NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
Translational Engineering in Health and Medicine

Evaluating the Diagnostic Performance of Large Language Models in Complex Clinical Cases: A Comparative Study

Postgraduate Diploma Thesis

Timotheos Kopsidas

Supervisor
Dr. Konstantina Nikita
Professor in School of Electrical and Computer Engineering
National Technical University of Athens

Co-Supervisors
Antonis Armoundas
Associate Professor of Medicine
Harvard Medical School
Konstantinos Mitsis, PhD
Vasilis Papakonstantinou
Vice Chairman, MIT Enterprise Forum Greece

Athens, July 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
Translational Engineering in Health and Medicine

Evaluating the Diagnostic Performance of Large Language Models in Complex Clinical Cases: A Comparative Study

Postgraduate Diploma Thesis

Timotheos Kopsidas

Supervisor
Dr. Konstantina Nikita
Professor in School of Electrical and Computer Engineering
National Technical University of Athens

Co-Supervisors
Antonis Armoundas
Associate Professor of Medicine
Harvard Medical School
Konstantinos Mitsis, PhD
Vasilis Papakonstantinou
Vice Chairman, MIT Enterprise Forum Greece

The postgraduate diploma thesis has been approved by the examination committee on July 4th, 2025.

1st member	2nd member	3rd member
Konstantina Nikita	Antonis Armoundas	Athanasios Voulodimos
Professor in School of Electrical and Computer Engineering, NTUA	*Associate Professor of Medicine Harvard Medical School*	*Assistant Professor in School of Electrical and Computer Engineering, NTUA*

Athens, July 2025

Timotheos Kopsidas

Graduate of the Interdisciplinary Postgraduate Programme,
"Translational Engineering in Health and Medicine",
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

Abstract

This thesis investigates the diagnostic performance of large language models (LLMs) in complex clinical scenarios, focusing on differential diagnosis generation across medical specialties. As LLMs gain attention for clinical decision support, this study offers a structured evaluation of their accuracy and diagnostic efficiency in real-world scenarios.

Eighty clinical cases were drawn from *The New England Journal of Medicine* (NEJM) clinicopathological series. Four LLMs were assessed: two commercial models (GPT-4o and o3-mini) and two open-source models (Qwen-2.5-32B and Qwen-QWQ-32B). Each model received the same case input and system instructions, and generated a primary diagnosis along with a ranked differential list, in zero-shot learning setup. To ensure consistency in scoring, a separate LLM, *Gemini Flash 2.0* was used to verify whether the correct diagnosis appeared in the output, and in which rank.

Evaluation metrics included Top-N accuracy (Top-1, Top-3, Top-5, Top-10) and a novel diagnostic efficiency score that considers both the rank of the correct diagnosis and the total number of suggestions. Comparative and statistical analyses were performed to assess the effects of reasoning capability, temperature variation, and model types (open-source vs. commercial) on the performance.

Results showed that reasoning-enabled models outperformed non-reasoning ones , and commercial models generally surpassed open-source alternatives, though direct comparison was limited by transparency gaps. Among all language models, OpenAI's o3-mini in multiple comparisons, consistently demonstrated the best performance, achieving both high accuracy and focused differential lists. Moreover, temperature had no impact on diagnostic accuracy but improved output consistency. Performance varied across specialties, indicating uneven generalization among models.

This study contributes a reproducible methodology for LLM diagnostic evaluation and highlights the importance of moving beyond single dimension accuracy metrics toward more holistic assessments of clinical reasoning. As LLMs approach clinical relevance, future work should prioritize benchmarking tools that capture reasoning quality, specialty-specific performance, and resource efficiency. Small, domain-adapted models and multimodal capabilities represent promising directions. Overall, while strict accuracy evaluations still offer valuable insight into model progress and advancements, responsible deployment will require deeper integration with clinical workflows and standards.

Keywords

Large language models, clinical diagnosis, differential diagnosis, prompt engineering, top-N accuracy, diagnostic reasoning, human–AI collaboration, evaluation frameworks

Acknowledgments

I would like to express my heartfelt gratitude to my supervising professor and Director of the M.Sc. Program, Konstantina Nikita, for the provided guidance in this assignment and in general for the opportunity to be part of the Program's community for the past 2 years.

Additionally, I want to express deep gratitude for the guidance and the collaboration I had with the Co-Supervisor of the Thesis Project, Mr. Vasilis Papakonstantinou, Dr. Antonis Armoundas and Dr. Konstantinos Mitsis. I am grateful for our collaboration, and their invaluable contribution. The time spent on the project became even more interesting and mind broadening, thanks to them.

Last but not least, I would like to thank my family and my partner Anastasia for their support and encouragement throughout my studies.

Table of Contents

List of Figures

5.2 Retrieved from McDuff, D., Schaekermann, M., Tu, T. et al. work. Comparison of the percentage of DDx lists that included the final diagnosis for AMIE versus GPT-4 for 70 cases. We used Med-PaLM 210, GPT-46 and AMIE as the raters—all resulted in similar trends. Points reflect the mean; shaded areas show ±1 s.d. from the mean across 10 trials.

5.3 Retrieved from HealthBench: Evaluating Large Language Models Towards Improved Human Health by Arora RK, Wei J, Hicks RS, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al HealthBench performance of OpenAI models over time.

Chapter 6

6.1 Transition for Phase 1 to Phase 2. Shifting from quiz questions towards complex real world scenarios. Shifting from one dimension evaluation scores towards advanced frameworks and quality evaluation with rubric criteria.

List of Tables

Key Glossary

Understanding the foundational terms used throughout this Thesis is essential for proceeding with the interpretation of the evaluation of LLMs in differential diagnosis tasks, as presented in the following chapters.

Below the key terms and concepts are presented, so that the reader can read now and/or use them as references during his reading.

- **Artificial Intelligence & Generative Artificial Intelligence**

Artificial Intelligence (AI) refers to computer programs capable of performing tasks mimicking human intelligence using algorithms and statistical models to process information and make decisions, often in real-time. Generative AI refers to the creation of various types of content, including text, images, audio, video, and synthetic data. Artificial General Intelligence (AGI) is when AI reaches a point where it possesses human-like abilities to learn, adapt to new problems, and apply generalized learned knowledge to other contexts similar to an average human [1].

- **Artificial Neural Networks**

Artificial Neural Networks (ANNs) are algorithms with an architecture inspired by the human brain and biological neural networks. ANNs consist of interconnected nodes, or artificial "neurons," forming an input layer, hidden layers for processing information, and an output layer. Neural networks can learn from examples of relationships in data and are capable of complex pattern recognition and decision-making [1].

- **Deep Learning**

Deep Learning (DL) is a subfield of ML that uses deep neural networks (DNNs) for data analysis, pattern recognition, and decision-making, with the distinct capability of autonomously learning crucial features for predictions with minimal human intervention. DNNs are multiple layers of weighted, interconnected nodes trained through supervised or unsupervised learning. The term "deep" refers to the number of layers within the neural network, ranging from several to thousands. DNNs can extract features from raw data using a learning algorithm called "backpropagation". As a result, they have the capacity to learn from mistakes and adjust their course over time to solve problems. DNNs are more powerful than traditional ML algorithms but require large amounts of data and computational resources [1].

- **Differential Diagnosis**

A comprehensive differential diagnosis (DDx) is a cornerstone of medical care that is often reached through an iterative process of interpretation that combines clinical history, physical examination, investigations and procedures [2].

- **Generative models**

Generative models are a type of ML model capable of learning the underlying probability distribution of data and are trained to generate similar, new unseen examples [1].

- **Language models**

A type of generative model that learns to produce new text from text data. Natural language generation (NLG) creates text similar to human-written text. LLMs are neural network-based models with billions of text-based parameters. Similar techniques are used in large multimodal models (LMMs), which can also incorporate additional inputs such as images or audio [1].

In the context of LLMs, the distinction between open-source and commercial models lies primarily in accessibility, transparency, and licensing.

Open-source models are publicly released with their architecture, weights, and training data (or data sources) made available for inspection, modification, and reuse, often under permissive licenses. Examples include *Meta's LLaMA* family, *Mistral*, and *Qwen*. On the other hand, commercial models, such as *GPT-4* by *OpenAI* or *Gemini* by *Google*, are proprietary systems accessed via platforms, with restricted visibility into their internal architecture, training data, fine-tuning processes and other technical and engineering characteristics, that are publicly available for the open-source ones.

Another distinction between reasoning and non-reasoning models refers to the model's ability to perform based on intermediate logical steps when generating responses to improve the accuracy and applicability of answers generated [3].

- **Machine Learning**

Machine Learning (ML) is a branch of AI that enables computer programs to learn from data by identifying patterns and improving performance through experience. ML often requires large datasets and benefits from iterative feedback and fine-tuning [1].

- **Model Temperature**

Temperature is a hyperparameter of LLMs, which is a factor that affects the randomness and originality of the LLMs' output. Lower temperature settings are associated with more prototypical and standard outputs, while higher temperature settings are associated with more creative and less predictable responses. Preferences for different temperature settings may be intuitive for certain use cases. For instance, for creative writing, one might prioritize higher temperature settings. For healthcare, however, it is not necessarily straightforward which setting will be most effective, and it may be that different clinical tasks for LLMs may require different settings [4].

- **Natural Language Processing**

Natural Language Processing (NLP) is a field of AI that analyzes and processes free text into structured language data. NLP tasks include, but are not limited to, translation, semantic analysis, automatic summarization, question-answering, and speech recognition [1].

- **Prompt Engineering**

The structured design of inputs to guide LLM behavior and performance on content generation [5].

- **Recurrent Neural Networks**

Recurrent neural networks (RNNs) are networks designed to process sequential data with temporal dependencies, such as time series or natural language, by utilizing internal loops that allow information to persist, essentially enabling a "memory." This enables the networks to make decisions based on the input sequence and previous context rather than just each input independently. The recurrent connections allow RNNs to develop complex temporal representations critical for sequence modeling tasks such as language translation, speech recognition, and time series forecasting. However, RNNs often struggle with capturing long-distance dependencies, meaning they may have difficulty relating information from earlier steps in the sequence when the gaps are too large [1].

- **Reinforcement Learning**

Reinforcement Learning (RL) is a type of unsupervised learning where the model learns to make decisions by receiving feedback as rewards or penalties in a problem-oriented environment. The goal is to find the optimal sequence of actions that maximizes results. RL uniquely maximizes reward signals instead of finding hidden structures like traditional ML models. Reinforcement learning from human feedback (RLHF) incorporates human feedback during the training process, known as "alignment" [1].

- **Tokenization**

Tokenization is an essential pre-processing step in LLM training that parses the text into non-decomposing units called tokens. Tokens can be characters, subwords, symbols, or words, depending on the tokenization process [6].

- **Transformer**

Transformer is a type of RNN architecture that employs self-attention, allowing it to focus on different parts of the input sequence simultaneously. This enables transformers to efficiently capture intricate relationships in input data and leverage parallelization, a technique that divides tasks into smaller subtasks executed concurrently across multiple processing units. This is particularly effective for NLP tasks involving lengthy sequences of speech/text [1].

- **Zero-Shot Learning**

Zero-shot learning is a machine learning scenario in which an AI model is trained to recognize and categorize objects or concepts without having seen any examples of those categories or concepts beforehand [7].

1 Introduction

This chapter introduces the reader to context and the environment around which this study has been designed and implemented. The research aim of this work, its significance and how this Thesis project will evolve in the following chapters will be presented. This introduction setup is crucial before moving onto understanding the status and findings of current research.

1.1 Background and Context

Artificial intelligence (AI) in medicine has evolved significantly in recent years, shifting from traditional machine learning algorithms focused on structured data analysis towards the integration of language models and large language models (LLMs). Initially, AI applications in healthcare relied on algorithms designed to analyze specific datasets for tasks like image recognition or predictive modeling. However, the rise of LLMs has opened new possibilities by enabling machines to process and understand unstructured text data, such as medical records, research papers, and patient dialogues [8]. In Fig. 1.1 it is displayed how much these topics have rapidly increased during the latest years in research publications in PubMed (https://pubmed.ncbi.nlm.nih.gov/), a searchable database provided by the National Library of Medicine (https://www.nlm.nih.gov/). This shift allows AI to engage in more complex tasks, including diagnostic reasoning, DDx generation, and patient communication, thus expanding its role from assisting with specific tasks to potentially augmenting broader aspects of clinical decision-making and patient care.

The emergence of Generative AI and LLMs has revolutionized natural language understanding and generation, enabling machines to perform complex cognitive tasks across domains. In the medical domain these models have shown potential to assist with clinical cases and diagnostic reasoning. The capacity of LLMs to generate structured differential diagnoses and synthesize case findings raises the possibility of using them as decision support tools in clinical practice, aimed to assist and support physicians' work.

Yet despite their promise, LLMs are not yet systematically validated in terms of diagnostic accuracy, generalizability across specialties, and their ability to reason in complex, real-world scenarios. Most prior work has focused on multiple question-answering tasks or benchmark datasets that may not reflect the complexity and ambiguity of actual clinical reasoning. Moreover, there is limited comparative analysis across different model types and parameters (e.g., reasoning-enabled vs non reasoning) and even less attention to how LLM performance ranges across medical specialties

This Thesis works on and presents an established way of evaluating how modern LLMs perform when tasked with solving complex, non trivial to the point of a multiple choice question diagnostic cases. The cases used are drawn from *The New England Journal of Medicine* (NEJM). Through a structured comparative framework, it aims to uncover the

strengths, limitations, and behavioral patterns of 4 representative models across a set of 80 diverse clinical cases.

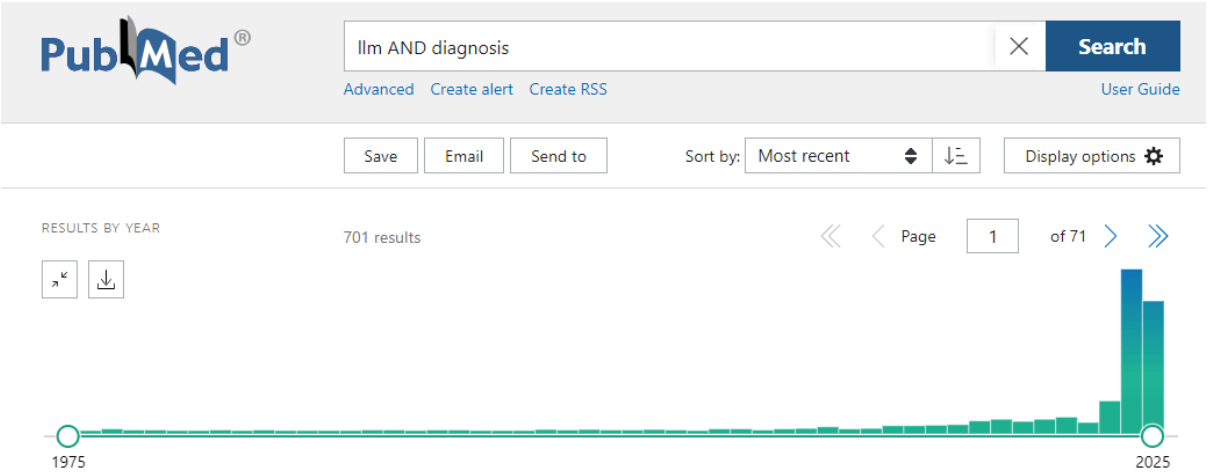

Figure 1.1: Trend of research publications on the topic of LLMs for diagnosis.

1.2 Research Aim and Objectives

The primary aim of this study is to evaluate the diagnostic performance of LLMs in complex clinical cases and to analyze how that performance varies across model architectures and clinical specialties. Performing this comparison in a systematic way, this study is not only aiming to distinguish engineering characteristics, parameters and language models that perform better than others, but also present a framework that can further enhance the establishment of systematic ways and metrics to compare models in future works.

To accomplish this, the study sets out the following objectives:

- To compare 4 LLMs — including both commercial (GPT) and open-source (Qwen) models — on their ability to generate accurate diagnoses across 80 real-world cases.
- To evaluate model performance, in regards to their accuracy and diagnostic focus scores as defined in later chapters.
- To examine the effect of model type (reasoning vs non-reasoning, commercial vs open-source) on diagnostic performance.
- To assess how the models' performance varies across different medical specialties.
- To explore the influence of prompting temperature on model behavior.
- To identify methodological limitations and propose recommendations for future diagnostic AI benchmarking.

1.3 Significance of the Study

This study contributes to the growing field of AI in healthcare by providing a rigorous, transparent, and domain-aware framework for evaluating LLMs in clinical diagnostic tasks. Its significance lies in several key areas, presented below:

- **Benchmarking across specialties:** Unlike prior studies limited to general QA formats or synthetic datasets, this research evaluates LLMs on specialty-specific, real-world cases that reflect clinical complexity and diagnostic ambiguity.
- **Comparative design:** By including both open-source and commercial models, and reasoning-enhanced vs baseline variants, the study sheds light on which architectural and functional features drive diagnostic performance.
- **Efficiency-focused evaluation:** Through the use of a novel efficiency metric, the research emphasizes diagnostic parsimony and focus — a key principle in real-world clinical decision-making.
- **Feasibility insights:** The study lays the groundwork for evaluating not just accuracy, but also computational realism, prompting strategies, and the potential for future integration into clinical workflows.

This thesis aims to inform future research directions in the crossing of the diagnostic domain of medicine and Generative AI and contribute towards the development of universal evaluation standards.

1.4 Structure of the Thesis

The remainder of this Thesis is organized in different chapters, that mainly present an introduction to the topic, the main aspects of the current research through several published works, the methodology framework that we implemented, its results and their interpretation. Finally, there is a discussion in regards to the future opportunities and challenges of the field, accompanied by a conclusion chapter.

In particular, these chapters are the following:

Chapter 2: Literature Review

Introduces reviews relevant work on LLMs in clinical diagnosis, and identifies gaps in current research and evaluation frameworks.

Chapter 3: Methodology

Details the dataset, language models, prompting strategies, scoring and statistical methods used in the analysis.

Chapter 4: Results

Presents the quantitative findings, including Top-N accuracy, efficiency scores, and subgroup analyses by model type, specialty, and prompt temperature.

Chapter 5: Discussion

Interprets the findings in light of existing research, outlines the strengths and limitations of the study, and proposes directions for future work.

Chapter 6: Conclusion

Summarizes the study's main contributions and ponders its broader implications.

2 Literature Review

This chapter provides a foundational overview of some key concepts, and prior research relevant to the evaluation of LLMs in healthcare and clinical diagnostics. It introduces the theoretical basis for diagnostic reasoning, outlines common performance evaluation metrics such as Top-N accuracy, and critically reviews the existing literature on LLMs in healthcare and DDx.

2.1 Theoretical Foundations of Diagnostic Reasoning

Before exploring the progress and applications of generative AI in clinical diagnosis, it is important to first establish common ground regarding the fundamental principles of diagnostic reasoning.

Understanding the complex nature of human diagnostic reasoning is fundamental to evaluating the integration of LLMs into clinical practice. The field of internal medicine, characterized by often blurred diagnostic boundaries, is an example of the flexible and adaptive discipline required for accurate diagnoses. Diagnostic reasoning in this context is an iterative, multi-stage process, critically dependent on continuous data collection, the generation of plausible illness scripts, and rigorous testing of diagnostic hypotheses [9].

Clinical practice is inherently challenging, in contrast to the controlled environment of laboratory research where only one variable is manipulated. In patient care, clinicians must simultaneously manage numerous interconnected factors and parameters, with many degrees of freedom. Unlike other natural sciences, clinical medicine operates with less certainty and greater variability, relying on fuzzy fields, statistical probability rather than definitive mathematical analysis. Diagnosis is a cornerstone of clinical medicine, making the principles of diagnostic reasoning a vital component of medical education. Physicians employ both analytical and intuitive approaches to clinical reasoning, with the latter being more prevalent among experienced practitioners. However, this intuitive approach, while efficient, is susceptible to cognitive biases, necessitating constant critical thinking and the application of de-biasing techniques to ensure diagnostic accuracy [9].

The necessity of formal instruction in clinical/diagnostic reasoning remains a point of debate. Some argue that this skill can only be acquired through direct patient interaction, independent of educators. However, experienced teachers play a crucial role in helping students recognize conceptual and causal links between seemingly disparate observations. They also foster metacognition, the conscious self-monitoring of one's thinking, by encouraging and guiding reflective feedback on students' thought processes. This process of open reflection and feedback is central to *deliberate practice*, a key theory in expertise development and maintenance particularly relevant to internal medicine. While deliberate practice is significant, it is not the sole determinant of expert performance. Individual capabilities, such as working memory capacity, which involves the efficient storage and retrieval of knowledge, are equally important [9].

In total, the in-depth understanding of human diagnostic processes and activities provides an essential benchmark against which the performance and reasoning mechanisms of AI systems can be critically assessed.

2.2 Evaluation Frameworks in Diagnostic AI

The systematic evaluation of automated diagnostic tools has a long history, with DDx generators representing an early iteration of computer programs designed to aid in clinical reasoning.

Back in 2012, a foundational study [10] sought to establish clear evaluation criteria for these systems, identifying key features such as input methods, filtering capabilities, consideration of various diagnostic factors (e.g., lab values, medications, demographics), and the inclusion of evidence-based medicine content. By applying these consensus criteria to a selection of four prominent DDx generators—Isabel, *DxPlain*, *Diagnosis Pro*, and *PEPID*—and testing their performance against challenging cases from the New England Journal of Medicine and Medical Knowledge Self Assessment Program (MKSAP), the research provided crucial insights into their effectiveness. The findings, which indicated mean scores for Isabel and DxPlain at 3.45 (on a 5-point scale) while *Diagnosis Pro* and *PEPID* performed lower, underscored the varying capabilities of these early systems and highlighted the importance of robust frameworks for assessing diagnostic AI**.**

This historical context is essential for understanding the evolution of evaluation methodologies and for designing comprehensive benchmarks for modern Generative AI and LLMs in complex clinical cases.

Recent advances in domain-specific LLMs have focused not only on model architecture but also on how clinical diagnostic performance is evaluated. *MedFound*, a 176B-parameter generalist medical LLM, demonstrated high diagnostic accuracy across both common and rare disease categories through an evaluation framework comprising eight clinical metrics [11]. This framework included not only diagnostic correctness but also medical reasoning, summarization, and risk management, offering a broader lens through which to assess real-world clinical utility. Such comprehensive evaluation strategies support a shift from static Top-N scoring toward context-aware benchmarking in diagnostic AI.

In addition, a notable study by Cabral et al. (2024) [12] directly compared the clinical reasoning abilities of GPT-4 with those of internal medicine residents and attending physicians using 20 structured case vignettes. Using the Revised-IDEA (R-IDEA) scoring system — a validated rubric for evaluating reasoning quality in clinical documentation — the LLM significantly outperformed both resident and attending cohorts in the synthesis of case information and problem representation. While diagnostic accuracy was similar across groups, GPT-4 demonstrated a slightly higher frequency of incorrect reasoning segments, highlighting the importance of evaluating not just outcomes but process quality. The study supports and verifies that LLMs must be assessed using multi-dimensional clinical reasoning metrics, not only accuracy-based benchmarks.

As LLMs are increasingly deployed in clinical and diagnostic settings, ensuring rigorous and transparent reporting has become a priority. The DEAL checklist (Development, Evaluation, Assessment of LLMs), recently proposed by Wang et al., offers a structured framework for documenting LLM research across two tracks: development-focused studies (DEAL-A) and applied evaluations (DEAL-B) [13]. Covering essential aspects such as model specifications, data usage, evaluation metrics, and transparency standards, the checklist is designed to promote reproducibility and methodological clarity for future work as well. This aligns with the goals of the present study, which applies a consistent and open evaluation framework across multiple LLMs and diagnostic domains.

As the capabilities of LLMs expand within healthcare, the development of robust and realistic evaluation frameworks becomes paramount to accurately assess their diagnostic performance and safety. As a result, more advanced evaluation frameworks suggest the rubric approach. One great example of this type of evaluation is the recently released HealthBench [14].

HealthBench offers an open-source benchmark specifically designed for multi-turn conversational evaluation in healthcare settings. Unlike traditional multiple-choice or short-answer benchmarks, HealthBench leverages 5,000 multi-turn conversations, with responses meticulously evaluated by 262 physicians using 48,562 unique rubric criteria that cover diverse health contexts—ranging from emergencies to global health—and critical behavioral dimensions such as accuracy, completeness, instruction following, and communication. This comprehensive approach allows for a more realistic and open-ended assessment of LLM performance, grounding progress in model development towards applications that genuinely benefit human health. The benchmark has already reflected steady initial progress, with notable improvements in newer and even smaller, more cost-effective models.

Rubric evaluation that is applied in HealthBench involves grading model responses based on conversation-specific criteria with point values from -10 to 10. Positive scores reward desired attributes, while negative scores penalize undesirable ones. Graders assess each criterion independently, awarding the points. The total score is the sum of points for met criteria, divided by the maximum possible score. As a result, a model's overall HealthBench score is the mean of its per-example scores, clipped to the range of 0 to 1.

According to the results, performance on HealthBench has shown steady initial progress (GPT-3.5 Turbo at 16% to GPT-4o at 32%) and more rapid recent improvements (o3 scores 60%). Smaller models have also significantly improved, with GPT-4.1 nano outperforming GPT-4o while being 25 times cheaper.

2.3 Strengths and Limitations of Existing Approaches

Existing approaches cover a wide range of evaluation frameworks designed, implemented and applied, language models assessed and methodologies followed. As a result, through these studies, it's valuable to overview their strengths that further advance the field or the limitations that are inevitably being introduced or not being challenged with the current State of the Art research.

2.3.1 Strengths

- **Promising potential**

LLMs demonstrate promising potential in various clinical applications, including clinical documentation, trial design, and diagnostic support, especially when integrated responsibly within healthcare frameworks [14].

- **Improving performance**

There have been numerous models and studies that have shown improving performance scores and evaluations. One of the most advanced and renown diagnostic LLMs to date, *Med-PaLM 2*, demonstrated state-of-the-art performance on multiple-choice and long-form medical question answering benchmarks [15]. The model was developed through domain-specific fine-tuning and tested against multiple human evaluators across several axes of clinical reasoning, including factuality, medical logic, and risk of harm. Notably, Med-PaLM 2 responses were often preferred over those of generalist physicians and, in some cases, even specialists. Moreover, a mid-scale evaluation of GPT-4's diagnostic performance on 70 NEJM-style clinicopathological cases found that the model listed the correct diagnosis in its differential 64% of the time, and ranked it first in 39% of cases [16]. The study also introduced a DDx quality score, with GPT-4 achieving a mean score of 4.2 out of 5—comparable or superior to existing DDx generators.

- **Range of applications**

The increasing capabilities of LLMs have led to their exploration in high-stakes clinical settings like Emergency Department (ED) triage, prompting a comparative study to assess their proficiency against human personnel [17]. This research evaluated the triage performance of various LLMs, including *GPT-4* based *ChatGPT*, *Llama 3 70B*, *Gemini 1.5*, and *Mixtral 8x7b*, using 124 anonymized case vignettes against a gold standard set by professional raters. The findings revealed that the best LLM models, specifically GPT-4-based ChatGPT, demonstrated substantial agreement with professional triage, performing comparably to untrained ED doctors.

2.3.2 Limitations

- **Methodology and framework**

Numerous researchers have pointed out methodology limitations despite the apparent continuous performance improvements [15,16,18,19]. Current assessment rubrics lack formal validation, and evaluations often conflate length or verbosity with quality. Most current diagnostic AI benchmarks remain constrained to multiple-choice or single-answer formats, which do not reflect the complex, layered reasoning required in clinical practice. These studies echoe a growing consensus that traditional QA benchmarks are saturated, reductive, and ill-suited for real-world deployment, and that the field urgently needs dynamic, context-rich evaluation frameworks to accurately measure diagnostic utility.

- **Regulations**

Despite their rapid advancement, existing approaches to evaluating and integrating generative AI in clinical contexts often lack the structural safeguards necessary for responsible adoption. A recent multidisciplinary review led by the Duke Clinical Research Institute highlighted that while LLMs hold promise in improving clinical documentation, trial design, and diagnostic support, their safe deployment is hindered by regulatory uncertainty, inconsistent evaluation standards, and insufficient stakeholder alignment [19]. This underscores the importance of developing transparent, reproducible, and clinically grounded frameworks, not only to assess LLM accuracy but also to ensure ethical integrity and data security.

- **Specialty specific**

LLMs applications shouldn't not just cover the wide range of clinical reasoning, but also start to become specialty-specific, possibly by taking advantage of smaller language models [19]. Additionally, it has been suggested that there is a high need for subspecialty-aware evaluation frameworks and more granular performance audits before LLMs can be responsibly applied in specialty-specific diagnostic support [20]. In detail, LLMs' proficiency in specialized medical subfields, such as nephrology, requires distinct evaluation to understand their precise strengths and limitations for clinical applications. A comparative study specifically investigated the medical knowledge of various LLMs by challenging them with 858 multiple-choice questions from the Nephrology Self-Assessment Program (nephSAP) . The results highlighted a significant performance disparity between open-source models (*Llama2-70B*, *Koala 7B*, *Falcon 7B*, *Stable-Vicuna 13B*, and *Orca-Mini 13B*), which scored between 17.1% and 30.6%, and proprietary models like *Claude 2* (54.4%) and *GPT-4* (73.3%).

- **Potential gaps between open-source and leading proprietary models**

Even though it's not feasible to compare commercial models that usually have their engineering specifications hidden from the public with open-source ones, it has been pointed out [20] that there are knowledge gaps between them. While leading proprietary models demonstrate considerable competence, the general landscape of LLMs still presents limitations relevant to their effective integration into specialized medical training and patient care.

- **Over-triage and under-triage**

As for the performance on triage tasks, there is a significant limitation, their inability to substantially improve untrained doctors' triage performance when used as a second opinion, and a consistent tendency of LLMs toward overtriage was observed, contrasting with the undertriage by untrained doctors. However, despite these current limitations in achieving gold-standard performance, there has been considerable performance enhancements in newer LLM versions, hinting at their evolving strengths and future potential in this critical medical domain [17].

- **Resources and costs**

The escalating scale of LLMs has introduced significant computational and memory challenges, directly impacting their accessibility and broad application in specialized fields like medical diagnostics. Achieving peak performance for these models often necessitates an extreme number of parameters, pushing them into the trillion-parameter range and thereby increasing resource demands. Addressing these limitations, research efforts have primarily focused on two strategic approaches: fine-tuning pre-trained models to attain state-of-the-art results for specific tasks, and developing methods to reduce operational costs or accelerate training without compromising accuracy. A systematic review [21] of 65 publications from 2017 to December 2023 highlights various optimization and acceleration strategies across LLM training, inference, and system serving, demonstrating practical methods to mitigate resource constraints while maintaining cutting-edge performance. These advancements are crucial for overcoming the practical barriers to deploying powerful LLMs in complex clinical scenarios, ultimately enhancing their utility and accessibility.

- **Overestimation of performance**

Additionally, a recent systematic review and meta-analysis of 83 studies spanning from 2018 to 2024 offers a broad evaluation of generative AI diagnostic performance compared to physicians [22]. The study reported a mean diagnostic accuracy of 52.1% for generative models, finding no statistically significant difference in performance compared to physicians overall or non-expert physicians. However, generative models performed significantly worse than expert clinicians ($p = 0.007$). These results underscore the current limitations in AI reliability for high-stakes diagnosis, especially in expert-level scenarios. While the findings support the use of generative AI for education and preliminary diagnostic support, they also highlight the risk of overestimating model readiness in the absence of rigorous, clinically grounded validation.

2.4 Human-AI Collaboration in Diagnostic Reasoning

The collaboration of human physicians and clinicians with AI is being studied in numerous research works during the latest years, to investigate whether this collaboration improves diagnostic performance, accuracy and reasoning, and also to perform comparisons between standalone performances.

A randomized clinical vignette study by DeFilippis et al. [23] investigated whether access to GPT-4 improves diagnostic reasoning among physicians compared to conventional resources. According to it, while GPT-4 alone outperformed both residents and attendings in diagnostic accuracy, its use as a decision-support tool did not significantly enhance physician performance or reduce diagnostic time. This suggests that simply introducing LLMs into clinical reasoning workflows may not yield synergistic benefits without improved interaction paradigms. The study highlights a limitation in current AI integration strategies and reinforces the need for workflow-sensitive evaluation frameworks.

Moreover, there is also the interaction of patients with LLMs. The physician-patient dialogue, with its emphasis on skillful history-taking, forms the solid ground of accurate

medical diagnosis and patient management, making the development of AI systems capable of diagnostic conversation a challenging field for increasing healthcare accessibility and quality. Addressing this challenge, a notable advancement is the introduction of AMIE (Articulate Medical Intelligence Explorer), an LLM-based AI system specifically optimized for diagnostic dialogue [24]. AMIE's innovative approach involves a self-play-based simulated environment coupled with automated feedback, enabling scalable learning across diverse disease conditions and clinical contexts. Evaluated against primary care physicians in a rigorous randomized, double-blind crossover study using text-based consultations with patient-actors, AMIE demonstrated superior diagnostic accuracy and outperformed human clinicians across a substantial majority of clinically meaningful performance axes as judged by both specialist physicians and patient-actors. While acknowledging the methodological limitation of using text-based chat, which is not yet typical in clinical practice, this research marks a significant milestone in advancing conversational diagnostic AI and profoundly informs the potential future of human-AI collaboration in complex clinical reasoning scenarios.

Furthermore, AMIE was also evaluated as part of LLM-clinician collaboration. Across 302 complex real-world medical cases, clinicians who received assistance from AMIE produced more complete differential diagnoses and demonstrated significantly higher Top-10 accuracy compared to those using conventional search tools or working unassisted [2]. The model also outperformed physicians when operating independently. These findings suggest that well-designed LLMs may offer a meaningful augmentation to physician diagnostic reasoning, particularly in high-complexity clinical contexts.

As LLMs are increasingly integrated into complex natural language processing tasks, particularly in classification contexts relevant to diagnostic reasoning, understanding the nuances of their operational parameters becomes critical. A recent study [25] specifically investigated the impact of 'temperature'—a key parameter controlling response randomness and creativity—on LLM performance in classification tasks, using Word Sense Disambiguation as a case study. Unlike previous explorations focused on text generation, this research highlighted that temperature significantly affects the accuracy of LLMs in classification scenarios, underscoring the necessity of a preliminary study to identify the optimal temperature setting for specific tasks. The findings further revealed varying degrees of performance consistency across different models, with GPT-3.5-Turbo and Llama-3.1-70B exhibiting notable performance shifts, while GPT-4-Turbo and Llama-3-70B demonstrated more stable results across different temperature settings. This insight is vital for understanding the methodological considerations and potential limitations when deploying LLMs for sensitive classification tasks in complex clinical environments.

Finally, prompt engineering has emerged as a key determinant of model performance across domains, including healthcare, pointing out how human-AI collaboration can further evolve and prosper. Recent reviews have highlighted some techniques for enhancing diagnostic inference, answer reliability, and robustness to adversarial prompts [5]. These techniques include simple ones, such as enabling role-playing for the model, giving it a specific role to play, such as a helpful assistant or a knowledgeable expert, or providing clear and specific guidelines. Other, more advanced ones are chain-of-thought

prompting (i.e. asking for *step-by-step* analyses) and generated-thought, where especially for commonsense reasoning, it allows the model to generate and utilize additional context that may not be explicitly present in the initial prompt.

2.5 Explainability and Transparency in LLM-driven Diagnosis

Successful integration of AI in medicine requires emulating doctors' reasoning for trust and better care, not just high accuracy. Automatic diagnosis should mirror clinical interaction: generating DDx, prioritizing severe conditions via exploration-confirmation, and explaining reasoning. While LLMs show diagnostic accuracy, their "black box" nature and lack of clear, interpretable DDx explanations hinder clinical adoption and trust.

To address this, a novel approach involves the development of tailored evaluation datasets and innovative methodologies specifically designed to elicit high-quality DDx explanations from LLMs. One such pioneering effort has led to the creation of the first publicly available DDx dataset, comprising expert-derived explanations for 570 clinical notes over nine distinct clinical specialties, called Open-XDDx [26]. This dataset copes with the absence of specialized evaluation datasets, since a significant barrier has been the pervasive lack of publicly available DDx datasets that are specifically annotated with detailed diagnostic explanations. This scarcity severely constrains the ability to develop, train, and rigorously evaluate models designed to generate such explanations. This dataset exhibits several key characteristics, like the extended size, clinical diversity, the data availability for further research and rich annotations for these notes too.

Alongside this dataset, there is a proposition of a novel framework, Dual-Inf, engineered to effectively harness LLMs for precise DDx explanation [26]. The fundamental design of Dual-Inf draws inspiration from the human diagnostic process of backward verification. Just as clinicians might reason forward from symptoms to formulate initial diagnoses and then reason backward from these potential diagnoses to confirm associated symptoms or findings, Dual-Inf enables bidirectional inference. This bidirectional approach is intended to enhance prediction correctness and the overall quality of explanations.

Moreover, it has been suggested that Deep Reinforcement Learning (DRL) frameworks, a field that combines DL and RL enhance the performance on explainability [27]. This paper proposes an innovative solution through a novel DRL framework, which incorporates an agent named CASANDE. In this case, DRL is used as it is particularly well-suited for this challenge. It excels in sequential decision-making problems where an intelligent agent learns optimal actions through iterative trial and error within a dynamic environment, receiving rewards for desirable outcomes. This paradigm closely mirrors the iterative and adaptive nature of human medical diagnosis, where doctors gather information, make preliminary assessments, and refine their understanding based on new data. This framework is designed to integrate three essential aspects of a doctor's reasoning: the generation of a DDx, an adaptive exploration-confirmation approach to gathering medical evidence, and the explicit prioritization of severe pathologies during the diagnostic process.

This focus on explicitly evaluating and enhancing the explainability of LLMs represents a crucial step towards bridging a critical gap in automated DDx, fostering greater transparency, and ultimately enhancing human clinical decision-making by moving beyond mere accuracy to verifiable understanding.

2.6 Ethical Dimensions and the Role of Patient Autonomy in LLM Deployment

As seen in previous sections, Human and AI collaboration and communication is not only about the dimension of clinicians and physicians. Evaluation frameworks and latest research work is also about the interaction of LLMs with the patient side.

Recent literature has begun to explore the broader societal implications of LLMs in medicine. For example, *Armoundas, A.A. and Loscalzo, J.* highlight and investigate the impact of LLMs on what is defined as patent agency; the patient's capacity to engage efficiently with, act on, and assume responsibility for their state of health. As noted, there is a critical shift in the way patients engage with these technologies, moving from clinician-facing tools to systems that increasingly influence patient-led health decisions [28].

This transformation raises pressing questions about autonomy, equity, and global disparities in access and understanding. This shift enhances precision medicine and the accessibility to medical care. Furthermore, the individuals' engagement with their health under improved health literacy that usually follows digital literacy helps achieve better understanding of their disease, the available types of treatment and individual decision-making in general.

3 Methodology

This chapter outlines the framework applied to evaluate the diagnostic performance of LLMs on complex clinical cases. The goal is to systematically compare the capabilities of 4 LLMs of different characteristics each, when applied to real-world diagnostic tasks drawn from a reliable, reputable and filled with challenging cases medical source, such as the NEJM.

3.1 Clinical Case Dataset

This section is about the presentation and analysis of the dataset used for the LLMs' evaluation, in terms of source and content too.

3.1.1 Source and Selection Criteria

The dataset used in this study consists of 80 clinical case reports sourced from the *Case Records of the Massachusetts General Hospital* series, published in the NEJM. The complete list of these cases used can be found online [29].

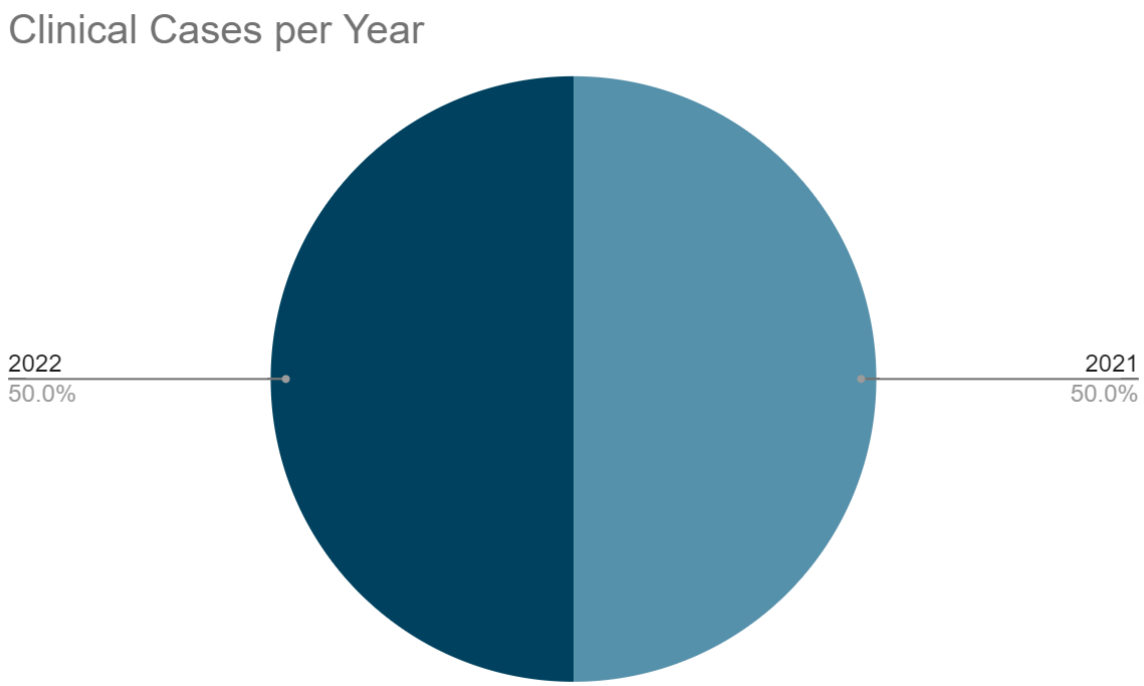

Figure 3.1: Clinical cases per year.

The dataset of cases belongs to the 2-years timespan of January 2021 to December 2022 (Fig. 3.1). This subset has been successfully used in other studies [16,18]. As of May 2025, there are 7119 clinical cases available, being dated from 1923 up to 2025. Other than these cases, NEJM offers a wide range of material that can be used by medical professionals or nowadays for scientific research in the biomedical field. *Images in Clinical*

Medicine, *Hospital Reports* and *Clinical Department* are namely just some of the data categories available.

As for the *Case Records of the Massachusetts General Hospital* series, these are widely regarded as gold-standard diagnostic exercises in internal medicine, featuring complex, real-world patient presentations and multidisciplinary clinical reasoning since the 1950s [18]. Cases were selected from publicly accessible NEJM archives, with the inclusion criteria being the presence of a clear final diagnosis for each case, and also the presence of it in other research studies so that any comparisons among them would be performed on the same data.

One important factor that was taken into account for exclusion of data would be the possible presence of these cases in the training dataset of some models. However, it has been shown that there is no significant difference in performance of LLMs before and after the pre-training cutoff date [18,20], pointing out that these cases were not explicitly included in the training data. Still, this is noted down as a potential limitation of this research study. The only way to cope with it, is to choose cases that are released after the cut off date of each model.

3.1.2 Specialty Distribution

According to NEJM, each case belongs to one or more medical specialties, so these labels serve like tags for each clinical case.

Specialties represented include neurology, infectious disease, rheumatology, cardiology, oncology, gastroenterology, and others. Some specialties are more common among others in the 2-year span that the 80 cases were spread. However, this dataset provides 5 categories present in at least 25 cases. Table 3.1 and Fig. 3.2 display all the medical specialties present in at least one case of the 80 available.

Medical Specialty	Count	Medical Specialty	Count
Hematology/Oncology	39	Endocrinology	9
Infectious Disease	35	Ophthalmology	7
Surgery	33	Nephrology	7
Emergency Medicine	30	Psychiatry	7
Neurology/Neurosurgery	26	Geriatrics/Aging	4
Rheumatology	20	Otolaryngology	4
Gastroenterology	17	Obstetrics/Gynecology	3
Pulmonary/Critical Care	16	Orthopedics	2
Genetics	15	Urology/Prostate Disease	1

Cardiology	13	Medical Ethics	1
Allergy/Immunology	10	Radiology	1
Pediatrics	10	Public Health	1
Dermatology	9		

Table 3.1: Medical specialties distribution in the *Case Records of the Massachusetts General Hospital* dataset.

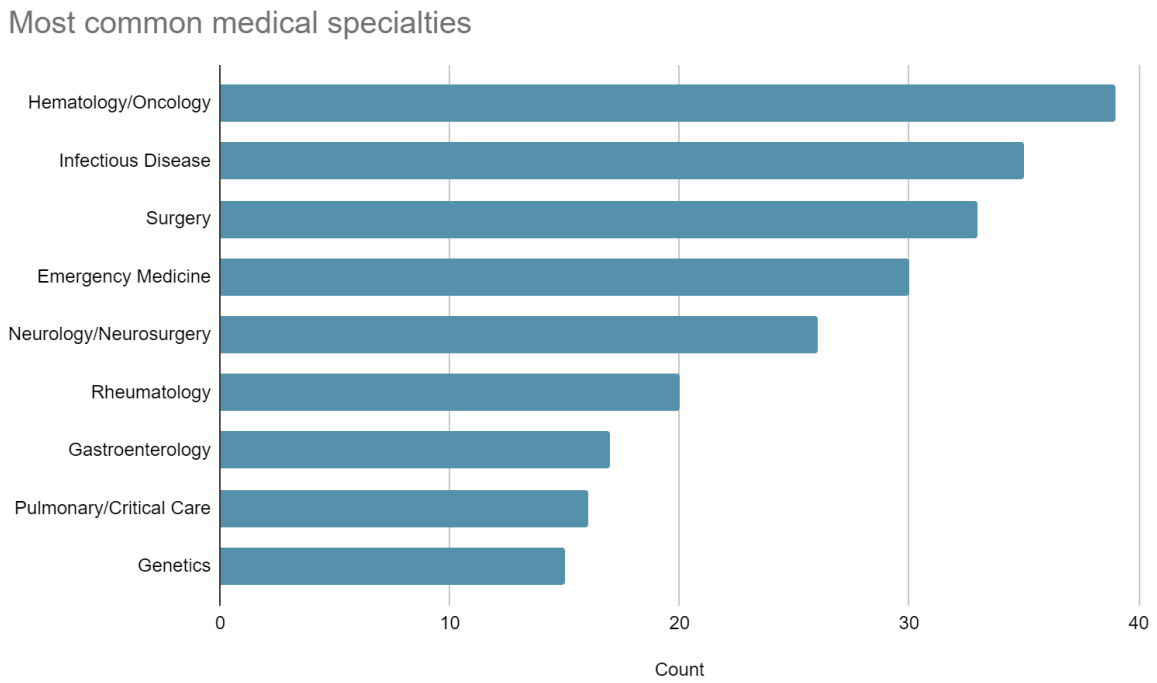

Figure 3.2: Most common medical specialties in the NEJM dataset used (Years 2021-2022).

So, the distribution of clinical specialties in the dataset is not uniform, with a higher representation of cases in domains such as *Hematology/Oncology* and *Infectious Disease*, and relatively few cases from specialties like *Radiology*. As a result, we will present performance comparisons only for the top 5 medical specialties, setting a threshold of 25 appearances, to find potential areas where LLMs perform their best or worst.

3.2 Overview of Research Design

This study adopts a comparative research design aimed at evaluating the diagnostic performance of LLMs across complex clinical cases. The investigation is structured as a performance benchmark using real-world clinical scenarios. Specifically, the study examines how different LLMs—varying in several characteristics like being reasoning-enabled or not and their temperature, perform in the task of DDx generation in general, and across several medical specialties too.

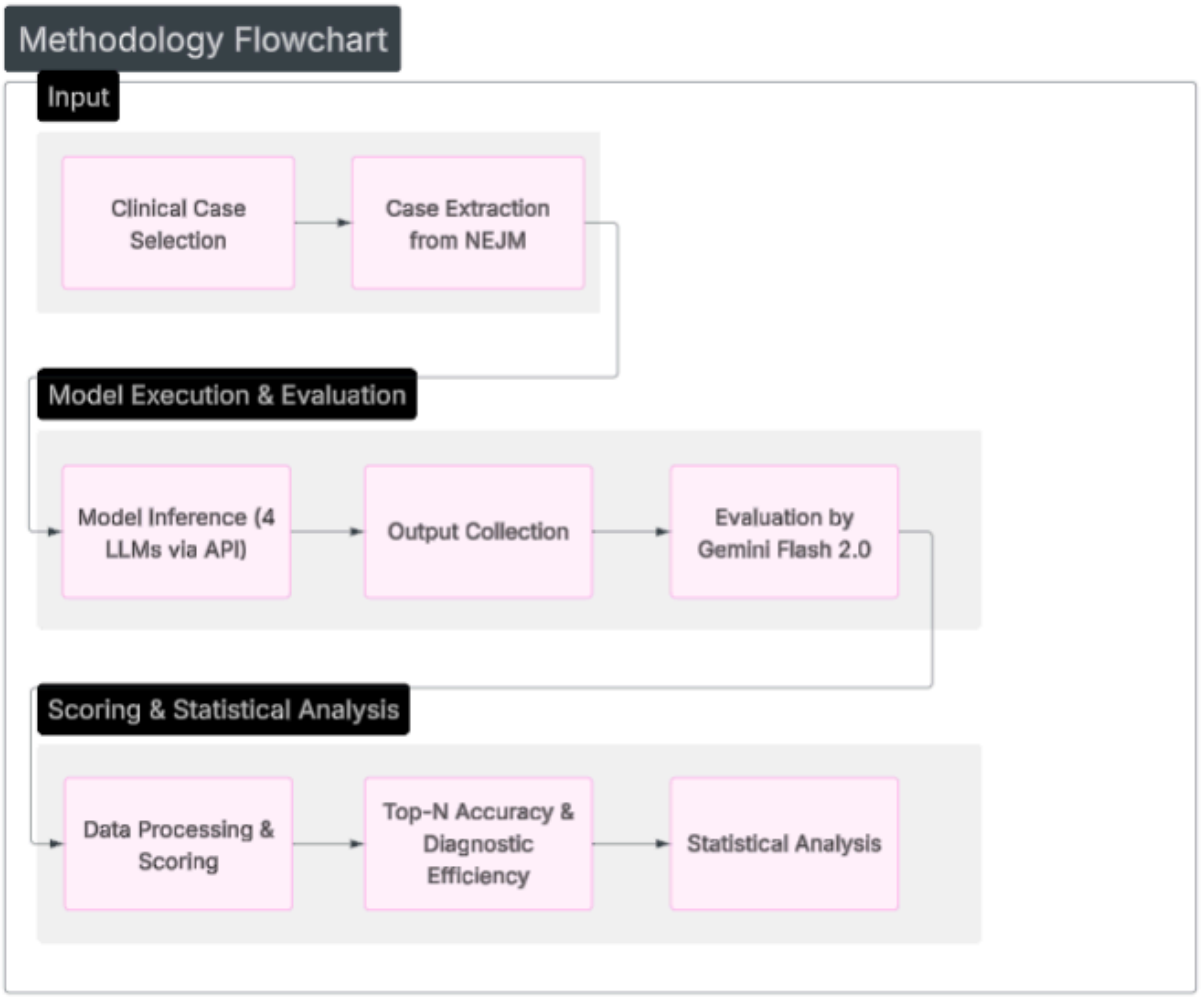

Figure 3.3: Flowchart that presents the overview of the evaluation framework implemented.

The design involves 4 LLMs, each tasked with providing diagnostic predictions for a dataset of 80 clinical case reports published in the NEJM. Each model processes the same set of case inputs, ensuring a within-subjects comparison that controls for variability in clinical content. The diagnostic outputs are subsequently evaluated on their accuracy using standardized metrics, with additional analyses addressing diagnostic focus and performance variation across several dimensions.

To ensure fair and consistent evaluation across models, all diagnostic outputs were assessed against the reference diagnoses provided at the conclusion of each NEJM case. To mitigate semantic variability (e.g., differences in terminology, contextual synonyms such as “tuberculous meningitis” vs. “central nervous system tuberculosis”), the correctness of each model’s response was reviewed by a secondary LLM (*Gemini 2.0 Flash*), to take into account medical synonymy and conceptual equivalence. This methodological choice was made to reduce evaluation bias introduced by non-expert human medical knowledge limitations, and automate the stage of the evaluation too. An alternative, more advanced approach with potential for more advanced capabilities would be to assign the evaluation to physicians. Both approaches have already been successfully applied in research [2,16,18].

Statistical analyses are conducted to determine whether observed performance differences across model types are statistically significant. Finally, the study also considers the ethical implications of using LLMs in clinical decision support contexts, particularly in reliability, transparency, and replicability.

An overview of the workflow followed with this evaluation framework is displayed in the flowchart of Fig. 3.3.

3.3 Language Models Evaluated

This study evaluates the diagnostic performance of 4 LLMs, selected to represent key dimensions of interest: reasoning-enabled vs. non-reasoning, open-source vs commercial and lower temperature vs higher. As for the temperature, the comparisons were performed between the default value, which was 1.00 and a lower one, that was set to 0.10. The assessment is under zero-shot learning, meaning that there were no examples or specific training provided to these models to perform the required tasks. The LLMs assessed are:

- **GPT-4o**

This model is a language model developed and released by *OpenAI* in May 2024. It's a non-reasoning model. For this model a comparison in regards to the temperature setting impact was conducted too (low temperature vs higher).

- **o3-mini**

This model is a language model developed and released by *OpenAI* in January 2025. It represents the most cost-efficient model in their reasoning series.

- **qwen 2.5-32B**

This model is an open-source language model released by *Alibaba Cloud* in September 2024. It's a non-reasoning model, with 32B parameters. For this model a comparison in regards to the temperature setting impact was conducted too (low temperature vs higher).

- **qwen QwQ-32B**

This model is an open-source language model released by *Alibaba Cloud* in March 2025. It's a reasoning model, with 32B parameters.

Each of these LLMs received the same clinical case input text and was prompted to respond with its DDx, with options ranked in order of likelihood. More details about prompting and output handling are available in Section 3.4.

To ensure consistent input/output handling and efficiency, all runs were made via API integration in *Google Colab* (https://colab.research.google.com/). The main parts of the code can be found in Appendices 8.1.1 and 8.1.2.

Parameters other than the temperature, such as top-p sampling, and max tokens were kept constant to the default ones across models to isolate the parameters under examination and limit other impacting factors.

Finally, these models responded using their knowledge after their training, without performing web searches that could lead to finding the right answer even by explicitly finding the solutions of these cases.

3.4 Prompting Strategies and Output Structuring

I am running an experiment on a clinicopathological case conference to see how your diagnoses compare with those of human experts. I am going to give you part of a medical case. These have all been published in the New England Journal of Medicine. You are not trying to treat any patients. As you read the case, you will notice that there are expert discussants giving their thoughts. In this case, you are "Dr. LLM", an AI language model that is discussing the case along with human experts.

A clinicopathological case conference has several unspoken rules. The first is that there is most often a single definitive diagnosis (though rarely there may be more than one), and it is a diagnosis that is known today to exist in humans. The diagnosis is almost always confirmed by some sort of clinical pathology test or anatomic pathology test, though in rare cases when such a test does not exist for a diagnosis the diagnosis can instead be made using validated clinical criteria or very rarely just confirmed by expert opinion. You will be told at the end of the case description whether a diagnostic test/tests are being ordered, which you can assume will make the diagnosis/diagnoses.

After you read the case, I want you to give two pieces of information.

The first piece of information is your most likely diagnosis/diagnoses. You need to be as specific as possible -- the goal is to get the correct answer, not a broad category of answers. You do not need to explain your reasoning, just give the diagnosis/diagnoses.

The second piece of information is to give a robust differential diagnosis, ranked by their probability so that the most likely diagnosis is at the top, and the least likely is at the bottom. There is no limit to the number of diagnoses on your differential. You can give as many diagnoses as you think, but they must be reasonable, so limit your answers accordingly to those that make sense. You do not need to explain your reasoning, just list the diagnoses. Again, the goal is to be as specific as possible with each of the diagnoses.

Figure 3.4: The system settings provided to the LLMs.

Each LLM received the full narrative text of the clinical case description (excluding doctors' diagnosis and the final diagnosis) as input, and was prompted to generate a diagnostic impression in a standardized format. Moreover, these cases included images, which were excluded since not all models had visual capabilities. Still, the narrative text of these cases included the description and the diagnostic findings of these medical images.

The system settings used for each of them are inspired and based on the corresponding system settings used in similar works [16] and are presented in Fig. 3.4.

Most Likely Diagnosis:

• Tuberculous meningitis

Differential Diagnosis (ranked by probability):

1. Tuberculous meningitis
2. Partially treated bacterial meningitis (e.g., pneumococcal meningitis with secondary vasculitic infarcts)
3. Listeria monocytogenes meningitis
4. Fungal meningitis (such as cryptococcal meningitis)
5. Varicella-zoster virus vasculopathy
6. Neurosyphilis

Figure 3.5: Structured Output from LLM.

The models were prompted to respond with their primary diagnosis, followed by a ranked DDx list, reflecting plausible alternative etiologies. This format mirrors clinical reasoning workflows and facilitates use of Top-N accuracy metrics. An illustrative model output is provided in Fig. 3.5.

3.5 Output Evaluation

Output evaluation has been performed using the *Gemini 2.0 Flash*. The prompt used for this evaluation is provided in Fig. 3.6.

I have a series of medical cases from The New England Journal of Medicine. For each case, I will provide:

The final diagnosis.
A list of predicted diagnoses.
Your task is to identify whether any of the predicted diagnoses correctly matches the final diagnosis based on meaning and clinical context, not exact wording, case sensitivity, or word order.

For example, if the final diagnosis is "Osteopenia", and option 2 says "Decreased bone density (Osteopenia)", you should respond: "Diagnosis predicted correctly in position 2"

even if the phrasing is not identical.

However, if none of the options capture the correct diagnosis in substance or intent, respond: "No correct answer"

Focus on the conceptual match rather than exact textual similarity.

Figure 3.6: Prompt provided to *Gemini Flash 2.0*, the LLM that worked as an evaluator.

3.6 Evaluation Metrics

The evaluation of model performance was based on two dimensions: The primary was the diagnostic accuracy. The secondary is diagnostic efficiency, as defined below. These metrics were chosen to reflect both the correctness but also the relevance and the efficiency of language model outputs in clinical diagnostic tasks. The primary goal was to assess how often each model correctly identified the true diagnosis—either as the top prediction or within its broader differential list—while also capturing how broad each diagnosis was.

3.6.1 Top-N Accuracy

The primary quantitative metric used to assess diagnostic performance was Top-N accuracy, a commonly applied measure in medical AI evaluations [18]. This metric captures whether the correct diagnosis appears within the top N ranked outputs returned by a model, assuming these are ordered by probability, descending. Specifically, this study computed Top-1, Top-3, Top-5 and Top-10 accuracy values for each model across all cases.

These metrics reflect clinically relevant thresholds, as real-world diagnostic support systems often aid clinicians by suggesting a list of leading candidates rather than a single definitive answer. High Top-N accuracy indicates the model's ability to generate a useful DDx, even when the top guess may be incorrect.

To qualify as a match, a diagnosis had to be conceptually equivalent to the reference diagnosis, either by exact match, accepted clinical synonym, or established diagnostic category (e.g., "acute myeloid leukemia" accepted as correct for "AML, M2 subtype"), as evaluated by the LLM *Gemini 2.0 Flash*.

3.6.2 Diagnostic Focus

In addition to diagnostic accuracy, this study evaluated the diagnostic efficiency of each language model—defined as a composite measure reflecting both the rank of the correct diagnosis and the total number of diagnostic hypotheses provided. This metric captures how focused and parsimonious the model's DDx is, which are important indicators of clinical usability.

For each case where the reference diagnosis appeared within the model's differential output, the model's rank position of the correct diagnosis was recorded. However, rather than assessing rank alone, efficiency and focus were contextualized by the total number of differential diagnoses generated. For example, a correct diagnosis appearing 3rd in a list of 5 options reflects higher quality of diagnosis than one appearing 3rd in a list of 10.

To quantify this, the diagnostic Smoothed Weighted Focus (SWF) was defined as:

$$SWF = \frac{1}{Rank} \div \sum_{i=1}^{Total\ Diagnoses} \frac{1}{i}$$

This formula gives higher scores to models that both:

- rank the correct diagnosis closer to the top, and
- generate a shorter, more targeted list of suggestions.

For every extra diagnosis that is added after the correct one, the score is getting lower, since the numerator is decreased.

The above formula is graphically depicted in Fig. 3.7 with a Heatmap and in Fig. 3.8, with a 3D plot.

In principle, it is inspired by the Discounted Cumulative Gain (DCG), a formula that is used in Information Retrieval systems, such as Search Engines. It rewards early correct answers and adds a penalty for later ones, similarly to the measurement we need to perform for the LLMs' focus score.

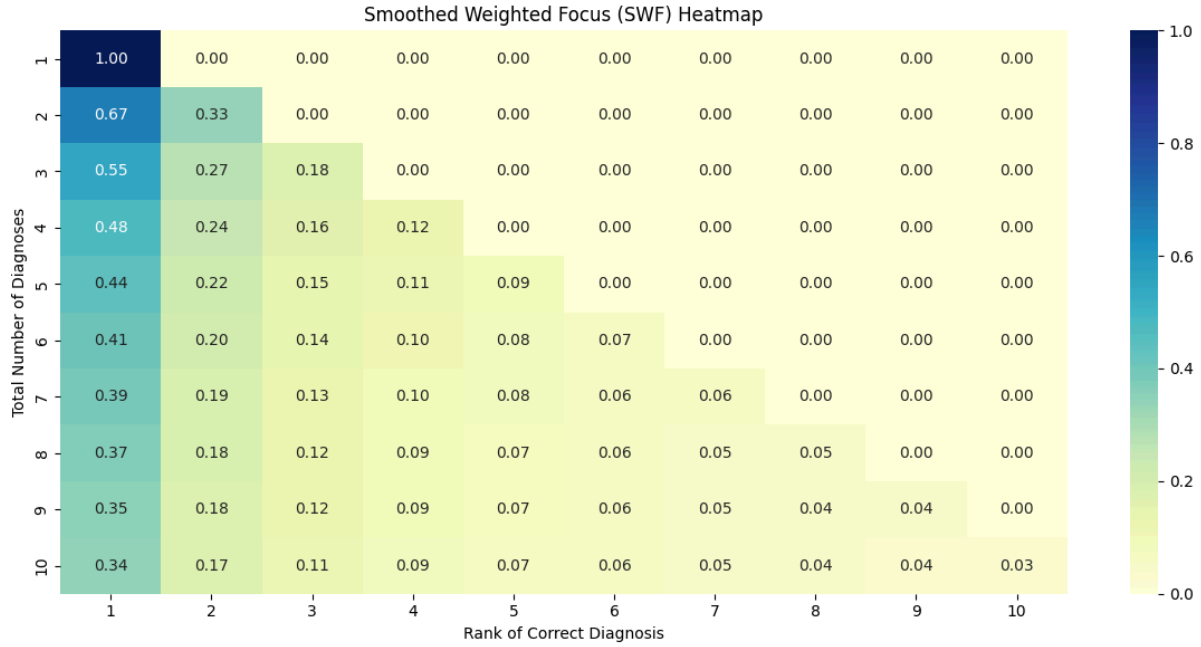

Figure 3.7: Smoothed Weighted Focus (SFW) Heatmap.

Some examples:

- The ideal scenario where a correct diagnosis ranked 1st out of 1, the score hits the maximum value, which is 1.
- A correct diagnosis ranked 1st out of 2, produces a score of approximately 0.67.
- A correct diagnosis ranked 3rd out of 6 produces a score of approximately 0.14.

- A missed diagnosis receives a score of 0.

This approach allows comparison not only of whether models are accurate, but also how directly they reach accurate conclusions.

Note, that for each model the average rank and also the average of total diagnoses were calculated. Since these numbers are decimals, rounding to the closest integer was required to be applied for the average total diagnoses of each model, causing minor differences, but without affecting the score and the comparison quality,

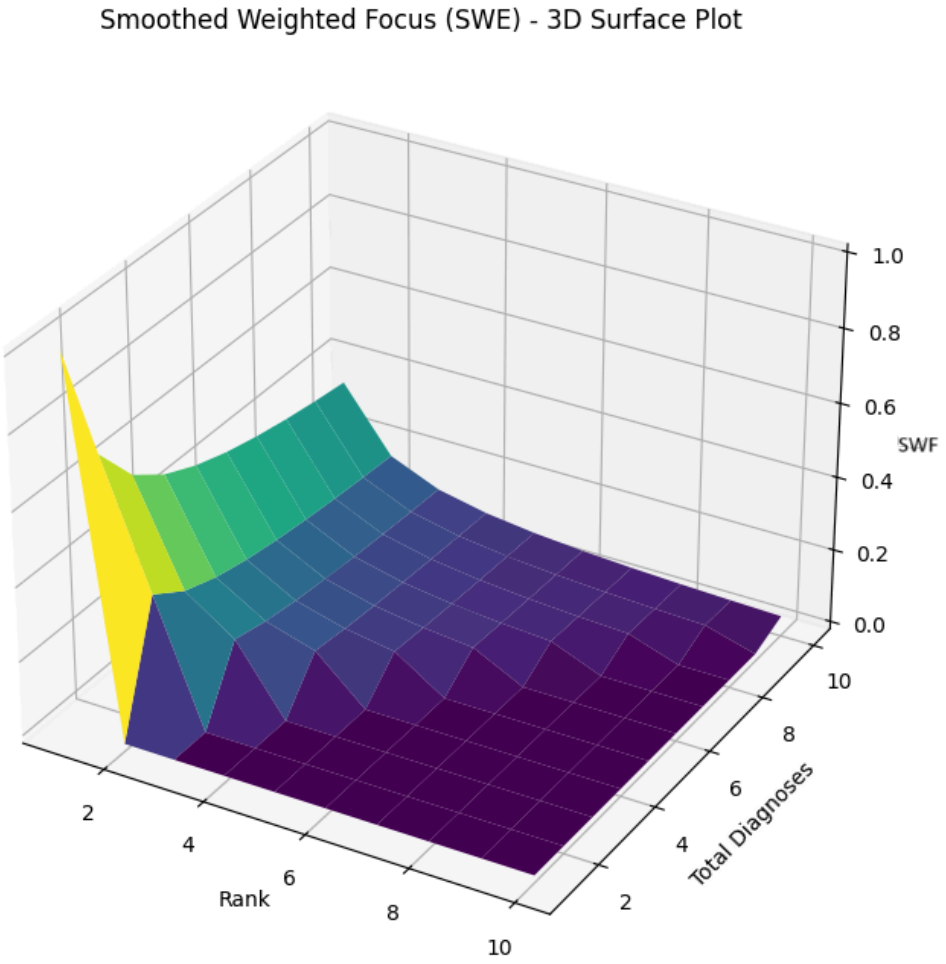

Figure 3.8: Smoothed Weighted Focus (SFW) 3D Surface Plot.

3.7 Statistical Analysis

Statistical analysis was performed to compare diagnostic performance across models and conditions.

To begin with all 4 models were compared at once, using *Friedman's Test*. This is a non-parametric statistical test used to investigate whether groups of three or more repeated measurements differ from each other. In our case, it was used to investigate whether the models' performance and focus differences are statistically significant or not.

For *Friedman's Test* it's a prerequisite to calculate means and medians based on common cases across the models. Therefore, it was applied in two different ways:

1. For the first run, only cases that were correctly diagnosed at any rank were included. This way, it was ensured that the cases used would be common across all models. However, this approach introduced the drawback of punishing and not letting models be evaluated on their full potential, if for example they got the right diagnosis on cases that others failed.

2. To resolve this issue mentioned in the previous paragraph, the second run included all the cases, introducing a constant value per model, for every case that it failed to diagnose. This value was set as the maximum rank of each model, increased by 5, meaning that we assumed that if each model had 5 extra attempts, it would have matched the golden standard diagnosis. The default values for each model as calculated, are displayed in Table 3.2, and they were assigned to all cases that were not correctly diagnosed, per model. This way, it was allowed to use *Friedman's Test* but without omitting cases of our dataset.

Both approaches have been used in the following runs of statistical tests, and each time the approach that is used is mentioned explicitly.

	gpt-4o	**o3-mini**	**qwen-2.5-32b**	**qwen-qwq-32b**
Default Rank	14	11	19	13

Table 3.2: Calculated default Ranks for missed diagnoses. Set to be as max rank plus 5.

Moreover, pairs of models and their runs have been compared using the *Wilcoxon Signed Rank Test*. This test is especially useful for ranked or ordinal data. It provides an excellent alternative for analyzing repeated measures or paired observations without requiring a normal distribution. In particular, the following comparisons have been performed:

- Several comparisons between models, to identify if any of them performs better than the rest.

- Comparisons of *GPT-4o* and *qwen 2.5-32B* runs, setting different temperature settings. (Low Temperature, set to 0.10 vs High Temperature, set to 1.00)

- Comparisons of *GPT-4o* versus *o3-mini* and *qwen 2.5-32B* versus *qwen QwQ-32B* to examine the factor of reasoning.

Furthermore, the presence or absence of a correct diagnosis within a given Top-N threshold (N = 1, 3, 5) was examined. *McNemar's Test* was used due to the paired nominal nature of the outcome (correct vs incorrect in the same case). This is a statistical test used to analyze paired nominal data, particularly in 2x2 contingency tables, to determine if there's a significant difference in proportions between two related groups

Finally, with the *Kruskal Wallis Test* it was investigated whether there was statistically significant difference on the models' performance across specialties. This is a

non-parametric statistical test used to determine if two or more independent samples originate from the same distribution.

All statistical tests were conducted using the web application https://www.statskingdom.com/index.html and also verified in *Microsoft Excel* or *Python* scripts run in *Google Colab*.

All confidence intervals were computed at the 95% level.

3.8 Environment Setup

To ensure reproducibility and consistency, all model evaluations were conducted within the controlled computational environment of *Google Colab*.

The *OpenAI* models were triggered via the company's API. As for the *Qwen* models at the time of the writing of this Thesis Project, they could be found in the *Groq* collection (https://groq.com/), and they were triggered via its API. For both cases, registration is required and also an API key creation too.

Each model was evaluated independently, without access to prior case results, and in different sessions. The code is available in Appendices 8.1.1 (*OpenAI* API) and 8.1.2 (*Groq* API).

3.9 Ethical Considerations

For this study no human or patient-identifiable data were used, since all data provided by NEJM *Case Records* are anonymous yet real life medical scenarios.

Key ethical aspects addressed in this study include:

- Clinical non-deployment: All models were used purely for research and evaluation purposes. No model was deployed in a real-time clinical setting or exposed to other live patient data. This study is about LLM comparison and benchmarking, not an endorsement of LLMs for clinical decision-making.

- Publication compliance: The use of NEJM content was limited to publicly available material and cited appropriately. No copyrighted images or other data were used.

4 Results

This chapter presents the results of the diagnostic performance evaluation of 4 LLMs across 80 complex clinical cases. The analysis is organized around statistical metrics such as mean and median, Top-N accuracy and diagnostic focus.

All analyses directly follow the methodology described in Chapter 3. Metrics like Top-N accuracy and SWF are reported per model. Furthermore, head to head statistical comparisons are included to assess significant differences between models.

4.1 Overall Model Performance

The mean and the median rank of correct diagnosis for each model is presented in Fig. 4.1, considering all cases that were correctly diagnosed for each model. SD for means are respectively 2.23, 1.53, 3.57, 2.58. Apparently, gpt-4o and o3-mini lead the performances, having the lowest means and medians.

For *Friedman's Test* it's a prerequisite to calculate means and medians based on the same cases. As a result, we applied it for 2 perspectives, as described in Section 3.7.

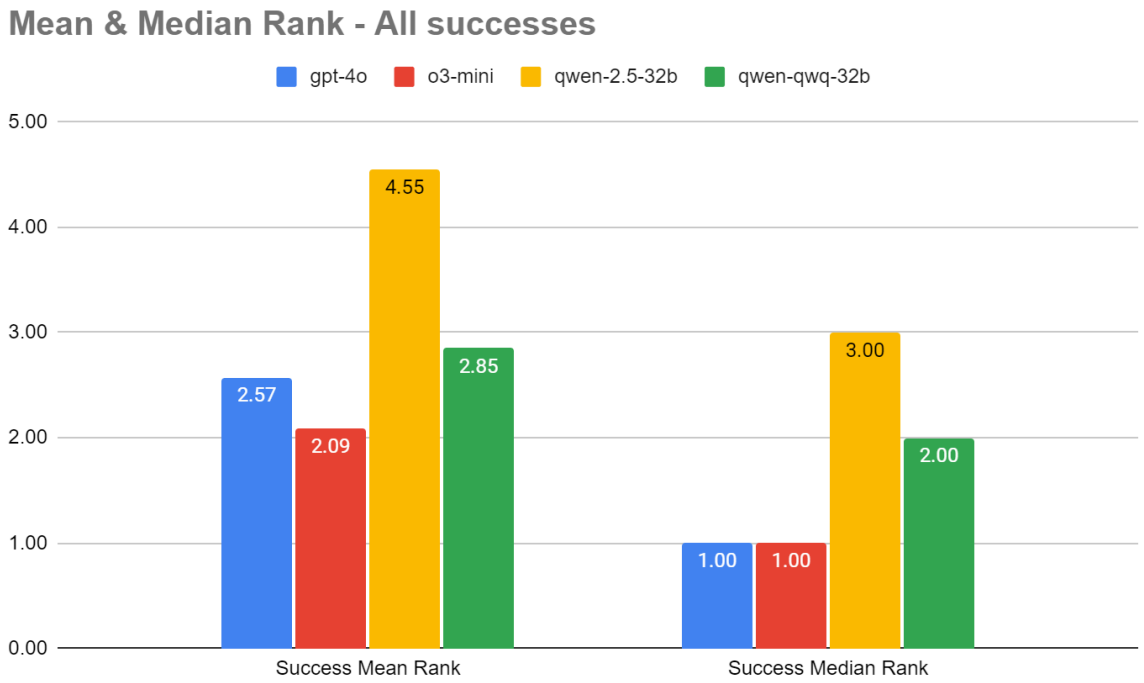

Figure 4.1: Mean and Median Rank, taking into account all cases that were correctly diagnosed per Model, at any Rank.

The first was to consider all cases correctly diagnosed by all models. The results are in Fig. 4.2. SD for means are respectively 1.65, 1.34, 2.74, 2.76. Having set the H_o as the hypothesis that all the ranks of the models have no statistically significant differences, the hypothesis was rejected (p=0.00008).

As a result, the Ranks of some models are not considered to be equal. In other words, the difference between the Ranks of some models is big enough to be statistically significant. Again, o3-mini and gpt-4o are performing the best.

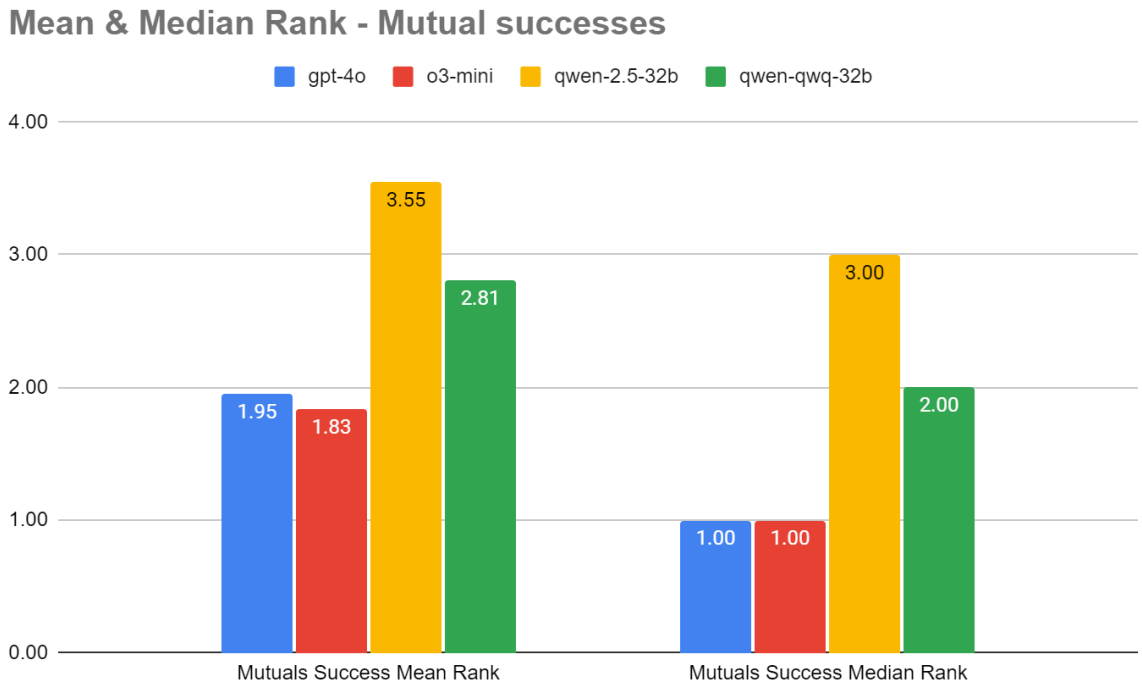

Figure 4.2: Mean and Median Rank, taking into account all cases that were correctly diagnosed by all language models, at any Rank. In other words, excluding cases that at least one model failed to diagnose correctly.

The second perspective was about taking into account all cases, setting a default Rank for cases where the diagnosis was missed, as described in Section 3.7.

The results are in Fig. 4.3. SD for means are respectively 5.10, 3.59, 6.41, 5.27. Having set the H_o as the hypothesis that all the ranks of the models have no statistically significant differences, the hypothesis was rejected again, with even higher power ($p=3.0649e^{-8}$). In result, the ranks of some models are not considered to be equal. In other words, the difference between the ranks of some models is big enough to be statistically significant. Similarly with the previous approach, o3-mini primarily and gpt-4o too, are performing the best.

It's important to notice that for the calculation of means and medians in Fig. 4.1, only cases in which the model produced the correct diagnosis within its output were included in the analysis. Similarly, for calculations in Fig. 4.2, all cases where at least one model failed to diagnose it correctly were excluded, to allow for statistical testing with *Friedman's Test*. On the other hand, introducing the default Ranks, even if these are calculated per model, according to their overall performance introduces another potential bias. As a result, Top-N accuracy metrics that are following in the next section, should and were used in parallel to capture diagnostic coverage and performance on the full case set.

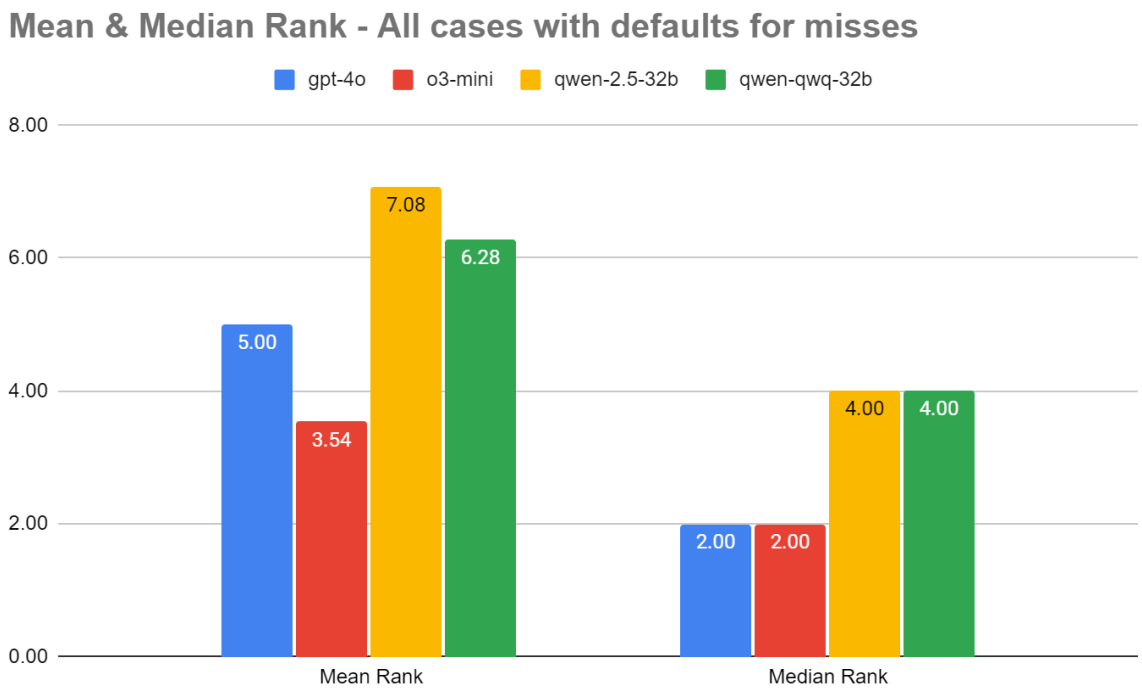

Figure 4.3: Mean and Median Rank, taking into account all cases and by setting default Ranks for misses.

Finally, according to these stats, there is an apparent improved performance by the *o3-mini* model. To investigate whether this is statistically significant, the *Wilcoxon Signed Rank Test* was performed for head to head comparison of *o3-mini* with the rest of the models. Again, 2 different tests we run, including all the cases with default Ranks for missed diagnoses in the first, and including only the mutually correct diagnoses in the second.

	gpt-4o	**qwen-2.5-32b**	**qwen-qwq-32b**
o3-mini	$p=0.02809 < 0.05$	$p=6.127e^{-7} < 0.05$	$p=5.685e^{-7} < 0.05$

Table 4.1: p-value for *Wilcoxon Signed Rank Test* comparisons between *o3-mini* and the rest of the models, including all cases, with default Ranks for missed diagnoses.

	gpt-4o	**qwen-2.5-32b**	**qwen-qwq-32b**
o3-mini	$p=0.5327 > 0.05$	$p=0.00017 < 0.05$	$p=0.02692 < 0.05$

Table 4.2: p-value for *Wilcoxon Signed Rank Test* comparisons between *o3-mini* and the rest of the models, including only cases where all models were correct.

According to these tests, results showed that the *o3-mini* better performance against any other model is statistically significant, as shown in Table 4.2 (using all the cases, with

default Ranks). Regarding the comparisons when only the cases with mutually correct diagnoses were included, again *o3-mini* surpassed both queen models, but not the *gpt-4o*, as shown in Table 4.3.

4.2 Top-N Accuracy

In this section, the results of the Top-N metric are presented, to approach a more complete description of performance for each model. Table 4.3 includes the accuracy for N={1, 3, 5, 10}. Then, detailed graphs are presented in Fig. 4.4, Fig. 4.5, Fig. 4.6, Fig. 4.7 and Fig. 4.8.

	Top-1	**Top-3**	**Top-5**	**Top-10**
gpt-4o	42.50%	57.50%	68.75%	78.75%
o3-mini	47.50%	63.75%	80.00%	83.75%
qwen-2.5-32b	18.75%	46.25%	57.50%	77.50%
qwen-qwq-32b	31.25%	48.75%	55.00%	65.00%

Table 4.3: Overall stats for Top-N accuracy of all 4 language models.

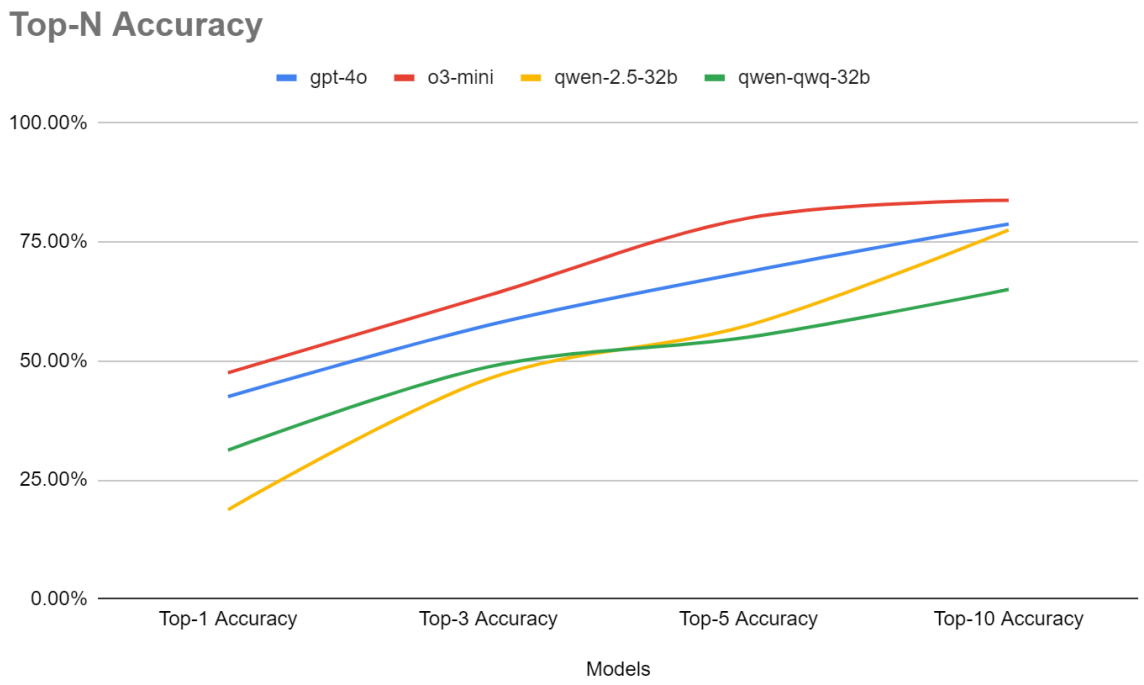

Figure 4.4: Overall stats for Top-N accuracy of all 4 language models.

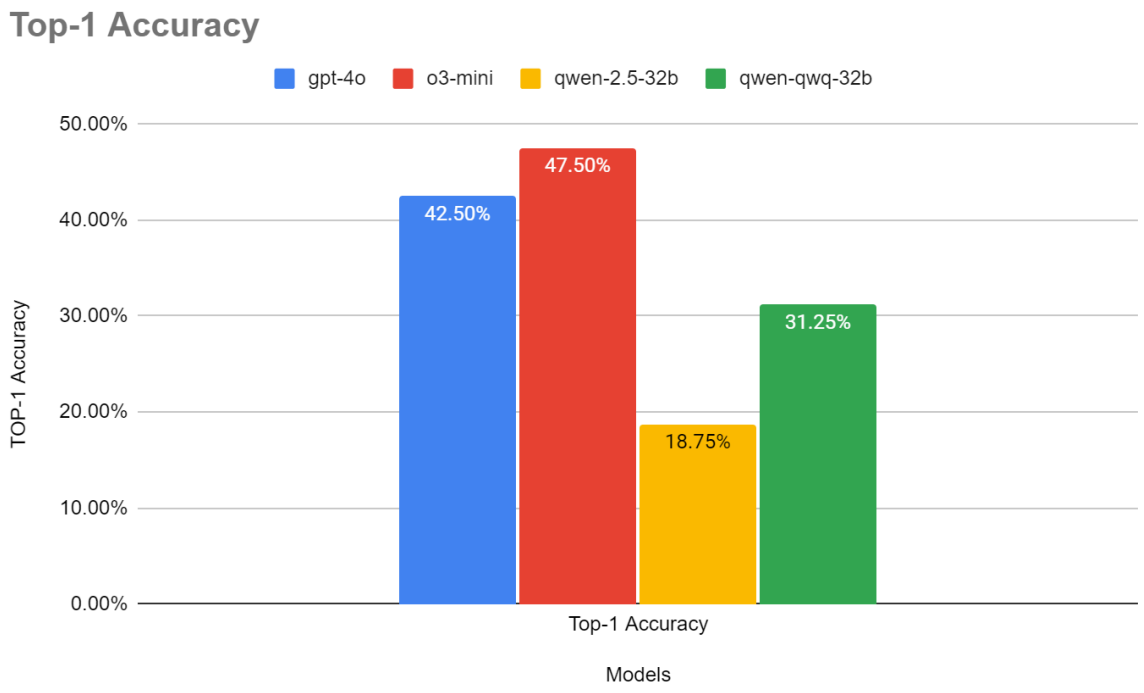

Figure 4.5: Top-1 accuracy of all 4 language models.

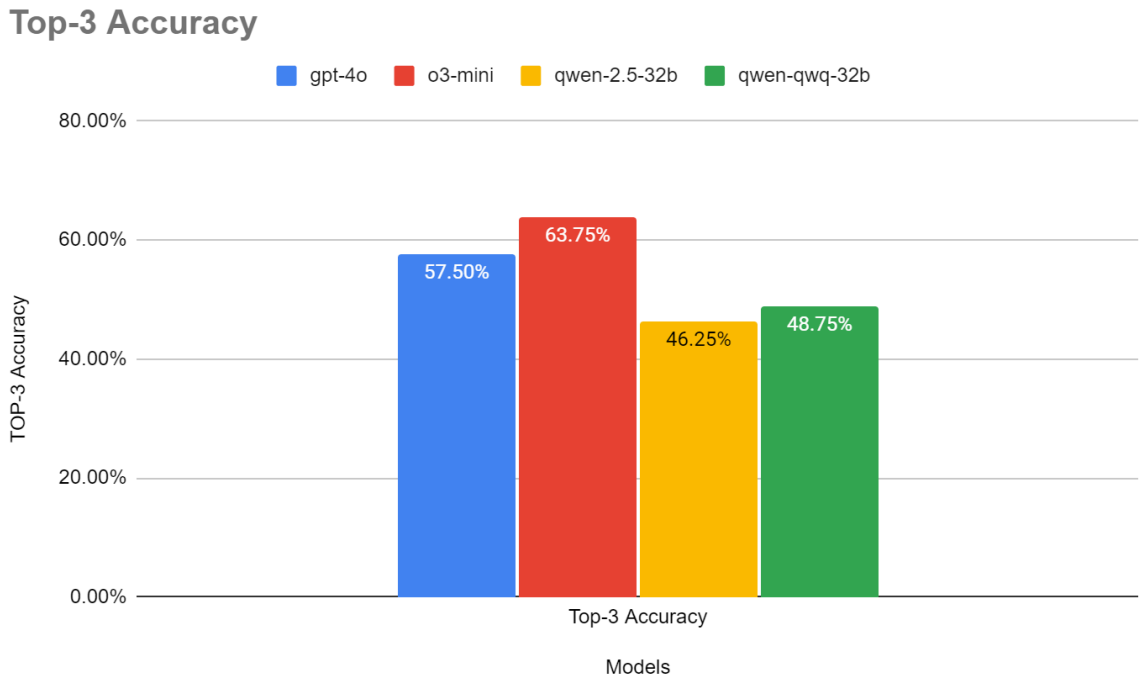

Figure 4.6: Top-3 accuracy of all 4 language models.

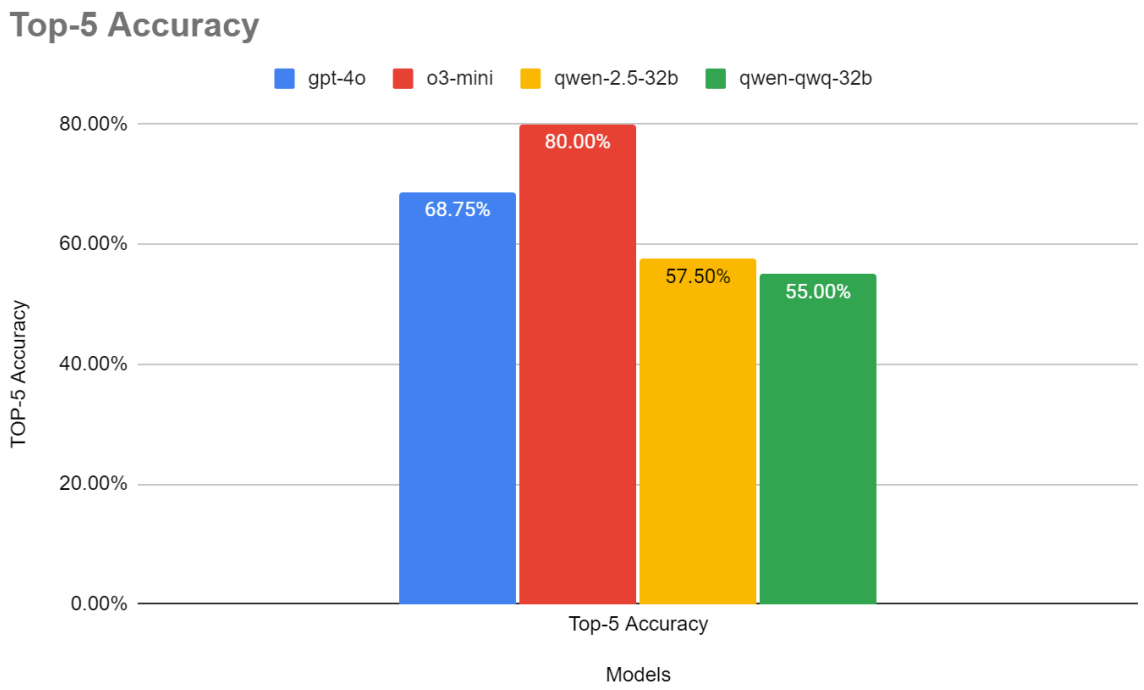

Figure 4.7: Top-5 accuracy of all 4 language models.

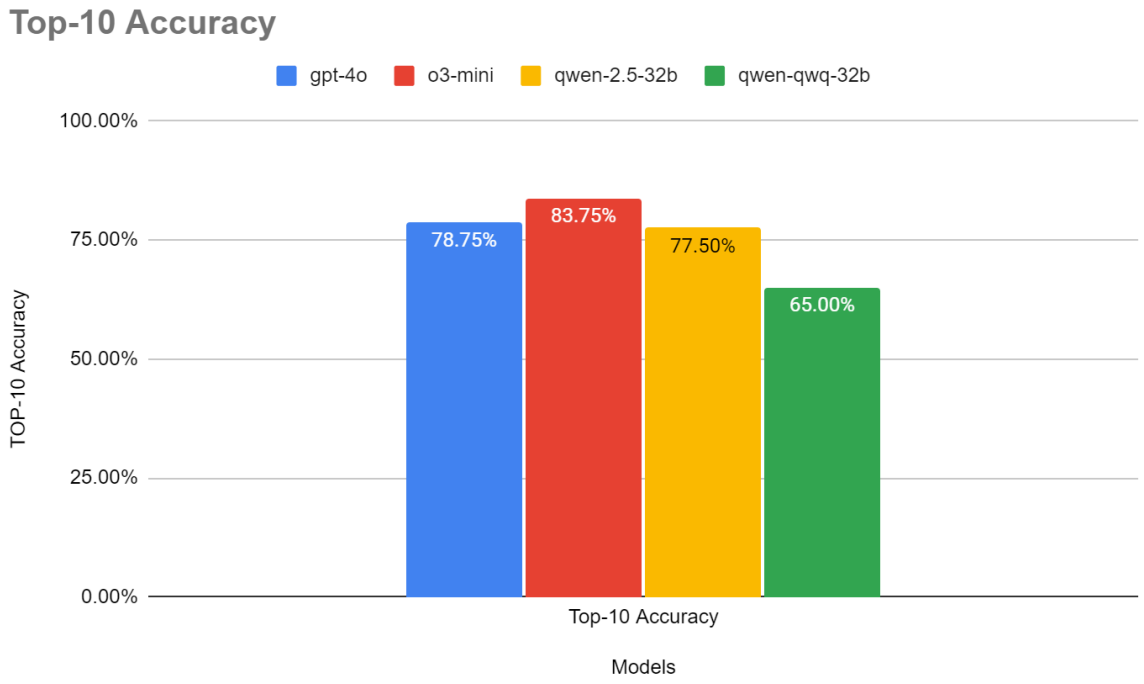

Figure 4.8: Top-10 accuracy of all 4 language models.

In all these metrics, *o3-mini* is apparently ahead by narrow or wide margins. The statistical significance of these differences were examined in detail using *McNemar's Test* for Top-1, Top-3 and Top-5 accuracy dimensions.

4.2.1 Top-1 Accuracy

- **o3-mini vs gpt-4o**

There were 11 cases where *o3-mini* hit the Top-1 limit while the *gpt-4o* didn't. On the other hand, there were 7 cases where the performance was reversed.

Therefore, according to *McNemar's Test* there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.3458) for the Top-1 accuracy dimension.

- **o3-mini vs qwen-2.5-32b**

There were 25 cases where *o3-mini* hit the Top-1 limit while the qwen-2.5-32b didn't. On the other hand, there were 3 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.00003) for the Top-1 accuracy dimension.

- **o3-mini vs qwen-qwq-32b**

There were 26 cases where *o3-mini* hit the Top-1 limit while the qwen-qwq-32b didn't. On the other hand, there were 3 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.00002) for the Top-1 accuracy dimension.

4.2.2 Top-3 Accuracy

- **o3-mini vs gpt-4o**

There were 14 cases where *o3-mini* hit the Top-3 limit while the *gpt-4o* didn't. On the other hand, there were 9 cases where the performance was reversed.

Therefore, according to *McNemar's Test* there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.2971) for the Top-3 accuracy dimension.

- **o3-mini vs qwen-2.5-32b**

There were 22 cases where *o3-mini* hit the Top-3 limit while the *qwen-2.5-32b* didn't. On the other hand, there were 2 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.00004) for the Top-3 accuracy dimension.

- **o3-mini vs qwen-qwq-32b**

There were 22 cases where *o3-mini* hit the Top-3 limit while the *qwen-qwq-32b* didn't. On the other hand, there were 2 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.01059) for the Top-3 accuracy dimension.

4.2.3 Top-5 Accuracy

- **o3-mini vs gpt-4o**

There were 15 cases where *o3-mini* hit the Top-5 limit while the *gpt-4o* didn't. On the other hand, there were 6 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.04953) for the Top-5 accuracy dimension.

- **o3-mini vs qwen-2.5-32b**

There were 25 cases where *o3-mini* hit the Top-5 limit while the *qwen-2.5-32b* didn't. On the other hand, there were 4 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.0001) for the Top-5 accuracy dimension.

- **o3-mini vs qwen-qwq-32b**

There were 22 cases where *o3-mini* hit the Top-5 limit while the *qwen-qwq-32b* didn't. On the other hand, there were 4 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.00042) for the Top-5 accuracy dimension.

4.3 Reasoning vs Non-Reasoning Models

In this section we will be comparing models of the same family, as for the impact of the reasoning factor on the Top-N accuracy.

In particular, we will be comparing *o3-mini* (reasoning enabled) versus *gpt-4o* (non reasoning) by *OpenAI and qwen-qwq-32b* (reasoning enabled) versus *qwen-2.5-32b* (non reasoning) by *Alibaba Cloud.*

4.3.1 Top-1 Accuracy

- **o3-mini vs gpt-4o**

This comparison has already been presented in Section 4.2.1. According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.3458) for the Top-1 accuracy dimension.

- **qwen-qwq-32b vs qwen-2.5-32b**

There were 14 cases where *qwen-qwq-32b* hit the Top-1 limit while the *qwen-2.5-32b* didn't. On the other hand, there were 4 cases where the performance was reversed.

Therefore, according to *McNemar's Test a significant difference was found* between the 2 models (p=0.01842) for the Top-1 accuracy dimension.

4.3.2 Top-3 Accuracy

- **o3-mini vs gpt-4o**

This comparison has already been presented in Section 4.2.2.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.3458) for the Top-3 accuracy dimension.

- **qwen-qwq-32b vs qwen-2.5-32b**

There were 13 cases where *qwen-qwq-32b* hit the Top-3 limit while the *qwen-2.5-32b* didn't. On the other hand, there were 15 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.7055) for the Top-3 accuracy dimension.

4.3.3 Top-5 Accuracy

- **o3-mini vs gpt-4o**

This comparison has already been presented in Section 4.2.3.

According to *McNemar's Test a significant difference was found* between the 2 models (p=0.04953) for the Top-5 accuracy dimension.

- **qwen-qwq-32b vs qwen-2.5-32b**

There were 16 cases where *qwen-qwq-32b* hit the Top-5 limit while the *qwen-2.5-32b* didn't. In contrast, there were 14 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a difference between the 2 models (p=0.715) for the Top-5 accuracy dimension.

4.4 Temperature Impact

In this section, we are investigating the impact of temperature for the non-reasoning models. The default temperature for *gpt-4o* and *qwen-2.5-32b* is 1.00. Another run has been completed for these models, changing only the temperature to 0.10, as described in Section 3.3.

Therefore, by performing head to head tests with *McNemar's Test*, the potential impact of temperature setting on the diagnostic performance is presented below.

4.4.1 Top-1 Accuracy

- **gpt-4o (low temperature) vs gpt-4o (high temperature)**

There were 4 cases where *gpt-4o* (low temp) hit the Top-1 limit while the *gpt-4o* (high temp) didn't. On the other hand, there were 6 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.5271) for the Top-1 accuracy dimension.

- **qwen-2.5-32b (low temperature) vs qwen-2.5-32b (high temperature)**

There were 4 cases where *qwen-2.5-32b* (low temp) hit the Top-1 limit while the *qwen-2.5-32b* (high temp) didn't. On the other hand, there were 3 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.7055) for the Top-1 accuracy dimension.

4.4.2 Top-3 Accuracy

- **gpt-4o (low temperature) vs gpt-4o (high temperature)**

There were 5 cases where *gpt-4o* (low temp) hit the Top-1 limit while the *gpt-4o* (high temp) didn't. On the other hand, there were 5 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=1.00) for the Top-3 accuracy dimension.

- **qwen-2.5-32b (low temperature) vs qwen-2.5-32b (default temperature)**

There were 6 cases where *qwen-2.5-32b* (low temp) hit the Top-3 limit while the *qwen-2.5-32b* (high temp) didn't. On the other hand, there were 12 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.1573) for the Top-3 accuracy dimension.

4.4.3 Top-5 Accuracy

- **gpt-4o (low temperature) vs gpt-4o (high temperature)**

There were 5 cases where *gpt-4o* (low temp) hit the Top-1 limit while the *gpt-4o* (high temp) didn't. On the other hand, there were 5 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.3173) for the Top-5 accuracy dimension.

- **qwen-2.5-32b (low temperature) vs qwen-2.5-32b (high temperature)**

There were 7 cases where *qwen-2.5-32b* (low temp) hit the Top-5 limit while the *qwen-2.5-32b* (high temp) didn't. On the other hand, there were 10 cases where the performance was reversed.

According to *McNemar's Test*, there is *not enough evidence* to suggest there is a significant difference between the 2 models (p=0.4669) for the Top-5 accuracy dimension.

4.5 Diagnostic Focus

The diagnostic focus as defined in Section 3.6.2 has been measured and displayed in Fig. 4.9 and Fig. 4.10. The first shows the mean and median total diagnoses for each model.

According to *Friedman's Test*, having set the H_0 as the hypothesis that all the diagnosis counts of the models have no statistically significant differences, the hypothesis was rejected (p=0). In result, the number of total diagnoses per model during the differential are not considered to be equal.

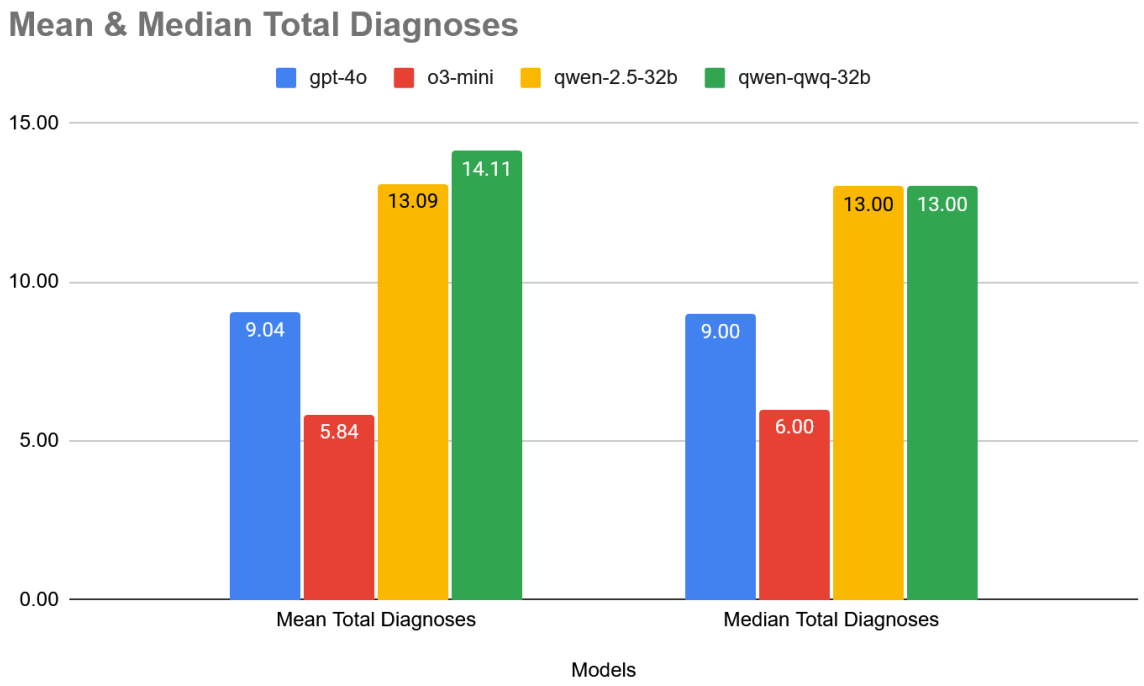

Figure 4.9: Mean Total Diagnoses of all 4 language models.

As for the second one, it reveals a trend towards having *o3-mini* as the model that combines high accuracy and the fewest options in its differential, offering focused and higher-quality differentials.

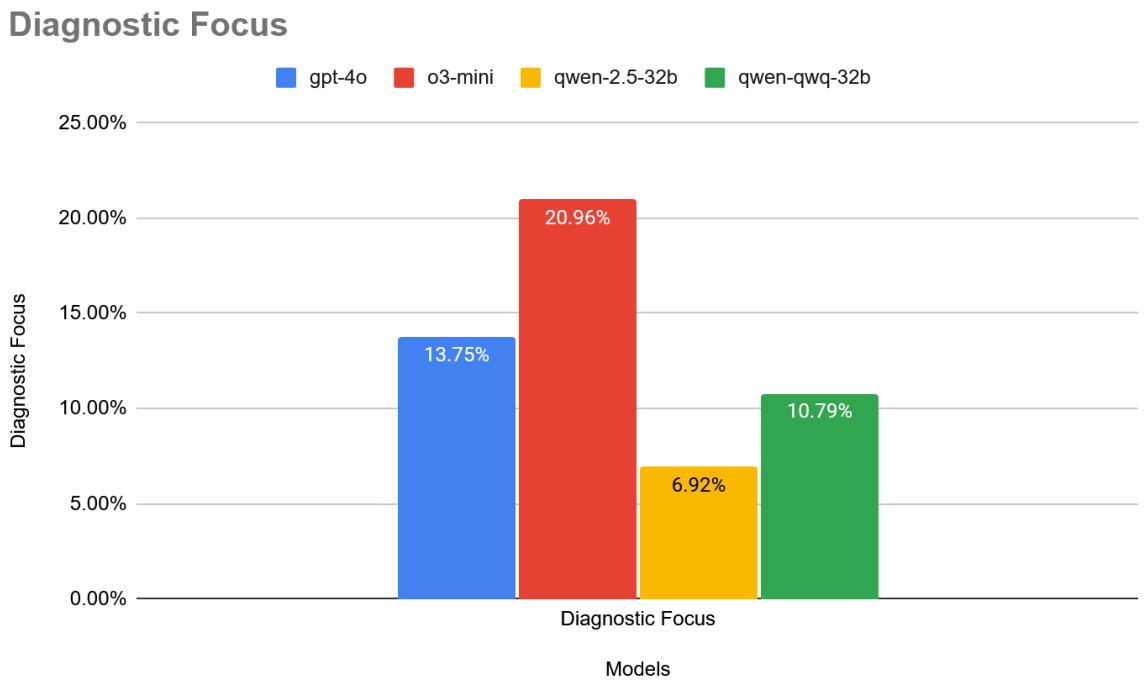

Figure 4.10: Diagnostic Focus of all 4 language models.

4.6 Specialty-Specific Performance

As described in Section 3.1.2 and in particular Table 3.1, the dataset is not balanced for the specialty dimension.

However, picking the top 5 specialties in terms of representation, the results for Top-N accuracy N={1, 3, 5} are presented in Table 4.4, Table 4.5 and Table 4.6.

Top-1 accuracy	gpt-4o	o3-mini	qwen-2.5-32b	qwen-qwq-32b
Hematology/Oncology	43.59%	51.28%	20.51%	35.90%
Infectious Disease	37.14%	48.57%	11.43%	25.71%
Surgery	36.36%	42.42%	12.12%	30.30%
Emergency Medicine	46.67%	46.67%	16.67%	43.33%
Neurology/Neurosurgery	34.62%	46.15%	11.54%	30.77%

Table 4.4: Top-1 accuracy of all 4 language models for the 5 well represented specialties of the dataset.

Top-3 accuracy	gpt-4o	o3-mini	qwen-2.5-32b	qwen-qwq-32b
Hematology/Oncology	58.97%	66.67%	43.59%	61.54%
Infectious Disease	51.43%	62.86%	45.71%	42.86%

Surgery	48.48%	63.64%	42.42%	39.39%
Emergency Medicine	63.33%	70.00%	53.33%	60.00%
Neurology/Neurosurgery	46.15%	65.38%	42.31%	38.46%

Table 4.5: Top-3 accuracy of all 4 language models for the 5 well represented specialties of the dataset.

Top-5 accuracy	gpt-4o	o3-mini	qwen-2.5-32b	qwen-qwq-32b
Hematology/Oncology	66.67%	79.49%	51.28%	61.54%
Infectious Disease	62.86%	80.00%	54.29%	54.29%
Surgery	57.58%	78.79%	54.55%	45.45%
Emergency Medicine	80.00%	90.00%	66.67%	73.33%
Neurology/Neurosurgery	57.69%	80.77%	57.69%	46.15%

Table 4.6: Top-5 accuracy of all 4 language models for the 5 well represented specialties of the dataset.

Kruskal Wallis	gpt-4o	o3-mini	qwen-2.5-32b	qwen-qwq-32b
p-value	0.5256	0.9718	0.8894	0.327

Table 4.7: *Kruskal Wallis Test* p-value for performance differences per model, per specialty showed statistically no significant difference in any of them.

Furthermore, using the *Kruskal Wallis Test*, there was *no statistically significant difference* in the performance of any of the 4 models, in any of the 5 specialties, as shown in Table 4.7. For this statistical test, the default Ranks, as calculated in Section 3.7 were used.

5 Discussion

This study presents a structured and comparative evaluation of LLMs on complex clinical diagnostic tasks, using real-world case data from the NEJM. In this chapter, there will be a discussion about its results, presented in the previous chapter.

Firstly, an interpretation of these results will be presented, with its core findings and potential conclusions. In addition, there will be some comparisons of this research results with corresponding results published on other research work. Finally, the strengths and the limitations of this project will be discussed, together with some interesting related work ideas that are recommended and should be considered to be performed soon.

5.1 Interpretation of Results

In this section, the most important conclusions drawn from the evaluation results will be presented. Where applicable, these will be grouped with their statistical tests and power.

5.1.1 o3-mini performance

According to multiple comparisons presented Chapter 4 *o3-mini* is the LLM that performs the best according to multiple metrics and statistical tests, in terms of performance and diagnostic focus as well.

Firstly, *o3-mini* had the lowest mean and median values across all models, using 3 different approaches, as shown in Table 5.1. As described in Section 3.7, the latest one is the most reliable one, since it allows for inclusion of cases where not all models were successful, and it will be used in the following results interpretations.

Approaches	Mean	SD	Median
All successfully diagnosed cases for each model included	2.09	1.53	1
All mutually successfully diagnosed cases for all models included	1.83	1.34	1
All cases using default Ranks included	3.54	3.59	2

Table 5.1: Overall Mean & Median scores for *o3-mini* were the lowest among all models in 3 different approaches.

Additionally, according to *Friedman's Test* the differences between the models' accuracy is statistically significant, with strong power, $p=3.0649e^{-8}$. Moreover, as shown in Table 4.1, *o3-mini* was the most accurate of all 4 models (*Wilcoxon Signed Rank Test*, p=0.02809, p=6.127e-7, p=5.685e-7).

As for the Top-N metric, again *o3-mini* had the highest scores across all N variants (1, 3, 5, 10) with its score being 47.50%, 63.75%, 80.00%, 83.75% respectively. Using *McNemar's Test* for N=1, 3, 5, it was shown that *o3-mini* out-performed with statistical significance both *Qwen* models in these N variants, and *gpt-4o* for N=5.

Finally, o3-mini hit the best scores in diagnostic focus. It had the lowest Mean (5.84) and Median (6) values for the total diagnoses their output had and also the highest score in the custom SFW metric that we defined in Section 3.7, with 20.96%, achieving something that was impressive: o3-mini had the **highest accuracy**, with **the fewest suggestions** in its differential diagnoses, among all LLMs.

5.1.2 Reasoning impact

The impact of having a model to be reasoning-enabled appears to be helping towards the performance improvement of LLMs.

The *Qwen* reasoning-enabled model (*qwen-qwq-32b*) outperformed the qwen non-reasoning model (*qwen-2.5-32b*) for Top-1 accuracy (*McNemar's Test*, p=0.01842) and the *o3-mini* which is reasoning-enabled outperformed the *gpt-4o* for Top-5 accuracy (*McNemar's Test,* p=0.04953).

On the other hand, there is no category or statistical test that showed the opposite impact for reasoning-enabled models.

5.1.3 Temperature impact

In Section 4.4 the impact of model temperature setting was studied and investigated across different values. By definition lower temperatures lead to less random results and more deterministic behaviors by LLMs. In healthcare and clinical diagnostics this is a desired outcome, since reproducibility is an important factor for any assistance provided by Generative AI.

According to our results, higher or lower temperature is not impacting the accuracy scores for Top-N metrics. In all the head to head comparisons between model variations with low (0.10) and the higher temperature (1.00) there was no statistically significant difference, using *McNemar's Test*, in any of the Top-N metrics (N=1, 3, 5). In fact, the statistical power of these tests never dropped below 15%, providing strong evidence that any differences noted should be considered random.

5.1.4 Per medical specialty performance

As for the LLMs performance per medical specialty, across the 5 ones that had adequate samples (threshold set to 25 cases) it was shown with the *Kruskal Wallis Test* that there was no statistically significant difference among them, for any of the LLMs. (gpt-4o - p=0.5256. o3-mini - p=0.9718, qwen-2.5-32b - p=0.8894, qwen-qwq-32b - p=0.327).

Still, even if the following were not justified in the statistical test, it was apparent that all LLMs performed their best in Emergency Medicine for almost every N. On the other

hand, in many cases Surgery and Neurology/Neurosurgery were the domains where LLMs failed the most.

In total, it's not fair to come to conclusions because of the imbalanced dataset for the medical specialty dimension, but the performance in some domains such as Emergency Medicine, Surgery and Neurology/Neurosurgery, might be hinting towards a direction for the performance of the state of the art language models, and for future research in clinical diagnosis too.

5.1.5 Open-source vs. Commercial performance

According to the results, the two models that had the best performance were those developed by *OpenAI*, which are the ones that have commercial license. On the other hand, those that are open source did not score as high. Given this, one could argue that the commercial LLMs outperform the open-source ones. However, this argument is quite a generalization and can not be justified under the conditions set in this study.

As pointed out in the Key Glossary, technical and engineering parameters are not always clearly and publicly available for commercial models. This limitation makes it difficult to fairly compare them with open-source models, since for such comparison it would be required that models of the same computing power and purpose are compared.

There are technical aspects, such as the number of parameters used for a model, that are absolutely crucial for its performance [21]. In this research for example, while for Qwen models it's documented that they are engineered with 32B parameters, the number of parameters for *OpenAI*'s models are not officially disclosed.

As a result, this topic is not addressed by this research, and it is considered open for future research.

5.2 Comparison with Previous Studies

In recent years, numerous research papers on the performance and impact of LLMs on DDx have been published. Actually, some of them have even influenced and inspired this Thesis Project.

To begin with, the dataset available by NEJM has already been used in numerous works [2,16,18]. The evaluation in these works [16,18] focuses mainly on the quality of the DDx, and it was performed by physicians. Therefore, a direct comparison of the results is not feasible. This is a different approach, focusing on evaluating the quality of the reasoning, and the context of the DDx, without focusing strictly on the accuracy.

The evaluation process that this Thesis Project has implemented is similar to the one presented in some other work [2] where one of the main evaluation metrics was the Top-N accuracy, automatically run by a non-participating LLM (*Med-PaLM 2*). The evaluated model in this case was the *Articulate Medical Intelligence Explorer* (*AMIE*), developed by the research team, whose metrics were compared against other sides,

such as clinicians, clinicians synergy with *AMIE* and clinicians assisted by Internet Search, but also against *GPT-4*.

A direct comparison of results is not fair and consistent since different LLMs fed with different prompts have been used as raters in these 2 works. In our case, the prompt used for the rater LLM *Gemini 2.0 Flash* was explaining the process in detail, as shown in Fig. 3.3. On the other hand, in *McDuff, D., Schaekermann, M., Tu, T. et al.* [2] work, the prompt was much more concise and strict as shown in Fig. 5.1.

As Fig. 5.2 displays, GPT-4 and AMIE were evaluated for their Top-N accuracy metric for N=1, 3, 5 starting from ~30% and increasing up to approximately 50-55% as N increases too. On the other hand, for this research *o3-mini* and *gpt-4o* that were proved to be the best performers LLMs, had their Top-N metric range respectively according to Table 4.3 from 42,50% up to even 80.00%, as N goes higher.

> Is our predicted diagnosis correct (y/n)? Predicted diagnosis: [diagnosis], True diagnosis: [label]
>
> Answer [y/n].

Figure 5.1: Prompt used for the evaluator LLM in *McDuff, D., Schaekermann, M., Tu, T. et al.*work.

However, primarily *o3-mini* and secondarily *gpt-4o* are more advanced, released later than *gpt-4*, and have been proven to perform better in health care tasks, evaluated on the HealthBench [14], as seen in Fig. 5.3. Therefore, results hint to potentially improved performance by the more advanced and State of the Art LLMs.

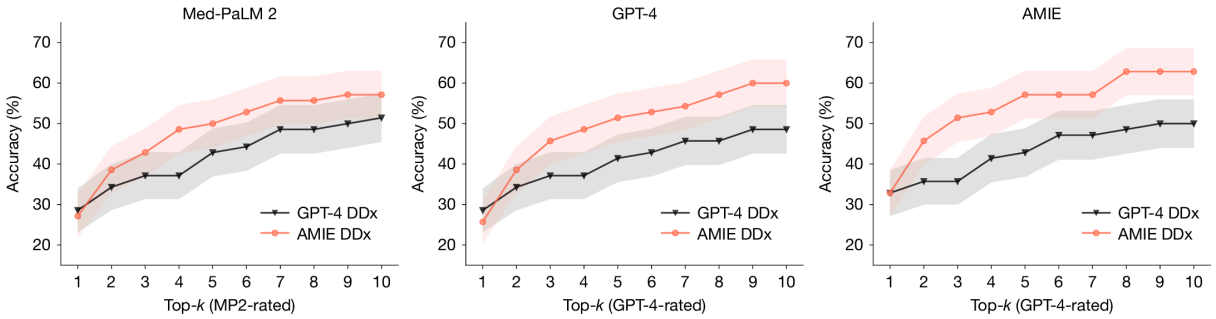

Figure 5.2: Retrieved from *McDuff, D., Schaekermann, M., Tu, T. et al.* work. Comparison of the percentage of DDx lists that included the final diagnosis for AMIE versus *GPT-4* for 70 cases. We used Med-PaLM 210, GPT-46 and AMIE as the raters—all resulted in similar trends. Points reflect the mean; shaded areas show ±1 s.d. from the mean across 10 trials.

Moreover, these studies did not include or mention the value of LLMs' diagnostic focus and efficiency, which was introduced in this work.

As for the temperature impact, in other research, it has been shown that it is a hyperparameter with low or no impact on accuracy scores [25]. Our research verified this outcome. Still, even if temperature appears to be an irrelevant setting since it's not

impacting the performance of any language model, it's important not to ignore its role by definition, the improvement of determinism and reproducible diagnoses.

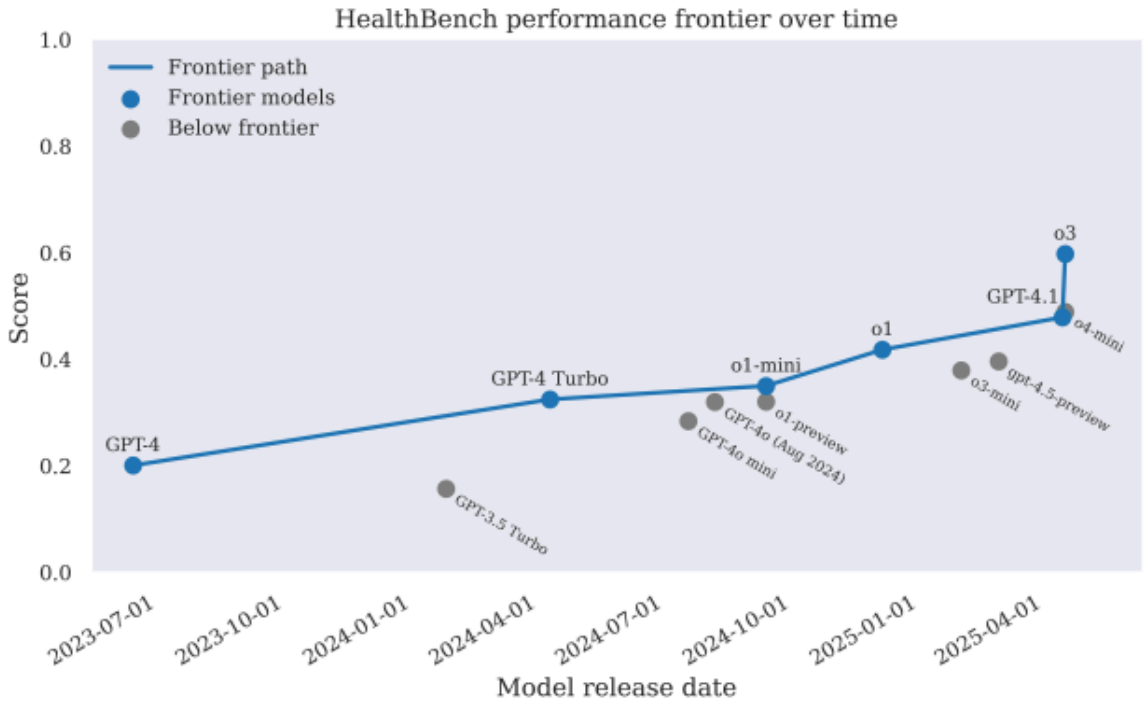

Figure 5.3: Retrieved from *HealthBench: Evaluating Large Language Models Towards Improved Human Health by Arora RK, Wei J, Hicks RS, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al* HealthBench performance of OpenAI models over time.

Finally, the reasoning process and diagnostic reasoning in principle is known to be crucial for Medical Doctors [9,30]. Similarly for LLMs, this feature that is intended to mimic the process followed by humans was found to improve the LLM performance on DDx.

5.3 Strengths and Limitations

For this research, there are a number of methodological strengths that enhance the robustness of the findings. First, the study design employed a set of 80 clinically validated cases spanning a wide range of specialties, ensuring both realism and diagnostic diversity. The inclusion of models of different characteristics allowed for meaningful insights on comparisons of several dimensions. Moreover, these language models are part of the State of the Art on Generative AI and LLMs, a field that has been progressing exponentially during the last years. Another strength is using meaningful and informative evaluation metrics. The combination of Top-N accuracy with the custom metric of SWE, allows for insights on both correctness and diagnostic focus as well. In addition, this research and the framework that was applied can be reproduced with different models under evaluation, since both the metrics and the dataset are publicly available and ready to be used.

On the other hand, several limitations must be acknowledged. First, even if the evaluation process covered several aspects of performance and avoided introducing major bias, it is still subject to potential secondary ones. To begin with, the requirements of some rank-based statistical tests (*Friedman's Test, Wilcoxon Signed Rank Test*) led to having to exclude cases with completely missed diagnoses by at least one model. This approach certainly masked some of the performance gaps among models, and this is why it was handled by allowing all the cases in these statistical tests, by setting a custom default value for all the missed diagnoses. This approach resolves the aforementioned issue, and it was preferred. Still, we need to point out the fact that introducing the calculated default Ranks introduces a potential minor bias in the evaluation process. Secondly, although specialty-specific performance was analyzed, uneven case distributions across specialties limited the ability to draw statistically robust conclusions in underrepresented domains. Moreover, this study focused exclusively on diagnostic generation and did not assess the quality of the reasoning by each LLM, which is critical for real-world cases. This aspect of evaluation would require at least one Medical Doctor to perform the evaluation of each model output.

5.4 Methodological Recommendations for Future Research

As the diagnostic capabilities of LLMs continue to evolve, so too must the methodologies used to evaluate their performance in clinical settings. While this study offers a structured and comparative analysis of several state-of-the-art models, it also highlights areas where future research can further improve and evolve the status of Generative AI in the field of healthcare and DDx.

Stratified Evaluation Using Balanced, Specialty-Specific Datasets

The current study demonstrated notable performance variability across clinical specialties. However, due to uneven representation in the dataset, certain specialties were excluded from comparative analysis. Future studies should aim to construct balanced specialty-specific datasets to enable more reliable evaluations of model behavior across medical domains. This could involve curating subsets of cases that are equated for complexity, length, and diagnostic ambiguity, thereby allowing researchers to isolate performance trends specific to specialties.

Scaling Up: Leveraging Full-Scope Datasets

Although 80 NEJM cases provided a robust foundation for this study, the full NEJM Case Records archive represents a far larger resource. Future studies should consider scaling to hundreds or thousands of cases, leveraging automated pipelines for prompt generation, response logging, and scoring. A larger dataset would not only increase statistical power but also allow for longitudinal evaluation, such as tracking performance across case types, publication years, or patient demographics.

Incorporating Visual Data into Diagnostic Evaluation

With the advent of multimodal LLMs, diagnostic reasoning is no longer limited to textual input. Many NEJM cases include key clinical images (e.g., medical imaging, pathology

slides, dermatological findings), which are essential to diagnostic accuracy in real-world settings. Future evaluations should incorporate these elements, allowing for visual-text multimodal prompting.

Developing a Diagnostic Reasoning Evaluation Framework

While Top-N accuracy and Weighted Efficiency metrics are informative, they do not fully capture the reasoning quality, explanatory coherence, or clinical appropriateness of LLM responses. There is a need to develop a standardized framework for evaluating the qualitative dimensions of diagnostic reasoning, including aspects such as the presence of hallucinations or unsafe recommendations, the logical coherence of differential construction or the justification of any action and medical exam requested Such a framework would most likely require human-in-the-loop evaluation, rubric-based scoring, or even new LLMs working as evaluators, trained specifically on clinical argumentation.

As LLMs move from experimental use to potential clinical integration, research methodologies must evolve to evaluate not only whether these systems are correct, but how, why, and under what conditions their reasoning is clinically trustworthy.

Computational Efficiency and Small Language Models Evaluation

Beyond diagnostic accuracy, the optimization of computational efficiency and cost for LLMs remains underexplored and lately starts to attract more and more research interest [31]. Many of the most accurate models are resource-intensive, requiring significant resources. Future work must address whether these models can realistically be deployed for clinical usage and how they might be optimized for real-time co-piloting with medical professionals. This leads also to evaluation of small language models and whether they can offer a balance in the performance - efficiency trade off, potentially by being trained and gaining expertise on specific specialties.

6 Conclusion

This chapter summarizes the findings from our methodology's implementation in Generative AI and LLM performance for DDx. It also offers final reflections on the subject, building upon the previous chapter's presentation of the implementation.

6.1 Summary of Findings

This thesis presented a structured and reproducible methodology for evaluating the diagnostic capabilities of LLMs in complex, real-world clinical scenarios. The evaluation framework combined traditional Top-N accuracy metrics with a novel diagnostic efficiency score, reflecting both the position and the value of correct diagnostic suggestions. This dual approach allowed for a more holistic assessment of diagnostic focus and clinical usability as well.

In general, the results revealed an improvement in LLM diagnostic performance and accuracy compared to earlier studies, reflecting ongoing advancements in model design and engineering. Notably, reasoning-enabled models consistently outperformed their non-reasoning counterparts, underlining the importance of structured thinking processes in clinical tasks, similarly with physicians and clinicians. On the other hand, temperature variation did not significantly influence diagnostic accuracy. Higher temperatures generally provide more consistent but not more accurate responses.

Among the models evaluated, proprietary language models by *OpenAI* outperformed open-source alternatives. However, due to the lack of transparency regarding model architecture, parameter count, and training data in commercial systems, direct comparisons remain partially constrained. Nevertheless, OpenAI's o3-mini model emerged as the most effective across both performance and diagnostic focus metrics, combining strong Top-N accuracy with concise, high-quality differential lists.

The analysis also revealed notable variation in model performance across medical specialties, suggesting that general-purpose LLMs do not generalize uniformly across clinical subdomains. This finding underscores the need for specialty-specific evaluations when considering real-world use in diverse clinical environments.

Finally, statistical analysis using non-parametric tests such as the *Friedman* test confirmed that the observed differences in model performance were statistically significant. These findings reinforce the necessity of rigorous, comparative evaluation frameworks when interpreting and applying generative AI technologies in healthcare and clinical diagnosis in particular.

6.2 Final Reflections

The diagnostic performance of LLMs continues to improve rapidly, with recent advances demonstrating increasingly competent reasoning in clinical tasks. However, as capabilities evolve, so must the methods used to assess them. This thesis reinforces the

need for the next step in advanced benchmarking tools that can automate and scale evaluation across diverse and complex medical domains. Encouragingly, initial steps in this direction have already been taken in current trending research, including structured grading rubrics, and more advanced evaluation frameworks.

The field is clearly shifting away from constrained, multiple-choice question benchmarks and toward more complex and realistic clinical challenges. This study contributed to that shift by evaluating model performance on open-ended DDx tasks derived from authentic clinical case reports. Future evaluations should continue to expand into even more advanced tasks, such as patient history-taking, effective selection and justification of diagnostic tests, and patient-specific management strategies. These tasks more closely resemble real-world medical ones and demand not just correctness but clinical judgment as well.

Another emerging trend is the move beyond single-dimensional accuracy metrics. While Top-N scores remain useful for establishing baseline performance, they fail to capture the full spectrum of qualities required in clinical AI tools—such as reasoning transparency, rubric-based quality, communication clarity, patient empathy, and questioning techniques. The development of multi-dimensional evaluation frameworks will be essential for responsibly guiding LLM integration into clinical workflows.

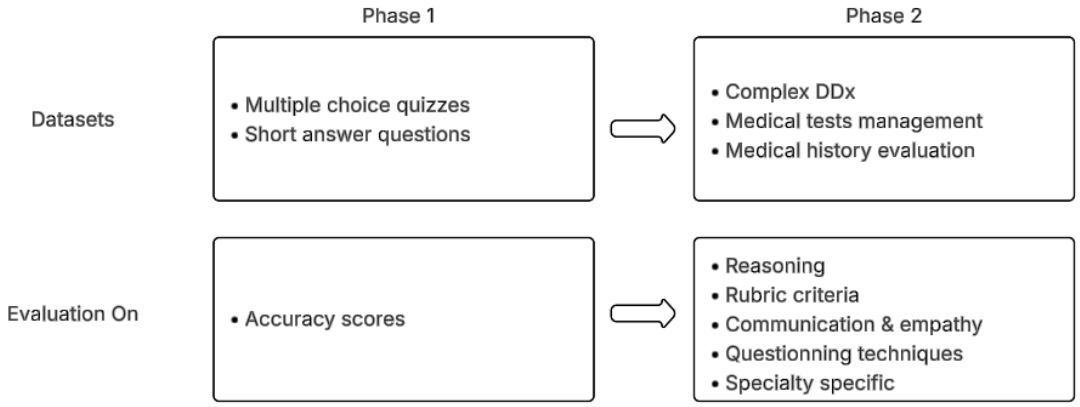

Figure 6.1: Transition for Phase 1 to Phase 2. Shifting from quiz questions towards complex real world scenarios. Shifting from one dimension evaluation scores towards advanced frameworks and quality evaluation with rubric criteria.

At the same time, the dominance of large proprietary models raises practical concerns about accessibility and computational cost. There is increasing interest in smaller, more efficient models that are fine-tuned for specific specialties or use cases. These lightweight models may offer a more sustainable and democratized path to clinical deployment, particularly in resource-limited settings.

Finally, despite the growing complexity of evaluation frameworks, simple zero-shot diagnostic tasks still hold value. They offer a quick and interpretable snapshot of the current status of generative AI capabilities. When combined with deeper evaluation

tools, these zero-shot assessments remain an important component of an evolving and layered benchmarking environment.

In summary, the future of diagnostic AI lies not only in more powerful models but also in the design of meaningful, transparent, and context-aware evaluation systems and frameworks. As models approach clinical utility, the emphasis must shift toward real-world applicability, responsible integration, and human-AI collaboration.

6.3 Data availability

Cases used, the models' responses, and their scoring per case are all available upon request.

7 Bibliography

1. Melnyk O, Ismail A, Ghorashi N, Heekin M, Javan R. Generative Artificial Intelligence Terminology: A Primer for Clinicians and Medical Researchers. Cureus. 2023 Dec 4;15.

2. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature [Internet]. 2025 Apr 9; Available from: https://doi.org/10.1038/s41586-025-08869-4

3. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2022. [NIPS '22].

4. Shahzad T, Mazhar T, Tariq MU, Ahmad W, Ouahada K, Hamam H. A comprehensive review of large language models: issues and solutions in learning environments. Discov Sustain. 2025 Jan 14;6.

5. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the potential of prompt engineering for large language models. Patterns. 2025 May;101260.

6. Naveed H, Khan A, Qiu S, Saqib M, Anwar S, Usman M, et al. A Comprehensive Overview of Large Language Models. 2023.

7. What Is Zero-Shot Learning? | IBM [Internet]. 2024 [cited 2025 Jun 2]. Available from: https://www.ibm.com/think/topics/zero-shot-learning

8. Scott I, Zuccon G. The new paradigm in machine learning - foundation models, large language models and beyond: a primer for physicians. Intern Med J. 2024 May 7;54.

9. Corazza GR, Lenti MV, Howdle PD. Diagnostic reasoning in internal medicine: a practical reappraisal. Intern Emerg Med. 2021 Mar;16(2):273–9.

10. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med. 2012 Feb;27(2):213–9.

11. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. 2025 Mar 1;31(3):932–42.

12. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour RE, et al. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. JAMA Intern Med. 2024 May 1;184(5):581–3.

13. Tripathi S, Alkhulaifat D, Doo F, Rajpurkar P, McBeth R, Daye D, et al. Development, Evaluation, and Assessment of Large Language Models (DEAL) Checklist: A Technical Report. NEJM AI. 2025 May 9;2.

14. Arora RK, Wei J, Hicks RS, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health [Internet]. 2025. Available from: https://arxiv.org/abs/2505.08775

15. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level

medical question answering with large language models. Nat Med. 2025 Mar 1;31(3):943–50.

16. Kanjee Z, Crowe B, Rodman A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. JAMA. 2023 Jul 3;330(1):78–80.

17. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. J Med Internet Res. 2024 Jun 14;26:e53297.

18. Brodeur P, Buckley T, Kanjee Z, Goh E, Ling E, Jain P, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. 2024.

19. Foote HP, Hong C, Anwar M, Borentain M, Bugin K, Dreyer N, et al. Embracing Generative Artificial Intelligence in Clinical Research and Beyond: Opportunities, Challenges, and Solutions. JACC Adv. 2025;4(3):101593.

20. Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology. NEJM AI. 2024 Jan 17;1.

21. K. Rostam Z, Szenasi S, Kertész G. Achieving Peak Performance for Large Language Models: A Systematic Review. IEEE Access. 2024 Jul 19;PP:96017–50.

22. Takita H, Kabata D, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, et al. A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians. Npj Digit Med. 2025 Mar 22;8(1):175.

23. Goh E, Gallo R, Hom J, Strong E, Weng Y, Kerman H, et al. Influence of a Large Language Model on Diagnostic Reasoning: A Randomized Clinical Vignette Study. medRxiv : the preprint server for health sciences. United States; 2024. p. 2024.03.12.24303785.

24. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature [Internet]. 2025 Apr 9; Available from: https://doi.org/10.1038/s41586-025-08866-7

25. Sumanathilaka D. Exploring the Impact of Temperature on Large Language Models: A Case Study for Classification Task based on Word Sense Disambiguation. Sumanathilaka D, Micallef N, Hough J, editors.

26. Zhou S, Lin M, Ding S, Wang J, Chen C, Melton GB, et al. Explainable differential diagnosis with dual-inference large language models. Npj Health Syst. 2025 Apr 24;2(1):12.

27. Tchango AF, Goel R, Martel J, Wen Z, Caron GM, Ghosn J. Towards Trustworthy Automatic Diagnosis Systems by Emulating Doctors' Reasoning with Deep Reinforcement Learning [Internet]. 2022. Available from: https://arxiv.org/abs/2210.07198

28. Armoundas AA, Loscalzo J. Patient agency and large language models in worldwide encoding of equity. Npj Digit Med. 2025 May 8;8(1):258.

29. Clinical Cases | The New England Journal of Medicine [Internet]. [cited 2025 Jun 2].

Available from: https://www.nejm.org/browse/nejm-article-category/clinical-cases

30. Yazdani S, Hoseini Abardeh M. Five decades of research and theorization on clinical reasoning: a critical review. Adv Med Educ Pract. 2019;10:703–16.

31. Hauna A, Yunus A, Fukui M, Khomsah S. Enhancing LLM Efficiency: A Literature Review of Emerging Prompt Optimization Strategies. Int J Robot Autom Sci. 2025 Mar 31;7:72–83.

8 Appendices

8.1 Source Code

8.1.1 OpenAI API integration

```
%%capture
!pip install openai httpx --upgrade --quiet
```

```
import asyncio
from openai import AsyncOpenAI
import os
import shutil

API_KEY = "sk****"
MODEL = "gpt-4o"
TEMPERATURE = 1.00
DRIVE_WORKSPACE = "/content/drive/MyDrive/ΕΜΠ/Master/My Thesis/PartB/colab/";
INSTRUCTIONS_PATH = DRIVE_WORKSPACE + "system_instructions.txt"
INPUT_DIR = DRIVE_WORKSPACE + "cases"
OUTPUT_FOLDER = "log"
OUTPUT_DIR = OUTPUT_FOLDER + "/" + MODEL

def clear_directory(directory_path):
    if os.path.exists(directory_path):
        for filename in os.listdir(directory_path):
            file_path = os.path.join(directory_path, filename)
            try:
                if os.path.isfile(file_path) or os.path.islink(file_path):
                    os.unlink(file_path)
                elif os.path.isdir(file_path):
                    shutil.rmtree(file_path)
            except Exception as e:
                print(f"Failed to delete {file_path}: {e}")
        try:
            os.rmdir(directory_path)
        except Exception as e:
            print(f"Failed to delete directory {directory_path}: {e}")


with open(INSTRUCTIONS_PATH, "r") as file:
    system_instructions = file.read()
```

```
clear_directory(OUTPUT_FOLDER)
os.makedirs(OUTPUT_DIR)
client = AsyncOpenAI(api_key=API_KEY)


async def main():
  for filename in os.listdir(INPUT_DIR):
    if not filename.endswith(".txt"):
      continue
    file_path = os.path.join(INPUT_DIR, filename)
    with open(file_path, "r") as file:
        case_description = file.read()
        base_filename = os.path.splitext(filename)[0]
        case_log_file = os.path.join(OUTPUT_DIR, f"{base_filename}.log")

        with open(case_log_file, "w") as log_file:
          content = [
              {"type": "text", "text": f"{case_description}\n"}
          ]

          response = await client.chat.completions.create(
              model=MODEL,
              messages=[
                  {"role": "system", "content": system_instructions},
                  {"role": "user", "content": content}],
              temperature=TEMPERATURE
          )

          answer_text = response.choices[0].message.content
          log_file.write(f"{answer_text}")

await main()
```

```
!zip -r /content/gpt-4o.zip /content/log/gpt-4o
from google.colab import files
files.download("/content/gpt-4o.zip")
```

8.1.2 Groq API integration

```
!pip install groq
```

```python
import asyncio
import os
import shutil
from groq import AsyncGroq

API_KEY = "sk****"
MODEL = "qwen-2.5-32b"
TEMPERATURE = 1.00
DRIVE_WORKSPACE = "/content/drive/MyDrive/ΕΜΠ/Master/My Thesis/PartB/colab/";
INSTRUCTIONS_PATH = DRIVE_WORKSPACE + "system_instructions.txt"
INPUT_DIR = DRIVE_WORKSPACE + "cases"
OUTPUT_FOLDER = "log"
OUTPUT_DIR = OUTPUT_FOLDER + "/" + MODEL


def clear_directory(directory_path):
    if os.path.exists(directory_path):
        for filename in os.listdir(directory_path):
            file_path = os.path.join(directory_path, filename)
            try:
                if os.path.isfile(file_path) or os.path.islink(file_path):
                    os.unlink(file_path)
                elif os.path.isdir(file_path):
                    shutil.rmtree(file_path)
            except Exception as e:
                print(f"Failed to delete {file_path}: {e}")
        try:
            os.rmdir(directory_path)
        except Exception as e:
            print(f"Failed to delete directory {directory_path}: {e}")


with open(INSTRUCTIONS_PATH, "r") as file:
    system_instructions = file.read()


clear_directory(OUTPUT_FOLDER)
os.makedirs(OUTPUT_DIR)


async def main():
  client = AsyncGroq(api_key=API_KEY)
  for filename in os.listdir(INPUT_DIR):
    if not filename.endswith(".txt"):
```

```
      continue
    file_path = os.path.join(INPUT_DIR, filename)
    with open(file_path, "r") as file:
        case_description = file.read()
        base_filename = os.path.splitext(filename)[0]
        case_log_file = os.path.join(OUTPUT_DIR, f"{base_filename}.log")

        with open(case_log_file, "w") as log_file:

          content = [
              {"type": "text", "text": f"{case_description}\n"}
          ]

          stream = await client.chat.completions.create(
              model=MODEL,
              messages=[
                  {"role": "system", "content": system_instructions},
                  {"role": "user", "content": content}
              ],
              temperature=TEMPERATURE,
              max_completion_tokens=4096,
              stream=True,
              stop=None
          )

          response_text = ""
          async for chunk in stream:
              response_text += chunk.choices[0].delta.content or ""

          log_file.write(f"{response_text}")

await main()
```

```
!zip -r /content/qwen-2.5-32b.zip /content/log
from google.colab import files
files.download("/content/qwen-2.5-32b.zip")
```