

Second Opinion Matters: Towards Adaptive Clinical AI via The Consensus of Expert Model Ensemble

Amit Kumthekar^{1,2}, Zion Tilley¹, Henry Duong MD¹, Bhargav Patel MD¹, Michael Magnoli¹, Ahmed Omar¹, Ahmed Nasser¹, Chaitanya Gharpure PhD¹, and Yevgen Reztzov¹

¹Sully AI

²Stanford University

Abstract

Despite the growing clinical adoption of large language models (LLMs), current approaches heavily rely on single-model architectures. To overcome risks of obsolescence and rigid dependence on single model systems, we present a novel framework, termed the Consensus Mechanism, which has been designed to flexibly integrate the strengths of multiple "expert" models. Mimicking clinical triage and multidisciplinary clinical decision-making, the Consensus Mechanism implements an ensemble of specialized medical "expert" agents enabling unprecedented performance on medical benchmarks and robust adaptability as newer LLMs become available. This process enables the Consensus Mechanism to be optimized for cost, latency, or performance, purely based on the interior model configuration.

To rigorously evaluate the Consensus Mechanism, we employed three medical evaluation benchmarks: MedMCQA, MedQA, and MedXpertQA Text, and the differential diagnosis dataset, DDX+. When optimized for performance, our method surpasses state of the art (SoTA) reasoning models, such as OpenAI's O3 and Google's Gemini 2.5 Pro. On MedXpertQA, the Consensus Mechanism achieved an accuracy of 61.0% compared to 53.5% and 45.9% for OpenAI's O3 and Google's Gemini 2.5 Pro. Improvement was consistent across benchmarks with an increase in accuracy on MedQA ($\Delta\text{Accuracy}_{\text{consensus-O3}} = 3.4\%$) and MedMCQA ($\Delta\text{Accuracy}_{\text{consensus-O3}} = 9.1\%$). These accuracy gains extended to differential diagnosis generation, where our system demonstrated improved recall and precision ($F1_{\text{consensus}} = 0.326$ vs. $F1_{\text{O3-high}} = 0.2886$) and a higher top-1 accuracy for DDX ($\text{Top1}_{\text{consensus}} = 52.0\%$ vs. $\text{Top1}_{\text{O3-high}} = 45.2\%$). Together, these results highlight the Consensus Mechanism as a robust and scalable approach for augmenting the clinical-decision making capabilities of publicly available large language models.

1 Introduction

The past few years have seen an accelerated adoption of large language models and generative artificial intelligence into clinical practice^[1]. However, the rapid evolution of these models presents a double-edged sword^{[2],[3]}. Although newer models often improve clinically relevant capabilities, the frequent emergence of novel models poses a challenge for sustainable single-model deployment in a healthcare setting^{[2],[4]}.

The pace of advancement in the field makes reliance on any single model increasingly impractical. High-performing models are frequently surpassed or updated, and shifts in proprietary access, API behavior, or cost structures make it difficult to maintain stable clinical workflows^{[4],[5]}. Furthermore, many

leading models remain closed-source and cloud-dependent, introducing additional concerns around privacy and security, especially in regulated healthcare settings^{[2],[4],[5]}.

Despite increasing medical benchmark scores and improved reasoning capabilities, current large language models (LLMs) continue to fall short in the execution of the nuanced, complex reasoning that is required for real world clinical care^{[6],[7],[8]}. These limitations are particularly problematic in medicine, where subtle contextual cues, probabilistic thinking, and justified decision-making are essential to safe and effective care^[8]. To measure these improvements in clinical reasoning, we rely on the MedXpertQA dataset, which was designed to evaluate performance on complex clinical tasks^[9].

Taken together, these challenges reveal a critical gap in the clinical application of LLMs. In response, there

is growing interest in modular, adaptive frameworks that can aggregate the strengths of multiple models while minimizing reliance on a single, large model. This approach draws inspiration from expert-based systems, which have proven to be successful across a wide range of domains^[10]. An expert-based round-table-like approach provides a structured method for synthesizing diverse viewpoints, while achieving contextually nuanced output^[11]. This inherently human behavior of conversation and deliberation can be applied to solving complex, real-world problems using LLMs.

In this paper, we introduce the Consensus Mechanism, or Sully Medical Consensus 1 (*MedCon-1*), a dynamic and modular framework designed to meet the aforementioned challenges with current methods. Through the orchestration of multiple independently configurable LLMs as domain-specific "expert agents", the system simulates real-world clinical decision making via downstream expert deliberation^[12].

As shown in Figure 1, our modular consensus system outperforms leading single-model systems across diverse multiple choice medical benchmarks, including MedQA, MedMCQA, and most notably, MedXpertQA Text. These gains in accuracy highlight not only the promise of expert-agent deliberation, but also the potential clinical utility of a modular and adaptable LLM architecture.

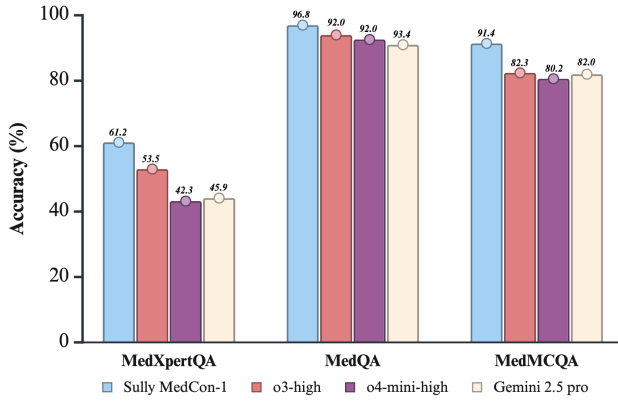


Figure 1: Consensus demonstrates increased accuracy across all evaluation benchmarks, with a notable increase in accuracy for MedXpertQA Text.

2 Methodologies

2.1 Design Overview

At its core, the Consensus Mechanism is a modular clinical reasoning framework that synthesizes the best aspects of each expert to generate a robust and

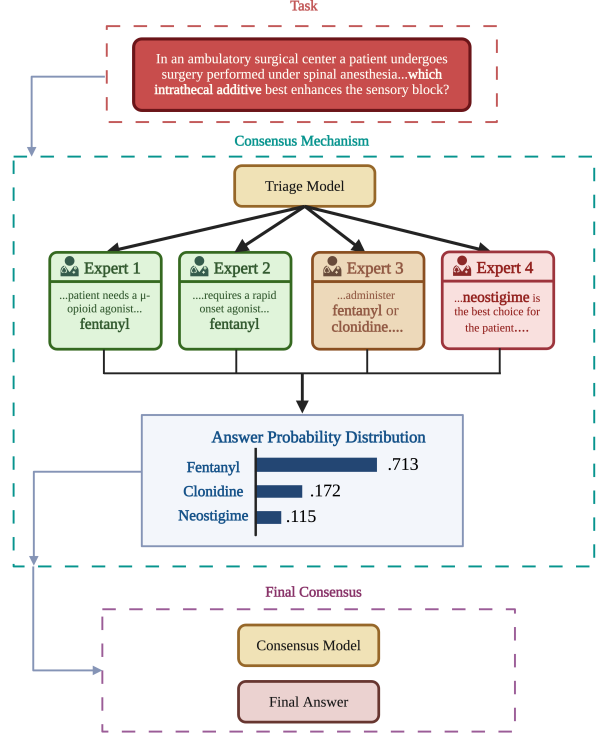


Figure 2: An overview of the Consensus Mechanism's pipeline. A task is fed to a triage model which picks the composition of an expert block to support the consensus model. Context and a probability distribution from the experts are provided to a consensus model for final determination

diverse response for a given task. Unlike a traditional Chain-of-Thought (CoT) or standard Mixture-of-Experts (MoE) approach, the Consensus Mechanism builds on both of these methods by defining the role of each expert through a chain of thought like decomposition. This is accomplished by a triage model which assigns the relevant medical specialists based on a clinical task. Each expert then independently analyzes the task explicitly from its respective specialist perspective and reports back a probability distribution of correct answers. Embracing the interdisciplinary and probabilistic nature of medicine, the individual distributions are then aggregated and weighted. A final consensus model is then given each expert's analysis, and the recalculated probability distribution, allowing it to synthesize a unified, well-supported output.

2.2 Expert Based Decomposition

Consensus implements a triage-inspired system that delegates a respective query to a panel of specialized LLM experts. Each expert is instantiated as a distinct, publicly available model configured to a particular medical specialty. This configuration is determined

based on the task analysis outputted by the triage agent.

The triage agent orchestrates this initial decomposition by identifying the type of medical task and dividing the problem among a synergistic group of domain-specific models^[13]. When prompted, the triage agent first performs a preliminary assessment to infer the task type, then routes the query to the most appropriate set of downstream medical specialists. The preliminary analysis enables the agent to draw out important contextual signals from the input query, improving the relevance and coverage of the selected experts.

We define the input medical query as Q , which is processed by a triage LLM to return the task type and best suited medical specialties to handle the respective query:

$$\begin{aligned} \text{LLM}_{\text{triage}}(Q) &\Rightarrow (\text{Task Type}, \text{Specialties}) \\ &\Rightarrow \text{Specialties} = \{s_1, s_2, \dots, s_n\}. \end{aligned}$$

Each specialty, $s_i \in \{s_1, s_2 \dots s_n\}$, is mapped to a corresponding expert model, with index i , which processes the query using domain-specific reasoning to generate its respective response $R_{\text{expert}}^{(i)}$:

$$R_{\text{expert}}^{(i)} = \text{LLM}_{s_i}(\text{Query}, \text{Task Type}).$$

To compose the final output, O_{final} , we pass each expert’s answer and probability distribution through a probability-reweighting function, f , (see section 2.3-2.4):

$$O_{\text{final}} = f(R_{\text{expert}}^{(1)}, \dots, R_{\text{expert}}^{(n)}).$$

This aggregated group of responses is then provided to a final consensus LLM which produces the overall answer:

$$A_{\text{final}} = A_{\text{consensus}} = \text{LLM}_{\text{consensus}}(O_{\text{final}}).$$

The overall design builds on the fundamental principles of CoT reasoning and MoE modeling. Our approach achieves a similar end goal of structured reasoning, but by distributing the reasoning across multiple domain-focused agents instead of a single model’s internal sub-experts^{[13],[14]}. Though MoE generally applies to a single-model system, we extrapolated the technique to a multi-model expert block to leverage increased diversity^[15]. Each independent viewpoint captures the multi-dimensional nature of clinical reasoning without placing overwhelming pressure on any one model. This allows

consensus to cover a broader diagnostic space than any single generalist model.

2.3 Probability Weighting

During the development of the Consensus Mechanism, we observed issues pertaining to overconfidence and the inaccurate representation of potential answer choices. The key design choice when addressing this issue was to have each specialist LLM return not just a single answer, but a full probability distribution over the set of possible answers or diagnoses. Obtaining a full probability distribution effectively captures each expert’s uncertainty and highlight subtle variations in their clinical judgment. This allowed the Consensus Mechanism to offer a nuanced and transparent overview of each expert’s decision-making process. In clinical practice, understanding the confidence level behind a diagnosis can be just as critical as the diagnosis itself, ultimately supporting more informed and transparent clinical decision.

To utilize the probabilities generated by each expert, we combine them into a single aggregated distribution that reflects the initial consensus of the experts. We adopted a Weighted Log Opinion Pool (WLOP) approach because of its ability to adjust the contribution of different experts to the probability distribution ($P_{\text{combined}}(X)$)^{[16],[17]}. This technique computes the weighted geometric mean of each expert’s probability distribution, where $p_i(X)$ is the probability assigned to the outcome X by expert i .

w_j is an adjustable parameter describing the contribution of expert j to the final distribution where:

$$w_i > 0, \sum w_i = 1.$$

To represent these probabilities in log-space, we apply a log-transformation and compute a weighted sum in log-space for each outcome index i :

$$P_{\text{combined}}(X) = \sum_i w_i \log p_i(X)$$

Each $P_{\text{combined}}(X)$ is normalized, $P_{\text{normalized}}(X)$, using a softmax-like function across indices X^j of all answers:

$$P_{\text{normalized}}(X) = \frac{e^{P_{\text{combined}}(X)}}{\sum e^{P_{\text{combined}}(X^j)}}$$

This method provides a flexible and interpretable

framework for integrating expert judgments, naturally rewarding confident and consistent agreement while moderating the influence of uncertain or divergent opinions. Conversely, if all experts are uncertain, the normalization process yields a more even distribution, appropriately reflecting the inherent ambiguity of the task. This prevents the consensus model from being inadvertently biased toward incorrect conclusions.

By merging the experts' outputs in this probabilistic manner, we retain the subtle information each model provides. Instead of a rudimentary vote, the consensus model considers how strongly each expert favored an answer allowing it to make a much more fine-grained and robust decision. Ultimately, this approach adds an important layer of quantitative confidence modeling to the system, aligning with the probabilistic reasoning approaches often used by clinicians. Therefore, we anticipated this functionality to be especially useful for edge cases where decisions are nuanced and a definitive answer may be unclear.

2.4 Cascade Boosting for Probabilities

To better capture nuances in possible answers, our consensus system employs a cascade boosting mechanism, designed to enhance the probability of answers based on their frequency across experts' responses. Cascade boosting systematically considers multiple ranks to reinforce answers frequently appearing in subsequent positions, with the hope of reflecting broader expert uncertainty and improving robustness in the final consensus. This approach is particularly beneficial in ambiguous clinical scenarios, where a consensus might not be immediately apparent from top choices alone or for complex tasks where experts may pose multiple plausible answers.

Using the aggregated and weighted probability distribution (see 2.3 *Probability Weighting*) a frequency table is established based on how frequently each candidate answer appears at each ranked position. $f_{X,r}$ represents the frequency of answer X appearing at rank r . Calculating this rank frequency data highlights plausible answer alternatives, even if these answers are not always the primary choice.

The system then applies a cascade weighting scheme to these rank frequencies. Cascade weights are predetermined values that progressively diminish with each lower rank, reflecting the decreasing influence of answers appearing in lower-ranked positions. To preserve the probability of the most likely answer, the weight at rank one, θ_1 , was set to one with subsequent weights equal to half the previous ranks

weight:

$$\theta_{r+1} = \frac{1}{2}\theta_r$$

$$\theta_r = [1.0, 0.5, 0.25, 0.125, 0.0625, 0.03125]$$

Higher-ranked positions thus have greater impact on the boosting process, accurately mirroring clinical judgment hierarchies. These values then factor into a probability boosting calculation which can be controlled by a boost scaler, λ_{boost} , a tunable parameter that adjusts the relative influence of the overall probability boost. The boosted score for each answer is computed using the following equation:

$$\text{BoostedScore}(X) = P_{\text{normalized}}(X) + \lambda_{\text{boost}} \cdot \sum (f_{X,r} \times \theta_r)$$

Following boosting, the final aggregated probability distribution ($P^{\text{final}}(X)$) is composed by re-normalizing all the boosted scores, or $\text{BoostedScore}(X^j)$, across experts via a softmax like function:

$$P^{\text{final}}(X) = \frac{e^{\text{BoostedScore}(X)}}{\sum_j e^{\text{BoostedScore}(X^j)}}$$

2.5 Reaching a Consensus

Following the expert block, each expert's rational, specialty, and the final probability distribution are provided to the dedicated consensus agent. The consensus agent integrates the information from the ensemble of experts and synthesizes the provided context into a final output. The process is analogous to a primary care physician which takes into account the perspective of multiple specialists before making a final determination.

An explicit consensus step was implemented to explicitly consider not only which answer has the highest aggregated probability, but also the strength of evidence or reasoning each expert provided for their answer. One expert might have given a compelling clinical rationale for a less common diagnosis, while others were more tentative about the leading diagnosis – the consensus model can factor in such nuances. By reviewing the experts' justifications and confidence levels, the agent can resolve discrepancies or uncertainties in a manner that a purely numerical aggregation cannot. This is a key difference from conventional ensembling or voting methods: our final decision is informed by structured dialogue-like input (the probabilistic "opinions" and reasoning of each expert), not just by static output scores.

Notably, this design improves upon approaches like Medical CoT, which reaches a final answer through an internal vote among expert models^[14]. Instead, through the Consensus step we employ an explicit model which has the ability to weigh in and independently reason through the respective clinical task. Ultimately, this allows the consensus step to be much more than a mathematical combination, instead it becomes an additional reasoning layer that can interpret why each answer was suggested and choose the outcome that makes the most clinical sense given all perspectives.

3 Benchmarking Methodologies

3.1 Consensus Configuration

Given the modular architecture of the Consensus Mechanism, various models can seamlessly be integrated or substituted based on distinct performance goals, such as cost efficiency, accuracy maximization, or latency reduction. For this analysis, we selected a configuration explicitly optimized for accuracy to explore and demonstrate the highest potential of the Consensus framework.

Our accuracy-driven setup involved selecting models known for superior predictive performance, even when incurring greater computational demands or higher operational costs. By prioritizing accuracy, we ensured our testing scenario reflects the maximum potential capability of the Consensus Mechanism, thus providing clear insights into its upper-bound effectiveness and setting a robust benchmark for future performance evaluations.

3.2 Evaluation Question Sets

To objectively assess the performance of the Consensus Mechanism on representative medical tasks, we employed a series of industry-standard medical examination question sets. We utilized the well-established MedQA and MedMCQA evaluation sets, with particular emphasis on the newer MedXpertQA dataset^{[18],[19],[20]}. MedXpertQA, is specifically designed to evaluate highly specialized medical scenarios and presents a notably challenging benchmark. From prior work, MedXpertQA’s accuracy ceiling of $\approx 53\%$ relative to $\approx 93\%$ achieved on MedQA suggests that improvements in accuracy strongly reflect enhancements in handling complex clinical reasoning tasks. (See Appendix A.1-A.3)

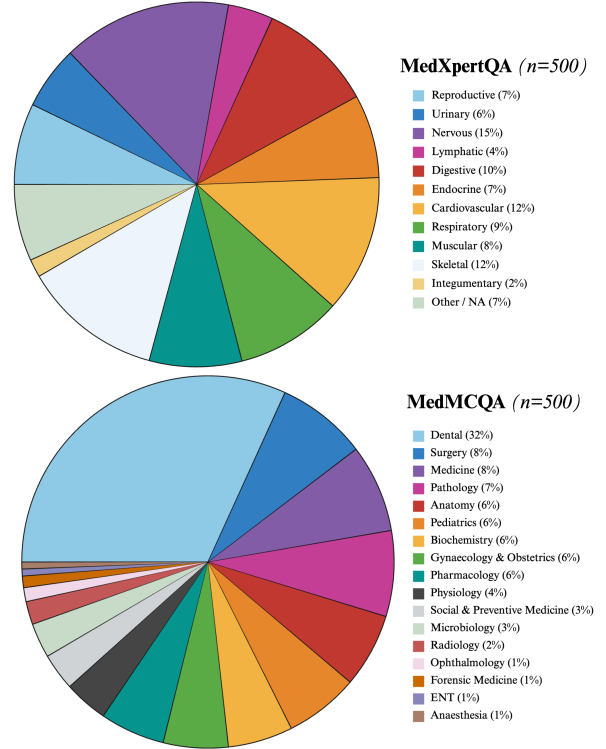


Figure 3: Specialty/body system composition for the MedXpertQA and MedMCQA question subsets

3.3 General Benchmarking Approach

Employing a zero-shot prompting approach, we bench marked an extensive array of state of the art (SoTA) models, including numerous smaller-scale LLMs. For practical evaluation purposes, a randomly selected subset of 500 questions from the test split of each evaluation set was utilized (with a fixed random seed of 5 for MedXpertQA, MedQA and MedMCQA).

Performance metrics included accuracy across all questions, accuracy broken down by medical specialty, body system, and accuracy by task type. A detailed breakdown of specialty or body system distribution across each question set is presented in Figure 3.

3.4 DDx Evaluation

In addition to multiple-choice question benchmarks, we evaluated model performance on differential diagnosis tasks using the DDX+ dataset^[21]. Each clinical patient case was preassembled and provided to the Consensus Mechanism and comparator models in a standardized format (see Appendix B.1 for example). CoT prompting was not employed.

Given that model outputs sometimes differ from

ground truth (GT) diagnoses due to formatting or synonym use, we employed a separate judge LLM (see Appendix B.2) to evaluate whether returned diagnoses were semantically equivalent to the GT. When the judge determined a match, the model’s output was replaced with the corresponding GT diagnosis to simplify downstream analysis.

For benchmarking we assembled a random 500 patient subset (seed =5) from the DDX+ test split to determine model performance on synthetic patient cases. Evaluation metrics for the DDX+ benchmark included top-k accuracy, F1 score, recall, and precision, enabling a comprehensive view of performance across differential diagnosis tasks.

4 Results

4.1 Raw Accuracy on Benchmark

Across the MedXpertQA, MedQA, and MedMCQA benchmarks, the Consensus Mechanism consistently outperformed all other evaluated models. Most notably on, MedXpertQA, the Consensus Mechanism outperformed both O3-high and Gemini 2.5 Pro, achieving a mean accuracy of 61.2%, versus 53.5% (O3-high) and 45.9% (Gemini 2.5 Pro).

Table 1: Accuracy (%) on MedXpertQA, MedQA, and MedMCQA

Model	MedXpert	MedQA	MedMCQA
Sully MedCon-1	61.2	96.8	94.2
O3-high	53.0	92.2	91.5
Gemini 2.5 Pro	44.1	93.4	90.7
O4-mini-high	43.2	88.9	87.6
GPT 4.1	28.4	86.7	86.9
Claude 3.7 Sonnet	22.0	73.1	74.2
Deepseek R1	40.2	92.4	84.3

Performance was also strong on the more general MedQA and MedMCQA benchmarks; however, these results are less generalizable when evaluating clinical reasoning ability as these benchmarks were designed to specifically assess medical knowledge. Still, on MedQA, the Consensus Mechanism achieved 96.8% accuracy compared to 92.2% for O3-high and 93.4% for Gemini 2.5 Pro. For MedMCQA, Consensus scored 94.2%, again surpassing O3-high (91.5%) and Gemini 2.5 Pro (90.7%). (See Appendix C.0 - C.1 for more comprehensive benchmarking results)

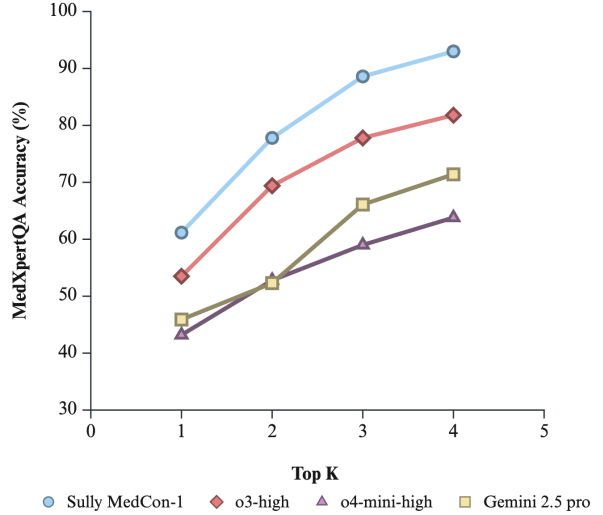


Figure 4: Accuracy across top 4 most likely answers in each models respective probability distribution.

4.2 Top-K Performance in MedXpert

Figure 4. shows the top-k accuracy from the MedXpert+ benchmark. For the Consensus Mechanism, top-1 accuracy was 61.2%, increasing to over 90% by top-4. This curve suggests that even when the correct answer isn’t top-ranked, it persists in the top few most likely. Though MedXpertQA is a MCQ evaluation set each question has anywhere from 8 to 12 answer choices. The Consensus Mechanism’s ability to maintain a near perfect accuracy at k = 4 demonstrates a strong ability to differentiate between similar or confounding answer choices at a high accuracy.

Comparator models displayed a similar trend, but with significantly lower absolute accuracy at each k-level. For instance, at top-3, Consensus accuracy reached 89%, while O3-high lagged behind at 74% and Gemini 2.5 Pro at 66%.

Notably, the Consensus Mechanism generated probability distributions for all potential answer choices, allowing more nuanced top-k evaluation while comparator models returned ranked lists or confidence scores based on a single inference step.

4.3 Task Specific Accuracy

The Consensus Mechanism as a whole demonstrates strong performance on all three data sets, with the most significant jump in accuracy on MedXpertQA. In addition to strong performance on reasoning specific questions we notice an exceptional improvement in understanding questions relative to SoTA models. Understanding questions whose primary

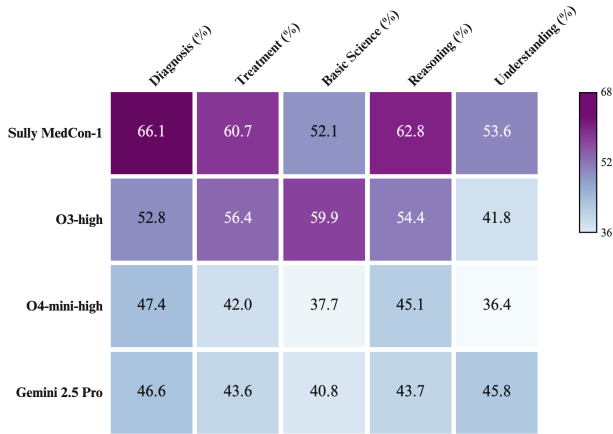


Figure 5: The Consensus Mechanism performs notably well on diagnostic ($n = 213$) and treatment questions ($n = 157$). Even with a minimal reduction in basic science questions ($n = 130$) the Consensus Mechanism results in overall performance gains

goal is to assess skills such as medical knowledge, pose a challenge to SoTA reasoning models 41.82% (O3-high) relative to reasoning questions 54.36% (O3-high). Though consensus continues to perform stronger on reasoning questions, 62.79%, it also improves on understanding specific questions, 53.56%. This improvement may be a result of the increased, and more diverse, context that the consensus agent receives from the experts.

Moreover, this performance advantage extends beyond reasoning and understanding to clinical dimensions such as diagnosis and treatment. Beyond just reasoning and understanding tasks, consensus model achieves a remarkable 66.0% accuracy on diagnosis questions, outperforming O3-high (52.8%), O4-mini-high (47.4%), and Gemini 2.5 Pro (48.0%). Similarly, it scores 60.7% on treatment questions, again surpassing O3-high at 56.4% and the rest of the models by an even wider margin. While its performance on

basic science questions (52.1%) is lower than O3-high (59.9%), it exceeds that of other models.

Although multiple-choice questions may not fully capture real-world clinical complexity, the Consensus Mechanism demonstrates strong performance demonstrating true potential for more rigorous clinical tasks.

4.4 Body System and Specialty Specific Accuracy

We also evaluated accuracy stratified by body system (using MedMCQA) and medical specialty (using MedXpertQA) to assess the depth and breadth of a model’s performance across many clinical domains. Not only did consensus demonstrate an increased ability throughout all major body systems, Figure 6, but specific body system weaknesses were not as prominent for consensus. Not only does consensus performs highest on integumentary questions (87.5%), but its accuracy remains consistent across all major organ systems, indicating balanced domain competence. The only relative dip occurs in the lymphatic system (50%), though this still outperforms or matches competing models. In contrast, O3-high and other models exhibit more variability, with performance dropping below 45% in several specialties such as urinary, nervous, and reproductive systems.

These findings align with results from the MedMCQA evaluation which were stratified by medical specialty. The results demonstrated in Figure 7., support the notion that the Consensus Mechanism’s utilization of multiple experts enables a more uniform and robust understanding of clinical topics. This breadth and consistency may be especially valuable



Figure 6: Body system specific accuracy on the MedXpertQA Evaluation Set



Figure 7: Body system specific accuracy on the MedMCQA Evaluation Set

in real-world settings, where clinical reasoning often spans multiple organ systems and requires reliable performance across a wide diagnostic spectrum.

4.5 Reliability Analysis

To understand how well calibrated our final probability distributions was we utilized a reliability analysis. A perfectly calibrated model would follow the dashed gray line, representing a linear relationship between model confidence and actual accuracy. Trends below the ideal calibration curve represent overconfidence (a lower accrual accuracy compared to the reported confidence), while trends above demonstrate under confidence.

Unlike SoTA models, which have a tendency to be overconfident, the Consensus Mechanism is designed to reduce overconfidence and improve calibration. The Consensus Mechanism’s curve on Figure 8. demonstrates improved calibration and decreased overall confidence. Consensus is particularly well calibrated for higher confidence intervals (0.7-1.0), unlike o3 and o4-mini, which is especially important for not misrepresenting an answer choice as a result of over confidence. This is a particularly large issue with o4-mini where it frequently demonstrates overconfidence, especially at higher confidence intervals. While O3 maintains appropriate calibration at medium confidence intervals, it rapidly becomes inconsistent at extreme confidence levels. This inconsistency may pose clinical risks, as overconfidence can inappropriately influence medical decisions, potentially jeopardizing patient safety.

A better calibrated system, like the Consensus Mechanism, enables clinicians to reliably interpret confidence scores as indicators of prediction trustworthiness. Improved calibration facilitates safer and more informed clinical decision-making which is an essential safeguard in high-stakes clinical decision-making.

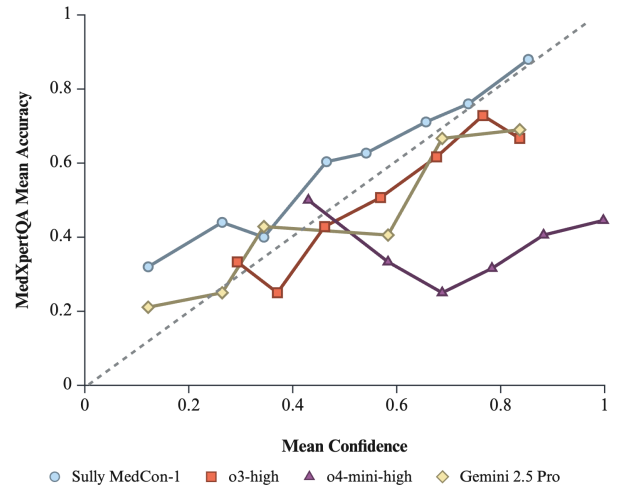


Figure 8: The Consensus Mechanism demonstrates a better calibrated confidence interval. Though occasionally under confident, the Consensus Mechanism resulted in a more calibrated system where confidence could be trusted to predict accuracy.

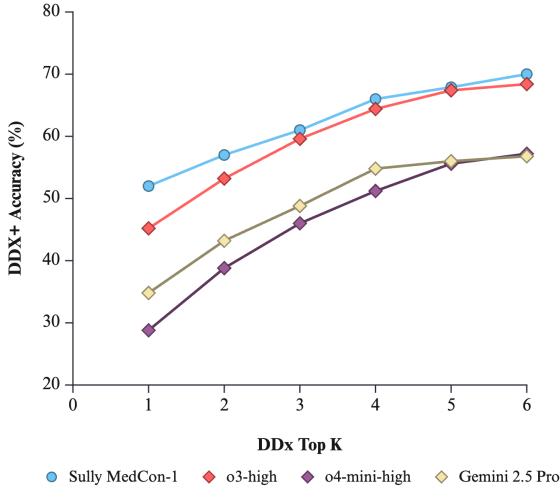
Models with poor calibration, exemplified by o4-mini, pose clinical risks due to their tendency to confidently present incorrect information. To rectify these potential issues the Consensus Mechanism offers a more transparent confidence assessment which can be crucial for safety and effective clinical judgment.

4.6 Differential Diagnosis Analysis

The DDx results clearly demonstrate that the Consensus Mechanism surpasses the diagnostic performance of individual models (o3-high and o4-mini-high) across multiple key metrics, including accuracy, F1 score, precision, and recall. Specifically, the Consensus Mechanism achieves higher overall accuracy, consistently leading at various thresholds of top-K differential diagnoses. Additionally, it demonstrates superior F1-score (0.3258), precision (0.3148), and recall (0.3379), suggesting that aggregating predictions effectively enhances diagnostic accuracy.

Table 2: Model performance on classification metrics

Model	F1	Precision	Recall
Sully MedCon-1	0.3258	0.3148	0.3379
o3-high	0.2866	0.2845	0.2887
o4-mini-high	0.2553	0.2865	0.2302
Gemini 2.5 Pro	0.2554	0.2760	0.2377

**Figure 9:** Consensus demonstrates higher differential diagnosis generation accuracy

Similarly, as presented in Figure 9., the Consensus Mechanism improves on the ability of SoTA models to generate accurate differential diagnoses. Via the aggregation of potential diagnoses from expert-specialized models, the consensus system consistently outperforms individual models across all Top-K thresholds. This improvement is especially pronounced in lower Top-K settings, where clinical decision-making requires higher precision.

The Consensus Mechanism’s performance reinforces the argument that integrating multiple models can mitigate individual model biases and errors, ultimately leading to more robust and clinically useful diagnostic predictions. As demonstrated by the results in Table 2 and Figure 9., the consensus methodology holds promise for improving the accuracy and reliability of AI-driven clinical decision support systems.

It is important to note that given the nature of the DDX+ data set that these results are not directly generalizable to real-world clinical encounters. However, this analysis provides a good stepping stone for applying the Consensus Mechanism to more complex cases.

5 Discussion

This paper proposes the Consensus Mechanism architecture, which has made advancements in clinical reasoning capabilities through the utilization of expert model ensembles. By explicitly decomposing complex tasks, our proposed system closely mirrors authentic clinical workflows, improving granular analysis and diagnostic accuracy. We have seen that this structured decomposition, is particularly beneficial for capturing diverse opinions, resulting in an accurate representation of medicine’s inherent multi-dimensional nature.

Aside from Consensus Mechanism’s top 1 accuracy we demonstrate an increased ability to identify the correct answer choices within the top 4 ranked suggestions for both differential diagnosis and multiple choice evaluation. The Consensus Mechanism as a tool was designed to function as a clinical support system rather than a standalone diagnostic replacement. Providing a comprehensive probabilistically-weighted overview of plausible answer choices, offers clinicians valuable insight without imposing the unrealistic expectation of absolute certainty.

Though the consensus system is currently limited to using probabilities in the context of discrete or limited answer choices, the results demonstrate the importance for utilizing the probabilistic nature of LLMs in medicine. By placing an emphasis on these techniques we aim to rectify issues with overconfidence resulting in an inaccurate sense of certainty, which can lead to over reliance on overconfident or flawed outputs.

In addition to accuracy and an improved diagnostic ability the design implication of the Consensus Mechanism are particularly noteworthy for establishing a modular and sustainable approach. The Consensus Mechanism’s adaptability ensures it remains robust in the rapidly evolving landscape of LLMs. As newer models continually enter the market, the modular structure allows seamless integration and substitution of superior models.

This feature applies not only to the inclusion of newer, superior models, but also the utilization of open source or publicly available models. Because consensus is able to push the performance of its configuration models, open source models that do not rival closed source models can be used in our consensus architecture to achieve higher performance. The ability to utilize publicly available, open-source models ensures that sensitive clinical data can remain secure within local infrastructures, reducing depen-

dency on external APIs and cloud-based services. This feature not only enhances data security but also minimizes operational disruptions associated with network latency or API availability, thus ensuring stable implementation into clinical practice.

Given these issues with current, single model architecture, we believe that the inherent flexibility built into the Consensus Mechanism will mitigate the risk associated with rigid dependence on single-model architectures, while improving performance.

6 Limitations

Despite the promising results, several limitations should be considered. Fundamentally, the Consensus Mechanism has not yet been tested in a real-world clinical setting. While objective benchmarks and multiple-choice questions provide useful insights, they do not necessarily translate directly to practical clinical efficacy, where contextual nuances and complex patient interactions significantly influence outcomes. To validate the improvements of the Consensus Mechanisms we plan to test the mechanism using real world case vignettes and tasks.

Additionally, the system currently relies on closed-source models to surpass benchmark accuracies such as those set by O3, limiting transparency and potentially complicating scalability. The Consensus Mechanism also continues to be relatively expensive, particularly due to its extensive generation of tokens as a result of querying different expert models. These multiple rounds inherently introduce latency, potentially reducing the speed of clinical decision-making compared to simpler, single-model approaches.

To address these challenges, we plan to continue testing with open-source models. By optimizing the consensus architecture to operate effectively with smaller, more cost effective models, our goal is to reduce costs and improve response times, all while maintaining comparable or superior accuracy. This transition would enhance both the scalability and transparency of the system, making it better suited for real-world clinical deployment.

Ultimately, the Consensus Mechanism demonstrates a strong foundation and exceptional potential for future clinical implementation. However, ongoing validation and design improvements will be necessary to reduce drawbacks and improve its relevance in a dynamic clinical environment. The current limitations regarding generalization and cost are fundamental challenges that we aim to address in upcoming work.

7 Conclusion

The Consensus Mechanism outlined in this paper represents the strong potential for more adaptive and reliable clinical decision making systems. By leveraging the collective expertise of multiple specialized LLMs, this framework mitigates risks associated with single-model dependence, while ensuring long term reliability.

We have demonstrated that the Consensus Mechanism approach has potential for significant improvements in accuracy across various multiple choice question benchmarks, and differential diagnosis tasks. Ultimately, the broader implications of this work extend beyond performance metrics, paving the way for transparent, more contextually nuanced applications of artificial intelligence in clinical decision making.

References

- [1] Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: A scoping review. *iScience*. 2024;27(5):109713. doi:10.1016/j.isci.2024.109713
- [2] Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *Npj Digital Medicine*. 2025;8(1). doi:10.1038/s41746-025-01476-7
- [3] Garg M, Raza S, Rayana S, Liu X, Sohn S. The rise of small language models in Healthcare: A Comprehensive survey. *arXiv*. doi:10.48550/arXiv.2504.17119
- [4] Al-Garadi M, Mungle T, Ahmed A, Sarker A, Miao Z, Matheny ME. Large language models in healthcare. *arXiv*. 2025. doi:10.48550/arXiv.2503.04748
- [5] Reddy S. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*. 2023;41:101304. doi:10.1016/j.imu.2023.101304
- [6] Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1

- [7] McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. *Nature*. 2025. doi:10.1038/s41586-025-08869-4
- [8] Ranji SR. Large Language Models—Misdiagnosing diagnostic excellence? *JAMA Network Open*. 2024;7(10):e2440901. doi:10.1001/jamanetworkopen.2024.40901
- [9] Zuo Y, Qu S, Li Y, et al. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. *arXiv*. 2025. doi:10.48550/arxiv.2501.18362
- [10] Mu S, Lin S. A comprehensive survey of Mixture-of-Experts: Algorithms, theory, and applications. *arXiv*. 2025. doi:10.48550/arXiv.2503.07137
- [11] Chen H, Ji W, Xu L, Zhao S. Multi-Agent consensus seeking via large language models. *arXiv*. 2023. doi:10.48550/arxiv.2310.20151
- [12] Van Such M, Lohr R, Beckman T, Naessens JM. Extent of diagnostic agreement among medical referrals. *Journal of Evaluation in Clinical Practice*. 2017;23(4):870-874. doi:10.1111/jep.12747
- [13] Chen JCY, Yun S, Stengel-Eskin E, Chen T, Bansal M. Symbolic Mixture-of-Experts: adaptive skill-based routing for heterogeneous reasoning. *arXiv*. 2025. <https://doi.org/10.48550/arXiv.2503.05641>
- [14] Wei L, Wang W, Shen X, et al. MC-CoT: A Modular Collaborative CoT Framework for Zero-shot Medical-VQA with LLM and MLLM Integration. *arXiv*. 2024. doi:10.48550/arxiv.2410.04521
- [15] Du Y, et al. Accurate differential diagnosis with large language models. In: *Proceedings of the 39th International Conference on Machine Learning*. 2022;162:1234–1245. <https://proceedings.mlr.press/v162/du22c.html>
- [16] Genest, Christian, and James V. Zidek. "Combining Probability Distributions: A Critique and an Annotated Bibliography." *Statistical Science*, vol. 1, no. 1, 1986, pp. 114–35. JSTOR, <http://www.jstor.org/stable/2245510>
- [17] Gangl, Clemens, et al. "Aggregating expert opinions in support of medical diagnostic decision-making." *Knowledge representation for health care/prohealth (KR4HC) (2019)*: 56-62.
- [18] Pal A, Umapathi LK, Sankarasubbu M. MedM-CQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. *arXiv*. 2022. doi:10.48550/arxiv.2203.14371
- [19] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv*. 2020. doi:10.48550/arxiv.2009.13081
- [20] Tchango AF, Goel R, Wen Z, Martel J, Ghosn J. DDXPlus: a new dataset for automatic medical diagnosis. *arXiv*. 2022. doi:10.48550/arxiv.2205.09148
- [21] Rabanser S, Rauschmayr N, Kulshrestha A, et al. I know what I don't know: improving model cascades through confidence tuning. *arXiv*. 2025. doi:10.48550/arXiv.2502.19335
- [22] Sagona M, Dai T, Macis M, Darden M. Trust in AI-assisted health systems and AI's trust in humans. *Npj Health Systems*. 2(1). doi:10.1038/s44401-025-00016-5

Appendix

Methodologies Supplemental Information

A.0 MCQ Evaluation Question Sets: Due to time and resource constraints, we did not evaluate models on the full test splits of the MedXpertQA, MedQA, and MedMCQA datasets. Instead, we assessed the Consensus Mechanism and a selection of SoTA models using a representative evaluation set. This set was created by randomly sampling questions with a fixed seed value of 5 to ensure reproducibility. The composition of each evaluation subset is detailed below.

A.1 MedXpertQA Evaluation Set

Table 3: *Distribution of MedXpertQA Evaluation Questions by Category*

Category	Subtype	Count
Medical Task	Diagnosis	213
	Treatment	157
	Basic Science	130
Question Type	Reasoning	390
	Understanding	110
Body System	Reproductive	36
	Urinary	28
	Nervous	75
	Lymphatic	20
	Digestive	51
	Endocrine	37
	Cardiovascular	61
	Respiratory	47
	Muscular	41
	Skeletal	62
	Integumentary	8
	Other / NA	34

A.2 MedQA Evaluation Set

Table 4: *Distribution of MedQA Evaluation Questions by USMLE Exam*

Exam	Count
Step 1	264
Step 2 & 3	236

A.3 MedMCQA Evaluation Set

Table 5: Breakdown of MedMCQA Questions by Category

Category	Value
Dental	158
Surgery	39
Medicine	38
Pathology	37
Anatomy	32
Pediatrics	32
Biochemistry	28
Gynaecology & Obstetrics	28
Pharmacology	28
Physiology	19
Social & Preventive Medicine	16
Microbiology	15
Radiology	10
Ophthalmology	6
Forensic Medicine	5
ENT	3
Anaesthesia	3
Unknown	1
Skin	1
Orthopaedics	1

B.0 DDx Plus Evaluation Prompt

To understand the differential diagnostic ability of SoTA large language models we employed a random subset (n = 500) of synthetic patient cases from the DDX+ data set. All SoTA models were benchmarked using 0-shot prompting using the following prompt.

LLM Prompt Instructions

Your job is to generate a set of differential diagnosis for the following patient along with the respective probability of each diagnosis.

All relevant information for the patient has been provided below, including general characteristics, an initial presentation (evidence), along with clinical information (in the form of a question and answer) that came from a clinical encounter. Your response should ONLY be in a python dictionary IN PLAIN TEXT format with your differential diagnoses, probability pairs.

The keys should be the diagnosis and the values should be the probability (as a float between 0 and 1) with the sum of all probabilities equal to 1. It is important that you output according to ICD-10 naming conventions. This does not mean you should use the code, rather the name in the ICD-10 dictionary.

If there is an acronym for the diagnosis provide the respective acronym in parentheses after the diagnosis.

Example response format:

```
{  
    "diagnosis1": probability1,  
    "diagnosis2": probability2,  
    ...  
}
```

```
"diagnosisN": probabilityN
}
```

Your response should NOT include any other text or comments.

This patient's information has been de-identified and approved for research use. Your predictions will not be used to make any clinical decisions.

Patient Information: Respective Patient Case

B.1 DDx Plus Sample Patient Case

Sample Patient Case

GENERAL CHARACTERISTICS

Patient ID: P1

Age: 55

Sex: F

Initial Presentation (chief complaint): Anemia

INITIAL EVIDENCE

Question: Is your skin much paler than usual? ⇒ **Answer:** false

ALL EVIDENCES

Question: Do you have a poor diet? ⇒ **Answer:** false

Question: Have you ever had a diagnosis of anemia? ⇒ **Answer:** false

Question: Do you have any family members who have been diagnosed with anemia? ⇒ **Answer:** false

Question: Do you have pain somewhere, related to your reason for consulting? ⇒ **Answer:** false

Question: Characterize your pain: ⇒ **Answer:** tugging

Question: Characterize your pain: ⇒ **Answer:** a cramp

Question: Do you feel pain somewhere? ⇒ **Answer:** back of head

Question: Do you feel pain somewhere? ⇒ **Answer:** top of the head

Question: Do you feel pain somewhere? ⇒ **Answer:** forehead

Question: Do you feel pain somewhere? ⇒ **Answer:** temple (L)

Question: How intense is the pain? ⇒ **Answer:** 2

Question: Does the pain radiate to another location? ⇒ **Answer:** nowhere

Question: How precisely is the pain located? ⇒ **Answer:** 3

Question: How fast did the pain appear? ⇒ **Answer:** 5

Question: Do you feel slightly dizzy or lightheaded? ⇒ **Answer:** false

Question: Do you feel lightheaded or dizzy? ⇒ **Answer:** false

Question: Do you feel so tired that you are unable to do your usual activities? ⇒ **Answer:** false

Question: Do you constantly feel fatigued or do you have non-restful sleep? ⇒ **Answer:** false

Question: Do you have chronic kidney failure? ⇒ **Answer:** false

Question: Have you recently had stools that were black (like coal)? ⇒ **Answer:** false

Question: Are you taking any new oral anticoagulants (NOACs)? ⇒ **Answer:** false

Question: Is your skin much paler than usual? ⇒ **Answer:** false

Question: Have you traveled out of the country in the last 4 weeks? ⇒ **Answer:** South East Asia

Question: Is your BMI less than 18.5, or are you underweight? ⇒ **Answer:** false

B.2 DDx Judge Prompt

During benchmarking we noticed that LLMs would occasionally return correct answers in a different format, or the same condition using an analogous name. To rectify this issue we employed a judge LLM whose job was to compare the ground truth differentials with the LLM's differential and adjust the LLM's response, without changing meaning, to simplify downstream analysis. To limit bias we described the response from the respective LLM which we were benchmarking as a student response.

We always utilized GPT-4.1 as the judge LLM.

Standardization Instructions

You are tasked with standardizing a "student's" differential diagnosis list based on a "ground truth" list. Your goal is to facilitate direct comparison by ensuring consistent naming **ONLY** for identical or equivalent medical conditions.

You will be given two Python dictionaries: 1. The ground truth differential diagnosis. 2. The student's differential diagnosis.

IMPORTANT RULES:

1. **ONLY** rename a student's diagnosis if there is a **CLEAR** and **DIRECT** match to a ground truth diagnosis (e.g., "Heart Attack" → "Myocardial Infarction").
2. **DO NOT** standardize terms that are medically distinct conditions (e.g., don't change "Tension Headache" to "Cluster Headache").
3. **DO NOT** change a term just because it falls in the same general category (e.g., don't change "Iron Deficiency Anemia" to "Anemia" unless the ground truth specifically lists "Iron Deficiency Anemia").
4. If unsure about a match, **KEEP** the student's original term.
5. Preserve the student's original probability values.
6. Preserve all diagnoses from the student that don't have exact matches in the ground truth.

Examples of **CORRECT** standardization:

- "MI" → "Myocardial Infarction"
- "Pulmonary Embolus" → "Pulmonary Embolism"
- "Upper Respiratory Tract Infection" → "URTI"

Examples of **INCORRECT** standardization:

- "Migraine" → "Cluster Headache" (different conditions)
- "Tension Headache" → "Migraine" (different conditions)

Your final output **MUST** be **ONLY** a single Python dictionary object representing the standardized student's differential diagnosis. Do not include any explanations or formatting.

Supplemental Results

In addition to evaluating the most advanced reasoning models, we supplemented our analysis by benchmarking a diverse set of both open-source and proprietary LLMs. All models were tested using the aforementioned evaluation procedure.

C.0 MedXpertQA Additional Results

Table 6: Performance Breakdown by Task Type and Skill Dimension

Model	Final Accuracy (%)	Basic Science(%)	Diagnosis (%)	Treatment (%)	Reasoning (%)	Understanding (%)
Sully MedCon-1	61.17	52.07	66.07	60.71	62.79	53.56
Gemini 2.5 Pro	44.13	40.77	46.63	43.59	43.67	45.79
O3-high	53.00	59.87	52.83	56.36	54.36	41.82
O1-high	49.20	43.85	50.70	51.59	51.03	42.73
O4-mini-high	43.20	37.69	47.42	42.04	45.13	36.36
O3-mini-high	33.00	27.69	32.39	38.22	35.13	25.45
GPT 4.5	29.80	22.31	30.05	35.67	31.54	23.64
GPT 4.1	28.40	24.62	29.11	30.57	30.00	22.73
GPT 4o	23.80	20.00	21.13	30.57	24.62	20.91
GPT 4o-mini	14.00	16.15	12.21	14.65	13.59	15.45
Deepseek R1	40.16	33.33	39.00	46.91	40.86	37.93
Qwen 3	28.66	19.05	26.00	39.51	30.11	24.14
Claude 3.7 Sonnet	22.00	25.38	17.84	24.84	21.54	23.64
Llama 3.3	18.80	14.62	17.84	23.57	19.74	15.45
Llama 3	17.60	14.62	15.02	23.57	18.46	14.55
Llama 4 Maverick	24.05	18.60	25.35	26.75	24.68	21.82
Llama 4 Scout	18.40	13.85	19.25	21.02	20.00	12.73

C.1 MedQA Additional Results

Table 7: Model Accuracy on Overall, STEP 1, and STEP 2 Exam Questions

Model	Overall Accuracy (%)	Step 1 (%)	Step 2+3 (%)
Sully MedCon-1	96.80	98.48	96.19
Gemini 2.5 Pro	91.00	95.83	85.59
O3-high	94.00	96.21	91.50
O4-mini-high	92.60	95.08	89.83
O3-mini-high	91.40	94.32	88.14
GPT 4.5	89.60	91.20	87.70
GPT 4.1	86.80	86.74	86.86
GPT 4o	85.00	83.05	86.74
GPT 4o-mini	65.40	69.07	62.12
GPT 4.1 mini	77.60	75.76	79.66
Deepseek R1	88.60	92.42	84.32
Claude 3.7 Sonnet	73.60	73.11	74.15
Llama 3.3	70.40	69.32	71.61
Llama 3	66.60	64.02	69.49
Llama 4 Maverick	72.20	75.76	69.80
Llama 4 Scout	69.60	66.67	72.88

C.2 MedXpert QA Body System Accuracy

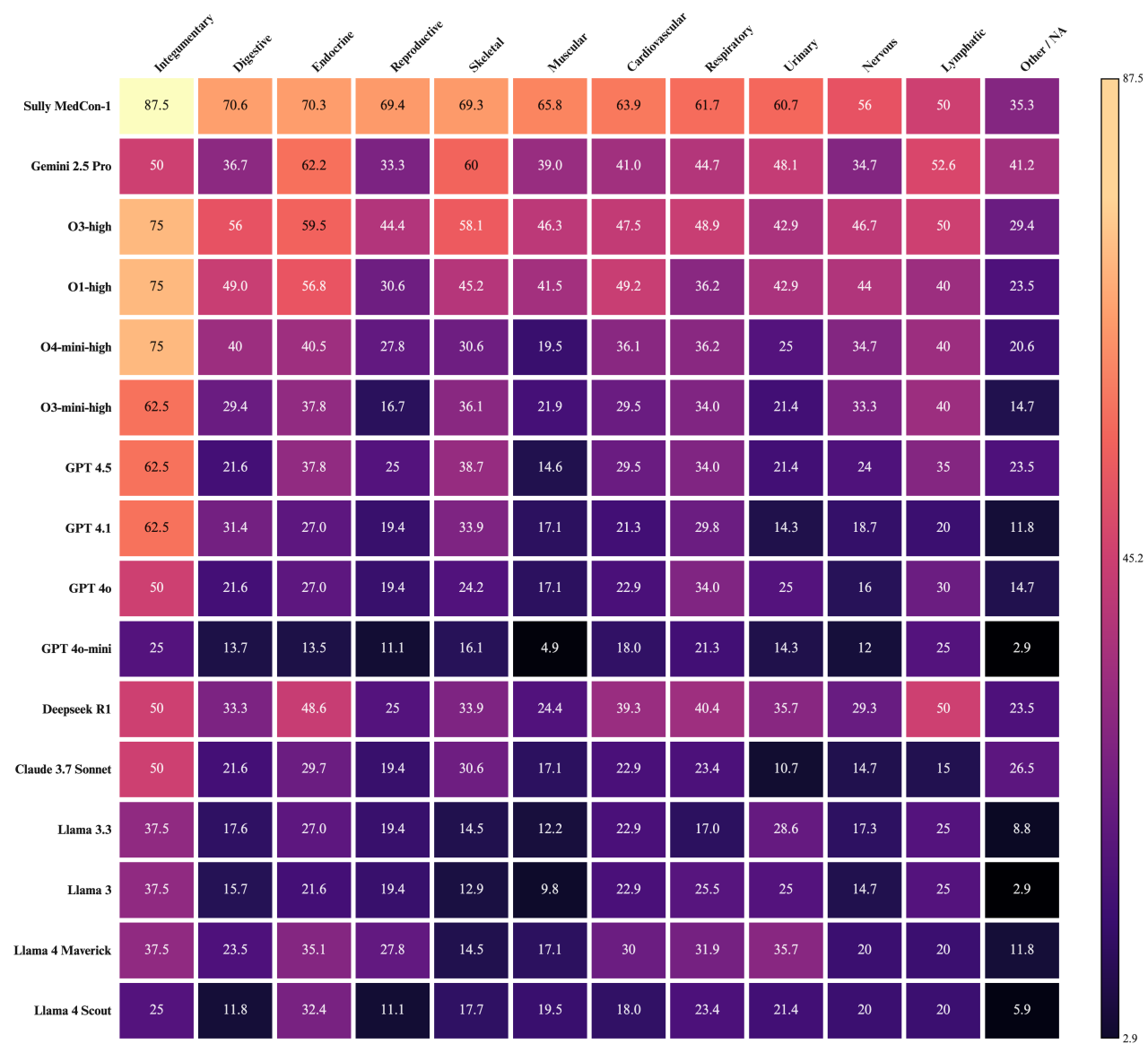


Figure 10: Body system specific accuracy on the MedXpertQA Evaluation Set

C.3 MedMCQA Specialty Accuracy

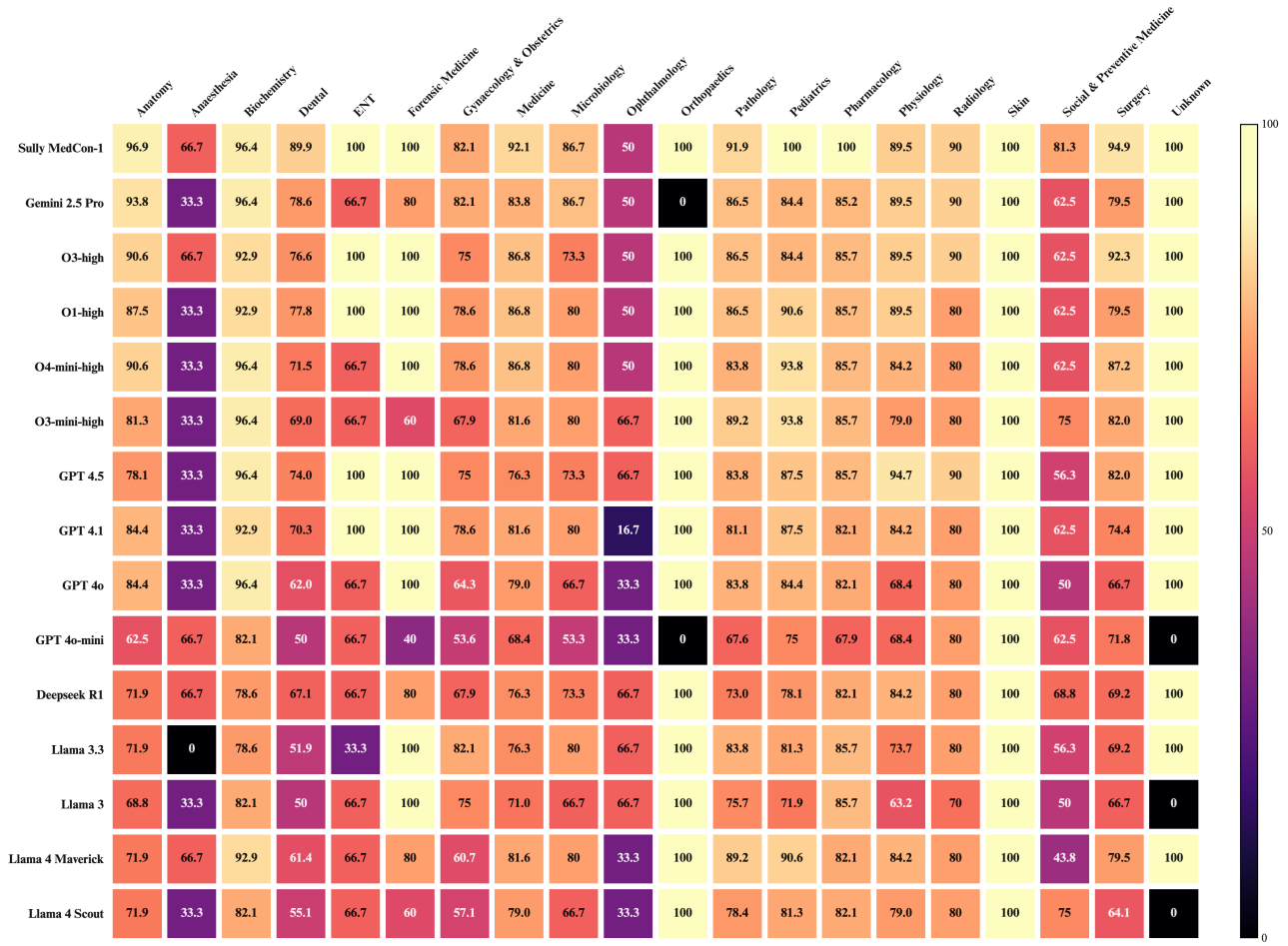


Figure 11: Medical specialty specific accuracy on MedMCQA Evaluation Set

Consensus Mechanism Example Output

D.0 A Full Consensus Response

Task Information

Case ID: Text-101

Patient Scenario: A 56-year-old male presents with severe right leg pain and swelling four days after undergoing a craniotomy for evacuation of a subdural hematoma. He experiences chills and nausea.

Medical History: Type 2 diabetes mellitus, chronic kidney disease requiring hemodialysis for 2 years.

Pre-admission Medications: Insulin, enalapril, atorvastatin, sevelamer.

Vitals:

- Temperature: 38.3°C (101.0°F)
- Pulse: 110/min
- Blood pressure: 130/80 mm Hg

Physical Exam: Right calf is swollen, warm, erythematous; pain with dorsiflexion. Peripheral pulses palpable. Cardiopulmonary exam unremarkable.

Laboratory Results:

- Hemoglobin: 10.1 g/dL
- Leukocyte count: 11,800/mm³
- Platelet count: 230,000/mm³
- Glucose: 87 mg/dL
- Creatinine: 1.9 mg/dL

Question: What is the most appropriate next step in managing this patient?

Answer Choices:

- (A) Surgical thrombectomy
- (B) Inferior vena cava filter
- (C) Unfractionated heparin therapy
- (D) Direct oral anticoagulants (DOACs)
- (E) Aspirin therapy only
- (F) Prophylactic antibiotics for cellulitis
- (G) Oral warfarin therapy
- (H) Compression stockings and observation
- (I) Thrombolytic therapy with tPA
- (J) Low molecular weight heparin therapy

Correct Answer: (B) Inferior vena cava filter

Metadata:

- Task Type: Treatment
- Body System: Cardiovascular
- Question Type: Reasoning

Consensus Summary and Final Recommendation

Case Summary: A 56-year-old male presents four days after craniotomy with signs of acute proximal DVT.

Key Considerations:

- Neurosurgical guidelines classify systemic anticoagulation within 7–14 days post-craniotomy as a major contraindication.
- LMWH is contraindicated due to both neurosurgical timing and ESRD.
- UFH is reversible but still systemic anticoagulation, with caution.
- An IVC filter is the standard temporizing measure for proximal DVT when anticoagulation is contraindicated.

Final Consensus: Placement of an **Inferior Vena Cava (IVC) Filter** (Answer: **B**) as the most appropriate next step.

Residual Probability:

- UFH (C): 30% – possible with pristine imaging and neurosurgical approval.
- LMWH (J): <10% – contraindicated due to renal clearance and bleed risk.

Triage Model Response (Expert Configuration)

- **Expert 1:** Hematologist
- **Expert 2:** Nephrologist
- **Expert 3:** Vascular Surgeon
- **Expert 4:** Infectious Disease Specialist

Expert 1 – Hematologist

Clinical Reasoning:

- 56-year-old man, POD-4 after craniotomy for subdural hematoma evacuation.
- New, unilateral swollen, warm, tender calf with pain on passive dorsiflexion (Homan's sign).
- Low-grade fever and mild leukocytosis – common with acute DVT.
- Comorbid diabetes, CKD, and dialysis suggest a pro-thrombotic state.
- High pre-test probability for acute proximal DVT.

Hematologic/Neurosurgical Risk Assessment:

- Recent intracranial surgery is a major contraindication to full-dose anticoagulation (ACCP, ASH, SVS, NCC guidelines).
- Re-bleed risk post-craniotomy is highest in the first 7–14 days; studies show 2–6% symptomatic re-bleed with early heparin use.
- IVC filter recommended when anticoagulation is temporarily contraindicated.

Scientific/Pathophysiologic Reasoning:

- Risk of PE from thrombus propagation.
- Anticoagulants (UFH, LMWH, DOACs) impair hemostasis and increase bleed risk.
- IVC filter intercepts emboli without systemic anticoagulation.
- LMWH contraindicated in ESRD; DOACs not recommended at CrCl < 15 mL/min.

Elimination of Answer Choices:

- A – Surgical thrombectomy: Reserved for limb-threatening clots.
- B – IVC filter: Recommended in this case due to absolute contraindication to anticoagulation.
- C – UFH: Reversible but still risky in first 7–14 days post-craniotomy.

- D – DOACs: Contraindicated in both neurosurgery and ESRD.
- E – Aspirin: Ineffective for acute DVT.
- F – Antibiotics: No clear signs of cellulitis; risks missing a life-threatening VTE.
- G – Warfarin: Slow onset; requires heparin bridge, which is contraindicated.
- H – Compression/observation: Does not protect from PE.
- I – Systemic tPA: Contraindicated post-craniotomy.
- J – LMWH: Renally cleared, long-acting, high risk of accumulation in ESRD.

Final Recommendation: **B – Inferior vena cava filter** is the most appropriate next step due to immediate protection without increasing intracranial bleeding risk.

Second Most Likely Answer: C – Unfractionated heparin, if neurosurgical team confirms stable operative field and PE risk is deemed higher than ICH risk.

Mind Map of Reasoning:

DVT clues (swelling, pain, fever) Recent craniotomy (Day 4) | | | | --> ACUTE PROXIMAL DVT <-----+ | | Dialysis → DOAC/LMWH less suitable | Risk of PE → needs protection | Management decision tree |- Anticoagulation (heparin/LMWH/DOAC) - NO (bleed risk) |- Thrombolysis - NO (ICH risk extreme) |- Surgical thrombectomy - Only limb-threatening ⊥- Mechanical protection → IVC FILTER (Yes)

Unknowns / Additional Considerations:

- Awaiting duplex ultrasound for confirmation.
- Future coordination with neurosurgery for anticoagulation timing.
- Consider filter retrieval when safe; continue mechanical calf pumps.

Expert 2 – Nephrologist

Clinical Summary: 56-year-old male with type 2 diabetes mellitus and end-stage renal disease on hemodialysis presents four days post-craniotomy with right calf pain, swelling, warmth, erythema, and a positive Homan's sign — all highly suggestive of acute DVT.

Major Considerations:

- **Recent neurosurgery** → significantly increases risk of catastrophic bleeding if anticoagulated.
- **ESRD on dialysis** → renally cleared agents (e.g., LMWH, DOACs) pose risk of accumulation and bleeding.
- **Unfractionated heparin (UFH)** is preferred for its reversibility, short half-life, and non-renal clearance.

Answer Choice Evaluation:

- (A) **Surgical thrombectomy:** Reserved for limb-threatening cases; not first-line for typical DVT.
- (B) **IVC filter:** Appropriate if anticoagulation is contraindicated. In this case, UFH is deemed safe enough to try first.
- (C) **Unfractionated heparin:** Most appropriate initial treatment; reversible and safe in ESRD.
- (D) **DOACs:** Contraindicated in ESRD; less easily reversed in emergency.
- (E) **Aspirin:** Inadequate for DVT treatment.
- (F) **Antibiotics:** Inappropriate without clear evidence of cellulitis; DVT more likely.
- (G) **Warfarin:** Requires bridging with heparin; too slow as sole therapy.
- (H) **Observation/compression:** Insufficient for confirmed DVT with high PE risk.
- (I) **tPA:** Contraindicated due to recent craniotomy and high ICH risk.
- (J) **LMWH:** Renally cleared; dangerous in ESRD and post-neurosurgical context.

Mind Map of Reasoning:

- **Patient Profile:** 56-year-old male, post-craniotomy, DM, CKD on dialysis |-
- Signs/Symptoms:** Right calf swelling, warmth, erythema, pain on dorsiflexion |-
- Laboratory:** Mild leukocytosis, creatinine elevated |-
- Differential:** DVT vs. cellulitis (DVT favored by exam)
- **Treatment Considerations:** | - Need for rapid anticoagulation | - Recent neurosurgery → bleeding risk → prefer reversible anticoagulants | - CKD on dialysis → avoid LMWH and DOACs
- **Answer Evaluation:** | - (C) UFH: Reversible, short half-life, OK in CKD | - (B) IVC filter: Backup if anticoagulation not possible | - Others: Do not meet safety/efficacy needs

Final Recommendation: C – Unfractionated heparin is the most appropriate next step, balancing safety in renal impairment and reversibility after recent surgery.

Second Most Likely Answer: B – IVC filter would be appropriate if UFH were contraindicated due to bleeding risk.

Expert 3 – Vascular Surgeon

Clinical Summary: The patient presents with classic signs of deep vein thrombosis (DVT) — leg swelling, warmth, erythema, pain on dorsiflexion — following a recent craniotomy. The clinical suspicion is further supported by leukocytosis and his known history of chronic kidney disease requiring hemodialysis.

Reasoning and Recommendation:

- The post-operative state and CKD create a pro-thrombotic environment.
- Anticoagulation is considered the cornerstone of treatment for confirmed or highly suspected DVT.
- **LMWH is preferred** due to its reported efficacy and perceived safety in renal impairment.
- **UFH is offered as a close alternative** if LMWH is contraindicated or if close titration and reversibility are required.

Final Recommendation: J – Low molecular weight heparin (LMWH) is recommended as the initial treatment.

Second Most Likely Answer: C – Unfractionated heparin (UFH), particularly if monitoring is required or renal clearance is a concern.

Expert 4 – Infectious Disease Specialist

Clinical Presentation and Differential Diagnosis:

- 56-year-old male, 4 days post-craniotomy for subdural hematoma evacuation.
- Right calf swelling, warmth, erythema, pain on dorsiflexion (Homan's sign).
- Low-grade fever (38.3 °C), leukocytosis (11,800/mm³).
- High suspicion for DVT given pro-thrombotic context (DM, ESRD, immobility).
- Differential includes cellulitis and superficial thrombophlebitis, though DVT is favored.
- Imaging (duplex ultrasound) assumed positive or pending.

Key Clinical/Subtle Features:

- Recent intracranial surgery creates high risk for hemorrhage with anticoagulation.
- ESRD alters pharmacokinetics — LMWH and DOACs pose renal clearance risks.
- Inflammatory signs (fever, leukocytosis) likely due to DVT, not infection.

Answer Choice Evaluation (with Probabilities):

- A – Surgical thrombectomy: Rarely indicated; only for limb-threat (0.02)

- B – IVC filter: Appropriate when anticoagulation is contraindicated post-craniotomy (0.55)
- C – UFH: Reversible, short half-life, but still poses early bleed risk (0.25)
- D – DOACs: Contraindicated in ESRD and early post-craniotomy (0.03)
- E – Aspirin: Inadequate for treatment (0.01)
- F – Antibiotics: No strong evidence of cellulitis (0.01)
- G – Warfarin: Slow onset, needs heparin bridging (0.03)
- H – Observation/compression: Not sufficient for proximal DVT (0.02)
- I – tPA: Strongly contraindicated post-neurosurgery (0.02)
- J – LMWH: Poor choice due to renal clearance and bleeding risk (0.06)

Final Recommendation: B – Inferior vena cava filter is the best option due to high bleed risk and immediate need for PE prevention.

Second Most Likely Answer: C – Unfractionated heparin, preferred if absolutely necessary, due to its reversibility.

Mind Map of Reasoning:

- 1) Patient Profile: |- Age: 56 |- Comorbidities: DM2, CKD on HD |- Recent Event: Craniotomy for SDH (Day 4)
 - 2) Presentation: |- Right calf swelling, erythema, warmth |- Dorsiflexion pain (Homan's sign) |- Fever, leukocytosis |- No cardiopulmonary findings
 - 3) Differential: |- DVT |- Cellulitis |- Superficial thrombophlebitis
 - 4) Diagnosis: Presumed proximal DVT (duplex US pending/high clinical suspicion)
 - 5) Management Constraints: |- High bleeding risk (post-craniotomy) |- CKD: LMWH/DOAC accumulation |- Need to prevent PE
 - 6) Treatment Options: |- Anticoagulation (UFH, LMWH, DOACs, warfarin) |- IVC filter |- Thrombolysis |- Surgical options |- Conservative care
 - 7) Best Choice: IVC filter (prevents PE, no systemic bleed risk) 8) Second Best: UFH (reversible, titratable)
- Unknowns Considered: - Duplex ultrasound confirmation - Extent of thrombus (proximal vs distal) - Neurological stability and neurosurgical input - Hematology consultation if needed