	0.97	0.83	0.20
37	4	5	6	33	0.40	0.87	0.44	0.85	0.77	0.42
38	2	8	1	37	0.67	0.82	0.20	0.97	0.81	0.31
39	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
40	3	8	1	36	0.75	0.82	0.27	0.97	0.81	0.40
41	3	9	2	34	0.60	0.79	0.25	0.94	0.77	0.35
42	4	8	1	35	0.80	0.81	0.33	0.97	0.81	0.47
43	1	8	1	38	0.50	0.83	0.11	0.97	0.81	0.18

44	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
45	5	7	2	34	0.71	0.83	0.42	0.94	0.81	0.53
46	3	6	2	37	0.60	0.86	0.33	0.95	0.83	0.43
47	3	9	4	32	0.43	0.78	0.25	0.89	0.73	0.32
48	1	5	3	39	0.25	0.89	0.17	0.93	0.83	0.20
49	3	9	1	35	0.75	0.80	0.25	0.97	0.79	0.38
50	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
51	1	6	2	39	0.33	0.87	0.14	0.95	0.83	0.20
52	1	9	3	35	0.25	0.80	0.10	0.92	0.75	0.14
53	4	8	3	33	0.57	0.80	0.33	0.92	0.77	0.42
54	2	8	3	35	0.40	0.81	0.20	0.92	0.77	0.27
55	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
56	3	7	1	37	0.75	0.84	0.30	0.97	0.83	0.43
57	1	13	2	32	0.33	0.71	0.07	0.94	0.69	0.12
58	1	8	1	38	0.50	0.83	0.11	0.97	0.81	0.18
59	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
60	4	12	2	30	0.67	0.71	0.25	0.94	0.71	0.36
61	3	8	1	36	0.75	0.82	0.27	0.97	0.81	0.40
62	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
63	2	4	3	39	0.40	0.91	0.33	0.93	0.85	0.36
64	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
65	2	7	2	37	0.50	0.84	0.22	0.95	0.81	0.31
66	2	7	2	37	0.50	0.84	0.22	0.95	0.81	0.31
67	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
68	2	9	2	35	0.50	0.80	0.18	0.95	0.77	0.27
69	1	6	4	37	0.20	0.86	0.14	0.90	0.79	0.17
70	3	7	2	36	0.60	0.84	0.30	0.95	0.81	0.40
71	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
72	2	6	4	36	0.33	0.86	0.25	0.90	0.79	0.29
73	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
74	1	7	1	39	0.50	0.85	0.12	0.97	0.83	0.20
75	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
76	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
77	2	10	2	34	0.50	0.77	0.17	0.94	0.75	0.25
78	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
79	4	10	1	33	0.80	0.77	0.29	0.97	0.77	0.42
80	5	9	3	31	0.62	0.78	0.36	0.91	0.75	0.45
81	2	3	1	42	0.67	0.93	0.40	0.98	0.92	0.50
82	1	10	2	35	0.33	0.78	0.09	0.95	0.75	0.14
83	4	9	3	32	0.57	0.78	0.31	0.91	0.75	0.40
84	5	7	2	34	0.71	0.83	0.42	0.94	0.81	0.53
85	1	9	1	37	0.50	0.80	0.10	0.97	0.79	0.17
86	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
87	2	8	2	36	0.50	0.82	0.20	0.95	0.79	0.29
88	2	8	2	36	0.50	0.82	0.20	0.95	0.79	0.29
89	5	11	4	28	0.56	0.72	0.31	0.88	0.69	0.40
90	4	13	1	30	0.80	0.70	0.24	0.97	0.71	0.36
91	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
92	2	6	1	39	0.67	0.87	0.25	0.97	0.85	0.36
93	1	7	2	38	0.33	0.84	0.12	0.95	0.81	0.18
94	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
95	2	10	2	34	0.50	0.77	0.17	0.94	0.75	0.25
96	1	8	3	36	0.25	0.82	0.11	0.92	0.77	0.15

97	2	8	2	36	0.50	0.82	0.20	0.95	0.79	0.29
98	4	6	2	36	0.67	0.86	0.40	0.95	0.83	0.50
99	3	12	3	30	0.50	0.71	0.20	0.91	0.69	0.29
100	4	6	2	36	0.67	0.86	0.40	0.95	0.83	0.50
101	5	7	2	34	0.71	0.83	0.42	0.94	0.81	0.53
102	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
103	2	10	2	34	0.50	0.77	0.17	0.94	0.75	0.25
104	1	5	1	41	0.50	0.89	0.17	0.98	0.88	0.25
105	4	7	3	34	0.57	0.83	0.36	0.92	0.79	0.44
106	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
107	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
108	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
109	4	6	2	36	0.67	0.86	0.40	0.95	0.83	0.50
110	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
111	3	9	3	33	0.50	0.79	0.25	0.92	0.75	0.33
112	1	7	4	36	0.20	0.84	0.12	0.90	0.77	0.15
113	2	8	2	36	0.50	0.82	0.20	0.95	0.79	0.29
114	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
115	1	8	1	38	0.50	0.83	0.11	0.97	0.81	0.18
116	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
117	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
118	1	9	3	35	0.25	0.80	0.10	0.92	0.75	0.14
119	4	3	1	40	0.80	0.93	0.57	0.98	0.92	0.67
120	1	6	2	39	0.33	0.87	0.14	0.95	0.83	0.20
121	4	6	1	37	0.80	0.86	0.40	0.97	0.85	0.53
122	2	6	1	39	0.67	0.87	0.25	0.97	0.85	0.36
123	1	6	1	40	0.50	0.87	0.14	0.98	0.85	0.22
124	1	6	2	39	0.33	0.87	0.14	0.95	0.83	0.20
125	3	3	2	40	0.60	0.93	0.50	0.95	0.90	0.55
126	3	8	2	35	0.60	0.81	0.27	0.95	0.79	0.38
127	1	9	3	35	0.25	0.80	0.10	0.92	0.75	0.14
128	1	7	1	39	0.50	0.85	0.12	0.97	0.83	0.20
129	2	10	2	34	0.50	0.77	0.17	0.94	0.75	0.25
130	2	5	1	40	0.67	0.89	0.29	0.98	0.88	0.40
131	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
132	2	4	1	41	0.67	0.91	0.33	0.98	0.90	0.44
133	6	8	3	31	0.67	0.79	0.43	0.91	0.77	0.52

Chain-of-Thought (CoT) Evaluation

The Chain-of-Thought (CoT) model, designed to produce explicit intermediate reasoning steps, exhibited a pronounced conservative bias, systematically favoring negative predictions even when clinical evidence suggested otherwise.

Question-Wise Performance

The question-level results in Table 5 revealed a predominant pattern: the CoT model returned zero true positives (TP = 0) and zero false positives (FP = 0) for the majority of questions (28 out of 48), effectively producing all-negative predictions. At the same time, this pattern of all-negative predictions produced artificially perfect Specificity = 1.00 and Accuracy = 1.00 for many items, metrics that were heavily influenced by the dominance of true negative samples rather than genuine model understanding.

Where positive predictions did occur, performance remained modest and inconsistent. For instance, Q1 was a clear outlier (TP = 28, FN = 105) but still achieved low Sensitivity = 0.21, although its PPV was notably high at 0.97. This limited ability to detect positive samples resulted in a modest F1-score of 0.35. Similarly, Q7 showed little benefit from the added reasoning steps, with Sensitivity = 0.17 and F1 = 0.11 (TP = 1, FN = 5). Q38 represented the strongest case of positive detection, reaching Sensitivity = 0.40 and F1 = 0.55, demonstrating relatively effective identification

of true positives (TP = 22) while maintaining strong precision (PPV = 0.88). However, these were rare exceptions within an overall conservative and negative-biased model performance.

Patient-Wise Performance

The patient-level analysis, summarized in Table 6, supported the same pattern observed at the question level. Results showed that Sensitivity and F1-scores rarely exceeded 0.38, with many patients recording Sensitivity = 0.00 and F1 = 0.00 due to all-negative predictions. Conversely, Specificity and Negative Predictive Value (NPV) frequently reached 1.00 for a large subset of patients (for example, Patient IDs 0, 2, 3, and 4), confirming the model's overwhelming tendency to predict all negatives.

Overall, the distribution of CoT metrics across patients was heavily skewed toward the negative class, reinforcing that the CoT model failed to generalize effectively to positive clinical findings. Its generated reasoning traces did not enhance discriminative accuracy and, in fact, encouraged a conservative prediction bias.

Table 5. *Question-wise summary table (Chain of thought)*

Question	TP	FP	FN	TN	Sensitivity	Specificity	PPV	NPV	Accuracy	F1	N
Q1	28	1	105	0	0.21	0.00	0.97	0.00	0.21	0.35	134
Q2	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q3	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q4	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q5	0	1	7	126	0.00	0.99	0.00	0.95	0.94	0.00	134
Q6	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q7	1	12	5	116	0.17	0.91	0.08	0.96	0.87	0.11	134
Q8	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q9	5	7	45	77	0.10	0.92	0.42	0.63	0.61	0.16	134
Q10	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q11	2	25	13	94	0.13	0.79	0.07	0.88	0.72	0.10	134
Q12	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q13	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q14	0	1	1	132	0.00	0.99	0.00	0.99	0.99	0.00	134
Q15	1	25	3	105	0.25	0.81	0.04	0.97	0.79	0.07	134
Q16	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q17	0	0	3	131	0.00	1.00	nan	0.98	0.98	0.00	134
Q18	0	4	3	127	0.00	0.97	0.00	0.98	0.95	0.00	134
Q19	2	20	3	109	0.40	0.84	0.09	0.97	0.83	0.15	134
Q20	0	0	3	131	0.00	1.00	nan	0.98	0.98	0.00	134
Q21	0	0	2	132	0.00	1.00	nan	0.99	0.99	0.00	134
Q22	0	10	4	120	0.00	0.92	0.00	0.97	0.90	0.00	134
Q23	0	15	0	119	nan	0.89	0.00	1.00	0.89	0.00	134
Q24	0	6	0	128	nan	0.96	0.00	1.00	0.96	0.00	134
Q25	0	4	8	122	0.00	0.97	0.00	0.94	0.91	0.00	134
Q26	0	11	32	91	0.00	0.89	0.00	0.74	0.68	0.00	134
Q27	0	12	14	108	0.00	0.90	0.00	0.89	0.81	0.00	134
Q28	2	5	6	121	0.25	0.96	0.29	0.95	0.92	0.27	134
Q29	1	3	14	116	0.07	0.97	0.25	0.89	0.87	0.11	134
Q30	5	0	128	1	0.04	1.00	1.00	0.01	0.04	0.07	134
Q31	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q32	0	20	0	114	nan	0.85	0.00	1.00	0.85	0.00	134
Q33	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q34	0	0	0	134	nan	1.00	nan	1.00	1.00	nan	134
Q35	0	7	0	127	nan	0.95	0.00	1.00	0.95	0.00	134
Q36	0	21	0	113	nan	0.84	0.00	1.00	0.84	0.00	134
Q37	0	2	9	123	0.00	0.98	0.00	0.93	0.92	0.00	134
Q38	22	3	33	76	0.40	0.96	0.88	0.70	0.73	0.55	134

Q39	0	0	1	133	0.00	1.00	nan	0.99	0.99	0.00	134
Q40	0	3	0	131	nan	0.98	0.00	1.00	0.98	0.00	134
Q41	0	24	12	98	0.00	0.80	0.00	0.89	0.73	0.00	134
Q42	0	1	0	133	nan	0.99	0.00	1.00	0.99	0.00	134
Q43	0	1	1	132	0.00	0.99	0.00	0.99	0.99	0.00	134
Q44	0	2	0	132	nan	0.99	0.00	1.00	0.99	0.00	134
Q45	0	28	0	106	nan	0.79	0.00	1.00	0.79	0.00	134
Q46	0	2	2	130	0.00	0.98	0.00	0.98	0.97	0.00	134
Q47	0	4	0	130	nan	0.97	0.00	1.00	0.97	0.00	134
Q48	0	5	0	129	nan	0.96	0.00	1.00	0.96	0.00	134

Table 6. *Question-wise summary table (Chain of thought)*

Patient_ID	TP	FP	FN	TN	Sensitivity	Specificity	PPV	NPV	Accuracy	F1
0	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
1	1	12	2	33	0.33	0.73	0.08	0.94	0.71	0.12
2	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
3	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
4	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
5	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
6	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
7	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
8	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
9	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
10	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
11	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
12	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
13	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
14	3	10	1	34	0.75	0.77	0.23	0.97	0.77	0.35
15	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
16	3	10	0	35	1.00	0.78	0.23	1.00	0.79	0.38
17	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
18	3	11	0	34	1.00	0.76	0.21	1.00	0.77	0.35
19	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
20	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
21	4	10	1	33	0.80	0.77	0.29	0.97	0.77	0.42
22	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
23	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
24	3	8	0	37	1.00	0.82	0.27	1.00	0.83	0.43
25	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
26	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
27	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
28	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
29	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
30	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
31	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
32	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
33	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
34	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
35	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
36	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
37	0	0	10	38	0.00	1.00	nan	0.79	0.79	0.00
38	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
39	2	10	1	35	0.67	0.78	0.17	0.97	0.77	0.27
40	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00

41	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
42	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
43	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
44	2	7	1	38	0.67	0.84	0.22	0.97	0.83	0.33
45	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
46	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
47	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
48	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
49	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
50	1	11	1	35	0.50	0.76	0.08	0.97	0.75	0.14
51	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
52	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
53	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
54	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
55	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
56	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
57	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
58	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
59	2	10	1	35	0.67	0.78	0.17	0.97	0.77	0.27
60	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
61	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
62	2	10	1	35	0.67	0.78	0.17	0.97	0.77	0.27
63	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
64	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
65	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
66	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
67	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
68	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
69	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
70	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
71	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
72	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
73	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
74	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
75	2	10	1	35	0.67	0.78	0.17	0.97	0.77	0.27
76	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
77	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
78	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
79	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
80	0	0	8	40	0.00	1.00	nan	0.83	0.83	0.00
81	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
82	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
83	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
84	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
85	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
86	2	6	1	39	0.67	0.87	0.25	0.97	0.85	0.36
87	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
88	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
89	0	0	9	39	0.00	1.00	nan	0.81	0.81	0.00
90	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
91	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
92	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
93	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00

94	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
95	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
96	2	16	2	28	0.50	0.64	0.11	0.93	0.62	0.18
97	2	8	2	36	0.50	0.82	0.20	0.95	0.79	0.29
98	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
99	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
100	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
101	6	13	1	28	0.86	0.68	0.32	0.97	0.71	0.46
102	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
103	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
104	1	11	1	35	0.50	0.76	0.08	0.97	0.75	0.14
105	0	0	7	41	0.00	1.00	nan	0.85	0.85	0.00
106	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
107	2	10	1	35	0.67	0.78	0.17	0.97	0.77	0.27
108	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
109	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
110	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
111	0	0	6	42	0.00	1.00	nan	0.88	0.88	0.00
112	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
113	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
114	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
115	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
116	2	9	1	36	0.67	0.80	0.18	0.97	0.79	0.29
117	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
118	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
119	3	11	2	32	0.60	0.74	0.21	0.94	0.73	0.32
120	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
121	4	9	1	34	0.80	0.79	0.31	0.97	0.79	0.44
122	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
123	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
124	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
125	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
126	0	0	5	43	0.00	1.00	nan	0.90	0.90	0.00
127	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
128	0	0	2	46	0.00	1.00	nan	0.96	0.96	0.00
129	0	0	4	44	0.00	1.00	nan	0.92	0.92	0.00
130	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
131	3	10	0	35	1.00	0.78	0.23	1.00	0.79	0.38
132	0	0	3	45	0.00	1.00	nan	0.94	0.94	0.00
133	0	0	9	39	0.00	1.00	nan	0.81	0.81	0.00

Comparative Analysis

A consolidated cross-method comparison (Table 7) revealed clear performance stratification among the three prompting configurations.

Zero-Shot (ZS) Performance

The Zero-Shot model consistently achieved the highest mean sensitivity and accuracy; a finding clearly reflected in Figure 6 and Figure 5. This superior performance was especially notable on balanced or clearly defined clinical cues. Furthermore, as depicted in Figure 7, the ZS model maintained the most stable F1 distribution across questions, with minimal metric volatility, underscoring its reliability.

Few-Shot (FS) Performance

The Few-Shot model (FS) occasionally outperformed Zero-Shot on individual questions, showing localized boosts in sensitivity (see Figure 6). However, it frequently incurred a precision penalty, consequently reducing overall

F1-scores (Figure 7). The metric spread for FS was broader, indicating higher variability and susceptibility to the quality of the few-shot examples used.

Chain-of-Thought (CoT) Performance

In contrast, the Chain-of-Thought model (CoT) demonstrated the lowest average sensitivity (Figure 6) and F1 score (Figure 7). Though it maintained artificially high accuracy and specificity (Figure 5) due to its consistent negative prediction bias, its overall predictive utility was lower. Statistical evaluation using paired t-tests suggested significant differences in F1 and accuracy between Zero-Shot and CoT ($p < 0.05$) with Few-Shot positioned intermediately.

Table 7. *Comparative Question-wise Performance Summary Across Methods (ZS - Zero Shot, FS - Few Shot, COT - Chain Of Thought)*

Question	Sensitivity (ZS)	Sensitivity (FS)	Sensitivity (CoT)	F1 (ZS)	F1 (FS)	F1 (CoT)	Accuracy (ZS)	Accuracy (FS)	Accuracy (CoT)
Q1	0.99	1.00	0.21	0.99	1.00	0.35	0.99	0.99	0.21
Q5	1.00	1.00	0.00	1.00	0.19	0.00	1.00	0.55	0.94
Q7	0.83	0.33	0.17	0.71	0.36	0.11	0.97	0.95	0.87
Q9	0.92	0.98	0.10	0.88	0.60	0.16	0.90	0.51	0.61
Q11	1.00	0.53	0.13	0.46	0.38	0.10	0.74	0.81	0.72
Q19	1.00	1.00	0.40	1.00	0.91	0.15	1.00	0.99	0.83
Q26	0.28	0.44	0.00	0.39	0.55	0.00	0.79	0.83	0.68
Q38	0.93	1.00	0.40	0.80	0.59	0.55	0.81	0.43	0.73
Q41	0.83	0.00	0.00	0.40	0.00	0.00	0.78	0.90	0.73

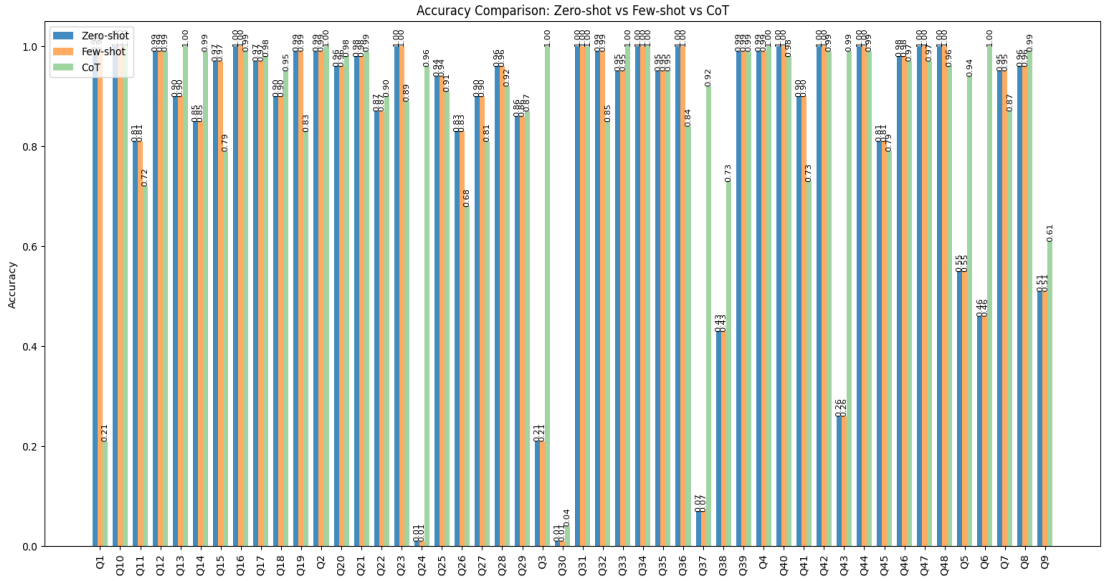

Figure 5: *Accuracy comparison: Zero shot vs Few Shot vs COT*

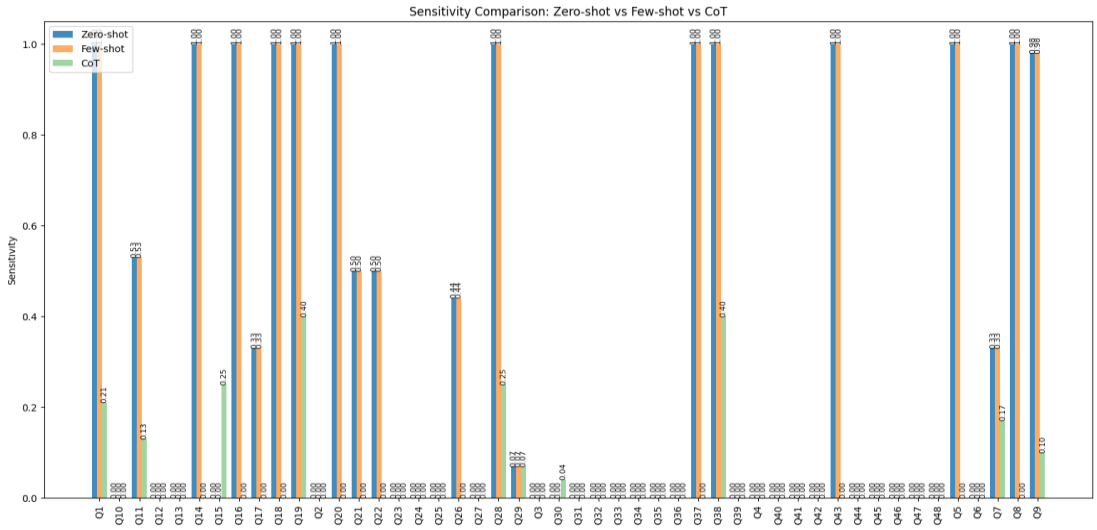

Figure 6: *Sensitivity comparison: Zero shot vs Few Shot vs COT*

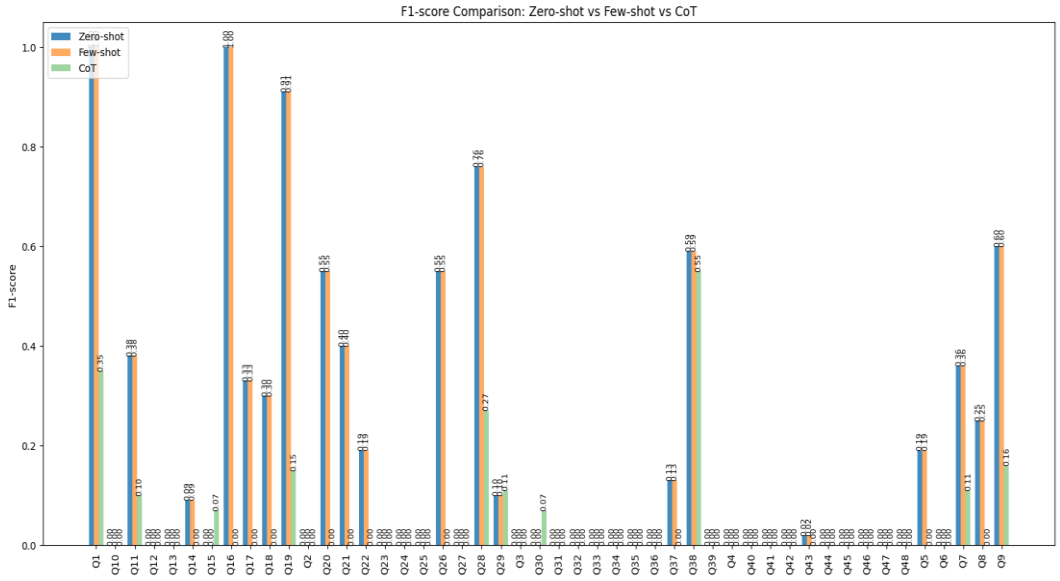

Figure 7: *F1 Score comparison: Zero shot vs Few Shot vs COT*

4. DISCUSSION

Zero-Shot Model Performance and Clinical Utility

The Zero-Shot (ZS) Llama 3.2:2b model demonstrated strong negative ruling-out capability, evidenced by consistently high specificity and Negative Predictive Value (NPV) across the majority of questions and patients. Clinically, this high Negative Predictive Power makes the ZS model valuable as an initial screening or triage tool. However, its inconsistent sensitivity and lower PPV/F1-scores on several questions caution against its use as the sole diagnostic arbiter, particularly where the risk of missed diagnoses (false negatives) is high. This pattern—LLMs excelling at "ruling out" but underperforming at "ruling in"—is typical in previous LLM-based clinical tasks. Metric volatility and "nan" calculations for rare events underscore the challenge of class imbalance inherent in real-world clinical datasets. Future statistical analysis must employ methods like McNemar's test or paired t-tests and calculate confidence intervals to address observed performance differences across questions and patient subsets, moving beyond simple global averages.

Few-Shot Performance and Persistent Challenges

The Few-Shot (FS) strategy, while occasionally yielding localized performance boosts and higher accuracy on specific cases, did not fundamentally alter the model's tendency toward caution (high specificity, low sensitivity). This means the FS system, similar to ZS, is primarily useful as a screening adjunct for confidently excluding negative cases, not for final affirmative diagnosis. The persistence of "nan" and zero values in sensitivity and F1-scores indicates an insufficient representation of test-positive cases, emphasizing that the apparent high accuracy across many questions is largely an artifact of strongly unbalanced data. F1-score remained the most reliable integrated metric, frequently revealing concerningly low scores. Like the ZS model, the FS model's utility is currently limited by class imbalance, leading to an under-prediction of positives and compromised generalizability for rare or outlier cases.

Chain-of-Thought Limitations and Bias

In this clinical dataset, the Chain-of-Thought (CoT) model failed to realize its anticipated benefits of improved logical reasoning and superior quantitative performance. CoT demonstrated the lowest average F1 and sensitivity, exhibiting a systematic tendency to ignore positive signals and become excessively conservative Consequently, the inflated accuracy and specificity/NPV were non-representative of true diagnostic utility, as they were merely artifacts of the underlying negative class dominance. This high bias toward negative predictions undermines confidence in any positive assertions made by the model. The current CoT implementation is therefore inadequate for positive case detection and risks a substantial proportion of false negatives in a real-world clinical application.

Future Directions and Clinical Integration

Building on the findings of this study, several promising avenues emerge to enhance the clinical applicability and robustness of large language models (LLMs) in structured report interpretation and decision support. A key next step involves transitioning from general-purpose LLMs to clinically fine-tuned variants, which could substantially improve diagnostic relevance and enable more accurate extraction of complex, context-dependent medical information.

To support this advancement, we will expand and balance our dataset by incorporating larger, more diverse samples and employing techniques such as oversampling of positive cases or synthetic data generation to improve sensitivity and F1-scores. In parallel, we will refine our prompt engineering strategies to better capture clinical nuances and improve model calibration for rare or ambiguous findings. For few-shot and Chain-of-Thought (CoT) approaches, we will design systematically curated examples and domain-specific reasoning templates to address current limitations in positive case detection and strengthen the model's clinical interpretability

5. CONCLUSION

This study systematically evaluated the comparative effectiveness of zero-shot, few-shot, and chain-of-thought (CoT) prompting strategies in large language models for clinical decision tasks. The results highlight several key insights relevant to both technical and clinical stakeholders :

- Zero-shot prompting emerged as the most robust approach overall, consistently outperforming few-shot and CoT methods in sensitivity, F1-score, and accuracy across the majority of questions and tasks. Its reliability makes it the current preferred baseline for general clinical inference and triage scenarios.
- Few-shot prompting demonstrated variable improvements in selected tasks, occasionally raising sensitivity or positive predictive value in domains with clear case signals. However, its performance was unstable across questions, occasionally introducing higher false positive rates and inconsistent F1-scores compared to zero-shot prompting. The effectiveness of few-shot learning appears highly dependent on the quality and relevance of the provided exemplars.
- Chain-of-thought prompts failed to yield anticipated performance gains—in most cases, CoT resulted in lower sensitivity and F1-score than either zero-shot or few-shot strategies, despite achieving moderate specificity and negative predictive values. The promise of intermediate stepwise reasoning did not translate into improved clinical accuracy, indicating that current CoT design and calibration require substantial refinement before deployment in high-stakes clinical settings.
- Statistical analyses confirmed the significance of these differences, with zero-shot models maintaining a clear advantage. CoT and few-shot approaches may have niche utility in special cases, but cannot yet replace the reliability of zero-shot inference for broad clinical application.
- Limitations identified include class imbalance, prompt quality, and lack of adaptation to specific clinical scenarios—highlighting the need for future work in prompt engineering, example selection, and model customization.

- For clinical and AI researchers, these findings underscore the necessity for rigorous prompt and workflow evaluation before adopting advanced prompting strategies in practice.

Funding

This research received **no specific grant from any funding agency** in the public, commercial, or not-for-profit sectors.

References

1. Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). Natural language processing in radiology: A systematic review. Radiology, 279(2), 329–343.
2. Spasić, I., Nenadić, G., & Krzyżanowski, M. (2020). Text mining of cancer-related information: Review of current status and future directions. International Journal of Medical Informatics, 83(9), 605–623.
3. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940.
4. Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388(13), 1233–1239.
5. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P.,... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.
7. Min, S., Lyu, X., Holtzman, A., Arora, M., Mirhoseini, A., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? arXiv Preprint, arXiv:2202.12837.
8. Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B.,... & Sui, Z. (2023). A survey on in-context learning. arXiv Preprint, arXiv:2301.00234.
9. Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1–11.
10. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.
11. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
12. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.,... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.
13. Love, C., Din, A. S., Tomas, M. B., Kalapparambath, T. P., & Palestro, C. J. (2003). Radionuclide bone imaging: An illustrative review. Radiographics, 23(2), 341–358.
14. Van den Wyngaert, T., Strobel, K., Kampen, W. U., Kuwert, T., van der Bruggen, W., Mohan, H. K.,... & Beheshti, M. (2016). The EANM practice guidelines for bone scintigraphy. European Journal of Nuclear Medicine and Molecular Imaging, 43(9), 1723–1738.
15. Chen, M. C., Ball, R. L., Yang, L., Moradzadeh, N., Chapman, B. E., Larson, D. B.,... & Langlotz, C. P. (2018). Deep learning to classify radiology free-text reports. Radiology, 286(3), 845–852.