Enhancing Fine-Tuning-Free Clinical Reasoning via Test-Time Scaling

Ji Young Byun
Johns Hopkins University
Baltimore, MD 21218
jbyun13@jhu.edu

Young-Jin Park
Massachusetts Institute of Technology
Cambridge, MA 02139
youngp@mit.edu

Navid Azizan
Massachusetts Institute of Technology
Cambridge, MA 02139
azizan@mit.edu

Rama Chellappa
Johns Hopkins University
Baltimore, MD 21218
rchella4@jhu.edu

Abstract

As a cornerstone of modern healthcare, artificial intelligence is expected to support diverse medical tasks, with large language models (LLMs) offering a promising path to enhanced capabilities. The proficiency of LLMs in text-based tasks has not yet translated to their widespread application for reasoning-based diagnosis in medical imaging. This gap is exacerbated by the impracticality of supervised fine-tuning for clinical reasoning tasks, owing to limited data availability and high annotation costs. In this work, we introduce a fine-tuning-free framework for medical image diagnosis that enhances reasoning through test-time scaling (TTS). Our approach operates in two stages: given either visual or textual inputs, candidate representations or reasoning steps are generated, and aggregated through a self-consistency decoding strategy to yield robust final predictions. This framework avoids the need for expensive supervision while leveraging additional inference-time computation to improve reliability. We provide a analytical justification—deriving scaling laws that characterize when and how TTS yields reliable gains—and a comprehensive empirical evaluation across medical benchmarks spanning textual and visual modalities. Results demonstrate consistent improvements over single-pass inference baselines, with performance gains of up to 30 percentage points, highlighting the potential of TTS as a practical pathway toward trustworthy medical reasoning without specialized reward models or domain-specific fine-tuning.

1 Introduction

Large language models (LLMs) and vision-language models (VLMs) show strong performance across diverse domains, including mathematics [Snell et al., 2024], robotics [Li et al., 2023a], autonomous driving [Qian et al., 2024], and scientific research [Xu et al., 2023, Roberts et al., 2024]. A central factor in these advances is their ability to perform *reasoning*. While conventional deep learning models are often treated as black-box predictors, outputting a final result without rationale, recent reasoning language models can articulate intermediate steps explaining how answers are derived.

In medicine, where decisions are safety-critical, transparent reasoning is particularly important. Beyond predicting class probabilities, models should justify their decisions in alignment with clinical workflows, and their outputs should be evaluated to mitigate risks of deviating from clinical standards

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

[Johnson et al., 2023]. To address these requirements, recent studies have investigated vision-language models (VLMs), which can generate intermediate reasoning in natural language. In particular, VLMs have been applied to medical image diagnosis, commonly referred as visual question answering (VQA) [Li et al., 2023b, Tu et al., 2024, Moor et al., 2023].

Explicit multi-step reasoning methods, such as chain-of-thought (CoT) prompting [Wei et al., 2022, Temsah et al., 2024], produce step-by-step explanations that enhance problem-solving ability. These approaches have proven effective in domains such as arithmetic and symbolic reasoning, but their use in medicine remains limited. Multi-stage reasoning aligns naturally with clinical practice, where clinicians sequentially observe, interpret, and diagnose. Early studies have applied CoT-style prompting to medical tasks [Liu et al., 2024, Tu et al., 2025], allowing models to explore multiple hypotheses before reaching a conclusion.

However, the effectiveness of multi-stage reasoning often depends on fine-tuning with large collections of annotated reasoning processes. In medicine, such annotations require domain experts and are costly to obtain. This scarcity motivates approaches that *do not rely on fine-tuning*, including zero-shot methods, which are promising for extending reasoning-capable language models (LMs) to medical domains without extensive supervision.

Zero-shot prediction with LMs, however, often yields suboptimal performance. To address this, *test-time scaling* (TTS) has recently emerged as a promising inference paradigm. The key idea is to allocate additional computation during inference to improve a model's reasoning ability. A common strategy is "parallel thinking", where multiple candidate outputs are sampled and aggregated, rather than relying on a single generated output (i.e., single-pass decoding) [Yao et al., 2023, Snell et al., 2024]. These approaches, ranging from majority voting [Wang et al., 2022] to verifier-based selection [Cobbe et al., 2021, Uesato et al., 2022, Wang et al., 2023, Lightman et al., 2024], have demonstrated strong performance in domains requiring complex reasoning, such as mathematics.

Directly transferring TTS methods, particularly those leveraging verifiers, to medical applications presents significant challenges. Verifiers, known as reward models, are often unavailable in the medical domain because their training requires vast amounts of labeled reward data. For example, Qwen-PRM [Zhang et al., 2025], a reward model used for mathematical reasoning, required 4.5 million labels for its training. Therefore, TTS in medicine have focused on reward-free inference schemes, such as self-consistency decoding [Singhal et al., 2025], mostly on textual benchmarks.

Despite these efforts, our understanding of TTS in medicine remains limited. Key open questions include: **under what conditions does TTS improve performance?** and **can these methods extend beyond text to multimodal medical VQA tasks?** Motivated by these limitations, this work presents a simple, effective fine-tuning-free framework integrating a TTS strategy to enhance clinical reasoning and support reliable medical diagnosis, without additional supervision or a specialized reward model.

Our key contributions are summarized as follows:

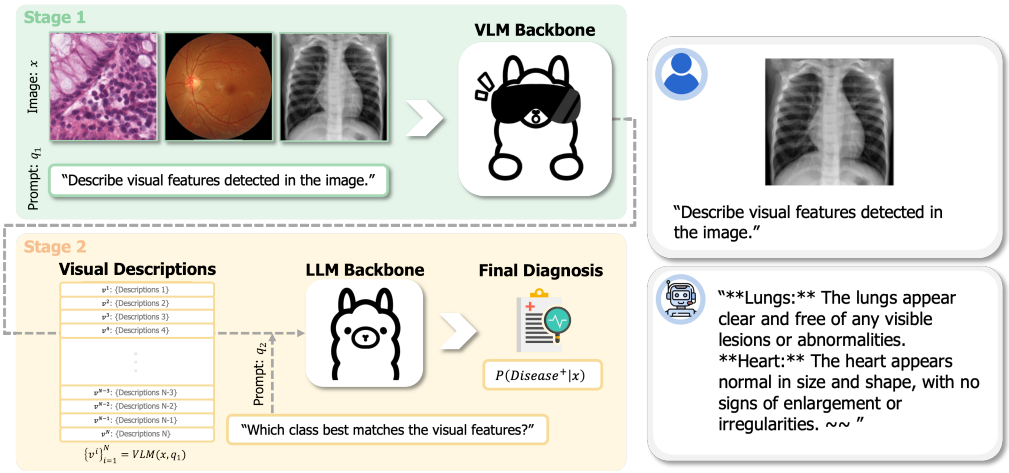
1. *Framework.* We investigate inference strategies (direct answering and CoT) and introduce a two-stage reasoning framework for medical VQA, where a VLM produces textual descriptions that are aggregated by an LLM for diagnosis.
2. *Empirical Validation.* We evaluate TTS on test-time inference strategies and show consistent improvements, with gains of up to 30.4 percentage points over single-pass baselines.
3. *Analytical Justification.* We present a characterization of TTS, deriving scaling laws that describe how performance improves with the number of samples and identifying conditions under which TTS yields reliable gains.

A detailed discussion of related work is in the Appendix A.

2 Methods

This paper presents an approach to modality-agnostic medical question answering (QA), with the goal is to generating accurate answers to clinically relevant questions based on given input data without requiring task-specific fine-tuning. Formally, a medical QA problem can be represented as a triplet $(\boldsymbol{x}, q, y)$, where:

- $\boldsymbol{x}$ denotes the *context*, which may take different modalities.



(a) (b)

Figure 1: Our proposed test-time-scaled reasoning framework for reliable zero-shot medical image diagnosis. Panel (a) details Stage 1, in which the VLM receives an image and a text prompt to generate N visual description samples, with Stage 2 applying a test-time scaling technique to determine the final diagnostic probability. Panel (b) shows a representative example of a healthy subject's chest X-ray paired with textual prompt and a generated visual description from VLM in Stage 1.

- q represents the *problem description*, expressed as a natural language query about the context.
- y is the *ground truth answer*, serving as the target output for the model.

This formalization provides a unified framework that accommodates a broad spectrum of medical QA tasks, ranging from text-based multiple choice to image-based VQA tasks. For example, in a medical VQA setting, the context $\boldsymbol{x}$ may correspond to a chest X-ray, the query q would be a natural language prompt such as "Does this patient have pneumonia?", and the ground truth answer y is a categorical label (e.g., 0 – normal, 1 – pneumonia).

2.1 Test-Time Inference Strategies

This section describes two widely adopted, single-step inference strategies for medical QA—zero-shot and CoT reasoning—and introduces a two-stage framework (illustrated in Figure 1) that explicitly decomposes the reasoning process into two distinct phases.

2.1.1 Direct Answering

Consider a language model (LM), such as Llama 3.2-Vision instruction-tuned model, that takes a textual prompt q and a context $\boldsymbol{x}$ and produces an answer a. A straightforward way to use such a model for diagnosis is to directly request a prediction of the target categorical label. For instance, we can set $q \leftarrow$ "`Given a pediatric chest X-ray image, classify it as 0 (normal) or 1 (pneumonia).`" The model directly provides a final answer without requiring any reasoning. We refer to this as the *Direct Answering* method.

2.1.2 Chain-of-Thought (CoT) Prompting

An alternative is to prompt the model to provide a step-by-step explanation before giving the final answer. This can be done by adding a phrase such as "`Let's think step by step.`" before the answer prompt [Kojima et al., 2022]. This method, commonly called chain-of-thought (CoT), encourages the model to reveal its reasoning process rather than just providing the final classification. We denote this approach as *one-stage CoT* method.

2.1.3 Two-Stage Reasoning for VQA

According to recent theoretical and experimental evidence [Abbe et al., 2025], the Transformer architecture often benefits when a complex task is decomposed into simpler sub-tasks. Motivated by this observation, we propose *describe-then-diagnose*, a two-stage approach to help the Transformer arrive at a more accurate diagnosis in a VQA setting and test-time compute scaling friendly.

Visual Description Generation. We first instruct the VLM to generate descriptions on visual features of the input image without directly querying for a diagnosis, illustrated in Figure 1b. Concretely, we prompt the VLM as follows: $v = \mathsf{VLM}(\boldsymbol{x}, q_1)$ where q_1 can be `"Describe visual features detected in the image"`.

Diagnosis from Descriptions. The generated visual descriptions v is then provided as input to a (potentially different) LLM that produces a final diagnosis. For example, we can construct the query: $q_2(v) :=$ "`Decide which class best matches the visual features described: 0 (normal) or 1 (pneumonia). **Features:** {features}`", where we substitute {`features`} with the previously generated v. Then, the diagnosis is obtained via:

$$a = \mathsf{LLM}\big(q_2(v)\big) = \underbrace{\mathsf{LLM}\Big(\underbrace{\mathsf{VLM}(\boldsymbol{x}, q_1)}_{\text{Describe}}, q_2\Big)}_{\text{Diagnose}} \tag{1}$$

2.2 Enhanced Clinical Reasoning via Scaling Test-Time Compute

General-purpose language models such as LLAMA or DEEPSEEK often struggle to provide accurate answers in medical QA, and finetuning is prohibitively expensive due to the scarcity of expert-annotated data. Therefore, we investigate the applicability of test-time scaling techniques—recently introduced in mathematical reasoning tasks [Yao et al., 2023, Snell et al., 2024]. We adopt self-consistency decoding [Wang et al., 2022] given the absence of reliable reward models in medicine.

One-Stage TTS. We estimate class probabilities by sampling N independent outputs from a reasoning language model RLM under randomized decoding (e.g., temperature scaling [Guo et al., 2017]). Let the label space be $\mathcal{Y} = \{1, \ldots, C\}$. For each draw $i \in \{1, \ldots, N\}$, the model produces an answer string $a^{(i)}$, which we map to a class via a parseϕ : text$\to\mathcal{Y}$ (e.g., extracting "A/B/C/D" or $\{1, \ldots, C\}$). Denote the parsed class by $\hat{a}^{(i)} = \phi\big(a^{(i)}\big) \in \mathcal{Y}$. Formally,

$$\{a^{(i)}\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mathsf{RLM}(\boldsymbol{x}, q), \qquad \hat{y}^{(i)} = \phi\big(a^{(i)}\big). \tag{2}$$

Each $\hat{y}^{(i)}$ can be viewed as a draw from the LM-induced predictive distribution over classes, $p(y \mid \boldsymbol{x}, q)$. We estimate these class probabilities by Monte Carlo:

$$\widehat{p}(y = c \mid \boldsymbol{x}, q) \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\big(\hat{y}^{(i)} = c\big). \tag{3}$$

The final prediction is the maximum-probability class under this estimate:

$$\hat{y} \;=\; \arg \max_{c \in \{1, \ldots, C\}} \widehat{p}(y = c \mid \boldsymbol{x}, q). \tag{4}$$

Two-Stage TTS. In a two-stage inference framework, we can apply TTS both in the description stage and in the diagnosis stage. Formally,

$$\{v^{(i)}\}_{i=1}^{N} \overset{\text{i.i.d.}}{\sim} \mathsf{VLM}(\boldsymbol{x}, q_1) \tag{5}$$

$$\{a^{(i,j)}\}_{j=1}^{M} \overset{\text{i.i.d.}}{\sim} \mathsf{RLM}(v^{(i)}, q_2). \tag{6}$$

where $v^{(i)}$ denotes the i-th description sampled from the vision–language model in the first stage, and $a^{(i,j)}$ is the j-th diagnosis from the language model given that description in the second stage.

Empirically, we observe that even under randomized decoding, the diagnosis $a^{(i,j)}$ remains unchanged for a fixed description $v^{(i)}$. This indicates that *the predictive uncertainty originates from the reasoning process (description stage) rather than from the decision-making process (diagnosis stage)*. Consequently, there is no measurable gain from scaling test-time compute in the second stage, and we therefore set $M = 1$. The final class probabilities and prediction are then estimated in the same way as in the single-stage case.

3 Results and Discussion

3.1 Datasets and Models

To provide an initial proof of concept for our framework, we first evaluate it on text-based medical QA tasks using the Massive Multitask Language Understanding (MMLU) benchmark [Hendrycks et al., 2020]. We focus on six medically relevant domains: clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine. Since these are multiple-choice questions, all answer options are included in the prompt along with the question.

To further assess generalizability across modalities and disease types in VQA tasks, we conduct three disease classification tasks using PneumoniaMNIST, PathMNIST, and RetinaMNIST from MedMNIST v2 [Yang et al., 2023]. Specifically, pneumonia detection is performed using PneumoniaMNIST, which consists of 390 pneumonia cases and 234 normal cases from frontal X-ray images. PathMNIST is utilized for colorectal cancer classification, containing 1,233 cases of colorectal adenocarcinoma epithelium and 741 cases of normal colon mucosa. Diabetic retinopathy (DR) classification is explored with RetinaMNIST, which includes 226 cases of referable (i.e., non-proliferative or proliferative DR) and 174 normal cases from fundus images. All images are standardized to a resolution of 224 × 224.

For the MMLU benchmark, we primarily evaluate LLAMA-3.1-8B-INSTRUCT [Touvron et al., 2023] and DEEPSEEK-R1-DISTILL-LLAMA-8B [DeepSeek-AI, 2025], as as well as additional results with LLAMA-3.2-1B-INSTRUCT and LLAMA-3.2-3B-INSTRUCT. For the VQA tasks, we employ LLAMA-3.2-11B-VISION-INSTRUCT [Touvron et al., 2023]. Since the second stage of our two-stage inference framework admits flexible model selection, we further experiment with smaller text-only Llama models (1B, 3B, and 8B) as well as the medical-domain–specific MED42-V2-8B model [Christophe et al., 2024].

The detailed prompts for each task and inference setting are provided in the Appendix C.

3.2 Comparison with Baselines

We begin by comparing the proposed framework against conventional baselines. For each dataset, we evaluate three test-time inference settings: (1) direct answering, (2) one-stage CoT, and (3) the proposed two-stage reasoning framework for VQA tasks. Then, each setting is further assessed both with and without the proposed TTS strategy.

Table 1 and Table 2 show that the proposed TTS strategy consistently delivers substantial performance gains across diverse tasks, models, and test-time inference strategies. While one-stage CoT without TTS often yields marginal or negative gains, TTS shows strong effects on vision-centric tasks through multi-sample aggregation. Our analysis further reveals several key observations:

- **Consistent gains across tasks and models.** On the MMLU dataset (Table 1), our TTS yields improvements across six medical knowledge areas, up to 17.5 percentage points (pp). Similarly, for MedMNIST tasks (Table 2), TTS consistently boosts AUC and AP scores across image modalities, with gains up to 30.4 pp. This indicates the advantage of TTS is not task- or model-specific, but generalizes across text- and vision-based medical tasks.
- **TTS outperforms CoT.** While prompt engineering alone often yields marginal or unstable effects, as reflected in the performance gap between direct answering and one-stage CoT, TTS consistently produces gains regardless of the prompting strategy. This highlights TTS as a more reliable mechanism for enhancing model performance than reformatting instructions.

Method	Clinical Knowledge	Medical Genetics	Anatomy	Professional Medicine	College Biology	College Medicine
Llama-3.1-8B-Instruct						
Direct Answering	0.71	0.75	0.62	0.73	0.75	0.64
Direct Answering (**+TTS**)	0.72 (↑ 1.3pp)	0.78 (↑ 3.5pp)	0.65 (↑ 2.8pp)	0.77 (↑ 3.6pp)	0.77 (↑ 2.3pp)	0.67 (↑ 3.1pp)
One-stage CoT	0.71	0.77	0.66	0.71	0.73	0.65
One-stage CoT (**+TTS**)	**0.80 (↑ 9.2pp)**	**0.84 (↑ 7.6pp)**	**0.72 (↑ 5.7pp)**	**0.85 (↑ 14.3pp)**	**0.84 (↑ 11.2pp)**	**0.77 (↑ 11.8pp)**
DeepSeek-R1-Distill-Llama-8B						
Direct Answering	0.52	0.55	0.48	0.46	0.54	0.47
Direct Answering (**+TTS**)	0.56 (↑ 3.8pp)	0.62 (↑ 6.8pp)	0.51 (↑ 3.4pp)	0.57 (↑ 10.6pp)	0.62 (↑ 8.1pp)	0.52 (↑ 5.0pp)
One-stage CoT	0.61	0.63	0.53	0.58	0.65	0.58
One-stage CoT (**+TTS**)	**0.73 (↑ 12.1pp)**	**0.80 (↑ 17.5pp)**	**0.64 (↑ 11.3pp)**	**0.72 (↑ 14.4pp)**	**0.80 (↑ 15.2pp)**	**0.74 (↑ 15.5pp)**

Table 1: Accuracy on six medical domains of MMLU using different prompting strategies. We compare **baselines** (direct answering and one-stage chain-of-thought (CoT)) with our **test-time scaling (TTS)** variants for $N = 64$. Across both LLAMA-3.1-8B-INSTRUCT and DEEPSEEK-R1-DISTILL-LLAMA-8B, applying TTS consistently improves performance over their respective baselines, with the largest gains observed for one-stage CoT, up to 17.5 percentage points (pp).

Method	Pneumonia		Colorectal Cancer		Diabetic Retinopathy	
	AUC	AP	AUC	AP	AUC	AP
Llama-3.2-11B-Vision-Instruct						
Direct Answering	0.50	0.62	0.56	0.65	0.61	0.63
Direct Answering (**+TTS**)	0.74 (↑ +24.2pp)	0.79 (↑ +16.9pp)	0.56 (↑ +0.5pp)	0.66 (↑ +0.3pp)	**0.71 (↑ +10.0pp)**	0.74 (↑ +11.1pp)
One-stage CoT	0.53	0.64	0.48	0.62	0.58	0.61
One-stage CoT (**+TTS**)	0.78 (↑ +24.9pp)	0.83 (↑ +18.8pp)	0.53 (↑ +5.4pp)	0.64 (↑ +2.4pp)	0.67 (↑ +9.2pp)	**0.74 (↑ +13.4pp)**
Two-stage Reasoning	0.52	0.63	0.54	0.65	0.57	0.61
Two-stage Reasoning (**+TTS**)	**0.82 (↑ +30.4pp)**	**0.86 (↑ +22.8pp)**	**0.65 (↑ +10.9pp)**	**0.75 (↑ +10.6pp)**	**0.71 (↑ +13.5pp)**	**0.74 (↑ +13.0pp)**

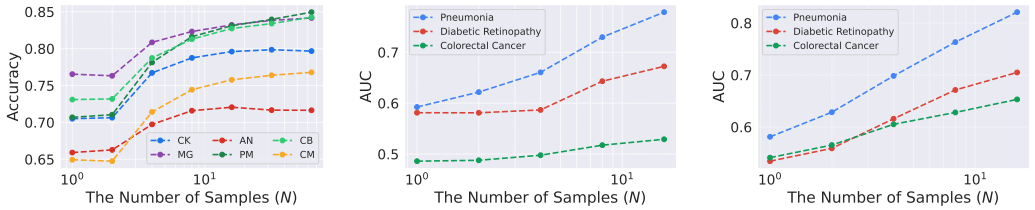
Table 2: Results on three MedMNIST datasets: PneumoniaMNIST (pneumonia detection), PathMNIST (colorectal cancer classification), and RetinaMNIST (diabetic retinopathy detection). We compare **baselines** (direct answering, one-stage CoT, and two-stage reasoning) with their **test-time scaling (TTS)** variants for $N = 16$. Across all datasets, applying TTS yields substantial and consistent improvements, with the two-stage reasoning + TTS method achieving the best overall performance. Largest gains observed for two-stage reasoning, up to 30.4 percentage points (pp). Metrics reported are AUC (area under the ROC curve) and AP (area under the precision–recall curve).

- **Strong effects on vision tasks.** Our TTS strategy achieves its most pronounced improvements on vision-centric tasks (Table 2). A single-pass VLM often overlooks subtle visual cues or produces ambiguous descriptions, whereas the multi-sample nature of TTS allows diverse perspectives to be aggregated into a more reliable representation. Importantly, these substantial improvements arise not from prompt design alone but from the synergy between structured reasoning and TTS, with scaling serving as the key driver of robustness in complex vision tasks.

3.3 Scaling Laws for Test-Time Compute

We analyze the effectiveness of TTS for medical image diagnosis in detail. As illustrated in Figure 2, we systematically vary the number of TTS samples, ranging from $N = 1$ (i.e., single-pass inference) up to $N = 64$ for text-based MMLU benchmarks and up to $N = 16$ for vision-based disease classification tasks, and evaluate the resulting model performance.

Figure 2 shows **performance scales monotonically with compute scale**, particularly up to medium sample sizes (e.g., $N \leq 16$): accuracy improves substantially as the number of samples increases. By



(a) One-stage CoT (Text) (b) One-stage CoT (Vision) (c) Two-Stage Reasoning (Vision)

Figure 2: The effect of sample size (N) in TTS setting. Increasing the sample size boosts performance across different datasets and inference methods, following a power law. LLAMA-3.1-8B-INSTRUCT and LLAMA-3.2-11B-VISION-INSTRUCT are used for text and vision tasks, respectively.

aggregating multiple complementary reasoning processes, the model reduces its reliance on a single, potentially flawed explanation. More comprehensive results are reported in Figure 4 in the Appendix.

This scaling behavior parallels our theoretical justifications in Corollary 1 in Section 3.6 as well as prior observations in Beeching et al. in language model reasoning tasks, where increasing test-time compute improves accuracy predictably. Importantly, the observed gains in medical image diagnosis suggest such scaling laws extend beyond mathematical reasoning to multimodal medical applications.

From a practical standpoint, this result underscores a critical lesson: **relying on a single model output is unreliable in medical tasks**, as LLMs can generate plausible yet misleading information in specialized contexts[1]. In contrast, test-time compute elevates diagnostic performance up to about 80% without requiring additional fine-tuning or retraining This highlights TTS as a promising avenue for improving both reliability and safety in real-world medical AI deployment.

3.4 Ablation on LLM Size in the Second Stage

We further note that the two-stage reasoning framework not only demonstrates empirical superiority but also provides practical advantages. By explicitly separating visual information from reasoning about clinical decisions, enabling flexible combinations of models across stages. In this setup, VLMs are more specialized in capturing and describing visual features, whereas LLMs are better suited for following instructions and providing structured reasoning. This separation allows practitioners to select a VLM best suited for image analysis while pairing it with an LLM tailored to the target clinical domain. To evaluate this flexibility, we replace the Llama model in Stage 2 with a medical domain-specific model, MED42-V2-8B; we observe it achieves slightly better performance (Figure 3).

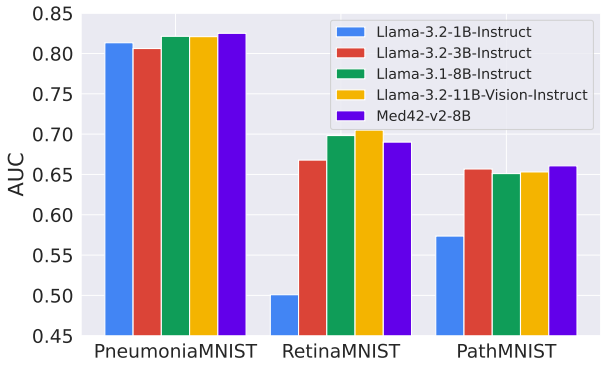


Figure 3: Proposed method's flexibility by replacing the Stage 2 LLM with varying model sizes. A 3B model performs comparably to an 11B model, demonstrating the benefit of the two-stage framework.

[1]For instance, across all test-time inference methods, single-sample inference occasionally yields AUCs of only 50–60% on disease classification tasks.

Method	Clinical Knowledge	Medical Genetics	Anatomy	Professional Medicine	College Biology	College Medicine
Llama-3.2-1B-Instruct						
Direct Answering	0.35	0.34	0.41	0.31	0.35	0.33
Direct Answering (**+TTS**)	**0.41** (↑ 6.2pp)	**0.39** (↑ 4.7pp)	**0.47** (↑ 6.5pp)	**0.38** (↑ 6.1pp)	**0.40** (↑ 4.4pp)	**0.39** (↑ 5.8pp)
One-stage CoT	0.16	0.16	0.21	0.15	0.18	0.17
One-stage CoT (**+TTS**)	0.04 (↓ 12.1pp)	0.02 (↓ 14.1pp)	0.08 (↓ 12.8pp)	0.01 (↓ 13.9pp)	0.05 (↓ 12.5pp)	0.07 (↓ 9.6pp)
Llama-3.2-3B-Instruct						
Direct Answering	0.60	0.65	0.57	0.68	0.65	0.54
Direct Answering (**+TTS**)	**0.64** (↑ 3.7pp)	**0.70** (↑ 4.4pp)	**0.64** (↑ 6.6pp)	**0.77** (↑ 8.7pp)	**0.70** (↑ 5.1pp)	**0.55** (↑ 0.8pp)
One-stage CoT	0.45	0.49	0.48	0.39	0.38	0.38
One-stage CoT (**+TTS**)	0.57 (↑ 12.4pp)	0.66 (↑ 17.1pp)	0.62 (↑ 14.5pp)	0.50 (↑ 11.2pp)	0.45 (↑ 7.8pp)	0.45 (↑ 7.2pp)

Table 3: Accuracy on six medical domains of MMLU using different prompting strategies. We compare **baselines** (direct answering and one-stage chain-of-thought (CoT)) with their **test-time scaling (TTS)** variants using $N = 64$. For LLAMA-3.2-1B-INSTRUCT, CoT prompting substantially degrades performance, and applying TTS further amplifies this degradation. In contrast, for LLAMA-3.2-3B-INSTRUCT, CoT also lowers baseline accuracy, but TTS recovers and yields consistent improvements across all domains. Overall, these results suggest that TTS is most effective when the underlying model achieves at least non-trivial accuracy (above random guessing, i.e., $\sim$25% for four-choice questions); otherwise, scaling may reinforce biased or uninformative reasoning.

Importantly, the modularity of this design enables adaptive scaling. Since smaller LLMs (e.g., 3B parameters) can provide sufficient reasoning capability, they can be employed in the second stage instead of the larger VLM used in the first stage (e.g., 11B), thereby reducing relative inference cost. To evaluate this property, we replace the original 11B LLM in Stage 2 with 1B, 3B, and 8B models. As shown in Figure 3, performance remains high even with a 3B model, closely matching its 8B and 11B counterparts. Notably, for pneumonia diagnosis, a 1B model achieves reasonable performance.

3.5 Model Capacity Matters for TTS

So far, we have observed that TTS consistently improves the zero-shot performance of reasoning models and exhibits stronger synergy with models that already possess sufficient reasoning ability (e.g., models of size 8B and above). To further investigate this trend, we conduct an ablation study on the MMLU dataset using smaller models (Table 3). These models reveal a critical finding: while TTS provides a modest boost to direct answering, applying a one-stage CoT prompt substantially degrades performance. This degradation is further amplified by TTS. For instance, with the 1B model, one-stage CoT cuts accuracy by more than half, and TTS on top of it drives accuracy to near-zero.

These results highlight the effectiveness of TTS depends critically on the baseline competence of the underlying models, and that naively introducing reasoning prompts can be counterproductive—as we show in Proposition 1 in subsequent Section 3.6. When a model exhibits non-trivial accuracy, TTS can enhance reasoning. Conversely, when a model struggles to reason, scaling may only reinforce biased or uninformative outputs, limiting its practical utility.

3.6 When Does TTS Help? An Analytical Justification

While self-consistency decoding has demonstrated strong empirical performance in medical applications [Singhal et al., 2023, 2025] and mathematical reasoning tasks [Beeching et al.], it remains underexplored whether TTS can be applied across different LLMs and how its scaling behavior unfolds (e.g., whether it converges quickly or grows monotonically). To address this gap, we first present a theoretical analysis of TTS based on self-consistency decoding. Proofs are in Appendix B.

Setup. Consider a C-class classification with true class $c^\star$. A single decode (vote) from the LM yields label $y \in \{1, \ldots, C\}$ with

$$\mathbb{P}(y = c^\star) = p, \qquad \mathbb{P}(y = j) = p_j \;\; (j \neq c^\star) \tag{7}$$

where $p + \sum_{j \neq c^\star} p_j = 1$. We draw N i.i.d. votes, let X_j be the number of votes for class j, and predict by majority vote $\hat{y}_{\mathrm{MV}} = \arg\max_j X_j$ (break ties uniformly at random). Define the strongest competitor $q := \max_{j \neq c^\star} p_j$.

Proposition 1 (Majority vote vs. strongest competitor)**.** *If $p > q$, then*

$$\mathbb{P}(\hat{y}_{\mathrm{MV}} \neq c^\star) \;\leq\; (C-1)\,\exp\Big(-\tfrac{N}{2}\,(p-q)^2\Big), \tag{8}$$

so the error decays exponentially in N, and improves as the margin $p - q$ grows (i.e., LM becomes more confident).

Conversely, if $q > p$, then

$$\mathbb{P}(\hat{y}_{\mathrm{MV}} = c^\star) \;\leq\; (C-1)\,\exp\Big(-\tfrac{N}{2}\,(q-p)^2\Big), \tag{9}$$

so $\hat{y}_{\mathrm{MV}}$ amplifies the wrong class as N grows.

Corollary 1 (Exponential scaling)**.** *If $p > q$, the error of $\hat{y}_{\mathrm{MV}}$ decays exponentially with N, and to achieve $\mathbb{P}(\hat{y}_{\mathrm{MV}} \neq c^\star) \leq \delta$ it suffices that*

$$N \;\geq\; \frac{2}{(p-q)^2}\,\log\Big(\frac{C-1}{\delta}\Big). \tag{10}$$

If $q > p$, then $\mathbb{P}(\hat{y}_{\mathrm{MV}} = c^\star)$ decays exponentially in N at the same rate.

Summary of the theoretical findings. Proposition 1 shows, if $p > q$, exponential decay of the error in N; if $q > p$, majority vote amplifies the wrong label. Hence: (1) self-consistent TTS improves with larger N *in regimes where the true class has the largest single-pass probability*, and (2) it is *effective only when the LLM is sufficiently confident*, in the sense of a nontrivial margin $p > q$.

4 Conclusion

In this work, we propose a fine-tuning–free framework that leverages test-time scaling (TTS) to enhance clinical reasoning in medical tasks. We demonstrate that TTS consistently improves performance over single-pass baselines, with gains up to 30.4 percentage points, and provide scaling laws to explain how and when such improvements emerge. Our analysis shows TTS is most beneficial when the underlying model possesses non-trivial baseline competence, as scaling amplifies informative reasoning. Our experiments also confirm that TTS generalizes beyond text, yielding strong effects on vision-centric tasks. This parallel thinking–based TTS method is significant as it improves zero-shot performance without costly supervision, addressing the scarcity of high-quality medical annotations for training verifiers or reward models. Future work includes exploring adaptive TTS strategies that dynamically allocate compute, investigating how TTS interacts with domain-specialized models, and conducting rigorous human-in-the-loop studies to assess integration with clinical workflows and ensure trustworthy decision-making.

References

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. *Advances in Neural Information Processing Systems*, 37:27850–27895, 2025.

Edward Beeching, Lewis Tunstall, and Sasha Rush. Scaling test-time compute with open models. URL `https://huggingface.co/spaces/HuggingFaceH4/blogpost-scaling-test-time-compute`.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025.

Douglas Johnson, Rachel Goodman, J Patrinely, Cosby Stone, Eli Zimmerman, Rebecca Donald, Sam Chang, Sean Berkowitz, Avni Finn, Eiman Jahangir, et al. Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*, pages rs–3, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14788–14798, 2025.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4542–4550, 2024.

Yuxiao Qu, Tianjun Zhang, Naman Garg, and Aviral Kumar. Recursive introspection: Teaching language model agents how to self-improve. *Advances in Neural Information Processing Systems*, 37:55249–55285, 2024.

Jonathan Roberts, Kai Han, Neil Houlsby, and Samuel Albanie. Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems*, 37:18695–18728, 2024.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Mohamad-Hani Temsah, Amr Jamal, Khalid Alhasan, Abdulkarim A Temsah, and Khalid H Malki. Openai o1-preview vs. chatgpt in healthcare: a new frontier in medical ai reasoning. *Cureus*, 16 (10), 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *CoRR*, abs/2312.08935, 2023. URL `https://doi.org/10.48550/arXiv.2312.08935`.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Related Work

A.1 Vision-Language Models in Medical Imaging

Conventional data-driven deep learning approaches parameterize a model with learnable parameters and train it on a dataset of image–label pairs. Such models are often treated as *black-box* predictors: they provide a final classification result but lack transparency, the reasoning underlying the diagnostic process non-interpretable. Given the safety-critical nature of medicine and the risk that generated content may deviate from clinical standards, rigorous evaluation is mandated to assess progress and mitigate harms [Johnson et al., 2023]. Addressing these interpretability and reliability concerns is paramount for the responsible deployment of AI in clinical settings.

On the other hand, the development of VLMs has rapidly progressed, enabling models to process both images and text for diverse applications, including robotics [Li et al., 2023a], autonomous driving [Qian et al., 2024], and scientific research [Xu et al., 2023, Roberts et al., 2024]. In the medical domain, early efforts focused on foundational tasks such as medical image captioning and VQA on datasets like VQA-RAD [Lau et al., 2018]. More recently, models like Med-PaLM [Tu et al., 2024], Med-Flamingo [Moor et al., 2023], and LLaVA-Med [Li et al., 2023b] have demonstrated strong performance in generating clinically-relevant text. These efforts are now being pushed further by more recent works such as VILA-M3 [Nath et al., 2025] and MedXpertQA [Zuo et al., 2025], which focus on more complex reasoning and comprehensive evaluations.

The use of multi-stage reasoning, a paradigm that breaks down a complex task into a series of intermediate steps, has gained significant traction as an alternative to end-to-end approaches. This aligns with the clinical workflow, where a clinician first observes an image and other patient information, analyzes the symptoms, and then formulates a diagnosis based on their observations and knowledge. For instance, recent works have explicitly incorporated multi-stage reasoning, such as CoT [Liu et al., 2024, Tu et al., 2025], to generate detailed diagnostic rationales and explain their decision-making process. More advanced methods have also emerged, including Tree-of-Thought (ToT) [Yao et al., 2023], which creates a tree-like structure of potential diagnostic paths and evidence. This allows the model to explore and evaluate multiple hypotheses simultaneously before reaching a final conclusion.

A.2 Test-Time Compute Scaling

TTS has become a prominent research area, offering a computationally efficient alternative to traditional retraining for enhancing model performance. By leveraging an increased computational budget at inference time, these strategies improve a model's accuracy and robustness without requiring any changes to its parameters or architecture. CoT [Wei et al., 2022, Temsah et al., 2024] is a notable example, where a model is prompted to generate a series of intermediate reasoning steps before arriving at the final answer. While effective, CoT can be sensitive to prompting and may not always yield consistent results. TTS, a related but distinct paradigm, further improves performance by moving beyond a single, deterministic output. Instead, TTS methods sample multiple candidate outputs and aggregate them to form a more robust and reliable final prediction.

A variety of TTS strategies have been explored, ranging from simple aggregation to more complex reasoning-based methods. Simple approaches like self-consistency [Wang et al., 2022] and majority voting rely on aggregating multiple generated outputs to improve reliability. More advanced techniques have significantly pushed performance boundaries on complex benchmarks. For instance, self-refinement [Qu et al., 2024, Madaan et al., 2023] is an iterative approach where a model critiques its own output and then revises it in a feedback loop. Similarly, verifier-based methods [Cobbe et al., 2021, Uesato et al., 2022] and process reward models [Wang et al., 2023] have achieved state-of-the-art results by training a separate model to select the best output. Recent works have validated these approaches on increasingly challenging benchmarks, such as MATH [Hendrycks et al., 2021], GSM8K [Cobbe et al., 2021], and the BIG-Bench Hard suite [Srivastava et al., 2023], demonstrating their strong performance in mathematical and symbolic reasoning tasks.

Although powerful AI techniques are promising, their application in medicine is still emerging. One direction for improving model performance is scaling "*deep thinking*" by increasing a model's computational budget for a single reasoning path, such as by expanding its token limit [Huang et al., 2025]. However, this approach faces significant challenges: it can lead to overthinking [Yang et al., 2025]. Consequently, "*parallel thinking*" strategies represent another important yet unexplored avenue in the medical domain. A key barrier to those advanced methods (e.g., Best-of-N, and beam

search [Snell et al., 2024]) is their reliance on reward models, which are often unavailable in medicine as they demand vast amounts of labeled data for training [2]. To this end, this paper explores the application of a reward-free TTS to medical image diagnosis by extending a majority voting strategy into a probabilistic framework that not only improves reliability.

B Omitted Proofs

Proof of Lemma 1.

Proof. For the first case, we bound the probability of error by applying a union bound over all possible failure modes. An error occurs if at least one competitor class $j \neq c^\star$ receives at least as many votes as the true class $c^\star$.

$$\begin{aligned}\mathbb{P}(\hat{y}_{\mathrm{MV}} \neq c^\star) &= \mathbb{P}\Big(\bigcup_{j\neq c^\star} \{X_j \geq X_{c^\star}\}\Big) \\ &\leq \sum_{j\neq c^\star} \mathbb{P}(X_j \geq X_{c^\star}).\end{aligned} \tag{11}$$

For each competitor j, let $D_j := X_{c^\star} - X_j$. The term $\mathbb{P}(X_j \geq X_{c^\star})$ is equivalent to $\mathbb{P}(D_j \leq 0)$. The quantity D_j is a sum of N i.i.d. random variables $V_i = \mathbf{1}\{y_i = c^\star\} - \mathbf{1}\{y_i = j\}$, where each $V_i \in \{-1, 0, 1\}$. The expectation is $\mathbb{E}[D_j] = N(p - p_j)$. By Hoeffding's inequality (with variable range $1 - (-1) = 2$):

$$\begin{aligned}\mathbb{P}(D_j \leq 0) &= \mathbb{P}\big(D_j - \mathbb{E}[D_j] \leq -N(p - p_j)\big) \\ &\leq \exp\left(-\frac{2(N(p-p_j))^2}{N \cdot 2^2}\right) \\ &= \exp\Big(-\tfrac{N}{2}(p - p_j)^2\Big).\end{aligned} \tag{12}$$

Since $q = \max_{k\neq c^\star} p_k$, we have $p - p_j \geq p - q$ for all $j \neq c^\star$. This implies $(p - p_j)^2 \geq (p - q)^2$. We can thus bound each term in the sum by the worst case:

$$\begin{aligned}\mathbb{P}(\hat{y}_{\mathrm{MV}} \neq c^\star) &\leq \sum_{j\neq c^\star} \exp\Big(-\tfrac{N}{2}(p - p_j)^2\Big) \\ &\leq \sum_{j\neq c^\star} \exp\Big(-\tfrac{N}{2}(p - q)^2\Big) \\ &= (C-1)\,\exp\Big(-\tfrac{N}{2}\,(p - q)^2\Big).\end{aligned} \tag{13}$$

For the second case, let $j^\dagger \in \arg\max_{j\neq c^\star} p_j$ so that $p_{j^\dagger} = q$. For the true class $c^\star$ to win, it must receive more votes than any other class, including the strongest competitor $j^\dagger$. Thus, the event $\{\hat{y}_{\mathrm{MV}} = c^\star\}$ is a subset of the event $\{X_{c^\star} > X_{j^\dagger}\}$.

$$\mathbb{P}(\hat{y}_{\mathrm{MV}} = c^\star) \leq \mathbb{P}(X_{c^\star} > X_{j^\dagger}). \tag{14}$$

Let $D := X_{c^\star} - X_{j^\dagger}$. The expectation is $\mathbb{E}[D] = N(p - q) < 0$. We bound $\mathbb{P}(D > 0)$. By Hoeffding's inequality:

$$\begin{aligned}\mathbb{P}(D > 0) &= \mathbb{P}\big(D - \mathbb{E}[D] > -N(p - q)\big) \\ &= \mathbb{P}\big(D - \mathbb{E}[D] > N(q - p)\big) \\ &\leq \exp\Big(-\tfrac{N}{2}(q - p)^2\Big).\end{aligned} \tag{15}$$

□

[2] This data is not merely correct answers but expert-annotated process supervision, where a model's step-by-step reasoning is evaluated. The high cost of clinical experts' time and the inherent complexity of medical judgment make acquiring this type of data prohibitively expensive and scarce.

C Prompts

For evaluation on the MMLU benchmark, we employed two primary prompt formats: direct answering and one-stage Chain-of-Thought (CoT). For medical image diagnosis on the MedMNIST dataset, we employed three prompt formats: direct answering, one-stage Chain-of-Thought (CoT), and our proposed two-stage reasoning.

C.1 Prompt Structure for MMLU Evaluation

C.1.1 Direct Answering Prompt

In the direct answering prompt, the model is instructed to select the correct letter choice without providing any intermediate explanation or reasoning. This setup evaluates the model's immediate knowledge of the subject matter. An example direct answering prompt is shown in the visualization below.

User Input

The following are multiple-choice questions (with answers) about {subject}. Provide your answer with "The answer is (X)" where X is the correct letter choice, with no additional explanation.

Question: {question}

Options: A. {o1}, B. {o2}, C. {o3}, D. {o4}

C.1.2 Chain-of-Thought (CoT) Prompt

The one-stage Chain-of-Thought (CoT) prompt instructs the model to produce step-by-step reasoning before selecting a final answer. We implement this by explicitly asking the model to "think step by step," then report the final letter choice. This setting assesses the model's reasoning ability in addition to its factual knowledge.

User Input

The following are multiple-choice questions (with answers) about {subject}. Think step by step and then finish your answer with "The answer is (X)" where X is the correct letter choice.

Question: {question}

Options: A. {o1}, B. {o2}, C. {o3}, D. {o4}

Assistant Response (Prefix)

Answer: Let's think step by step.

C.2 One-Stage Prompt Structure for MedMNIST

In medical imaging tasks with one-stage inference (i.e., direct answering and one-stage CoT), we use a direct-instruction format: the model receives a single-turn system prompt that specifies the classification task and the required output format. For one-stage CoT, we append the cue "Let's think step by step." to the prompt.

C.2.1 Pneumonia Detection

User Input

Your task is binary-class classification of ‘pneumonia’ against ’normal’. Given a given gray-scale pediatric chest X-Ray image, classify it as 0 (normal) or 1 (pneumonia). Make sure to put the answer (and only answer) inside `\boxed{}`.

C.2.2 Colorectal Cancer

User Input

Your task is binary-class classification of ‘malignant: colorectal adenocarcinoma epithelium’ against ’normal’. Given a hematoxylin & eosin stained histological image, classify it as 0 (normal) or 1 (malignant). Make sure to put the answer (and only answer) inside `\boxed{}`.

C.2.3 Diabetic Retinopathy

User Input

Your task is binary-class classification of ‘diabetic retinopathy (DR)’ against ’normal’. Given a retina fundus image, classify it as 0 (normal) or 1 (DR). Make sure to put the answer (and only answer) inside `\boxed{}`.

C.3 Two-Stage Prompt Structure for MedMNIST

In the two-stage reasoning setup for medical image diagnosis, the prompt is structured into two phases: (1) a general instruction asking the model to summarize the visual features from a given image, and (2) a task-specific prompt that provides detailed guidelines and questions, combined with the summary generated in the first stage (referred to as the note).

C.3.1 Stage 1 Prompt for All Tasks

User Input

Summarize the list of key observable features detected in the image using bullet points.

C.3.2 Stage 2 Prompt for Pneumonia Detection

User Input

You are a healthcare professional to provide accurate pneumonia diagnosis.
Task:
- You will receive a report describing a patient's pediatric chest X-Ray image.
- Your goal is to classify:
- 0 = normal
- 1 = pneumonia

Guidelines:
1. Carefully read the note.
2. Decide which class (0 or 1) best matches the clinical features described. Assume that all of the relevant details have been explained in the text.
3. Provide your final answer enclosed in `\boxed{}` with no additional explanation, e.g., `\boxed{1}`.

IMPORTANT:
- Strictly adhere to the format by outputting only the final grade inside `\boxed{}` and nothing else.

Note:
{note}

—

Question:
Based on the above note, what is the correct pneumonia diagnosis? Please consider that all necessary details have been provided in the text above. Remember to provide only the class (0 or 1) inside `\boxed{}`.

C.3.3 Stage 2 Prompt for Colorectal Cancer

User Input

You are a pathologist to provide an accurate colorectal adenocarcinoma epithelium diagnosis.
Task:
- You will receive a report describing a patient's hematoxylin & eosin stained histological image.
- Your goal is to classify the tissue type:
- 0 = normal
- 1 = malignant (colorectal adenocarcinoma epithelium)

Guidelines:
1. Carefully read the note.
2. Decide which class (0 or 1) best matches the clinical features described. Assume that all of the relevant details have been explained in the text.
3. Provide your final answer enclosed in `\boxed{}` with no additional explanation, e.g., `\boxed{1}`.

IMPORTANT:
- Strictly adhere to the format by outputting only the final grade inside `\boxed{}` and nothing else.
Note:
{note}

—

Question:
Based on the above note, what is the correct tissue type? Please consider that all necessary details have been provided in the text above. Remember to provide only the class (0 or 1) inside `\boxed{}`.

C.3.4 Stage 2 Prompt for Diabetic Retinopathy

User Input

You are an ophthalmologist to provide accurate diabetic retinopathy (DR) diagnosis.
Task:
- You will receive a report describing a patient's retina fundus image.
- Your goal is to classify:
- 0 = normal
- 1 = referrable

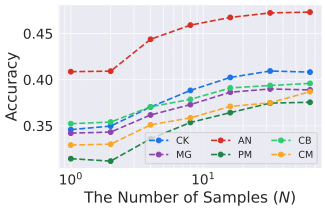
Guidelines:
1. Carefully read the note.
2. Decide which class (0 or 1) best matches the clinical features described. Assume that all of the relevant details have been explained in the text.
3. Provide your final answer enclosed in `\boxed{}` with no additional explanation, e.g., `\boxed{1}`.

IMPORTANT:
- Strictly adhere to the format by outputting only the final grade inside `\boxed{}` and nothing else.

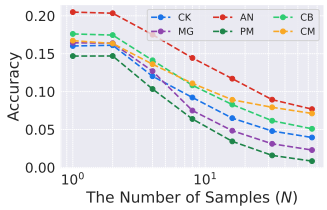
Note:
{note}

—

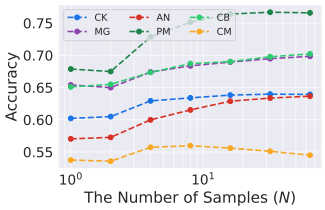
Question:
Based on the above note, what is the correct diabetic retinopathy (DR) diagnosis? Please consider that all necessary details have been provided in the text above. Remember to provide only the class (0 or 1) inside `\boxed{}`.



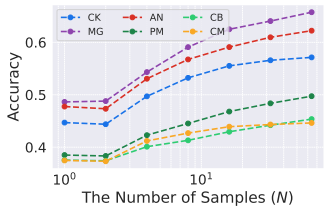
(a) Direct Answering (Text) w/ Llama-1B



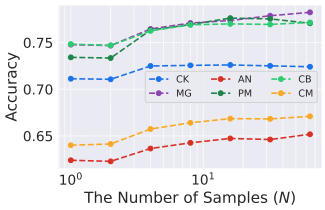
(b) One-stage CoT (Text) w/ Llama-1B



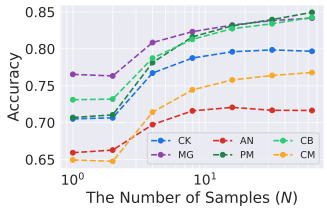
(c) Direct Answering (Text) w/ Llama-3B



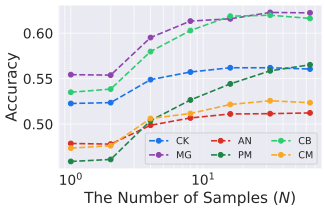
(d) One-stage CoT (Text) w/ Llama-3B



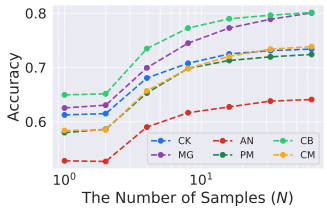
(e) Direct Answering (Text) w/ Llama-8B



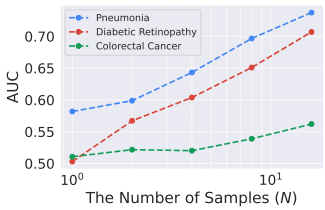
(f) One-stage CoT (Text) w/ Llama-8B



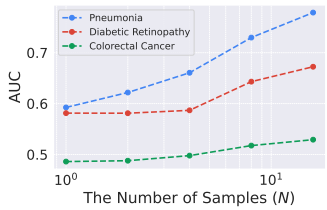
(g) Direct Answering (Text) w/ DeepSeek-8B



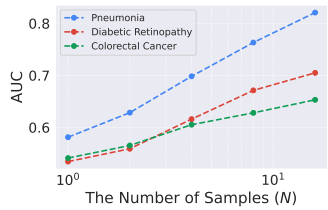
(h) One-stage CoT (Text) w/ DeepSeek-8B



(i) Direct Answering (Vision) w/ Llama-11B-Vision



(j) One-stage CoT (Vision) w/ Llama-11B-Vision



(k) Two-Stage Reasoning (Vision) w/ Llama-11B-Vision

Figure 4: A study examining the effect of sample size (N) in TTS setting. Increasing the sample size boosts performance across different datasets and inference methods, following a power law.