Declaration of Originality

Unit Name: Advanced Artificial Intelligence Research Topics COMP6011

I/We declare that:

- The above information is complete and accurate.
- The work I am/We are submitting is entirely my/our own, except where clearly indicated otherwise and correctly referenced.
- I/We have taken (and will continue to take) all reasonable steps to ensure my/our work is not accessible to any other students who may gain unfair advantage from it.
- I/We have not previously submitted this work for any other unit, whether at Curtin University or elsewhere, or for prior attempts at this unit, except where clearly indicated otherwise.

I/We understand that:

- Plagiarism and collusion are dishonest and unfair to all other students.
- Detection of plagiarism and collusion may be done manually or by using tools (such as Turnitin).
- If I/We plagiarise or collude, I/We risk failing the unit with a grade of ANN ("Result Annulled due to Academic Misconduct"), which will remain permanently on my academic record. I/We also risk termination from my/our course and other penalties.
- Even with correct referencing, my/our submission will only be marked according to what I/we have done myself/ourselves, specifically for this assessment. I/We cannot re-use the work of others, or my/our own previously submitted work, in order to fulfil the assessment requirements.
- It is my/our responsibility to ensure that the submission is complete, correct, and not corrupted.

Date: May 4, 2025

Jain	**Hardik Ashish**	**21657791**	

REPORT 2: Assessing the clinical diagnosis accuracy of LLMs

Hardik Ashish Jain, *Master of Computing, 21567791*

Abstract—The use of large language model has become mainstream due to strong generalisation capabilities that it possess across disciplines. Its initial commercial applications however faced significant litigation due to factual inaccuracies, inconsistencies and bias contained in the generated output. These inaccuracies were understood to be a trade off which has to be made in data driven approaches. Recent advancements exploring the reasoning capabilities in LLMs however has the potential to significantly overcome these limitations and improve factual accuracy which would enable large scale adoption of LLMs in disciplines such as healthcare, finance and legal services where factual accuracy and model's consistency are key. In this study, we evaluate the performance of recent state of the art models having reasoning capabilities against models without reasoning capabilities on the task of clinical diagnosis of disease. Ten clinical assessments were passed to the models GPT-4o, Gemini-Pro-2.5, Grok 3, Deepseek R1, o4-mini with suitable prompt and the output was scored using three categories: fully correct, partially correct and incorrect using a majority voting compared against human specialist diagnosis. The models Qwen 2.5 Max, Claude 3.7 Sonnet (thinking), and Llama 4 (Maverick 13.7B) were used for scoring the generated results closeness to actual diagnosis. For judgement models, an attempt was made to select state of the art reasoning LLMs which aren't similar to the ones being evaluated or designed by same company to avoid any unintended bias in judgement which occurs due to similar training data and/or strategy usage. Our assessments show that reasoning models overall beat their non-reasoning counterparts on specialised tasks like clinical assessments with GPT-4o in the lead, followed by Grok3, o4-mini, Gemini Pro 2.5 and Deepseek R1 in that order. Reasoning models consume significantly more resources than their non-reasoning counterparts and have significantly higher associated costs. Based on our estimates, the per query cost ranges from as low as 0.12 cents for Deepseek R1 to as high as 2.25 cents (approx. 18 times more expensive) for Grok3. Their inference time is also very high with 34.6s taken on average per query with smaller models such as o4-mini looking promising at only 6s/query.

Index Terms—Reasoning, LLM, Clinical diagnosis.

◆

1 INTRODUCTION

THE race for developing better large language models (LLMs) has been ongoing since the introduction of transformers for machine translation. Since then, LLMs have been demonstrated to perform exceptionally well on categories of natural language processing (NLP) tasks such as text summarisation, sentiment analysis, text classification, question answering, named entity recognition, semantic text similarity and many more. These developments were enabled by the strong generalisation and scaling capabilities which the transformer architecture offered. Many foundation models were created using unsupervised techniques which can be further refined for downstream tasks after some supervised fine-tuning. However, the introduction of commercial applications such as ChatGPT provided more impetus to grow interest in LLMs. ChatGPT when initially released was based on GPT3.5 model which showcased that multiple tasks can be handled by a single model. Although the model had strong creativity, it had limited factual accuracy, consistency and arithmetic capability. While this isn't a problem in many fields like creative writing, in other fields like healthcare, legal services, and research, these capabilities are non-negotiable. Many individuals and organisations also registered complaints and legal cases for incorrect factual information[1]. This led to development of reasoning capabilities in LLMs which makes the model think by breaking down a task into multiple intermediate steps before responding. The chain-of-thoughts concept in this regard is demonstrated to significantly enhance model performance on many challenging tasks such as competitive programming, Math Olympiad and surpassing human PhD benchmark in Sciences.

The field of health care requires the model's output to be explainable and factually correct/consistent so that a physician can trust the system and adopt it. In this regard the chain-of-thoughts process which breaks down a larger task into multiple intermediate steps before responding with a final output is useful as it lets a layperson trace how a model came to a particular response. To test this hypothesis, we select five state of the art reasoning models: Gemini-Pro-2.5, Grok3, Deepseek R1, GPT-4o, o4-mini which currently are leading in Chatbot Arena ranking (22nd April). Most of these models were released in 2025. These models' capabilities for clinical diagnosis would be estimated by performing a small experiment with ten cases published in New England Journal of Medicine (NEJM) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. These cases had the case description and diagnosis provided which will be used as input X and expected output y. The model's generated output $\hat{y}$

- *Hardik Jain is a Master of Computing, AI Major at Curtin University. E-mail: h.jain2@student.curtin.edu.au*

Report started April 5; submitted May 4.

[1] This portal provides a comprehensive list of most of the notable cases: https://sustainabletechpartner.com/topics/ai/generative-ai-lawsuit-timeline/

would be stored and evaluated against y using a scoring function $f(y, \hat{y})$. The text generation and evaluation process is mostly automated and more details on the exact process in described in section 4. Besides the scoring function, we also techniques such as BertScore which uses negative cosine similarity between expected output embedding and generated output embedding for measuring closeness between the two.

In this study our primary contributions are:

- Understanding the capability of zero-shot generalised reasoning LLMs for clinical diagnosis task
- Estimating the cost and inference speed per query for clinical diagnosis case
- Identifying the limitations, and improvements that can be made in future studies

2 LITERATURE REVIEW

2.1 NLP, Transformers and LLMs

The field of NLP has evolved significantly since its early days as shown in Figure 1. The first models, used statistical approaches such as n-gram developed in 1948 by Shannon and Weaver [11]. In 1957, linguists like Noam Chomsky [12] proposed rule-based systems relying on manually crafted grammars and dictionaries. These models excelled at structured tasks but were limited by their inability to handle linguistic variability. In 1970, hidden Markov models by Baum et al. [13] leveraged probabilistic approaches for improving robustness; however, they struggled with long-range dependencies. Neural approaches began with Word2Vec in 2013 [14], offering dense word embeddings that captured semantic relationships, though lacking sequential context. Recurrent neural networks (RNNs), notably Long Short Term Memory (LSTM) which were proposed earlier in 1997 [15], didn't face these issues as they modelled sequences. LSTM significantly improved context retention, yet faced scalability issues due to vanishing gradients. The idea of attention, was first proposed by Bahdanau et al. in 2014 [16] as a modification to RNNs for machine translation which allowed models to focus on relevant input parts, enhancing performance but retaining sequential limitations. In 2017, Vaswani et al. [17] proposed the transformer architecture in now famous work: "Attention is All You Need". Transformer design proposed removing all sequential components like RNNs and CNNs and rely solely on self-attention, transformers enabled parallel processing, significantly boosting scalability and performance in tasks like machine translation.

The transformer architecture (Figure 2) processes an input sequence X through a series of mathematical operations. Initially, X is passed to a tokenizer, converting text into a sequence of tokens, which are then embedded into a high dimensional vector space $\mathbb{R}^d$. Positional encodings, sinusoidal functions of position, are added to these embeddings to address the lack of order in self-attention, providing sequence information critical for language understanding. The encoder, comprising N identical blocks, transforms these vectors using multi-head self-attention, defined as $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$, where Q, K, and V are query, key, and value matrices derived from the input, and d_k is the key dimension. This is followed by a feed-forward network, with residual connections and layer normalisation enhancing training stability. The decoder, also with N blocks, includes an additional encoder-decoder attention layer, generating output auto regressively and predicting each token based on prior tokens using masked self-attention to prevent attending to future tokens. The encoder-decoder structure predates transformers, standard in sequence-to-sequence tasks like translation, enabling input-output mapping. Primary innovations in the transformer paper include a) multi-head attention, b) positional encodings, and c) subword tokenization. Training time complexity for self-attention is $O(n^2 \cdot d)$ over sequence length n, while inference time complexity is $O(n \cdot d)$. With 65 million parameters, the original transformer achieved a BLEU score of 28.4 on WMT 2014 English-to-German translation, outperforming RNNs and CNNs in experiments [17]. It used cross-entropy loss, the Adam optimiser, a custom learning rate schedule, and the WMT corpus.

The transformer's success spawned derivatives like BERT and GPT, which adapted its architecture for specialised tasks. BERT, introduced in 2018 [18], uses only the encoder, trained bidirectionally on masked language modelling, excelled in comprehension tasks like text classification, sentiment analysis, question answering due to its contextual depth. Subsequent works explored BERT model variants with different subject specific corpus [19], [20], [21], [22]. GPT [23] was also introduced in the same year and employed the decoder, trained auto-regressively for generation, leveraging its sequential prediction strength for coherent text output. Recent works address transformers' quadratic complexity. The decoder heavy model had strong text generation capability and excelled in multiple tasks simultaneously. Almost all modern LLMs are inspired from this design. The Reformer [24] uses locality-sensitive hashing, approximating attention to $O(n \log n)$, while the Linformer [25] projects keys and values to lower dimensions, achieving $O(n)$ complexity. Sparse transformers [26] apply sparse attention patterns, reducing memory demands. These advancements enhance efficiency, critical for scaling LLMs. However, LLMs often lack multi-step reasoning and thus perform significantly worse in multiple disciplines, necessitating further research.

2.2 Reasoning capability in LLMs

LLMs are mostly data driven approaches which do not have an inherent reasoning capability which OpenAI aimed to address when the released their model o1 in September 2024. Among its many impressive accomplishments, o1 ranked in the 89th percentile for competitive programming questions on Codeforces compared to 11th percentile for GPT4o, scored 13.9/15 in the Math Olympiad qualifier (AIME) which places it within the top 500 US students against 1.8/15 scored by GPT4o, and surpassed human PhD-level accuracy on the GPQA science benchmark[2]. Other companies soon followed this trend with Google working on

---

[2] OpenAI o1 blog: https://openai.com/index/learning-to-reason-with-llms/

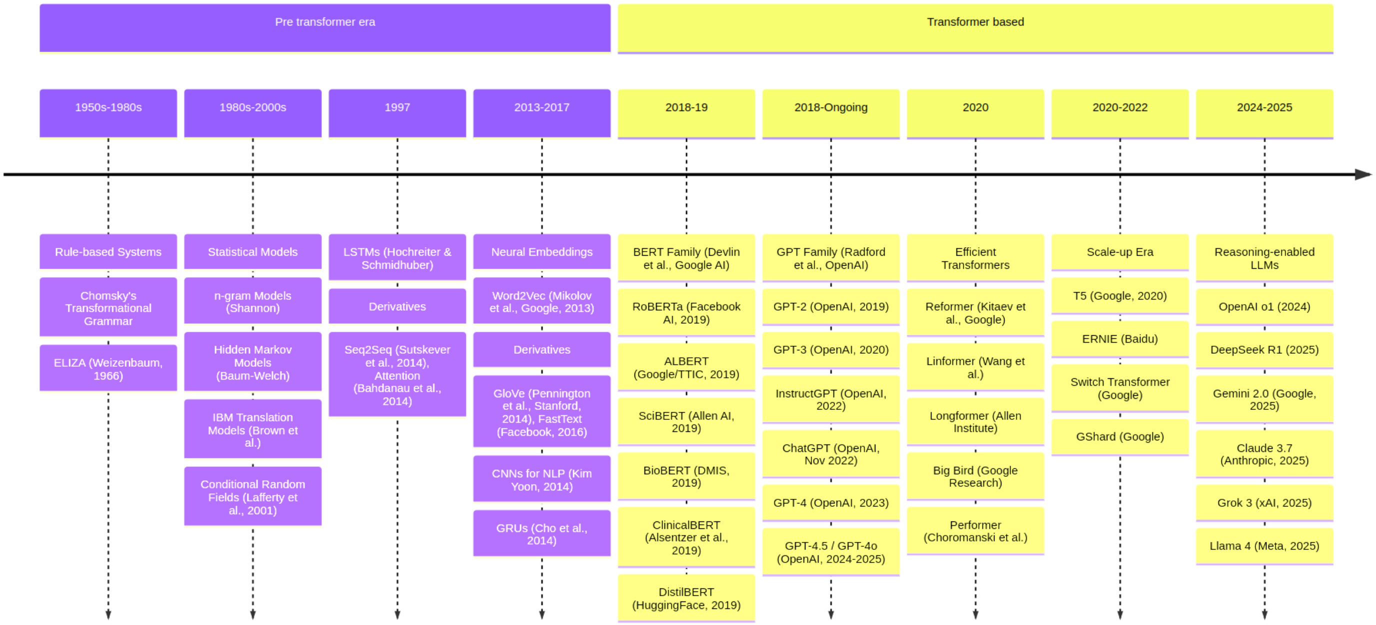
Pre transformer era
Transformer based
1950s-1980s
1980s-2000s
1997
2013-2017
2018-19
2018-Ongoing
2020
2020-2022
2024-2025
Rule-based Systems
Chomsky's Transformational Grammar
ELIZA (Weizenbaum, 1966)
Statistical Models
n-gram Models (Shannon)
Hidden Markov Models (Baum-Welch)
IBM Translation Models (Brown et al.)
Conditional Random Fields (Lafferty et al., 2001)
LSTMs (Hochreiter & Schmidhuber)
Derivatives
Seq2Seq (Sutskever et al., 2014), Attention (Bahdanau et al., 2014)
Neural Embeddings
Word2Vec (Mikolov et al., Google, 2013)
Derivatives
GloVe (Pennington et al., Stanford, 2014), FastText (Facebook, 2016)
CNNs for NLP (Kim Yoon, 2014)
GRUs (Cho et al., 2014)
BERT Family (Devlin et al., Google AI)
RoBERTa (Facebook AI, 2019)
ALBERT (Google/TTIC, 2019)
SciBERT (Allen AI, 2019)
BioBERT (DMIS, 2019)
ClinicalBERT (Alsentzer et al., 2019)
DistilBERT (HuggingFace, 2019)
GPT Family (Radford et al., OpenAI)
GPT-2 (OpenAI, 2019)
GPT-3 (OpenAI, 2020)
InstructGPT (OpenAI, 2022)
ChatGPT (OpenAI, Nov 2022)
GPT-4 (OpenAI, 2023)
GPT-4.5 / GPT-4o (OpenAI, 2024-2025)
Efficient Transformers
Reformer (Kitaev et al., Google)
Linformer (Wang et al.)
Longformer (Allen Institute)
Big Bird (Google Research)
Performer (Choromanski et al.)
Scale-up Era
T5 (Google, 2020)
ERNIE (Baidu)
Switch Transformer (Google)
GShard (Google)
Reasoning-enabled LLMs
OpenAI o1 (2024)
DeepSeek R1 (2025)
Gemini 2.0 (Google, 2025)
Claude 3.7 (Anthropic, 2025)
Grok 3 (xAI, 2025)
Llama 4 (Meta, 2025)

Fig. 1. This figure was created by us and explores the most significant early techniques and present state of the art models/approaches and covers seven decades of progress in the field of natural language processing.

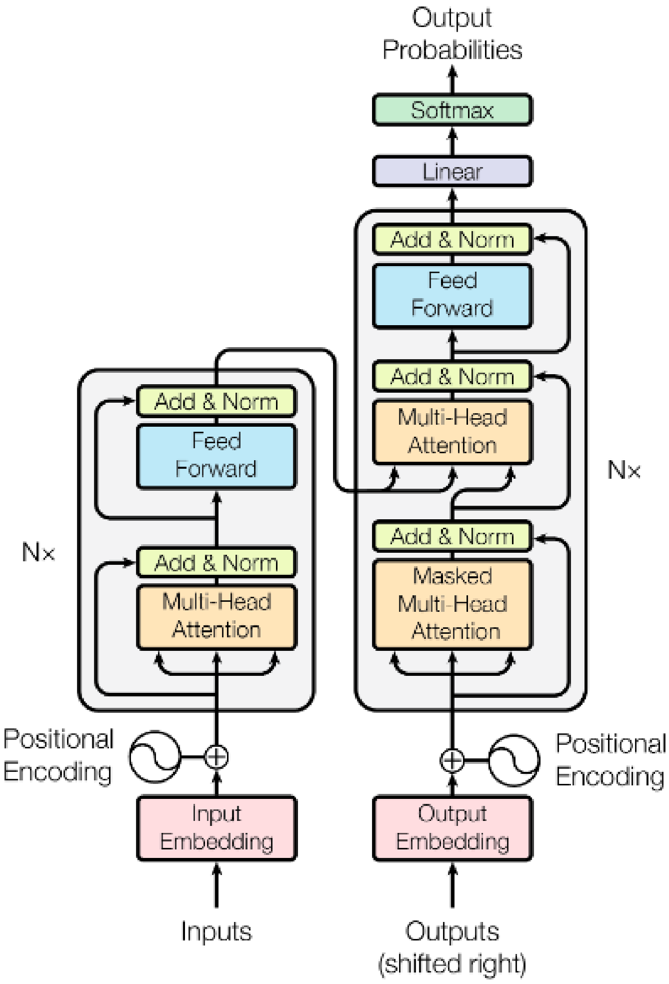

Fig. 2. Transformer model architecture proposed by Vaswani et al. [17]

Gemini 2.5[3], Anthropic working on Claude 3.7[4] and xAI developing Grok3[5] to have thinking capabilities. Amongst open source models, Deepseek R1 [27] was the first, followed by Google's Gemma 3 [28] and Meta's Llama 4[6] more recently. These advancements were achieved in something now known as post-training steps. Reinforcement learning algorithms are used to teach models how to think productively. In OpenAI o1 blog, it was mentioned that the performance of the model consistently improved with more reinforcement learning (train-time compute) and with more time spent thinking (test-time compute) as shown in Figure 3. The use of additional time by model when solving difficult reasoning problems was termed as **chain-of-thought** (CoT) by OpenAI with the analogy given that humans take good time thinking before responding to a difficult question.

2.3 chain-of-thought

CoT is a means of demonstrating intermediate steps before a model can arrive to a final response (see Appendix 3). It has been under discussion for some time now since a **prompt based strategy** was first proposed in 2022 [29]. In this strategy, multiple human annotators created a small scale dataset as shown in Table 1 which was then passed to the model for few shot learning. The model then was asked to solve all arithmetic problems based on similar strategy by breaking down the output into sequences of intermediate steps. This approach was further improved in 2023 when the benefit of having multiple reasoning paths

[3] Google blog: https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/
[4] Anthropic blog: https://www.anthropic.com/claude/sonnet
[5] xAI blog: https://x.ai/news/grok-3
[6] Meta blog: https://ai.meta.com/blog/llama-4-multimodal-intelligence/

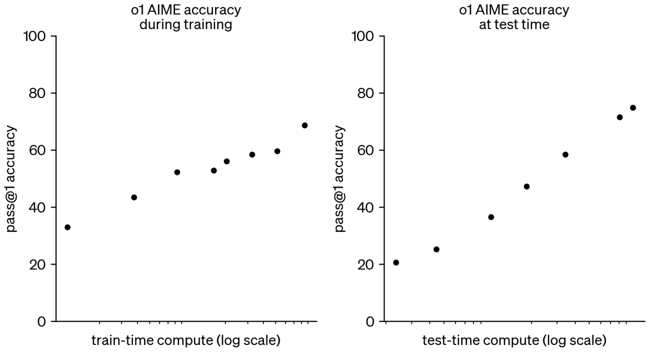

Fig. 3. Performance improvement for the o1 model based on time spent in training and during inference. The figure was derived from OpenAI o1 blog.

was demonstrated, termed as self-consistency [30]. In self-consistency, the model is once again trained on a small subset similar to Wei et al. [29], followed by sampling a set of candidate outputs from the LLM's decoder. In doing this, there are multiple candidate output which the authors state represents multiple reasoning pathways. The answers from these candidates is then aggregated by choosing the answer which is most consistent among these candidates. While self-consistency gave substantial improvements in performance it also increased computation requirements several times.

2.4 Reinforcement learning based reasoning

OpenAI specified that RL helps the model in refining the strategies that can be employed and most importantly recognising and correcting its mistakes by breaking down tricky steps into simpler ones (CoT). However the technical specifics of how a reasoning model can actually be implemented wasn't described by OpenAI. Multiple studies have previously attempted doing the same to attain marginal performance improvements but was shown to be limited by 1) **behavioural collapse** and 2) **distribution shift** [31], [32], [33], [34]. As per Kumar et al. [35][7], distribution shift can be resolved by shifting from off-policy to on-policy behavioural collapse cannot and multiple consecutive attempts will only lead to a point where the model doesn't change its response anymore. Kumar et al. [35] thus proposed a new technique called SCoRe, designed to explicitly encourage self-correction behaviour. This two-stage approach begins with Stage I, where the model is initialised by training it to optimise second-attempt accuracy while keeping the first attempts similar to the base model. In Stage II, both attempts are jointly optimised, with a bias in the reward to ensure the model progresses through self-correction rather than collapsing to a direct solution.

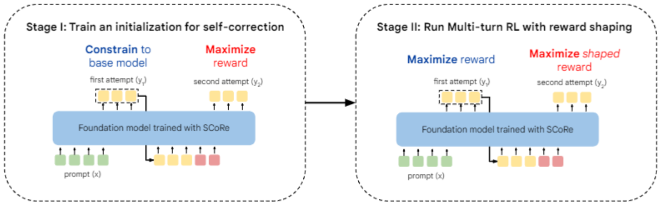
Stage I: Train an initialization for self-correction
Constrain to base model
Maximize reward
first attempt (y1)
second attempt (y2)
Foundation model trained with SCoRe
prompt (x)
Stage II: Run Multi-turn RL with reward shaping
Maximize reward
Maximize shaped reward
first attempt (y1)
second attempt (y2)
Foundation model trained with SCoRe
prompt (x)

Fig. 4. Overview of SCoRe approach [35]. SCoRe trains a model in two stages: Stage I: Instead of using supervised fine tuning (which can amplify biases) to start RL training, a good initial model is created that can produce high-reward responses on the second attempt while mimicking the base model's initial response on the first attempt. Stage II: We optimise both attempts together. The second attempt uses a shaped reward to encourage finding a self-correction strategy, rather than just making minor edits to the first response.

2.5 Reinforcement Learning techniques

[8] Large language models (LLMs) are commonly pre-trained on vast corpora, such as web pages or books, using cross-entropy loss to predict the next token in a sequence. This supervised fine-tuning (SFT) approach optimises the model to maximise the likelihood of the training data, producing readable and fluent outputs. The cross-entropy loss for a sequence $X = (x_1, x_2, \ldots, x_n)$ is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{n} \log p(x_t | x_{<t}) \tag{1}$$

where $p(x_t|x_{<t})$ represents the model's predicted probability of token x_t given prior tokens. While this method excels at generating coherent text, it does not inherently align outputs with human preferences or task-specific goals, prompting exploration into further optimisation strategies. To enhance alignment with human preferences, an initial idea is to fine-tune the model solely on human-labelled responses, abandoning cross-entropy loss. However, this approach often results in unreadable outputs, as the model overfits to the limited labelled dataset and loses the linguistic structure acquired during pre-training. This underscores the challenge of balancing alignment with readability, driving the development of RL-based techniques.

2.5.1 ORPO

One effective solution is **Odds Ratio Preference Optimization (ORPO)** [36], which was also used in TULU 3 model [37]. ORPO integrates cross-entropy loss with an odds ratio term to favour chosen responses over rejected ones, using a dataset containing correct (SFT-like), chosen, and rejected answers. The ORPO loss function is:

$$\mathcal{L}_{\text{ORPO}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \log \left(\frac{p(\text{chosen}|x_{<t})}{p(\text{rejected}|x_{<t})} \right) \tag{2}$$

where λ balances the objectives. The cross-entropy term preserves readability, while the odds ratio adjusts the probability distribution to align with human preferences, offering a practical improvement over naive fine-tuning.

---

[7] This was a Google Deepmind paper which was released just a week after OpenAI's o1 blog post, highlighting the intense competition that currently exists as companies attempt to develop new techniques at break neck speed.

[8] We acknowledge the use of resources by Trelis Research, an educational content creation company for understanding the logic and math behind many of the techniques in this section https://trelis.com/.

TABLE 1
The below is a question taken from GSM8k dataset and CoT based model's answer derived from Wei et al. [29] Table 8, Appendix D.1

QUESTION: Brandon's iPhone is four times as old as Ben's iPhone. Ben's iPhone is two times older than Suzy's iPhone. If Suzy's iPhone is 1 year old, how old is Brandon's iPhone?

MODEL ANSWER: Brandon's iPhone is 4 times as old as Ben's iPhone. Ben's iPhone is 2 times older than Suzy's iPhone. So Brandon's iPhone is 4 x 2 = 8 times older than Suzy's iPhone. Suzy's iPhone is 1 year old. So Brandon's iPhone is 8 x 1 = 8 years old. The answer is 8.

2.5.2 TRPO

Another approach, **Trust Region Policy Optimization (TRPO)** [38], uses KL divergence to ensure the updated policy π_θ remains close to the original $\pi_{\theta_{\text{old}}}$, preserving readability. Its objective is:

$$\mathbb{E}\left[\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)}A(s,a)\right] - \beta \cdot \mathrm{KL}(\pi_{\theta_{\text{old}}}||\pi_\theta) \tag{3}$$

where $A(s,a)$ is the advantage function, and β controls the constraint strength. By limiting distribution shifts, TRPO maintains linguistic coherence while optimizing for human feedback.

2.5.3 PPO

Proximal Policy Optimization (PPO)[9] refines TRPO by clipping policy updates within a small range defined by ϵ, simplifying computation while retaining benefits. The PPO objective is:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}\left[\min\left(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right] \tag{4}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$, and A_t is the advantage. This makes PPO a practical choice for LLM optimization.

Before applying RL for post-training optimization, key components required are:

- A pre-trained model, typically fine-tuned via SFT.
- A dataset or reward function defining desirable and undesirable behaviours.
- A value model (a neural network) to estimate future rewards, enabling the advantage calculation $A = r - V$, where r is the reward and V is the value estimate.

2.5.4 DRPO

In 2024, a year before introducing their R1 model, Deepseek published a paper called DeepSeekMath which focused on improving LLM accuracy on mathematical problem [39]. This paper introduces **Deep Reinforcement Preference Optimization (DRPO)** algorithm which is current state of the art open source technique in RL based LLM optimisation. DRPO eliminates the value model requirement in ORPO, TRPO and PPO by using a group-based baseline. Unlike ORPO's pairwise comparison, DRPO optimises across a group (e.g., eight answers in Deepseek R1), calculating the advantage as:

$$A = r_{\text{this}} - \frac{1}{7}\sum_{i=1}^{7} r_i \tag{5}$$

where r_{this} is the current answer's reward, and r_i are the others' rewards. Larger groups improve stability but increase VRAM demands.

Rewards are typically composite, summing accuracy and format compliance:

$$r = r_{\text{accuracy}} + r_{\text{format}} \tag{6}$$

Accuracy is measured via numerical comparison (using ground truth or another LLM), while format compliance may involve regex checks for tags like $\langle$think$\rangle$.

The main steps in DRPO involve:

1) Sampling a group of responses.
2) Assigning rewards based on accuracy and format.
3) Backpropagating using the group's reward baseline to compute the advantage.

This leverages collective performance for stable optimization.

While performing post-training on LLMs using DRPO isn't feasible out of the box unless one has access to high end GPUs like A100 or H100 due to high memory requirements, there exist few popular implementations by groups such as Unsloth[10] which significantly reduce memory requirements and enables one to run DRPO even in colab environment. Unsloth DRPO incorporates memory optimizations (e.g., gradient checkpointing or mixed precision) to enable reasoning model creation from standard LLMs, though specifics are limited. It supports larger group sizes in DRPO, enhancing training efficiency. Post-training, reason-based models optimised via RL (e.g., ORPO, DRPO) do not differ from standard LLMs during inference. Both generate text token-by-token autoregressively. However, RL fine-tuning enhances their reasoning and format adherence, reflected in output quality rather than inference mechanics.

2.6 Clinical diagnosis and use of LLM

Clinical diagnosis is a critical process in medicine where physicians identify a patient's condition based on signs, symptoms, medical history, and diagnostic tests. This process typically begins with a detailed patient interview and physical examination, followed by laboratory tests, imaging, or other diagnostic procedures as needed. Physicians rely on their clinical expertise, pattern recognition, and differential diagnosis, i.e. a systematic method to narrow down possible conditions by comparing and contrasting clinical findings. To standardise diagnoses globally, the World Health Organization (WHO) maintains the International Classification of

---

[9] Just observed that Alec Radford the first author of GPT-1 paper is also a co-author in this paper.

[10] Interestingly, the two individuals behind Unsloth are Australian though they are currently based in California, they were previously associated with UNSW.

Diseases (ICD)[11], currently in its 11th revision (ICD-11). The ICD provides a comprehensive framework for classifying diseases, disorders, and health conditions, ensuring consistent terminology and coding across healthcare systems. This classification aids in epidemiological tracking, healthcare management, and clinical decision-making, though it does not replace the nuanced judgement required in individual patient care.

State-of-the-art approaches leveraging artificial intelligence have significantly advanced clinical diagnosis. Initial approaches mostly focused on medical imaging. In 2016, Gulshan et al. from Google Research developed a deep learning model using Inception-v3, achieving an AUC of 0.99 for diabetic retinopathy detection on a dataset of 128,175 retinal images [40]. Advancements leveraging NLP in clinical setting include Clinical-T5, which a T5-based model trained on MIMIC-III and IV dataset. Additionally, in 2023, Chen et al. [41], proposed DxFormer, a transformer model trained on three public real-world medical dialogue datasets: Dxy, MZ-4 and MZ-10, excelling in multi-disease prediction with an accuracy of 87.5%. However, these models face critical limitations. For instance, Gulshan's model requires extensive labelled data, unavailable for less common conditions, and thus would struggle with generalisation across diverse imaging modalities. LLMs are also known to exhibit a performance drop on out-of-domain datasets, and DxFormer's opaque reasoning hampers clinician trust, as noted in a 2023 evaluation [42]. Bias remains a concern, with lower accuracy for underrepresented demographics. A recent differential diagnosis approach shows promise where the authors demonstrate a model's top 10 accuracy to have outperformed human physician accuracy on clinical diagnosis [43]. However, from an operational perspective it would also be helpful if a model's top1 accuracy can also perform comparable or better than human physician performance. These shortcomings highlight the potential of reason-enabled LLMs to enhance transparency and adaptability in diagnostic applications.

3 BENCHMARKING RESULTS

3.1 Metrics

BLEU: Bilingual Evaluation Understudy computes the precision of n-gram overlaps between candidate and reference texts, applying a brevity penalty to penalise overly short outputs. **ROUGE-L**: Recall-Oriented Understudy for Gisting Evaluation-LCS measures the longest common subsequence between candidate and reference, emphasising recall of contiguous matching sequences. **BERTScore**: Leverages contextual token embeddings from pretrained Transformer models to compute soft cosine similarity between candidate and reference tokens, capturing semantic and syntactic alignment. The original bertscore uses `bert-base-uncased` but since this is clinical setting, we further evaluate it using `emilyalsentzerBio_ClinicalBERT` checkpoint. This checkpoint was trained on MIMIC discharge summary.

It is difficult to assess and rank LLMs on traditional metrics such as BLEU, ROUGE, etc. This is because LLMs are very creative and thus the semantic similarity of generated output with the input should be measured instead. Semantic similarity was measured using negative cosine similarity. Furthermore, LLMs output is often looked into for granular details such as factual correctness, readability, fluency, relevance, bias and fairness with the use of automated LLMs or human annotators. Since we had limited time and resources, we use independent LLMs for judging accuracy of generated diagnosis against reported diagnosis by a three step scoring criteria which gives full score (10) for complete alignment, partial score (5) for partial alignment and no score otherwise.

3.2 Qualitative performance

The diagnosis as reported in the NEJM, and as generated by the models GPT-4o, Gemini-Pro-2.5, Deepseek R1, Grok3, o4-mini are shown in Tables 2 and 3. They were split into two tables due to space constraints. It can be observed that most models produce and output, irrespective of whether sufficient details are present or not to make an assessment. Gemini differed in this, as it would specify couldn't be determined in such cases.

3.3 Quantitative performance

We first performed the evaluation on popular NLP metrics such as BLEU, Rouge-L, and BertScore (Table 3.3). We noticed significant limitations as the task was to predict one or two words describing the clinical diagnosis which is challenging to be evaluated by these metrics. BLEU which works by use of n-grams completely failed and resulted in zeroes. Rouge-L still performed significantly better and its results more or less align with independent LLM judgement performed in Table 3.3. BertScore also provided aligns with independent LLM judgement but shows lesser variability compared to ROUGE-L. ClinicalBertScore shows almost no variation and has practically no use.

The generated diagnosis were evaluated by independent LLM as shown in Table 3.3 with GPT-4o achieving the best score at 30, followed by Grok3 and o4-mini at 25. Gemini (15) and Deepseek (10) performed significantly worse for this task. We also experimented evaluation of generated output by similar architecture and observed these LLMs score similar design output better which could indicate possible bias. We also include evaluation by disabling thinking (few models didn't support it) and note slightly worse performance as shown below.

3.4 Inference speed and cost per query

In terms of practical usage, inference speed and costs associated per query are also significant influences besides accuracy (Table 6). We observed that costs vary widely with Deepseek R1's cost being the lowest at 0.12 cents and Grok3's cost of 2.15 cents being the highest. Models such as o4-mini which offer good accuracy (25) at relatively low cost (0.44 cents) might be a good trade-off between the two.

---

[11] https://www.who.int/standards/classifications/classification-of-diseases

TABLE 2
The reported clinical and final diagnosis in New England Journal of Medicine, GPT-4o and Gemini-Pro-2.5 preview model generated diagnosis for each of the ten cases.

Reported		GPT-4o		Gemini	
Clinical	Final	Clinical	Final	Clinical	Final
Metastatic melanoma	Pulmonary histoplasmosis	Recurrent metastatic melanoma	Disseminated histoplasmosis	Suspected metastatic melanoma	Paracoccidioidomycosis
Not available	Human immunodeficiency virus type 2 infection and cerebral toxoplasmosis	Cerebellar metastasis	Cerebral toxoplasmosis	Intracranial Mass Lesions and AIDS	Can’t be determined
Not available	Borrelia miyamotoi infection	Autoimmune encephalitis or CNS vasculitis	Chronic Toxoplasma gondii meningoencephalitis	Meningoencephalitis	Primary CNS Lymphoma (PCNSL)
Not available	Vitamin D deficiency	Hypocalcemia with seizure, possibly nutritional rickets	Nutritional rickets due to vitamin D deficiency	Hypocalcemic Seizure	Can’t be determined
Not available	Systemic primary amyloidosis	Transient monocular vision loss, possible giant cell arteritis (GCA) or orbital mass	IgG4-related disease with orbital and hepatobiliary involvement	Suspected Giant Cell Arteritis with orbital mass	Systemic Sarcoidosis
Cutaneous bacterial infection	Erysipelothrix rhusiopathiae infection	Necrotizing fasciitis	Group A Streptococcus necrotizing fasciitis	Cellulitis with lymphangitis	Aeromonas hydrophila infection
Not available	Inflammatory bowel disease (Crohn’s disease)	Reactive arthritis	Crohn’s disease	Systemic inflammatory response with arthritis and enterocolitis.	Crohn’s disease.
Not available	Aspiration pneumonia	Community-acquired pneumonia	Invasive pulmonary aspergillosis	Infectious Bronchiolitis and Pneumonia	Invasive Pulmonary Aspergillosis
Not available	Granulomatosis with polyangiitis	Community-acquired pneumonia	Anti-glomerular basement membrane disease	Pulmonary-renal syndrome	Can’t be determined
Sarcoidosis or tuberculosis	Sarcoidosis	Uveitis	Sarcoidosis	Tuberculous uveitis	Can’t be determined

4 METHODOLOGY

4.1 Models selection

ChatBot Arena (also known as LMArena) [45] has emerged as a notable platform for ranking LLMs. This platform features anonymous, randomised battles where LLMs are released by some of the biggest companies before their release and compete against each other in a crowd-sourced manner. The rankings are determined using the Elo rating system, which is widely used in competitive games like chess. Users can contribute by asking questions and voting for their favourite responses, making the evaluation process highly interactive and community-driven. The leaderboard is periodically updated to reflect the latest performance metrics, ensuring that the rankings remain current and relevant. At the time when we selected the models in second week of April, the models Gemini-Pro-2.5, GPT-4o, Deepseek R1, o4-mini, Grok3 were leading models. These models also have suitable chat interfaces which are made accessible by their respective companies for usage. Besides these we use the models Llama 4 Maverick, Qwen 2.5 max and Claude 3.7 Sonnet as judges for model generated diagnosis.

4.2 Prompt engineering

4.3 Response parsing

Each model produced responses in slightly varying format. In fact the same model was observed to produce responses in varying format for different cases. Due to this limitation, we did not create an automated approach and instead manually cropped the response to relevant diagnosis from the overall text.

4.4 Automated judgement

4.5 Implementation and deployment considerations

Our clinical-diagnosis pipeline would be designed for horizontal scalability and resilience. Each logical component (data ingestion, de-identification, inference,

TABLE 3
The generated clinical and final diagnosis by the models Deepseek R1, Grok3, o4-mini for each of the ten cases.

DeepSeek		Grok		o4-mini	
Clinical	Final	Clinical	Final	Clinical	Final
Immune checkpoint inhibitor pneumonitis	Immune-related sarcoid-like reaction	Recurrent metastatic melanoma with pulmonary involvement.	Metastatic melanoma with BRAF V600E mutation, pulmonary metastases confirmed.	Pulmonary metastatic melanoma	Pembrolizumab-induced sarcoid-like granulomatous pneumonitis
Cerebellar mass lesion (suspected neoplasm)	CNS lymphoma (AIDS-associated)	Cerebellar tumor, possible metastatic disease.	Cerebellar tumor, possible metastatic disease.	Metastatic cerebellar neoplasm (brain metastases of unknown primary)	HIV-associated diffuse large B-cell lymphoma
Meningoencephalitis (autoimmune/infectious)	Neurosarcoidosis	Encephalitis, possible autoimmune or infectious.	Primary central nervous system lymphoma.	Aseptic lymphocytic meningoencephalitis	Primary CNS lymphoma (vitreoretinal large B-cell lymphoma)
Hypocalcemic seizure secondary to suspected vitamin D deficiency.	Vitamin D-dependent rickets (type likely confirmed post-testing).	Hypocalcemic seizure	Vitamin D deficiency rickets	Hypocalcemic seizure (suspected rickets)	Nutritional vitamin D–deficiency rickets
Giant Cell Arteritis	IgG4-Related Disease	Suspected giant cell arteritis	Metastatic carcinoma	Giant cell (temporal) arteritis	IgG4-related orbital inflammatory disease
Cellulitis with bullous formation aThe diagnosis as reported in the NEJM, andnd lymphangitis.	Necrotizing cellulitis due to Streptococcus pyogenes.	Cellulitis	Streptococcal cellulitis	Necrotizing fasciitis	Group A Streptococcus necrotizing fasciitis
Lyme disease	Lyme arthritis with gastrointestinal involvement	Inflammatory Bowel Disease	Crohn's Disease	Suspected Lyme disease	Crohn's disease
Opportunistic fungal pneumonia (suspected)	Invasive pulmonary aspergillosis	Pneumonia	Invasive Pulmonary Aspergillosis	Recurrent pneumonia in an immunocompromised host	Subacute invasive pulmonary aspergillosis
Pulmonary-renal syndrome (suspected)	Anti-glomerular basement membrane (anti-GBM) disease (Goodpasture syndrome)	Pneumonia with suspected glomerulonephritis.	Pulmonary-renal syndrome (suspected Goodpasture's syndrome).	Community-acquired pneumonia	Anti-GBM (Goodpasture) syndrome
Uveitis (suspected tuberculous etiology)	Tuberculous Panuveitis	Uveitis	Sarcoidosis	Bilateral granulomatous anterior uveitis	Sarcoidosis (ocular sarcoidosis)

TABLE 4
tab:Model performance on BLEU, Rouge-L and BertScore

Model	BLEU	ROUGE-L	BERTScore	ClinicalBERTScore
GPT-4o final	0	**0.313**	**0.454**	0.645
Gemini final	0	0.14	0.391	**0.619**
Deepseek final	0	**0.033**	**0.446**	**0.647**
Grok final	0	0.271	0.45	0.636
o4-mini final	0	0.192	0.41	0.636

post-processing) is containerised with Docker and orchestrated via Kubernetes, allowing pods to scale automatically in response to workload (e.g. using Horizontal Pod Autoscalers on CPU/GPU metrics). All services expose RESTful API endpoints over HTTPS, with TLS 1.2 encryption ensuring data-in-transit confidentiality and integrity. RESTful API would also enable easier integration with EHR records. Incoming patient data is de-identified and pseudonymised at the edge—direct identifiers are tokenised or stripped before storage—complying with privacy regulations. A dynamic model-selection strategy routes low-complexity cases to a lightweight, on-premises LLM fine-tuned for rapid inference, while high-risk or ambiguous cases are escalated to a more powerful, albeit slower and more resource-intensive

TABLE 5
Model generated diagnosis was evaluated compared to reported diagnosis using cross examination by independent LLM and self-examination by similar LLM. The cross examination was performed using Qwen 2.5 Max, Claude 3.7 Sonnet and Llama 4 Maverick which are state of the art reasoning models. The self examination was done by similar architectures such as GPT4o-mini, Gemini Flash 2.5 and Microsoft Copilot quick response.

Models	GPT-4o	Gemini-Pro-2.5	Deekseek-R1	Grok3	o4-mini
Qwen 2.5 Max judgement	b b c b c c a c c a (35.0)	c b c c c c a c c c (15.0)	c b c b c c c c c c (10.0)	c c c a c c a c c a (30.0)	c b c a c c a c c a (35.0)
Claude judgement	b c c b c c a c c a (30.0)	c c c c c c a c c c (10.0)	c c c b c c c c c c (5.0)	c c c b c c a c c a (25.0)	c c c b c c a c c a (25.0)
Llama judgement	b c c b c c a c c a (30.0)	c b c c c c a c c c (15.0)	c b c b c c c c b c (15.0)	c a b b c c a c b a (45.0)	c c c b c c a c c a (25.0)
Majority cross voting	b c c b c c a c c a (30.0)	c b c c c c a c c c (15.0)	c b c b c c c c c c (10.0)	c c c b c c a c c a (25.0)	c c c b c c a c c a (25.0)
Majority self voting	b b c a c c a c c a (40.0)	c b c b c c a c b c (25.0)	c b c c c c c c c c (5.0)	c c b b c c a b b b (35.0)	c c c b c c a c c a (25.0)
Mean score of the two	35	20	7.5	30	25
Same model without thinking	25	-	5	20	-

TABLE 6
Inference speed and cost per query for the models under evaluation. We estimated the speed and cost for the first case only. The number of parameters for most of these models is not officially known and was specified in third party estimates or official partner reports.

Models	Inference time in seconds	Speed	Parameters	Estimated price per query (in cents)
ChatGPT 4o	40	Decent	200 billion [44]	0.63
Gemini 2.5 Pro Experimental	24	Fast	>500 billion[12]	2.15
DeepSeek R1	43	Decent	671 billion (37b/token) [27]	**0.12**
Grok 3 Think Preview	**60**	Slow	**2.7 trillion**[13]	**2.25**
o4-mini Preview	**6**	Very fast	-	0.44
Average	34.6			1.118

TABLE 7
Prompt which is prepended to the case description for instructing the LLM on how to produce the desired response. The same prompt was used for all 10 cases and models.

Given the clinical case details, identify the clinical diagnosis and the final diagnosis based on the following criteria. Both the diagnosis need to be in very few words.

- **Clinical Diagnosis:** This is the initial diagnosis based on the patient's symptoms, medical history, physical examination, and preliminary test results. Keywords to consider include symptoms (e.g., fever, cough, weight loss), medical history (e.g., previous cancers, chronic diseases), and initial imaging or lab results (e.g., CT scan findings, blood test results).

- **Final Diagnosis:** This is the definitive diagnosis after all diagnostic tests, imaging, biopsies, and specialist consultations have been completed. Keywords to consider include biopsy results, advanced imaging findings, molecular profiling, and response to initial treatments.

The case is: ...{case description}

TABLE 8
Prompt used for scoring model's output based on majority voting by three independent reasoning models.

I used an LLM model to diagnose a clinical case and received the response {model's diagnosis}. Do you think that is correct to the actual diagnosis in the dataset {known diagnosis}. So what do you think, is the assessment provided by you: a) fully correct, b) partially correct but missed details/less specific, c) incorrect (not good for actual use).

model. This approach balances throughput, cost, and diagnostic accuracy, and all model diagnosis and chain-of-thoughts (see Appendix 3) are logged for auditability.

Beyond core inference, we will employ DevSecOps best practices to maintain reliability and security. Continuous Integration pipelines build and version Docker images, while Continuous Deployment with blue–green or canary releases would enable zero-downtime updates. Kubernetes health checks and autoscaling policies ensure service level agreement (SLA) compliance under variable loads. A centralised logging and monitoring stack (e.g. Prometheus, Grafana, ELK) collects metrics, traces, and alerts on anomalous behaviours or security threats. Model governance is enforced via a registry that tracks versioned models, training-data provenance, performance benchmarks, and bias audits. Inference results below a confidence threshold or flagged by automated fairness checks trigger a human-in-the-loop review, ensuring accountability and patient safety. Finally, strict role-based access controls, regular security-patching of containers, and encrypted storage at rest complete the end-to-end secure deployment.

5 AI ETHICS

Our proposed clinical-diagnosis pipeline was designed in alignment with Australia's AI Ethics Principles [46]. Below we critically examine each principle in light of our results and propose some mechanisms for mitigating issues identified:

5.1 Human, Social and Environmental Wellbeing

By improving diagnostic accuracy beyond non-reasoning baselines (e.g.GPT-4o achieved a mean score of 30 vs. 10 for Deepseek R1), our system has the potential to enhance patient outcomes and reduce diagnostic errors. However, post-training steps incurs additional energy use and associated emissions and inference time thinking also leads to higher resource consumption (average inference time 34.6 s/query). Mitigation strategies include model distillation, mixture of experts (adopted by Deepseek R1 and Llama 4) and scheduling inference on renewable-powered infrastructure to offset the carbon emissions.

5.2 Human–Centred Values

Chain-of-thought explanations [29], [30] improve the interpretability of model outputs, allowing clinicians to trace reasoning steps. Our manual review showed that reasoning models like Grok 3 produced intermediate rationales that aligned with clinical workflows. Future work should include end-user studies to quantify usability in real settings.

5.3 Fairness

While overall accuracy improved, we did not evaluate performance across demographic subgroups. Given known disparities in LLM performance on underrepresented populations [43], a dedicated fairness audit-measuring metrics such as equalised odds is required to ensure equitable care.

5.4 Privacy Protection and Security

We processed only de-identified NEJM cases and hosted data securely, compliant with the Australian Privacy Act 1988 and Australian Privacy Principles [47]. During deployment in production server, all data would be first de-identified using a rule-based strategy and encrypted during transit between API endpoints using protocols such as TLS 1.2.

5.5 Reliability and Safety

Our multi-model majority-voting scheme (using Qwen 2.5 Max, Claude 3.7 Sonnet, Llama 4 Maverick) reduced bias associated with single-model evaluation, achieving consensus scores up to 30/100. This is only slightly lower than human physician performance reported at 33.6/100 [43]. Regular monitoring and rollback mechanisms would be employed for guarding against anomalous outputs [42].

5.6 Transparency and Explainability

We log chain-of-thought traces and maintain versioned prompts to reconstruct each inference. All models and prompts used are documented in our repository which would allow in reproducing the results. The chain-of-thought associated with the models provides insights into models' thinking strategy.

5.7 Contestability

Clinicians can flag suspect cases via an audit interface; disputed inferences are re-evaluated by alternative models or human experts. We propose extending this with a formal appeals workflow.

5.8 Accountability

Responsibility for model selection, deployment, and monitoring is assigned to the clinical AI governance board. Post deployment to production system, we would maintain an immutable audit trail of inferences and user feedback.

Challenges and Mitigation Strategies: The primary trade-off encountered was between accuracy and environmental impact. We propose (1) distilling open sourced reasoning models into smaller student networks specialised for clinical diagnosis, (2) batching inference to improve GPU utilisation, and (3) scheduling non-urgent queries during periods of low-carbon-intensity grid supply.

6 DISCUSSION

6.1 Alignment with the UN Sustainable Development Goals

In 2015, United Nations proposed 17 [48] designed to achieve a better and more sustainable future for all by 2030. These goals address a wide range of global challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice. Three of these goals listed below are suitably addressed by this work:

- **SDG 3 (Good Health and Well-Being)** via enhanced diagnostic accuracy in clinical cases.

- **SDG 9 (Industry, Innovation and Infrastructure)** by advancing AI methodologies for healthcare.
- **SDG 13 (Climate Action)** through our carbon-footprint analysis and proposed energy-efficient strategies.

6.2 Data security and privacy management

During deployment, we will adhere to the Australian Privacy Act 1988 and the Australian AI Ethics Principles' privacy tenets, employing:

- End-to-end encryption (TLS 1.2),
- Role-based access controls and audit logs,
- De-identification and pseudonymisation of clinical text.

6.3 Scalability and integration

Our production pipeline shall be containerised via Docker and orchestrated with Kubernetes, enabling horizontal scaling across multiple servers. Inference latencies range from 6s (o4-mini) to 60s (Grok 3), which can be optimised via model-parallel deployments with easier cases sent to high speed low cost model and difficult cases sent to low speed high cost models. The determination of which case is easier or difficult shall be done by our on-premises LLM. Integration into electronic health record systems will be facilitated by a RESTful API layer.

6.4 Limitations

- **Sample size:** Only ten NEJM cases were evaluated, thus our evaluation has limited statistical significance. Subsequent study should focus on evaluation many more cases.
- **Demographic bias:** No subgroup analyses were conducted which is possible again with larger sample size.
- **Manual parsing:** Diagnosis extraction was semi-automated, introducing potential human error.
- **Environmental metrics:** Indirect estimates of carbon footprint lack real-time measurement as most models are closed source.

6.5 Future directions

To strengthen and extend this work, we recommend:

- Expanding to larger, more diverse clinical datasets.
- Automating response parsing with robust NLP extraction pipelines.
- Conducting formal user-experience studies with clinicians.
- Incorporating differential privacy guarantees for patient data.
- Implementing dynamic model selection based on carbon-intensity forecasts.

7 CONCLUSION

We evaluated five state-of-the-art reasoning LLMs (GPT-4o, Gemini-Pro 2.5, Grok 3, Deepseek R1, o4-mini) on ten NEJM clinical cases. Reasoning models consistently outperformed non-reasoning counterparts, with GPT-4o leading (mean score 35/50). However, gains in accuracy come at the cost of higher inference latency (up to 60s/query) and increased carbon emissions. Our findings suggest that chain-of-thought–enabled LLMs are promising for clinical decision support, provided that ethical, environmental, and privacy considerations are proactively managed.

APPENDIX 1 - VIDEO PRESENTATION

In this section, you are asked to provide a link to a video presentation that you made specifically for this task, for example

https://echo360.net.au/media/
30dad243-0ed2-4a08-a8ab-b1faf7be2f38/public

The video should explain your approach to the task and must not exceed 5 minutes.

APPENDIX 2 - OTHERS

All code for reproduction has been provided in well documented notebooks accessible in my GitHub repository: https://github.com/nepython/COMP6011.

APPENDIX 3 - CHAIN OF THOUGHTS

Based on the provided clinical case details:

REFERENCES

[1] L. G. Winston, M. L. Winkler, A. Kheterpal, and J. A. Villalba, "Case 36-2021: a 22-year-old man with pain and erythema of the left hand," *New England Journal of Medicine*, vol. 385, no. 22, pp. 2078–2086, 2021.

[2] K. A. Armstrong, J. V. Cohen, J.-A. O. Shepard, E. E. Folch, M. K. Mansour, and J. A. Stefely, "Case 16-2020: a 47-year-old woman with recurrent melanoma and pulmonary nodules," *New England Journal of Medicine*, vol. 382, no. 21, pp. 2034–2043, 2020.

[3] G. S. Gottlieb, J. M. Rosenberg, R. G. Gonzalez, and R. T. Gandhi, "Case 27-2020: A 53-year-old woman with headache and gait imbalance," *New England Journal of Medicine*, vol. 383, no. 9, pp. 859–866, 2020.

[4] S. S. Mukerji, K. L. Ard, P. W. Schaefer, and J. A. Branda, "Case 32-2020: a 63-year-old man with confusion, fatigue, and garbled speech," *New England Journal of Medicine*, vol. 383, no. 16, pp. 1578–1586, 2020.

[5] Y. V. Virkud, N. D. Fernandes, R. Lim, D. M. Mitchell, and W. T. Rothwell, "Case 39-2020: A 29-month-old boy with seizure and hypocalcemia," *New England Journal of Medicine*, vol. 383, no. 25, pp. 2462–2470, 2020.

[6] S. H. Unizony, H. R. Kelly, E. K. O'Donnell, and S. Sadigh, "Case 3-2021: A 48-year-old man with transient vision loss," *New England Journal of Medicine*, vol. 384, no. 4, pp. 363–372, 2021.

[7] K. Zanger, K. Nimkin, E. J. Israel, and A. R. Shih, "Case 3-2022: A 14-year-old boy with fever, joint pain, and abdominal cramping," *New England Journal of Medicine*, vol. 386, no. 4, pp. 375–383, 2022.

[8] R. S. Knipe, G. A. Alba, J. M. Harvey Barnes, and L. P. Hariri, "Case 15-2022: A 57-year-old man with persistent cough and pulmonary opacities," *New England Journal of Medicine*, vol. 386, no. 20, pp. 1933–1944, 2022.

[9] M. C. Ottolini, R. Shailam, A. Sharma, and I. A. Rosales, "Case 21-2022: A 17-year-old girl with fever and cough," *New England Journal of Medicine*, vol. 387, no. 2, pp. 168–176, 2022.

TABLE 9

The below table shows the CoT for Gemini-Pro-2.5 experimental with the final output generated by the model showing Clinical Diagnosis as Meningoencephalitis and Final Diagnosis: Primary CNS Lymphoma (PCNSL). The CoT provides an insight into the model's inner working.

Here's a thinking process to arrive at the clinical and final diagnoses:
Understand the Goal: The request asks for a Clinical Diagnosis and a Final Diagnosis based on a provided clinical case summary (Case 32-2020). Both diagnoses should be concise (very few words). The definitions of each type of diagnosis are given, emphasizing the information used for each (initial symptoms/tests vs. definitive tests/biopsies).
Analyze the Case for Clinical Diagnosis Clues:
Presenting Symptoms: Fever, confusion, headache, garbled speech, fatigue, vision changes (flashes, blurriness, floaters), photophobia, sonophobia, neck stiffness, staring spells, weakness (transient), facial droop (transient), weight loss, cachexia, bitemporal wasting, somnolence, perseverative thoughts, memory issues, word-finding difficulties, limited upward gaze, intention tremor, hyperreflexia, ankle clonus, myoclonic jerks. Medical History: Primary membranous nephropathy (treated with prednisone, cyclosporine, rituximab, cyclophosphamide - immunosuppression is key), hypertension, dyslipidemia, CAD (PCI, CABG), color blindness, colonic polyps, BPH, smoking history (quit). Family history of sarcoidosis. Initial Investigations (before/early in admission at this hospital): CSF (initial): Lymphocytic pleocytosis, elevated protein. MRI/CT head (previous hospital): Initially normal, then reported subacute infarct in pons (questionable given transient symptoms and later findings?). EEG (previous hospital): Focal left temporal slowing. Blood tests (previous hospital): Mostly normal, ruling out many common causes (Lyme, hepatitis, RPR, ANA, RF, ANCA, complement). CT head (this hospital, initial): No hemorrhage, acute infarct, or mass. Opacified right maxillary sinus. MRI head (this hospital, initial): Sulcal FLAIR hyperintensity with subtle enhancement, subarachnoid/subdural restricted diffusion foci (no enhancement), nonspecific white matter changes. MRA normal. MRI spine: Degenerative changes only. CSF (this hospital, early): Yellow, turbid, high WBC (lymphocytic predominance), high protein, low glucose (relative to serum). Many infectious tests negative (bacteria, fungi, common viruses like HSV, VZV, Enterovirus, West Nile, EBV, CMV, Lyme, Cryptococcus, Listeria). Autoimmune panel ordered. Ophthalmology: Bilateral optic nerve edema, tortuous vessels, posterior vitreal detachment. Later, vitreous cells noted. Blood tests (this hospital, early): Leukocytosis initially, then normalized. Low IgG, IgA, IgM (hypogammaglobulinemia). Negative for many specific infections and autoimmune markers.
Synthesize Clinical Diagnosis:
The patient presents with a constellation of neurological symptoms (encephalopathy, seizures/staring spells, focal deficits, headache, vision changes) and systemic symptoms (fever, fatigue, weight loss, cachexia). Key findings point towards inflammation of the central nervous system: CSF pleocytosis (lymphocytic), high protein, low glucose, optic nerve edema, vitreous cells (uveitis/vitritis). The MRI findings (sulcal FLAIR hyperintensity/enhancement, restricted diffusion foci) suggest inflammation/infection involving the meninges and potentially small infarcts or inflammatory lesions. The patient is significantly immunosuppressed due to treatment for membranous nephropathy (rituximab, cyclophosphamide). This increases the risk of opportunistic infections and certain malignancies (like lymphoma). Many common infectious causes have been ruled out by CSF and blood tests. Autoimmune causes are being investigated. Malignancy (like CNS lymphoma or carcinomatous meningitis) is also a concern given the immunosuppression, cachexia, and CSF findings. Therefore, the clinical diagnosis revolves around inflammation/infection of the CNS in an immunosuppressed host. Terms like "Meningoencephalitis" or "Encephalopathy of unknown origin" fit. Given the CSF findings and symptoms, "Meningoencephalitis" seems more specific. The chronicity and immunosuppression also raise suspicion for specific types (e.g., opportunistic infection, lymphoma). Analyze the Case for Final Diagnosis Clues:
The provided text stops before definitive results like biopsies, detailed molecular profiling, or response to specific targeted treatments are mentioned. Key tests pending or ordered that would lead to a final diagnosis include: CSF cytologic examination (crucial for lymphoma). CSF autoimmune encephalopathy panel. Further specialized infectious disease testing if initial screens are negative. Brain biopsy (if non-invasive tests are inconclusive). The low immunoglobulin levels (hypogammaglobulinemia) are noted, likely secondary to rituximab treatment. The PET-CT was negative for systemic malignancy/inflammation focus. The ophthalmology findings evolved to include vitreous cells, confirming intraocular inflammation (uveitis/vitritis). Synthesize Final Diagnosis:
Since the provided text doesn't give the results of the truly definitive tests (like CSF cytology, specific rare infection results, or biopsy), a final diagnosis cannot be definitively determined from the provided text alone. However, based on the entire clinical picture presented, the differential diagnosis has been significantly narrowed. The combination of B-cell depleting therapy (rituximab), neurological symptoms, CSF findings (lymphocytic pleocytosis, high protein, low glucose, negative common infectious workup), uveitis/vitritis, and hypogammaglobulinemia strongly points towards a specific diagnosis often seen in this context. Self-Correction/Refinement: While I can't know the actual final diagnosis from the text snippet which cuts off, the prompt asks for a diagnosis based on the criteria and the given case details. The text describes the full workup up to day 5, including many negative results, the persistent neurological picture, the ophthalmological findings, and the crucial history of immunosuppression (especially rituximab). A common complication in this scenario is CNS lymphoma, often primary CNS lymphoma (PCNSL), which can present exactly like this (meningoencephalitis picture, uveitis/vitritis, specific CSF findings) and is notoriously difficult to diagnose without specific cytology or biopsy. Other possibilities include unusual opportunistic infections not yet identified or an autoimmune process. Revisiting the prompt instructions: "Final Diagnosis: This is the definitive diagnosis after all diagnostic tests, imaging, biopsies, and specialist consultations have been completed." The text doesn't provide the results of all these tests (e.g., CSF cytology is pending). Therefore, I cannot give a definitive final diagnosis based solely on the provided text. Alternative interpretation: Perhaps the question expects the most likely diagnosis given the extensive negative workup and classic presentation in an immunosuppressed patient, even if the definitive test result isn't explicitly stated. Primary CNS Lymphoma is a very strong possibility given the rituximab history, CSF profile, negative infectious workup, and uveitis/vitritis. Decision: Acknowledge the limitation. State the most likely diagnosis suggested by the presented data, but qualify that the text doesn't provide the definitive proof (like cytology). However, the prompt asks for the final diagnosis based on the criteria applied to the case. Let's assume the case implicitly leads to a standard diagnosis often featured in such Case Records (which this format resembles). PCNSL is the classic diagnosis fitting this picture, especially the combination with uveitis/vitritis in a rituximab-treated patient. Let's re-evaluate the clinical diagnosis. "Meningoencephalitis" covers the inflammation. Perhaps "Suspected CNS infection or malignancy" is more accurate initially? Or just "Encephalopathy with meningitis"? Let's stick with Meningoencephalitis as it captures the core findings well. For the final diagnosis, given the strong indicators (immunosuppression, specific CSF profile after negative infectious workup, uveitis), Primary CNS Lymphoma is the most probable conclusion these case studies often build towards. Format the Answer: Provide the two diagnoses concisely as requested.
Clinical Diagnosis: Meningoencephalitis Final Diagnosis: Primary CNS Lymphoma (PCNSL) (or state inability to confirm from text if strict interpretation is preferred). Let's go with PCNSL as the likely intended answer in this type of case presentation format.

[10] E. M. Cohen, L. Sobrin, M. G. Figueiro Longo, D. B. Pier, D. R. Brown, and A. M. Stagner, “Case 33-2022: An 11-year-old girl with redness of the eyes,” *New England Journal of Medicine*, vol. 387, no. 17, pp. 1598–1607, 2022.

[11] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[12] N. Chomsky, *Syntactic structures*. Mouton de Gruyter, 2002.

[13] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains,” *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.

[15] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[19] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019.

[20] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Biobert: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[22] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, “Publicly available clinical bert embeddings,” *arXiv preprint arXiv:1904.03323*, 2019.

[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, “Improving language understanding by generative pre-training,” 2018.

[24] N. Kitaev, Ł. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” *arXiv preprint arXiv:2001.04451*, 2020.

[25] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: Self-attention with linear complexity,” *arXiv preprint arXiv:2006.04768*, 2020.

[26] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating long sequences with sparse transformers,” *arXiv preprint arXiv:1904.10509*, 2019.

[27] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.

[28] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, “Gemma 3 technical report,” *arXiv preprint arXiv:2503.19786*, 2025.

[29] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[30] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” in *International Conference on Learning Representations (ICLR)*, 2023.

[31] S. Welleck, X. Lu, P. West, F. Brahman, T. Shen, D. Khashabi, and Y. Choi, “Generating sequences by learning to self-correct,” *arXiv preprint arXiv:2211.00053*, 2022.

[32] A. Singh, J. D. Co-Reyes, R. Agarwal, A. Anand, P. Patil, X. Garcia, P. J. Liu, J. Harrison, J. Lee, K. Xu *et al.*, “Beyond human data: Scaling self-training for problem-solving with language models,” *arXiv preprint arXiv:2312.06585*, 2023.

[33] Y. Qu, T. Zhang, N. Garg, and A. Kumar, “Recursive introspection: Teaching language model agents how to self-improve,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 249–55 285, 2024.

[34] Y. Zhou, A. Zanette, J. Pan, S. Levine, and A. Kumar, “Archer: Training language model agents via hierarchical multi-turn rl,” *arXiv preprint arXiv:2402.19446*, 2024.

[35] A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs *et al.*, “Training language models to self-correct via reinforcement learning,” *arXiv preprint arXiv:2409.12917*, 2024.

[36] J. Hong, N. Lee, and J. Thorne, “Orpo: Monolithic preference optimization without reference model,” *arXiv preprint arXiv:2403.07691*, 2024.

[37] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu *et al.*, “T\" ulu 3: Pushing frontiers in open language model post-training,” *arXiv preprint arXiv:2411.15124*, 2024.

[38] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *International conference on machine learning*. PMLR, 2015, pp. 1889–1897.

[39] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.

[40] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[41] W. Chen, C. Zhong, J. Peng, and Z. Wei, “Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations,” *Bioinformatics*, vol. 39, no. 1, p. btac744, 2023.

[42] O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, and A. Freitas, “Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making,” *Artificial Intelligence*, vol. 316, p. 103839, 2023.

[43] D. McDuff, M. Schaekermann, T. Tu, A. Palepu, A. Wang, J. Garrison, K. Singhal, Y. Sharma, S. Azizi, K. Kulkarni *et al.*, “Towards accurate differential diagnosis with large language models,” *Nature*, pp. 1–7, 2025.

[44] A. B. Abacha, W.-w. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, and T. Lin, “Medec: A benchmark for medical error detection and correction in clinical notes,” *arXiv preprint arXiv:2412.19260*, 2024.

[45] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez *et al.*, “Chatbot arena: An open platform for evaluating llms by human preference,” in *Forty-first International Conference on Machine Learning*, 2024.

[46] S. Australian Government, Department of Industry and Resources, “Australia’s ai ethics principles,” https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles, 2021, accessed: May 4, 2025.

[47] A. Government, “Privacy act 1988 (cth),” https://www.legislation.gov.au/Series/C2004A03712, 1988, accessed: May 4, 2025.

[48] “The sustainable development goals,” https://sdgs.un.org/goals, n.d., accessed: May 4, 2025.