Health and Technology (2026) 16:387–400
https://doi.org/10.1007/s12553-025-01041-y

ORIGINAL ARTICLE

Can reasoning LLMs enhance clinical document classification?

Akram Mustafa[1] · **Usman Naseem**[2] · **Mostafa Rahimi Azghadi**[1,3]

Received: 2 August 2025 / Accepted: 20 December 2025 / Published online: 7 January 2026

Abstract

Background Clinical document classification is a critical process in healthcare, converting unstructured medical texts into standardized ICD-10 diagnoses. This process faces challenges due to the complex and varied nature of medical language, which includes domain specific terminology, abbreviations, and unique writing styles across institutions. Additionally, privacy regulations and limited high quality annotated datasets hinder the development of robust models. LLMs have emerged as a transformative technology in healthcare, improving the efficiency and accuracy of tasks like clinical document classification by leveraging advanced natural language understanding.

Objective The objective of this study is to evaluate the performance and consistency of LLMs in binary classification clinical discharge summaries based on ICD-10 codes. By leveraging both reasoning and non-reasoning LLMs, the study aims to determine how effectively these models can identify and classify clinical patterns in a binary context, providing insights into their potential for improving automated clinical coding accuracy and enhancing decision support in healthcare settings.

Methods This study used a balanced subset of the MIMIC-IV dataset, comprising 3,000 discharge summaries including 150 positive and 150 negative samples for each of the top 10 ICD-10 codes. The summaries were tokenized using cTAKES, which converted clinical narratives into structured SNOMED codes, capturing contextual details such as affirmation or negation. Eight LLMs, including four reasoning (Qwen QWQ, Deepseek Reasoner, GPT o3 Mini, Gemini 2.0 Flash Thinking) and four non-reasoning models (Llama 3.3, GPT 4o Mini, Gemini 2.0 Flash, Deepseek Chat), were evaluated over three experimental runs. Final predictions were determined using majority voting across the runs to assess accuracy, F1 score, and consistency.

Results Among the eight evaluated LLMs, reasoning models demonstrated superior performance in ICD-10 classification, achieving an average accuracy of 71% and an F1 score of 67%, compared to 68% accuracy and 60% F1 score for non-reasoning models. Gemini 2.0 Flash Thinking achieved the highest accuracy at 75% and F1 score at 76%, while GPT 4o Mini had the lowest performance 64% accuracy, and 47% F1 score. Consistency analysis revealed that non-reasoning models exhibited higher stability of 91% average consistency than reasoning models of 84%. Performance variations across ICD-10 codes highlighted strengths in identifying well defined conditions but challenges in classifying abstract diagnostic categories.

Conclusion The evaluation of reasoning and non-reasoning LLMs in ICD-10 classification highlights a trade-off between accuracy and consistency. Reasoning models achieved higher classification accuracy and F1 scores, excelling in complex clinical cases, while non-reasoning models demonstrated superior stability across repeated trials. These findings suggest that a hybrid approach, leveraging the strengths of both model types, could optimize automated clinical coding by balancing accuracy and reliability. Future research should explore multi-label classification, domain specific fine tuning, and ensemble modeling to enhance performance and generalizability in real-world healthcare applications.

Keywords Reasoning · Large language model · ChatGPT · DeepSeek · Clinical coding · Gemini · Llama · Qwen

✉ Mostafa Rahimi Azghadi
mostafa.rahimiazghadi@jcu.edu.au

[1] College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

[2] School of Computing, Macquarie University, Sydney, NSW 2113, Australia

[3] Centre for AI and Data Science Innovation, James Cook University, Townsville, Australia

1 Introduction

Clinical document classification faces significant challenges due to the complex nature of medical narratives. The global clinical coding market size was valued at more than $37 Billion USD in 2024 [1]. The clinical coding process faces significant challenges due to the complex, varied nature of medical narratives. These documents use heavy domain specific terminologies, abbreviations, and unique writing styles that vary by institution and practitioner. Moreover, in addition to healthcare strict privacy regulations, the limited availability of large, high quality annotated datasets slows down the development of robust classification models. The frequent imbalance in class distributions and the inherent ambiguity in clinical language further complicate model training and evaluation. To address these issues, specialized Natural Language Processing (NLP) frameworks such as the clinical Text Analysis and Knowledge Extraction System (cTAKES) have been developed to better handle the unique characteristics of clinical texts [2].

Clinical document classification is one of the processes that the healthcare industry relies on to convert unstructured medical documents to International Classification of Diseases, 10th Revision (ICD-10) diagnoses [3]. ICD-10 coding has become an important process of clinical practice by providing a standardized language for documenting diagnoses and procedures. This systematic classification facilitates efficient communication among healthcare providers, ensuring that patient information is accurately recorded and easily shared across different institutions and systems [4, 5]. The consistent application of ICD-10 codes not only enhances clinical documentation and billing processes but also plays a pivotal role in epidemiological research, quality assurance, and public health reporting [6]. By enabling comprehensive tracking of disease trends and treatment outcomes, ICD-10 coding informs health policy and resource allocation, ultimately contributing to improved patient care and safety [7, 8].

Large language models (LLMs) have rapidly emerged as a transformative technology in healthcare, driven by their advanced natural language understanding and reasoning capabilities [9]. These models, trained on extensive and diverse corpora that include medical literature and clinical narratives, can parse complex clinical language, infer relationships among diverse medical concepts, and generate contextually nuanced insights [10, 11]. Such reasoning abilities are particularly beneficial for automating tasks like clinical document classification and diagnosis coding, where they can support decision-making and improve the efficiency and accuracy of healthcare delivery. Recent developments in LLM architectures, including techniques like chain-of-thought processing, have further enhanced their capacity to handle the intricacies of clinical data, marking a significant shift from traditional NLP approaches [12, 13].

Comparing reasoning versus non-reasoning large language models addresses a crucial research gap in applying these systems to complex clinical tasks. While models enhanced with explicit reasoning have shown promise in capturing intricate medical relationships and nuances in clinical narratives [14, 15], their benefits over standard non-reasoning models have not been systematically quantified in clinical document classification. This comparison is essential to determine whether the additional computational complexity and resource demands of reasoning-based approaches translate into significantly improved diagnostic coding accuracy and consistency, ultimately informing more effective and interpretable model deployment in healthcare settings.

This study addresses several critical research questions that bridge the gap between reasoning and non-reasoning LLMs in binary clinical document classification. First, it investigates how reasoning LLMs compare with their non-reasoning counterparts in classifying ICD-10 coded discharge summaries for specific binary outcomes, exploring whether the enhanced reasoning abilities translate to improved handling of clinical narratives. Second, the research evaluates the overall binary classification accuracy of LLMs on tokenized discharge summaries, providing quantitative insights into their performance on a challenging and balanced dataset. Third, the study examines the consistency of classification results across multiple experimental runs for each model, thereby assessing the robustness and reliability of these approaches. Collectively, these research questions aim to illuminate the strengths and limitations of current LLM methodologies in binary clinical document classification, offering valuable contributions toward the optimal deployment of AI in clinical settings.

2 Related work

2.1 Clinical text classification

Recent advances in clinical text classification have progressively shifted from traditional machine learning techniques to more nuanced deep learning methods that better capture the complexity of clinical language. For instance, the work by Haoran Shi et al. [16] introduces a hierarchical deep learning model with an attention mechanism specifically designed for the automated assignment of ICD diagnostic codes from written diagnosis descriptions. Their approach leverages character-aware neural language models to generate detailed hidden representations of both clinical narratives and ICD codes, effectively addressing the inherent mismatch between the number of descriptions and their corresponding codes. The significant performance gains, marked by an F1 score of 53% and an AUC of 90%, underscore the potential

of deep learning frameworks over conventional methods that often rely on manual feature engineering.

Gehrmann et al. [17] explored the potential of deep learning for patient phenotyping which is a clinical text classification task closely related to automated ICD coding, by leveraging discharge summaries from the MIMIC-III dataset. Their study compared convolutional neural networks (CNNs) with traditional n-gram models and cTAKES based approaches, demonstrating that CNNs could automatically learn salient textual features, thereby significantly outperforming the other methods across 10 distinct phenotyping tasks. With an average F1 score of 76%, their findings highlight the advantages of deep learning in capturing complex clinical information, reducing reliance on manually crafted annotation rules, and improving interpretability. While the focus of their work was on patient phenotyping, the methodological insights and demonstrated performance gains underscore the relevance of deep neural architectures for ICD coding tasks as well, further cementing the MIMIC dataset's role as a critical benchmark in clinical NLP research.

2.2 Tokenization & preprocessing in clinical NLP

In clinical NLP, the preprocessing of unstructured medical texts is essential for ensuring that downstream models can effectively interpret clinical narratives. In his work on classifying medical notes into standard disease codes [18], Karmakar emphasizes a systematic approach to note preprocessing that includes standard steps such as lowercasing, removal of special characters, and tokenization. Tools like cTAKES are particularly relevant in this context because they provide domain-specific processing capabilities tailored to the nuances of clinical language. cTAKES not only streamlines tokenization and sentence boundary detection but also offers robust named entity recognition that can identify medical terms, abbreviations, and concepts despite variations in clinical documentation. This targeted preprocessing is crucial for managing the inherent variability of discharge summaries and ultimately contributes to the effectiveness of models, such as CNNs and Long Short Term Memory models (LSTMs), when classifying complex clinical data into ICD codes.

2.3 Large language models for clinical coding

Li et al.'s work [19] provides a study exploring the use of multiple LLM agents collaborating to assign ICD codes to clinical notes. In this framework, distinct agents were tasked with reading clinical notes, generating relevant ICD codes, and then engaging in a discussion to reach a consensus. By structuring the interaction as a multi-agent dialogue, the authors aimed to simulate expert deliberation and improve coding accuracy. Their findings demonstrated that this collaborative LLM approach significantly enhanced performance, highlighting the potential of agent-based coordination in complex medical classification tasks..

Mustafa et al.'s work [20] is closely aligned with the ongoing exploration of LLMs for clinical coding. By assessing ChatGPT 3.5 and 4 compared to traditional machine learning and SNOMED mapping methods, their study investigates the challenge of accurately classifying ICD-10 codes in complex medical records. They demonstrate that, while human coders still achieve superior accuracy, ChatGPT 4's ability to match the median human performance and exhibit improved consistency suggests that advanced LLMs can play a valuable role in clinical coding.

These studies highlight both the potential and challenges of using LLMs for clinical coding. While they show promising performance, issues such as hallucination and handling complex imbalanced code sets remain. Approaches that incorporate reasoning, domain knowledge, and structured workflows offer a path forward. In this research, we investigate if reasoning LLMs can enhance document classification.

3 Materials and methods

3.1 Dataset preparation

The MIMIC IV dataset, which contains a diverse range of discharge summaries annotated with ICD-10 diagnoses, was used. These diagnoses were classified into two clinically significant categories: the Top 5 Primary Diagnoses, directly extracted from the discharge summaries and representing the main clinical focus of each case, and the Top 5 Secondary Diagnoses, capturing additional clinical details that, while not the primary focus, are important for patient care and outcomes (see Table 1).

The analysis incorporated both primary and secondary diagnoses, forming a set of the top 10 ICD-10 codes. This dual list approach was selected to capture a more comprehensive spectrum of clinically relevant conditions. By combining both primary and secondary diagnoses, the study benefits from a richer diagnostic landscape, which is essential for robust evaluation of classification performance across diverse clinical scenarios. To ensure class balance within the dataset, for each of these 10 ICD-10 codes, 150 discharge summaries confirmed as positive (i.e., diagnosed with the respective ICD-10 code) were extracted. An equal number of 150 negative discharge summaries, not associated with the given ICD-10 code, were randomly sampled from the set of ICD codes, excluding the 10 selected codes. This approach yielded a total of 300 discharge summaries per ICD-10 code, resulting in an aggregate dataset of 3,000 discharge summaries.

The balanced nature of this dataset, along with the inclusion of both primary and secondary diagnoses, forms the

Table 1 Top 10 ICD codes used in this study. The top part of the table shows 5 codes with the highest number of primary diagnoses cases, while the bottom part shows the top 5 codes with the maximum total cases in the MIMIC IV dataset

Diagnosis	ICD-10 Code	Total Cases	Primary Diagnosis Cases
Sepsis	A41	7,430	4,830
Myocardial infarction	I21	5,735	2,722
Other medical care	Z51	6,919	2,370
Chronic ischaemic heart disease	I25	38,157	2,302
Hypertensive heart and renal disease	I13	8,366	2,114
Disorders of lipoprotein metabolism and other lipidemias	E78	49,310	9
Essential hypertension	I10	43,574	76
Long term drug therapy	Z79	40,393	0
Personal history of other diseases and conditions	Z87	40,255	0
Place of occurrence of the external cause	Y92	35,297	0

foundation for the subsequent experiments evaluating the performance and consistency of various large language models in the classification task. It is important to note that due to the significant computational and financial costs associated with running multiple large language models, the dataset was intentionally limited to 3,000 discharge summaries. This decision ensured that the experiments remained feasible within the available resources while still providing a robust evaluation framework for the classification tasks.

3.2 Preprocessing

cTAKES was utilized to convert unstructured discharge summaries into structured, machine-interpretable data by transforming the clinical narratives into lists of standardized SNOMED codes. These codes represent a comprehensive array of clinical entities such as diseases, lab tests, procedures, medications, symptoms, anatomy, and events. During the tokenization process, cTAKES not only extracted these terms but also determined their contextual status categorizing each as either affirmed or negated. This automated pipeline significantly reduced the length of the original documents by eliminating unnecessary text, which in turn improved downstream processing efficiency and cost. After processing, the extracted SNOMED items were aggregated into two separate sections based on their contextual designation, and the frequency of each term's occurrence was recorded. This comprehensive approach provided a robust dataset of tokenized discharge summaries, enriched with detailed frequency counts and contextual annotations, which served as the foundation for subsequent classification experiments using large language models.

Table 2 Large Languages Models used

Model	Reasoning Type	Platform
Deepseek Reasoner	Reasoning	Deepseek
Gemini 2.0 Flash Thinking	Reasoning	Google
GPT o3 Mini	Reasoning	OpenAI
Qwen QWQ 32B	Reasoning	Groq
Deepseek Chat	Non-Reasoning	Deepseek
Gemini 2.0 Flash	Non-Reasoning	Google
GPT 4o Mini	Non-Reasoning	OpenAI
Llama 3.3 70B Versatile	Non-Reasoning	Groq

3.3 Experimental setup

A total of eight large language models were chosen to evaluate the classification of clinical discharge summaries based on ICD-10 codes. The models were categorized into two groups: reasoning and non-reasoning, as shown in Table 2. The deliberate selection of these eight models enabled a comprehensive comparison between architectures that incorporate explicit reasoning and those that do not, thereby addressing the research question concerning the impact of reasoning on clinical document classification.

For the experimental procedure, each of the eight models was tasked with classifying a dataset of 3,000 tokenized discharge summaries that had been preprocessed using cTAKES. To ensure the robustness and repeatability of the results, each model was run three times, with each run independently processing the entire dataset. In every experimental iteration, the models generated classifications indicating whether each discharge summary belonged to its corresponding ICD-10 group. The prompt used for all models was:

"Discharge Summary: [Report Text] Does this summary contain the diagnosis associated with ICD-10 code [ICD10 Code] or any of its specific subcategories, Answer with Yes or No only?"

The final classification outcome for each summary was determined using a majority vote across the three repetitions, meaning that the label assigned by at least two out of three runs was taken as the definitive result. This methodological approach not only ensured consistency in the classification outcomes but also provided a reliable measure of each model's performance in distinguishing between the ICD-10 groups. Figure 1 illustrates this process. In addition, in favour of reproducible research, our code and data are shared publicly on GitHub: https://github.com/asmgx/LLMs.

A key consideration in evaluating LLM consistency is the role of the temperature parameter, which controls the randomness and creativity of the model's outputs. A higher temperature such as greater than 0.7 increases diversity by sampling from a broader distribution of likely tokens, while a lower temperature as less than 0.3 makes the model more deterministic, typically choosing the most probable token. Setting the temperature to 0 would, in theory, lead to 100% consistency for a given

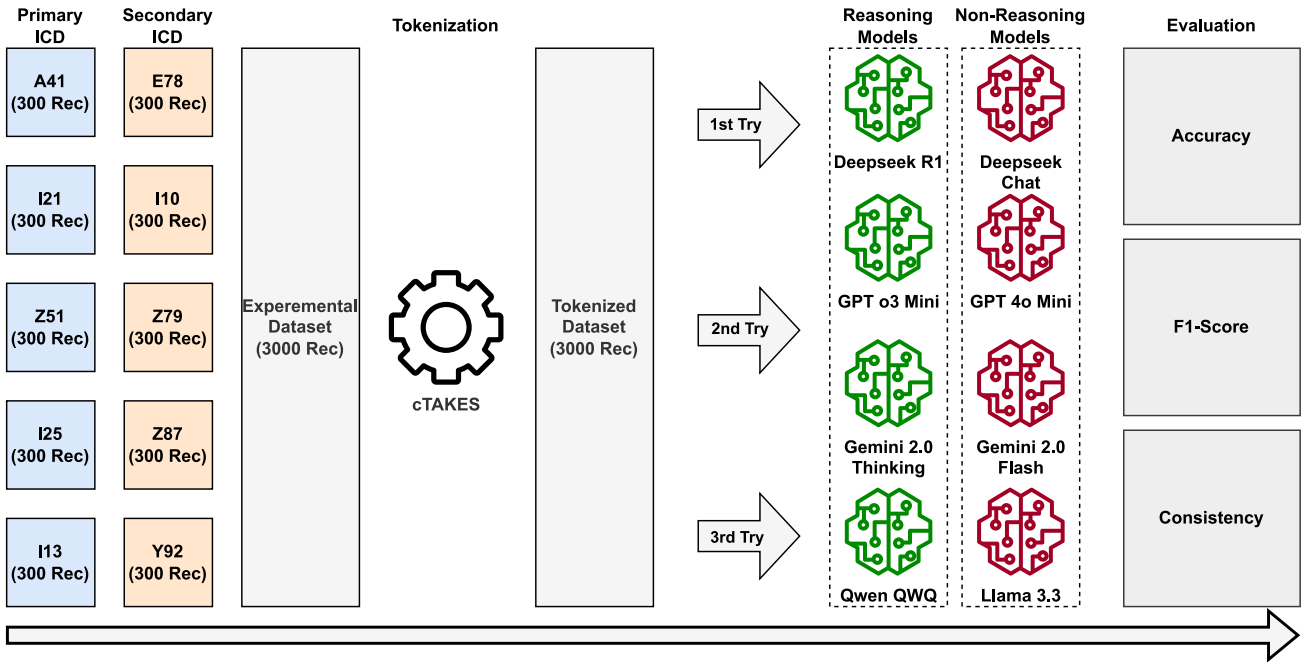

Fig. 1 Workflow of cTAKES based Tokenization and Repeated LLM Classification for ICD-10 coded Discharge Summaries

prompt and model version. In this study, we intentionally used the default temperature settings provided by each model's API (which are typically non-zero) for all experimental runs. This "out-of-the-box" evaluation strategy was chosen to reflect the typical usage scenario for practitioners who may leverage these models without extensive parameter tuning. While this means our consistency metric captures inherent variability under standard conditions, it is important to note that consistency is not an absolute property of the model but a function of its configuration. The observed differences in consistency between models therefore reflect their default behavior and stability in a practical, plug-and-play context. For applications requiring maximum reliability, users could likely improve consistency by manually setting a lower temperature parameter, albeit potentially at the cost of some creativity or nuance in reasoning.

The prompt was intentionally designed to be short and strict ("Answer with Yes or No only") to evaluate the base capability of each model with minimal guidance and to ensure output consistency for automated parsing. This approach simulates a zero-shot, instruction-following scenario and provides a conservative baseline of model performance without the potential performance gains of advanced prompt engineering techniques.

3.4 Evaluation metrics

Once final classifications were established, overall accuracy was computed to evaluate the models. The F1 score provided a more robust assessment of classification accuracy than accuracy alone. Together, these metrics offered deeper insight into the models' predictive reliability.

Lastly, we assessed how consistently each model classified the same discharge summaries across its three runs. Specifically, for each discharge summary, we recorded whether the model's prediction (positive or negative) remained the same in all three trials. The proportion of summaries that received identical classifications in all runs was then computed as an indicator of consistency. This measure offers insight into the reliability of the model's decision-making process, revealing whether fluctuations in output are minimal (high consistency) or more pronounced (low consistency) over repeated evaluations.

4 Results

4.1 LLMs performance

Among the eight LLMs evaluated, Gemini 2.0 Flash Thinking, a reasoning model, achieved the highest overall accuracy at 75% and the highest F1 score at 76%. In contrast, GPT 4o Mini, a non-reasoning model, had the lowest performance among all models, with an accuracy of 64% and an F1 score of 47%. Among the non-reasoning models, Gemini 2.0 Flash recorded the highest accuracy at 72% and the highest F1 score at 71%, while Qwen QWQ had the lowest accuracy and F1 score among reasoning models, with 65% accuracy and 55% F1 score. Table 3 presents the detailed accuracy and F1 scores for each model.

Overall, the reasoning models outperformed the non-reasoning models in terms of average accuracy and F1 scores.

Table 3 Average performance of various LLMs over 3 runs in classifying 3000 clinical discharge summaries from the top 10 ICD-10 codes within the MIMIC IV dataset

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	73.83%	69.91%	82.83%
Gemini 2.0 Flash Thinking	**75.30%**	**75.50%**	78.43%
GPT 3o Mini	71.37%	63.59%	**95.47%**
Qwen QWQ 32B	65.23%	54.83%	78.07%
Deepseek Chat	69.80%	59.48%	95.40%
Gemini 2.0 Flash	71.93%	70.50%	90.67%
GPT 4o Mini	63.77%	46.69%	88.73%
Llama 3.3 70B Versatile	68.33%	60.15%	88.30%

The reasoning models achieved an average accuracy of 71% compared to 68% for the non-reasoning models. Similarly, the reasoning models recorded an average macro F1 score of 67%, whereas the non-reasoning models achieved a lower average macro F1 score of 60%. Figure 2 shows the average performance of the four state-of-the-art reasoning versus four non-reasoning LLMs in classifying clinical documents. These results suggest that, in general, the reasoning models provided slightly better classification performance than the non-reasoning models.

4.2 Consistency analysis

The consistency of model performance across the three experiments reveals a varied level of reliability among the different LLMs. Results are shown in Table 3. Models such as GPT o3 Mini and Deepseek Chat at 95%, and Gemini 2 Flash at 91% demonstrate high consistency, with performance largely stable across all three trials. This suggests that these models are robust in classifying discharge summaries, maintaining similar results across multiple runs. On the other hand, models like Qwen QWQ and Gemini 2 Flash Thinking, with a consistency rate of 78%, show significantly lower consistency, indicating greater variability in their performance across the experiments. This variability could potentially indicate issues with stability or sensitivity to different input data or model configurations. Other models such as Deepseek Reasoner at 83%, Llama 3.3 at 88%, and GPT 4o Mini at 89% fall in between, suggesting a moderate level of consistency. Overall, the high consistency observed in certain models like Deepseek Chat and GPT o3 Mini suggests they are more reliable for repeated use in clinical document classification tasks. An analysis of the average consistency based on model reasoning, reveals that non-reasoning models exhibit a higher average consistency of 91% compared to reasoning models, which show an average consistency of 84%. Figure 2 illustrates these findings in detail.

4.3 Detailed performance analysis of various LLMs

The performance analysis of the eight investigated LLMs in classifying ICD-10 codes is presented in two different ways. Tables 4 to 13 list the accuracy, F1 score, and consistency of all LLMs on each specific code. Additionally, Figs. 3 to 12

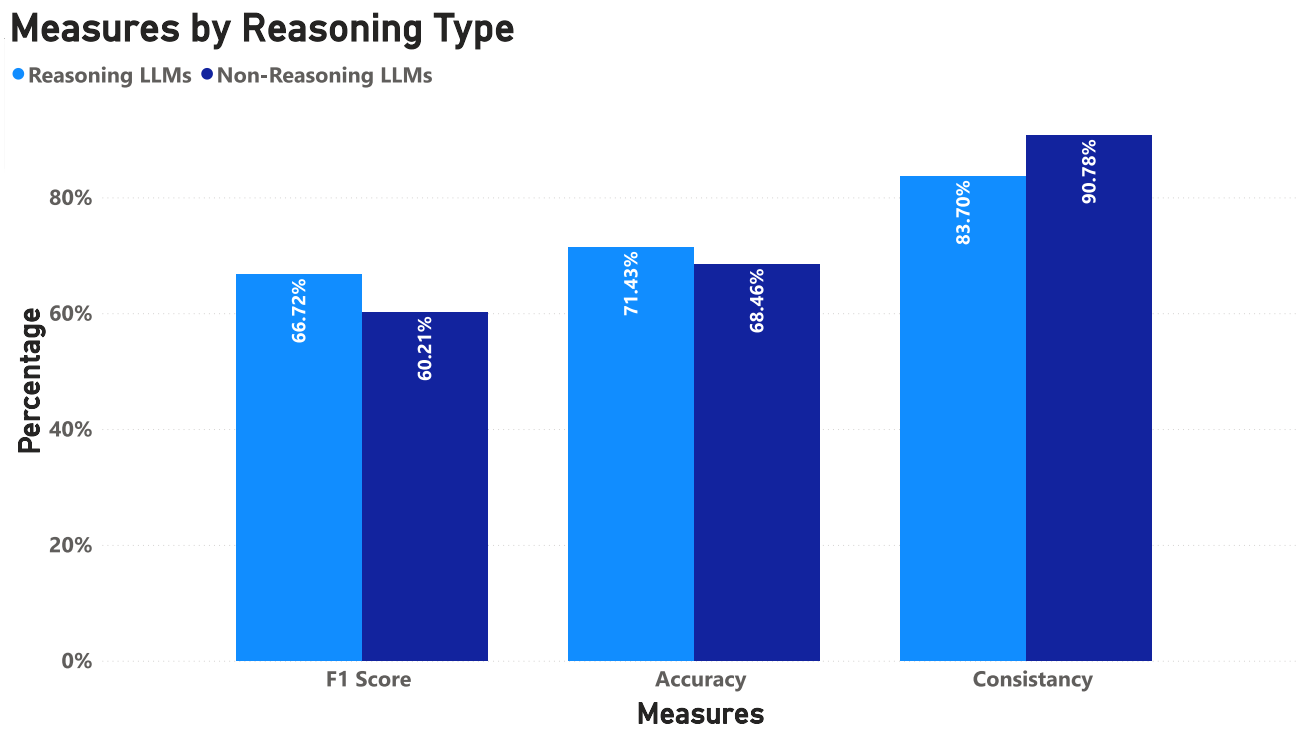

Fig. 2 Average performance of 4 reasoning vs 4 non-reasoning LLMs in classifying clinical documents

demonstrate the F1 score of all LLMs vs their consistency, for each code. These results assist in classifying the codes into three categories.

The first category includes codes for which all models demonstrate strong performance in both consistency and F1 scores, as shown in Figs. 3, 5, 7, and 8, exhibiting similar trends and levels of accuracy. This category comprises ICD-10 codes A41, I10, I21, and I25. For these codes, consistency ranges from 67% to 100%, while F1 scores range from 63% to 96%, as shown in Tables 4, 6, 8, and 9. Reasoning models, particularly GPT 3o Mini, stood out in terms of consistency, achieving the highest levels for 3 out of the 4 codes. In terms of F1 scores, both reasoning and non-reasoning models led, depending on the specific ICD code. Notably, GPT 4o Mini achieved the highest F1 score of 96%.

The second category includes codes where consistency remained high and stable across all models, but F1 scores varied significantly, as shown in Figs. 4, 6, 9, and 10. This suggests that while the models consistently made similar predictions, their accuracy in identifying true positive cases varied widely. ICD-10 codes in this category include E78, I13, Y92, and Z51. Consistency scores ranged from 64% to 100%, while F1 scores showed a much wider spread—ranging from 23% to 79% for E78, 30% to 86% for I13, 0% to 20% for Y92, and 4% to 87% for Z51, as shown in Tables 5, 7, 10, and 11. GPT 4o Mini showed notably poor performance in this category, recording the lowest F1 scores for E78, I13, and Y92 at 23%, 30%, and 0%, respectively.

The third category includes codes Z79 and Z87, where both consistency and F1 scores varied greatly across models, indicating inconsistent performance in identifying positive discharge summaries, as shown in Figures as shown in Figs. 11, 12. For Z79, consistency ranged from 33% to 100%, while F1 scores ranged from 0% to 57%. For Z87, consistency ranged from 36% to 100%, and F1 scores ranged from 0% to 62%, as detailed in Tables 12 and 13. GPT 3o Mini demonstrated 100% consistency for both codes but scored 0% on F1, indicating that it consistently misclassified all records as negative. Similarly, Deepseek Chat and GPT 4o Mini also failed to identify any positive cases for Z87, both recording an F1 score of 0%, although their consistency levels were slightly lower.

5 Discussion

5.1 Interpretation of findings

The findings of this study indicate that reasoning LLMs generally outperform non-reasoning models in ICD-10 classification tasks, achieving higher accuracy and F1 scores on average. This aligns with an initial hypothesis that explicit reasoning capabilities would enhance the ability of LLMs to capture complex clinical relationships within discharge summaries. However, the results also reveal unexpected patterns. While reasoning models demonstrated superior classification performance, they exhibited lower consistency across repeated experiments than non-reasoning models. This suggests that while reasoning models may improve accuracy, they could also introduce greater variability in results, possibly due to sensitivity to input variations or internal model heuristics. Additionally, while some reasoning LLMs excelled in identifying well-defined medical conditions such as sepsis and myocardial infarction, they struggled with more abstract ICD-10 codes related to patient history and external causes, reinforcing concerns about their contextual understanding. These findings highlight both the potential and limitations of reasoning LLMs, underscoring the need for further refinement to enhance their reliability in clinical applications.

One explanation for the lower consistency of reasoning models is their reliance on multi-step generative reasoning pathways, which, while enhancing their ability to capture nuanced relationships, also introduce variability in outputs due to stochastic sampling and sensitivity to prompt phrasing. Unlike non-reasoning models that rely more on direct pattern recognition, reasoning models may reach different intermediate rationales across runs, resulting in fluctuations in final predictions. Furthermore, the observed variation in performance across ICD-10 codes reflects differences in how easily the clinical concepts can be identified in discharge summaries. Codes representing explicit and well-documented conditions such as sepsis (A41) or myocardial infarction (I21) provide strong lexical and contextual cues, enabling both reasoning and non-reasoning models to perform well. In contrast, more abstract or administrative codes such as long-term drug therapy (Z79) or place of occurrence of external cause (Y92) are less directly referenced in clinical text, requiring contextual inference that is inconsistently handled by current LLMs. This disparity highlights the challenge of aligning reasoning capabilities with the varied granularity of ICD coding

Additionally, to contextualize the performance of LLMs in clinical document classification, we refer to Mustafa et al.'s study [9] that compared ChatGPT and human coders on a similar task involving challenging discharge summaries. In that study, human coders exhibited a median accuracy of 22% and F1 scores ranging from 12.59% to 50.60% across various ICD-10 codes, with significant inter-rater variability. Notably, ChatGPT-4 achieved a median accuracy of 22%, matching the median human performance, while also demonstrating F1 scores closely aligned with human averages such as 11.76% vs. 12.59% for A41, and 52.63% vs. 50.60% for Z51. These results suggest that even state-of-the-art LLMs can perform on par with human coders in

complex classification scenarios, particularly when dealing with ambiguous or challenging cases. However, the high variability among both human and model outputs underscores the inherent difficulty of clinical coding and the need for robust, hybrid approaches that leverage both AI and human expertise to enhance reliability and accuracy.

Furthermore, the role of prompt engineering must be considered when interpreting these results. Our study employed a minimal, strict prompt to establish a baseline performance level under conditions of minimal guidance. It is well-established that more sophisticated prompt design, such as employing few-shot examples, chain-of-thought reasoning, or providing detailed, task-specific context, can substantially improve LLM performance on complex tasks [21, 22]. Therefore, the accuracy and consistency metrics reported here should be viewed as a lower bound of potential performance. The observed struggles with abstract codes such as Z87, and Z79 might be particularly amenable to improvement through carefully engineered prompts that provide explicit definitions and coding rules. Future work should systematically explore the impact of different prompt engineering strategies on clinical classification tasks to optimize the trade-off between simplicity, cost, and accuracy.

5.2 Clinical relevance

The optimization of ICD-10 coding workflows and the broader integration of AI in clinical documentation are of significant clinical relevance. The superior accuracy of reasoning LLMs suggests that they could serve as valuable tools for assisting clinical coders in accurately classifying discharge summaries, potentially reducing manual workload and minimizing coding errors. However, the observed variability in their consistency raises concerns about reliability, highlighting the importance of model calibration and validation before deployment in real world settings. Additionally, the inability of some models to correctly classify abstract or context-dependent ICD-10 codes underscores the need for further refinements, such as enhanced domain-specific training and better handling of nuanced medical contexts. If properly fine-tuned and integrated, these AI-driven coding assistants could streamline the ICD-10 classification process, improve coding accuracy, and ultimately support more efficient clinical documentation and billing workflows, leading to better healthcare data quality and decision making.

5.3 Future work

Future research could expand the scope of ICD-10 classification by incorporating all ICD-10 codes mentioned in discharge summaries, including both primary and secondary diagnoses. This would provide a more comprehensive evaluation of model performance across a broader range of clinical scenarios, capturing complex multi-diagnostic relationships that reflect real-world coding challenges. Additionally, exploring the performance of LLMs on multi-label classification tasks, where a single discharge summary may correspond to multiple ICD-10 codes, could enhance the generalizability of the models in handling overlapping clinical patterns, serving as a continuation of previous research in clinical document classification [20]. Further improvements could include fine tuning the models with domain specific data to enhance their understanding of clinical language and improve performance on nuanced cases. Moreover, conducting experiments with larger datasets and incorporating real-world clinical notes from multiple healthcare institutions could improve model robustness and generalizability. Comparing the performance of ensemble models that combine outputs from both reasoning and non-reasoning models may also provide a more balanced approach to complex classification tasks, leveraging the strengths of both model types to enhance coding accuracy and consistency.

Beyond the avenues mentioned above, future work should pursue two parallel tracks. The first is the near-term development and validation of the ensemble models discussed in this paper, which provide a readily testable framework based on our empirical results. The second, more ambitious track is the design of a fully-agentic AI coding system. Such a system would leverage advanced reasoning capabilities to dynamically retrieve information from the ICD-10 classification manual and other clinical guidelines such as via retrieval-augmented generation (RAG), simulate multi-step coding deliberations, and provide auditable rationales for its decisions. Research into near-term ensemble techniques will provide crucial insights into failure modes and integration logic that are essential for eventually building these more complex, reliable, and transparent agentic systems.

6 Conclusion

The comprehensive evaluation of reasoning and non-reasoning LLMs in classifying ICD-10 coded discharge summaries provides key insights into model performance and consistency. The results reveal that reasoning models outperformed non-reasoning models in terms of overall accuracy and F1 scores, with reasoning models achieving an average accuracy of 71% and a macro F1 score of 67%, compared to 68% and 60%, respectively, for non-reasoning models. However, non-reasoning models demonstrated higher consistency, averaging 91% compared to 84% for reasoning models, suggesting greater reliability across repeated trials. These results indicate that while reasoning models exhibit strength in handling complex clinical cases and achieving higher classification accuracy, non-reasoning models provide more stable and consistent performance in structured ICD-10 classification tasks.

A. Appendix

A.1. Tables

Table 4 A41 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	**87.00%**	86.22%	**96.67%**
Gemini 2 Flash Thinking	86.67%	**87.34%**	85.67%
GPT 3o Mini	86.00%	85.11%	96.33%
Qwen QWQ	76.00%	71.43%	78.00%
Deepseek Chat	86.67%	86.39%	93.67%
Gemini 2 Flash	76.33%	74.18%	94.00%
GPT 4o Mini	71.67%	65.86%	78.00%
Llama 3.3	71.00%	62.66%	85.67%

Table 5 E78 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	**80.67%**	**78.52%**	98.33%
Gemini 2 Flash Thinking	80.00%	78.26%	97.00%
GPT 3o Mini	80.33%	78.23%	**98.67%**
Qwen QWQ	69.33%	60.00%	79.33%
Deepseek Chat	64.33%	47.80%	92.67%
Gemini 2 Flash	74.00%	73.83%	91.33%
GPT 4o Mini	56.33%	23.39%	91.00%
Llama 3.3	70.67%	72.33%	74.33%

Table 6 I10 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	82.67%	84.88%	99.00%
Gemini 2 Flash Thinking	**83.00%**	**85.22%**	99.33%
GPT 3o Mini	**83.00%**	**85.22%**	**100.00%**
Qwen QWQ	75.67%	75.75%	67.00%
Deepseek Chat	77.33%	76.55%	93.67%
Gemini 2 Flash	81.67%	83.58%	96.67%
GPT 4o Mini	69.67%	65.66%	82.33%
Llama 3.3	77.00%	77.08%	87.67%

Table 7 I13 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	70.33%	58.60%	75.00%
Gemini 2 Flash Thinking	**84.33%**	**86.30%**	87.33%
GPT 3o Mini	82.33%	80.59%	63.67%
Qwen QWQ	62.33%	40.21%	67.67%
Deepseek Chat	83.33%	80.31%	89.67%
Gemini 2 Flash	**84.33%**	85.89%	**91.67%**
GPT 4o Mini	58.33%	30.17%	80.00%
Llama 3.3	80.67%	79.14%	77.33%

Table 8 I21 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	86.67%	84.73%	95.00%
Gemini 2 Flash Thinking	86.67%	85.07%	97.33%
GPT 3o Mini	86.67%	84.85%	**98.33%**
Qwen QWQ	86.00%	83.97%	89.33%
Deepseek Chat	88.33%	86.99%	93.00%
Gemini 2 Flash	85.00%	82.76%	**98.33%**
GPT 4o Mini	80.33%	76.31%	85.67%
Llama 3.3	**95.00%**	**94.98%**	94.00%

Table 9 I25 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	93.67%	94.04%	**99.67%**
Gemini 2 Flash Thinking	92.00%	92.59%	96.33%
GPT 3o Mini	94.00%	94.34%	**99.67%**
Qwen QWQ	94.00%	94.34%	95.67%
Deepseek Chat	95.33%	95.54%	99.00%
Gemini 2 Flash	90.67%	91.46%	97.00%
GPT 4o Mini	**96.00%**	**96.08%**	92.67%
Llama 3.3	87.33%	88.76%	91.00%

Table 11 Z51 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	85.00%	83.27%	72.00%
Gemini 2 Flash Thinking	**85.33%**	**86.83%**	65.67%
GPT 3o Mini	51.33%	6.41%	98.00%
Qwen QWQ	46.33%	3.59%	80.33%
Deepseek Chat	51.00%	3.92%	**98.67%**
Gemini 2 Flash	70.67%	63.93%	83.33%
GPT 4o Mini	55.33%	19.28%	83.33%
Llama 3.3	52.33%	8.92%	90.00%

Table 12 Z79 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	**55.00%**	51.61%	39.67%
Gemini 2 Flash Thinking	54.00%	**57.14%**	32.67%
GPT 3o Mini	50.00%	0.00%	**100.00%**
Qwen QWQ	49.33%	3.80%	83.67%
Deepseek Chat	51.67%	7.64%	94.00%
Gemini 2 Flash	52.33%	36.44%	85.67%
GPT 4o Mini	50.00%	2.60%	96.00%
Llama 3.3	50.33%	2.61%	93.00%

Table 10 Y92 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	50.00%	2.60%	94.33%
Gemini 2 Flash Thinking	**51.33%**	7.59%	86.67%
GPT 3o Mini	50.00%	0.00%	**100.00%**
Qwen QWQ	50.00%	0.00%	98.67%
Deepseek Chat	50.00%	0.00%	**100.00%**
Gemini 2 Flash	51.00%	**19.67%**	91.67%
GPT 4o Mini	50.00%	0.00%	**100.00%**
Llama 3.3	50.00%	0.00%	99.67%

Table 13 Z87 Accuracy, F1 score & Consistency by LLM

Model	Accuracy	F1 Score	Consistency
Deepseek Reasoner	47.33%	26.17%	58.67%
Gemini 2 Flash Thinking	49.67%	55.46%	36.33%
GPT 3o Mini	50.00%	0.00%	**100.00%**
Qwen QWQ	43.33%	39.72%	41.00%
Deepseek Chat	50.00%	0.00%	99.67%
Gemini 2 Flash	**53.33%**	**62.37%**	77.00%
GPT 4o Mini	50.00%	0.00%	98.33%
Llama 3.3	49.00%	2.55%	90.33%

A.2. Charts

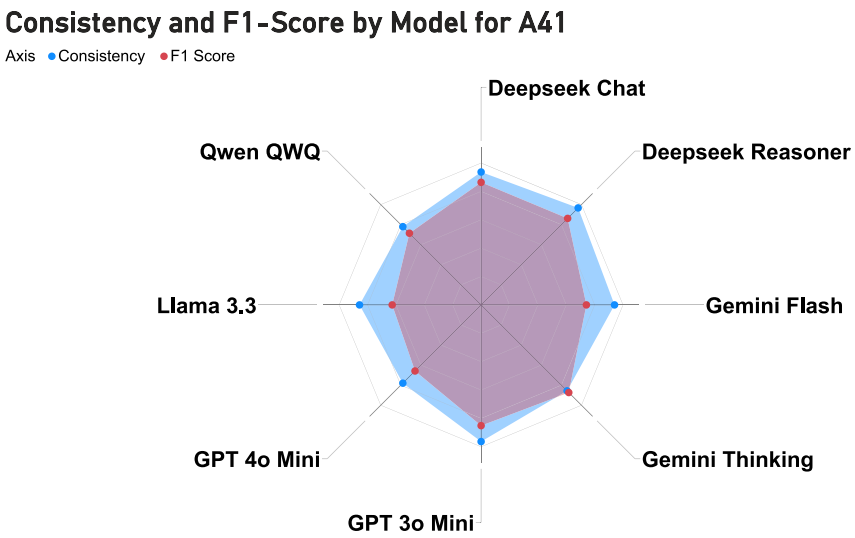

Fig. 3 Consistency and F1 Score by Model for A41

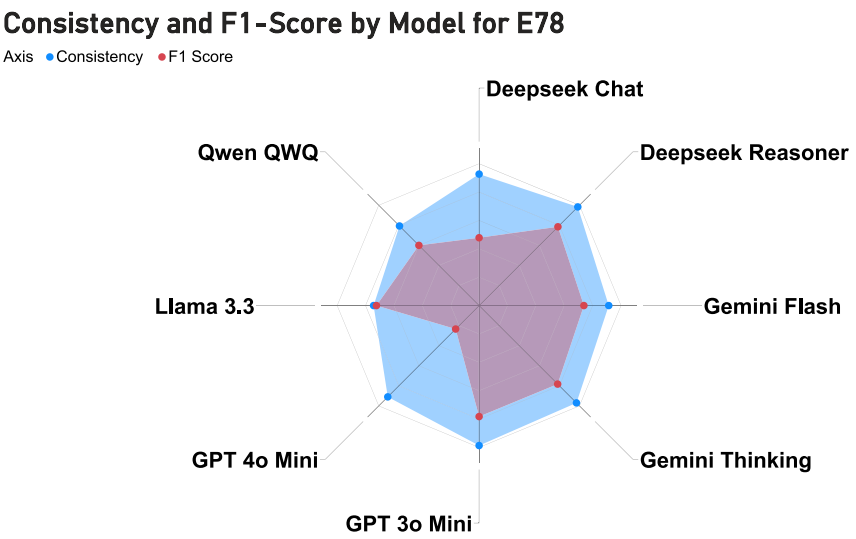

Fig. 4 Consistency and F1 Score by Model for E78

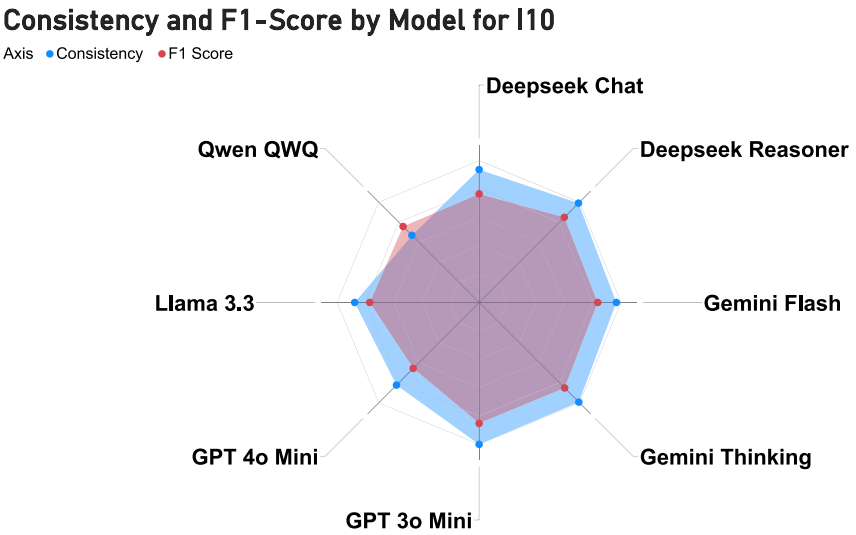

Fig. 5 Consistency and F1 Score by Model for I10

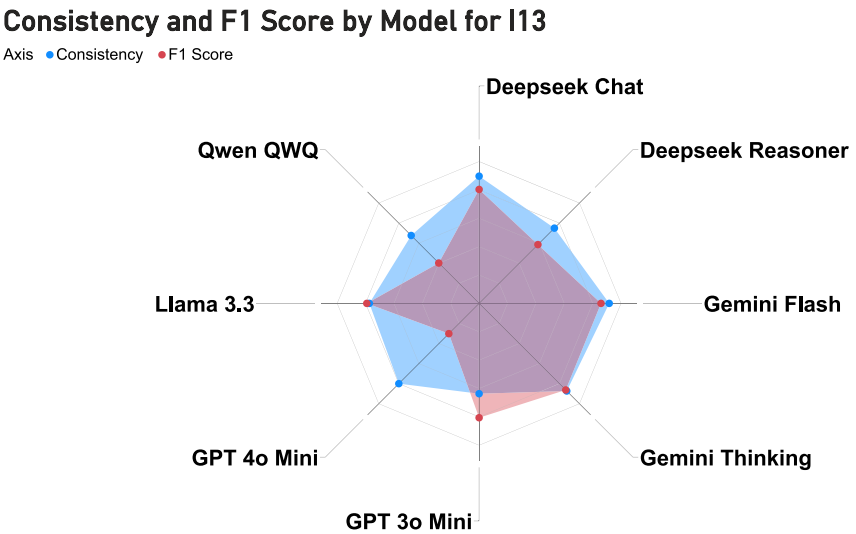

Fig. 6 Consistency and F1 Score by Model for I13

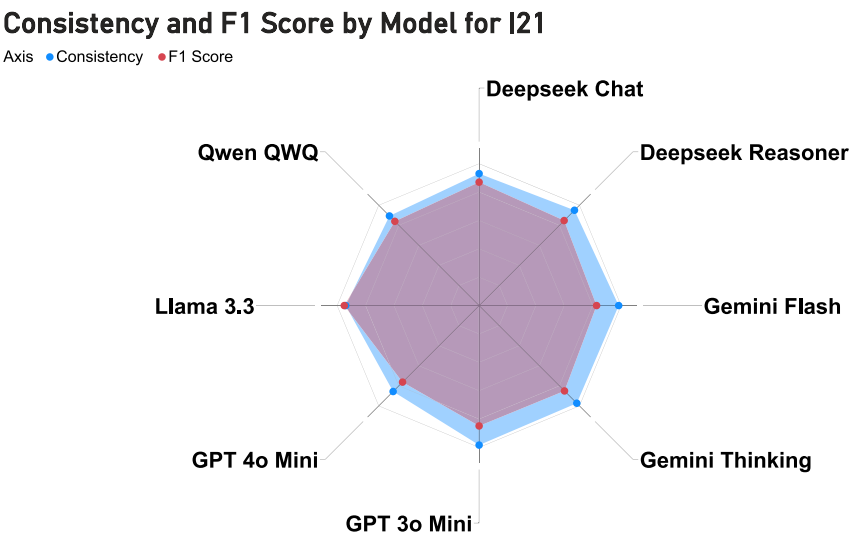

Fig. 7 Consistency and F1 Score by Model for I21

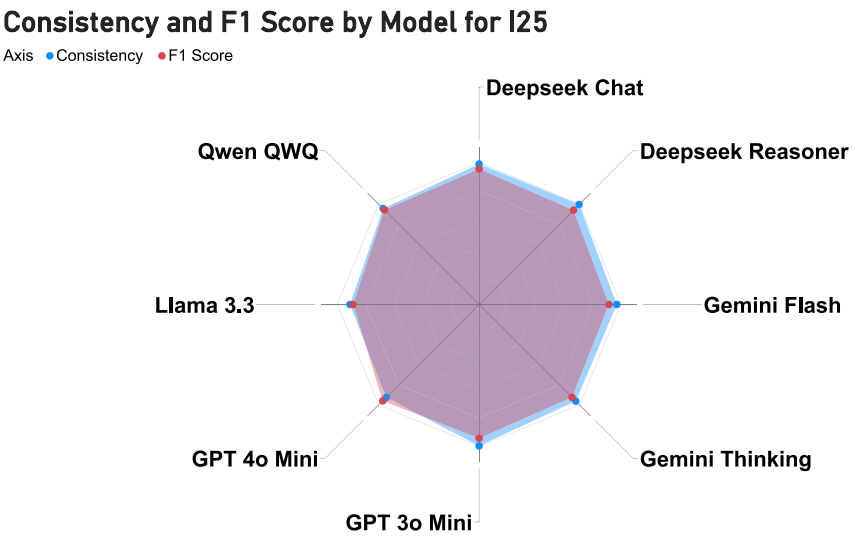

Fig. 8 Consistency and F1 Score by Model for I25

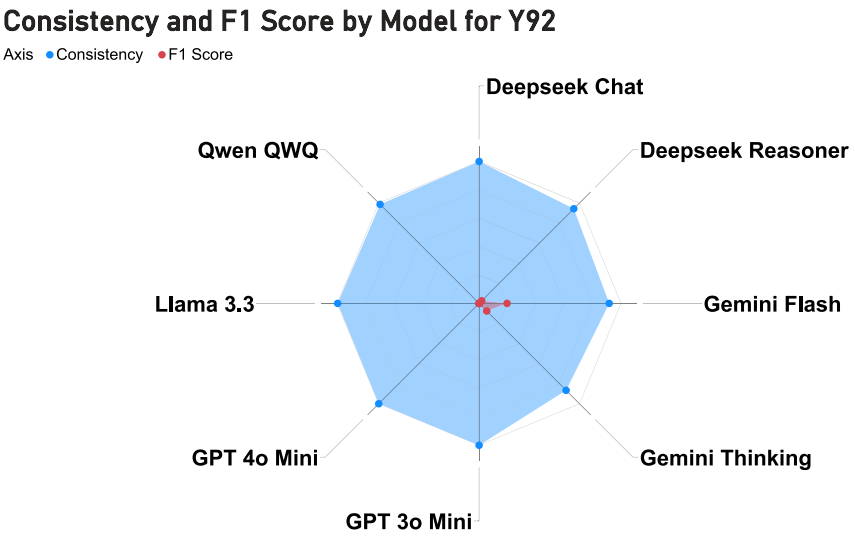

Fig. 9 Consistency and F1 Score by Model for Y92

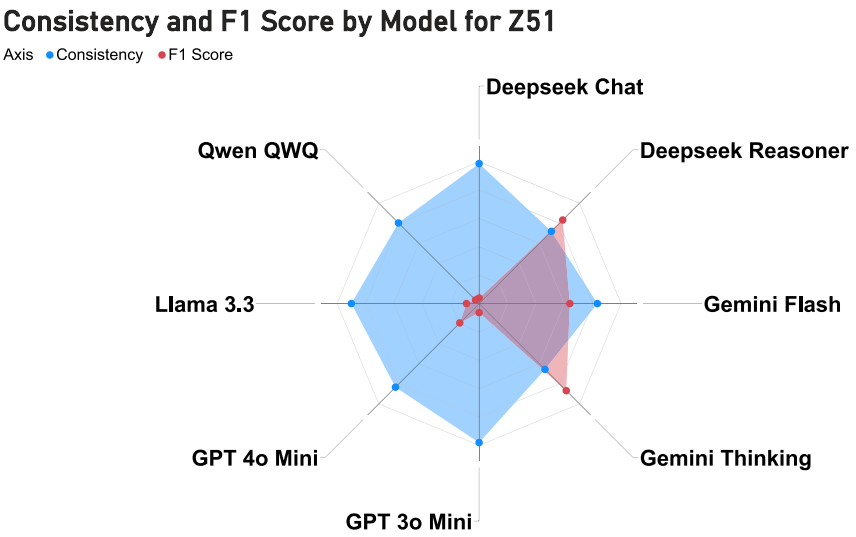

Fig. 10 Consistency and F1 Score by Model for Z51

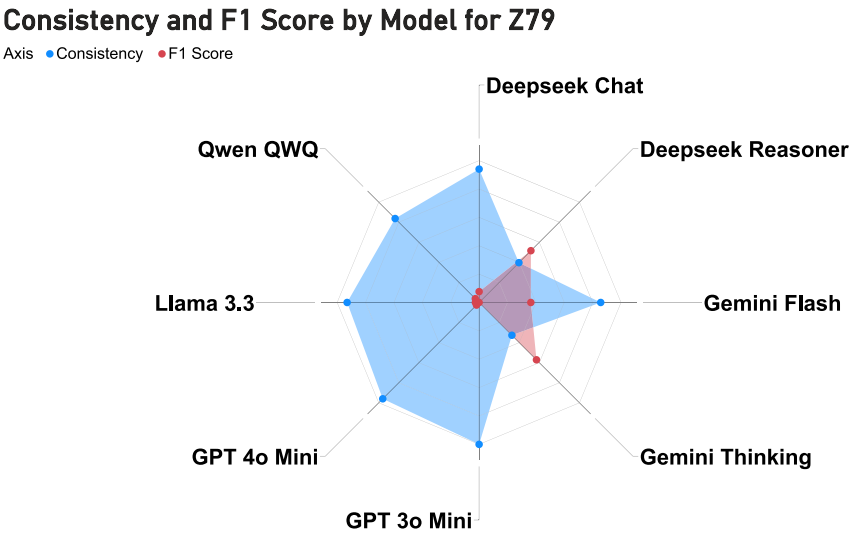

Fig. 11 Consistency and F1 Score by Model for Z79

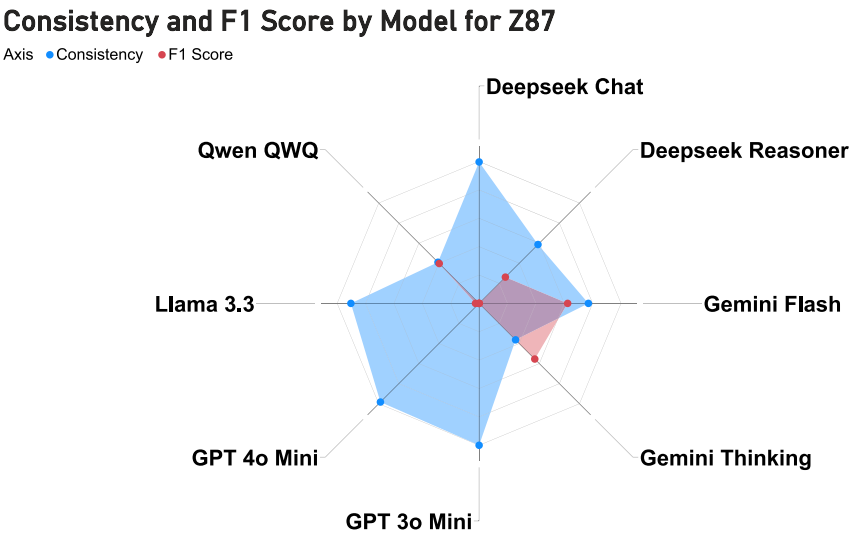

Fig. 12 Consistency and F1 Score by Model for Z87

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research was supported by the Australian Government Research Training Program.

Code availability The source code used for data processing and analysis is available at https://github.com/asmgx/LLMs.

Availability of Data and Material The MIMIC-IV dataset used in this study contains protected health information and cannot be shared publicly, in accordance with the data use agreement. Researchers may obtain access to MIMIC-IV through PhysioNet after completing the required training and agreements.

Declarations

Conflicts of interest The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Nester R. Medical coding market size & share, growth trends 2037. https://www.researchnester.com/reports/medical-coding-market/5899, accessed: 2025-03-16 2024.
2. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. J Am Med Inf Assoc. 2010;17(5):507–13.
3. Gao C, Goswami M, Chen J, Dubrawski A. Classifying unstructured clinical notes via automatic weak supervision. In: Machine Learning for Healthcare Conference, PMLR 2022;673–690.
4. Meyer H. Coding complexity: Us health care gets ready for the coming of icd-10. Health Aff. 2011;30(5):968–74.
5. Atutxa A, de Ilarraza AD, Gojenola K, Oronoz M, Perez-de VO. Interpretable deep learning to map diagnostic texts to icd-10 codes. Int J Med Inf. 2019;129:49–59.
6. Sammani A, Bagheri A, van der Heijden PG, Te Riele AS, Baas AF, Oosters C, et al. Automatic multilabel detection of icd10 codes in dutch cardiology discharge letters using neural networks. NPJ Digital Med. 2021;4(1):37.
7. Mclaughlin A, Hardt J, Canavan J, Donnelly M. Diagnosis-related group-based reimbursement is unrealistic for icus. Crit Care. 2009;13(Suppl 1):P485.
8. Kusnoor SV, Blasingame MN, Williams AM, DesAutels SJ, Su J, Giuse NB. A narrative review of the impact of the transition to icd-10 and icd-10-cm/pcs. JAMIA open. 2020;3(1):126–31.
9. Mustafa A., Naseem U., Azghadi MR. Large language models vs human for classifying clinical documents. Int J Med Inf. 2025;105800.
10. Nazi Z, Peng W. Large language models in healthcare and medical domain: A review. Informatics. 2024;11:57.
11. Meskó B. The impact of multimodal large language models on health care's future. J Med Int Res. 2023;25:e52865.
12. Far A, Bastani A, Lee A, Gologorskaya O, Huang C-Y, Pletcher MJ, Lai JC, Ge J. Evaluating the positive predictive value of code-based identification of cirrhosis and its complications utilizing gpt-4. Hepatology 10–1097.
13. Kwon T, Ong KT-i, Kang D, Moon S, Lee JR, Hwang D, Sohn B, Sim Y, Lee D, Yeo J. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales. In: Proceedings of the AAAI conference on artificial intelligence 2024;38:18417–18425.
14. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of thought utilization in large language models and application in nephrology. Medicina. 2024;60(1):148.
15. Bhatia S. Next-generation healthcare information systems: Integrating chain-of-thought reasoning and adaptive retrieval in large-scale document analysis, Available at SSRN 5029973 2024.
16. Shi H, Xie P, Hu Z, Zhang M, Xing EP. Towards automated icd coding using deep learning, 2017. arXiv:1711.04075
17. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, Foote Jr J, Moseley E, Grant DW, Tyler PD, et al. A comparison of rule-based and deep learning models for patient phenotyping, preprint, arxiv. org 2017.
18. Karmakar A. Classifying medical notes into standard disease codes using machine learning, 2018. arXiv:1802.00382.
19. Li R, Wang X, Yu H. Exploring llm multi-agents for icd coding, 2024. arXiv:2406.15363.
20. Mustafa A, Rahimi Azghadi M. Clustered automated machine learning (caml) model for clinical coding multi-label classification. Int. J. Mach. Learn. Cybernet. 2024;1–23.
21. Wever M, Tornede A, Mohr F, H¨ullermeier E. Automl for multi-label classification: overview and empirical evaluation. IEEE transactions on pattern analysis and machine intelligence. 2021;43(9):3037–54.
22. Li M, Zhou H, Yang H, Zhang R. Rt: a retrieving and chain-of-thought framework for few-shot medical named entity recognition. J Am Med Inform Assoc. 2024;31(9):1929–38.