m1: Unleash the Potential of Test-Time Scaling for Medical Reasoning with Large Language Models

Xiaoke Huang[1], **Juncheng Wu**[1], **Hui Liu**[2], **Xianfeng Tang**[2], **Yuyin Zhou**[1]
[1] UC Santa Cruz [2] Amazon Research
https://github.com/UCSC-VLAA/m1

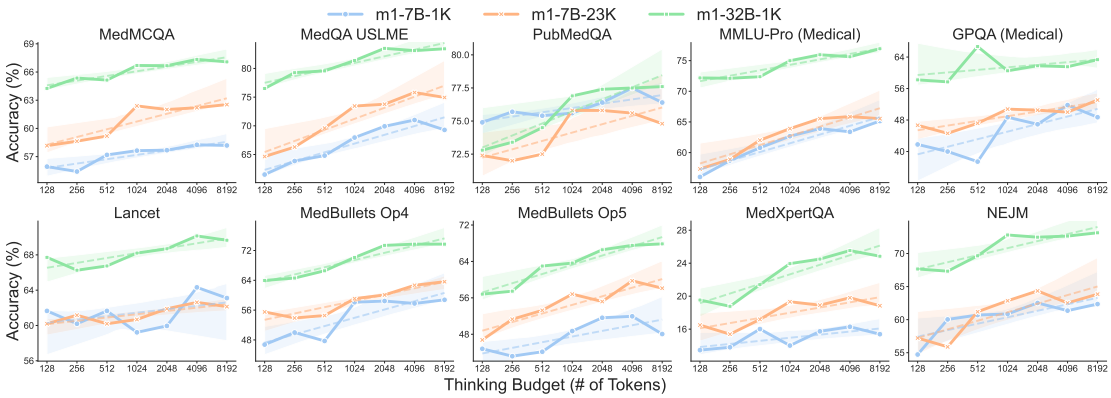


Figure 1: **Test-time scaling of `m1` series.** Each plot shows accuracy (%) vs. reasoning token budget for different `m1` model variants on various medical QA datasets. All models improve steadily as the thinking length increases, with the 32B model reaching the best accuracy. The linear regression lines are marked in dot line with their 95% confidence interval.

Abstract

Test-time scaling has emerged as a powerful technique for enhancing the reasoning capabilities of large language models (LLMs). However, its effectiveness in medical reasoning remains uncertain, as the medical domain fundamentally differs from mathematical tasks in terms of knowledge representation and decision-making processes. In this paper, we provide the first comprehensive investigation of test-time scaling for medical reasoning and present **m1**, a simple yet effective approach that increases a model's medical reasoning capability at inference. Our evaluation across diverse medical tasks demonstrates that test-time scaling (by increasing the "thinking" token budget) consistently enhances medical reasoning, enabling lightweight fine-tuned models under 10B parameters to establish new state-of-the-art performance, while our 32B model achieves results comparable to previous 70B-scale medical LLMs. However, we identify an optimal reasoning token budget of approximately 4K, beyond which performance may degrade due to overthinking. Budget forcing, which extends test-time computation through iterative prompts (e.g., appending "Wait"), helps models double-check answers but does not necessarily improve the overall medical QA performance and, in some cases, even introduces errors into previously correct responses. Taken together, our case-by-case analysis further identifies insufficient medical knowledge as a key bottleneck that prevents further performance gains through test-time scaling. To overcome this constraint, we find that increasing data scale, improving data quality, and expanding model capacity consistently enhance medical knowledge grounding,

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

enabling continued performance improvements—particularly on challenging medical benchmarks where smaller models reach saturation. These findings underscore fundamental differences between medical and mathematical reasoning in LLMs, highlighting that enriched medical knowledge, other than increased reasoning depth alone, is essential for fully realizing the benefits of test-time scaling.

1 Introduction

Test-time scaling has emerged as a promising direction to enhance LLM reasoning by enabling models to "think more" during inference [31]. OpenAI's o1 [9] demonstrated that significantly extending an LLM's chain-of-thought can yield remarkable gains in problem-solving ability in both STEM fields and the medical domain [16, 28], but the exact methodology was not disclosed, spurring many replication efforts. Among the most successful replication attempts is the open-source s1 method [16], which achieved remarkable results through a surprisingly simple approach. By fine-tuning a 32B parameter model on just 1K carefully curated examples with reasoning traces and implementing an inference control mechanism via a "Wait" token, s1 enabled the model to effectively double-check its work. This simple approach produced state-of-the-art results on challenging mathematical benchmarks, *e.g.*, outperformed OpenAI's o1-preview by up to 27%.

Despite these advances, applying test-time scaling to the medical domain remains largely underexplored [10]. The medical domain presents unique challenges for LLMs: questions often involve multi-step logical reasoning, accurate recall of medical knowledge, and careful consideration to avoid unsafe or harmful answers [4]. As the medical field fundamentally differs from mathematical tasks in terms of knowledge representation and decision-making processes, the effectiveness of test-time scaling for medical reasoning remains uncertain.

While advanced proprietary models like GPT-4 [8] and Med-PaLM [21] have achieved expert-level scores on medical exams [32], open-source medical LLMs still struggle to reliably solve complex medical problems. Improving reasoning in these models is critical, as healthcare applications demand not just factual accuracy but robust diagnostic and therapeutic reasoning capabilities. Existing medical reasoning LLMs such as HuatuoGPT-o1 [4] typically rely on computationally intensive methods like reinforcement learning with verification mechanisms. This raises a key question: *Can a simple test-time scaling strategy, with minimal fine-tuning, also unlock strong medical reasoning?*

In this paper, we answer in the affirmative by presenting `m1`, a lightweight methodology that adapts the test-time scaling paradigm to medical QA tasks. Our approach is straightforward: we curate a high-quality set of medical questions with detailed step-by-step solutions (only 1K / 23K examples), fine-tune open LLMs on this data, and at inference use test-time controls to ensure the model fully "thinks through" problems before answering. Figure 1 illustrates the outcome: as we allow the model to generate longer chains of thought (x-axis increasing), accuracy on various medical benchmarks consistently improves for our `m1` models. Notably, even our 7B-parameter model fine-tuned on 1K examples shows significant gains with more reasoning steps, and our 32B model achieves the highest scores across the board.

To better understand the impact of test-time scaling on medical reasoning in LLMs, we conduct a fine-grained study, systematically examining the effects of **thinking budgets, inference techniques, data curation**, and **model capacity**. While increasing the token budget consistently improves performance, we identify an optimal reasoning threshold of approximately 4K tokens, beyond which accuracy declines due to overthinking. In addition to increasing the token budget (Figure 1), reasoning can also be extended through budget forcing, wherein the model iteratively prolongs its thought process during inference [16]. However, unlike in mathematical reasoning—where iterative refinement often enhances accuracy—forcing additional reasoning in medical QA yields limited benefits and, in some cases, even degrades performance. This occurs when models with erroneous knowledge reconsider correct responses during extended reasoning, ultimately arriving at incorrect conclusions.

A closer analysis of failure cases reveals that this bottleneck stems from deficiencies in essential medical knowledge, which cannot be resolved merely by increasing the thinking budget. Consequently, extending the reasoning window reaches a fundamental limit, and budget-forcing techniques offer negligible benefits, as models lacking foundational knowledge remain anchored to incorrect assumptions. Even with additional reasoning steps, these models still struggle to retrieve accurate information. In such cases, improving data quality and increasing model capacity provide more

effective avenues for improvement. Our thorough ablation of data filtering strategies, dataset size, and model scaling demonstrates that when scaling thinking budget reaches its bottleneck, further performance gains can be achieved by enhancing data quality and scaling the model. Specifically, larger, difficulty-filtered, and diversity-sampled datasets consistently improve performance, while larger models further enhance scalability. This is because larger-capacity models or those fine-tuned on more extensive, high-quality datasets inherently possess richer medical knowledge, leading to higher accuracy. In conclusion, test-time scaling alone is insufficient for enhancing medical reasoning in LLMs—it needs to be complemented by scaling model size and improving knowledge grounding through high-quality data.

Our 7B model fine-tuned on 23K examples (`m1-7B-23K`) attains new state-of-the-art accuracy of 60.32% among in-domain and out-domain medical exam datasets, surpassing previously established specialized models of similar scale such as HuatuoGPT-o1-7B/8B (trained with complex RL on 40K instances) [4] and UltraMedical-8B (trained on hundreds of thousands of medical instructions) [32]. Furthermore, our larger 32B model trained with only 1K fine-tuning samples (`m1-32B-1K`) achieves performance comparable to 2X bigger resource-exhausted models (around 70B parameters with high training costs), underscoring the efficiency of our test-time scaling approach. All data, code, and models are publicly available to encourage future exploration in optimizing inference strategies in clinical AI applications.

2 Related Works

Test-time scaling for LLMs. There is a growing interest in techniques that enhance an LLM's reasoning without altering its weights, by allocating more computation at inference time [9, 15, 7]. A basic form is chain-of-thought prompting, e.g. instructing the model to "think step by step," which often improves performance on complex tasks [25]. More explicit approaches include generating multiple solutions and using majority voting or self-consistency to pick an answer, or employing search-based strategies with verifiers and lookahead. These methods trade extra inference passes for accuracy gains. In contrast, sequential test-time scaling keeps a single reasoning thread but makes it longer. OpenAI's o1 model hinted at the power of simply extending the reasoning length [9]. [16] formalized this by fine-tuning an LLM to utilize special "Wait" tokens, which allow controlling response length during inference. Their budget forcing method (described below) proved more effective than parallel voting strategies. Other recent research has proposed optimizing the allocation of test-time compute, for example finding an optimal stopping length per problem to avoid overthinking [31]. Our work builds directly on the simple test-time scaling idea [16, 2] by extending thinking traces with "wait" – we apply it to a new domain (medicine) and confirm its benefits in a very different setting. We focus on single-trace sequential reasoning, noting that it is complementary to orthogonal advances like tool use or retrieval augmentation.

Medical LLMs. The success of GPT-4 in medical exams [32] has spurred numerous open efforts to train medical domain LLMs [14, 29]. Early approaches centered on domain-specific pre-training: e.g. [26] and [19] pre-trained Llama models on medical text corpora (MIMIC-III [13], PubMed [1], etc.) to inject medical knowledge. While this improves knowledge recall, the gains on reasoning-heavy tasks were limited [10, 27]. More recent projects emphasize instruction tuning and reinforcement learning specialized for medicine. For example, OpenBioLLM was fine-tuned with expert-validated instructions and Direct Preference Optimization, and reportedly outperformed GPT-4 and Med-PaLM-2 on several biomedical QA benchmarks [18]. Med42 is another open model suite that achieved impressive results, even exceeding GPT-4.0 on many multi-choice medical QA tasks [5]. To push reasoning ability further, some works incorporate explicit reasoning supervision or verification. HuatuoGPT-o1 introduced verifiable medical problem-solving: they constructed 40K problems with known solutions and used a two-stage training (SFT + RL with a verifier) to train a 70B model [4]. This model achieved new state-of-the-art results on medical reasoning benchmarks, outperforming both general and prior medical LLMs [32]. UltraMedical built a massive dataset of 410K mixed manual/synthetic instructions for biomedicine [10], and fine-tuned Llama-3 models with supervised and preference learning. The 70B UltraMedical model reached 86.5% accuracy on MedQA, nearly matching Med-PaLM2 [22] and GPT-4 [1]. In contrast, our approach remains lightweight as we do not introduce new RL or verification components, and our dataset size (1K–23K) is relatively small.,

[1] `https://pubmed.ncbi.nlm.nih.gov/`

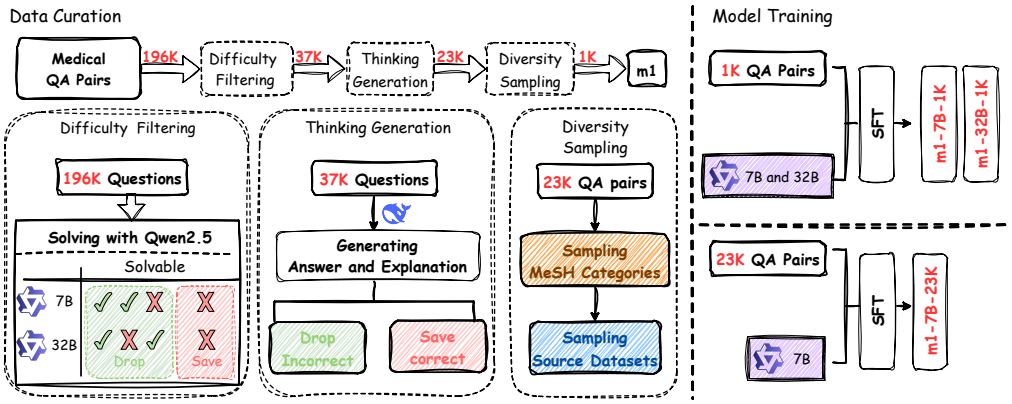


Figure 2: **An overview of our data curation and training pipeline.** We start with 196K raw medical QA examples, apply difficulty filtering (retaining 37K that Qwen2.5-7B-Instruct [30] or its 32B version cannot solve), then use DeepSeek-R1 [6] to generate reasoning and keep correct solutions (`m23K`). We perform diversity sampling to select a 1K high-quality subset (`m1K`). These datasets are used to fine-tune base models (Qwen2.5 7B and 32B Instruct) via Supervised Fine-Tuning (SFT), resulting in the m1 models (`m1-7B-1K`, `m1-7B-23K`, `m1-32B-1K`).

yet through test-time scaling, we achieve competitive results with these state-of-the-art models. We hope this encourages more exploration of inference-time techniques as an efficient alternative for domain-specific LLMs.

3 Method

Our approach consists of three parts: 1) **Data curation:** selecting and generating a high-quality set of medical QA examples with detailed reasoning (Section 3.1), 2) **Model training:** Supervised Fine-Tuning (SFT) of base LLMs on this data (Section 3.2), and 3) **Inference:** test-time control of the model's reasoning length (Section 3.3). Figure 2 provides an overview of the full pipeline.

3.1 Data curation

Initial collection. To construct the training data for `m1` through a multi-step refinement process, we begin with a large pool of approximately 196K medical QA samples compiled from public datasets: MedMCQA [17], MedQA-USMLE [11], HeadQA [23], and PubMedQA [12]. These include multiple-choice questions from medical exams as well as open-ended research questions. All samples are decontaminated against the evaluation data in Section 4.1. More details are presented in Appendix A.

Difficulty filtering. Following s1 [16], we identify a subset of solvable yet non-trivial problems by performing difficulty filtering using two strong base models. Specifically, we use Qwen2.5-Instruct [30] (an open general LLM) of 7B and 32B parameters to attempt each question. We filter a question if either Qwen-7B or Qwen-32B answers it correctly. This heuristic retains questions that are challenging to solve, eliminating those that are too easy for either models. Difficulty filtering pruned the dataset from 196K down to 37K samples.

Thinking generation. We employ DeepSeek-R1 [6], a state-of-the-art open reasoning LLM, to generate a chain-of-thought and final answer for each of the 37K questions. DeepSeek-R1 was chosen for its robust reasoning capability (it's comparable to OpenAI's o1 [9] in multi-step problem solving). For each question, we prompt DeepSeek-R1 to produce a detailed solution explanation ending in a definitive answer. We then apply solution validation: we only keep those instances where DeepSeek-R1's final answer is correct (matching the ground-truth). This yields a set of 23K high-quality "thinking - answers", where each question now paired with a verified-correct reasoning

process. This step ensures our training data predominantly consists of valid reasoning, while incorrect chains are discarded. More details are presented in Appendix A and B.1.

Diversity sampling. We design a diversity sampling strategy to construct a well-balanced and enriched subset for training our primary model. This process highlights two key components: domain balance and dataset balance. First, we ensure domain balance by annotating each sample with Medical Subject Headings (MeSH) categories[2] (See Appendix for the details), enabling systematic coverage across medical specialties (e.g., cardiology, neurology) and question types. Second, we address dataset imbalance through stratified sampling, first selecting domains, then source datasets, and finally individual samples (see Appendix Tables 3 and 4 for distributions). We perform stratified sampling at the dataset level to address the imbalance in sample counts across datasets (Appendix, Table 3, 4). Specifically, we first sample a domain, then sample a dataset, and finally roll-out a sample. The process is repeated until there are 1K samples (`m1K`), which will be served as our core training set for `m1`. The remaining 23K difficult samples (`m23K`) can be used to augment training or for ablations. We provide summary statistics of the final data in Appendix A.

3.2 Model Training

We fine-tune three model variants, corresponding to two model sizes (7B and 32B) and two training set sizes (1K and 23K). For each, we use the pre-trained Qwen2.5-Instruct model as the initialization. Qwen2.5 is a recent high-performance open LLM [30]; using it as our base ensures strong general language ability and allows us to focus on injecting medical reasoning. We format each training example in a "question $\rightarrow$ reasoning $\rightarrow$ answer" style. This format teaches the model to produce a coherent reasoning process and then give the answer. Using this data, we perform SFT for each model:

• `m1-7B-1K`: Fine-tuned on the 1K `m1K` dataset using the Qwen2.5-7B-Instruct. This represents the minimal training scenario.

• `m1-7B-23K`: Fine-tuned on the full 23K filtered dataset using Qwen2.5-7B-Instruct. This lets us examine the effect of more training data (23K vs 1K) at the same model size.

• `m1-32B-1K`: Fine-tuned on the 1K dataset using the larger Qwen2.5-32B-Instruct. This shows the effect of a larger model with minimal data.

3.3 Inference

At inference time, we employ test-time scaling by managing the model's generation of the chain-of-thought. Specifically, we define a thinking budget: a maximum number of tokens the model is allowed to generate before producing a final answer. By allocating a larger budget, we give the model more "thinking space" to potentially reason through the problem. If the model would naturally finish its reasoning early, we intervene to use the budget fully. We also apply budget forcing technique to extend the thinking process of the model: when the model outputs `end-of-think` token indicating the end of thinking before reaching the token budget, we replace it with "Wait." and force the model to keep the generation of the thinking traces. According to [16], this method often leads the model to double-check or refine its initial answers for math problems.

4 Experiments

4.1 Evaluation Settings

Datasets. We evaluate on nine medical QA benchmarks, grouped into In-Distribution and Out-of-Distribution tests. We measure accuracy for all datasets. 1) In-Distribution tests: The accompanying test splits from our training data, include MedMCQA [17] (MedMC); MedQA-USMLE [11] (MedQA), and PubMedQA [12] (PubMed). 2) Out-of-Distribution tests: The datasets are not included in training and stylistically distinct, assessing `m1`'s reasoning generalization: medical related questions from MMLU-Pro [24] (MMLU-P) and GPQA [20], small QA sets from Lancet and the

[2] `https://www.ncbi.nlm.nih.gov/mesh/`

Model	MedMC	MedQA	PubMed	MMLU-P	GPQA	Lancet	MedB (4)	MedB (5)	MedX	NEJM	Avg.
	In-Distribution Test			Out-of-Distribution Test							
$<$ 10B LLMs											
MedLlama3-8B-v1	34.74	55.07	52.70	27.43	30.77	42.23	38.31	33.77	11.04	49.25	37.53
MedLlama3-8B-v2	59.34	59.39	75.50	55.11	36.41	52.43	48.38	39.94	13.46	54.56	49.45
OpenBioLLM-8B	54.63	55.30	70.10	49.32	41.03	52.43	41.23	32.47	14.29	54.23	46.50
MMed-8B	52.71	54.28	63.40	48.27	34.87	53.40	41.23	35.39	13.73	54.39	45.17
MMedS-8B	47.29	57.19	77.50	33.55	22.05	55.10	54.22	55.84	17.39	53.40	47.35
MMed-8B-EnIns	58.09	60.33	63.80	51.60	45.90	55.34	59.09	56.17	18.56	62.35	53.12
Med42-8B	56.35	59.78	76.00	55.64	48.21	59.47	44.81	46.75	14.63	62.69	52.43
UltraMedical-8B-3	59.22	71.09	71.10	61.50	50.00	61.89	54.22	52.27	15.25	64.51	56.11
UltraMedical-8B-3.1	63.78	75.73	79.20	64.30	48.72	67.23	64.61	55.19	17.39	66.83	60.30
HuatuoGPT-o1-7B	63.47	71.56	78.60	67.23	47.95	62.14	52.92	50.65	15.11	65.17	57.48
HuatuoGPT-o1-8B	63.97	74.78	80.10	63.71	55.38	64.32	58.44	51.95	16.84	64.84	59.43
Qwen2.5-7B-Instruct	56.56	61.51	71.30	61.17	42.56	61.17	46.75	40.58	12.15	59.04	51.28
+CoT	56.11	64.49	72.60	62.15	52.56	60.68	50.97	42.86	13.18	58.54	53.41
m1-7B-1K	58.26	71.01	77.50	65.15	51.79	64.32	58.77	51.95	16.29	62.52	57.76
m1-7B-23K	62.54	75.81	75.80	65.86	53.08	62.62	63.64	59.74	19.81	64.34	60.32
$>$ 10B LLMs											
Qwen2.5-72B-Instruct	66.60	74.55	70.80	66.06	62.05	66.50	57.14	53.57	14.91	68.99	60.12
+CoT	66.15	76.43	71.30	69.77	63.85	65.78	60.06	54.22	14.84	69.15	61.16
Med42-70B	62.28	51.14	78.10	54.53	50.77	54.61	45.78	37.99	16.29	56.05	50.75
OpenBioLLM-70B	74.23	75.10	79.30	71.92	50.77	68.93	58.44	54.55	21.33	67.83	62.24
UltraMedical-70B-3	72.94	83.90	80.00	73.94	58.72	75.49	72.08	64.61	21.67	73.13	67.65
HuatuoGPT-o1-70B	75.23	86.80	81.40	76.09	66.67	72.82	72.08	68.51	26.36	74.13	70.01
HuatuoGPT-o1-72B	76.76	88.85	79.90	80.46	64.36	70.87	77.27	73.05	23.53	76.29	71.13
Qwen2.5-32B-Instruct	64.83	75.26	68.00	74.72	63.85	66.02	60.39	52.92	13.87	66.67	60.65
+CoT	64.33	74.86	68.90	74.72	64.87	66.75	60.39	54.22	14.56	66.33	60.99
m1-32B-1K	67.34	83.50	77.60	76.94	66.67	70.15	73.70	67.86	25.53	73.13	68.24

Table 1: **Baseline Comparisons.** We report accuracy (%) on each evaluation dataset for various models. Our m1 models (in bold) are shown in the $\leq 10B$ group (m1-7B variants) and $> 10B$ group (m1-32B). "+CoT" indicates using chain-of-thought prompting at inference for that base model. We mark Green color within each parameter group: the deeper the color, the higher the accuracy. For header abbreviations, please refer to Section 4.1.

New England Journal of Medicine (NEJM); 4 Options (MEdB (4)) and 5 Options (MedB (5)) splits from the MedBullets platform [3]; and MedXpertQA [33] (MedX).

LLM baselines. We compare our models against a variety of general and specialized medical LLM baselines: 1) general base instruct models Qwen2.5-7B, tested both as-is and with chain-of-thought prompting (+CoT); 2) specialized medical models including MedLlama3 [3], OpenBioLLM [18], MMed-Llama [19], Med42 [5], UltraMedical [32]; 3) medical reasoning model HuatuoGPT-o1 [4], which undergoes complex RL training.

Additionally, we compare to state-of-the-art large open medical LLMs (>10B), including Med42-70B [5], OpenBioLLM-70B [18], UltraMedical-70B [32], HuatuoGPT-o1-70B/72B [4], and baseline Qwen2.5 models (32B, 72B) with their respective +CoT versions. HuatuoGPT-o1 and UltraMedical employ complex training strategies involving reinforcement learning or expert feedback; therefore, matching or exceeding their performance with our simpler test-time scaling method underscores its effectiveness.

4.2 Results

Test-time scaling with different thinking budgets. We first evaluate how increasing the chain-of-thought token budget at inference affects performance on various medical QA datasets. As illustrated by the upward trajectories in Figure 1, our m1 approach gains consistent accuracy improvements as the thinking budget grows, demonstrating the efficacy of simple test-time scaling. Despite simplicity, Table 1 presents that our `m1-7B-23K` achieves an average accuracy of 60.32% amongst in-distribution and out-of-distribution sets, which exceeds complex RL tuned HuatuoGPT-o1-7B by 2.84%, and matches the large-scale SFT-tuned UltraMedical-8B. Notably, beyond 4K tokens, the improvements begin to saturate, indicating limited additional benefit from extremely long reasoning.

[3] `https://huggingface.co/johnsnowlabs/`

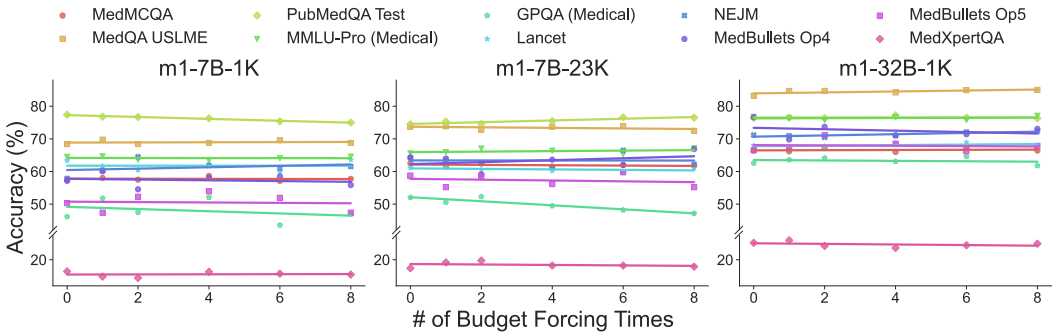


Figure 3: **Force thinking for different evaluation datasets.** Accuracy vs. number of budget forcing times (iterations of injecting "Wait") for each m1 model (`7B-1K`, `7B-23K`, `32B-1K`). A value of 0 means the model's first answer is taken without forcing, while higher values mean the model was compelled to reconsider up to that many times (within a 2048-token limit).

Data Filtering	MedMC	MedQA	PubMed	MMLU-P	GPQA	Lancet	MedB (4)	MedB (5)	MedX	NEJM	Avg.
	In-Distribution Test			Out-of-Distribution Test							
1K Scale											
Random	57.26	66.46	73.60	61.63	44.62	62.62	54.55	53.57	14.70	58.54	54.75
Hard Random	56.99	68.66	73.70	63.26	46.41	62.14	56.82	44.81	16.36	61.03	55.02
Hard Domain	**58.88**	66.38	74.00	**64.95**	45.38	63.11	54.55	49.68	16.36	**61.19**	55.45
Hard Domain Dataset	57.97	**70.23**	**76.10**	64.23	**49.74**	**62.14**	**57.79**	**50.97**	**17.12**	59.20	**56.55**
23K Scale											
Random	60.41	71.64	**76.50**	**67.43**	48.72	**62.38**	60.06	54.22	15.32	61.86	57.85
Hard	**62.01**	**73.76**	75.80	65.54	**50.51**	61.89	60.06	**55.19**	**18.91**	**64.34**	**58.80**

Table 2: **Data Filtering Ablation: difficulty filtering ("Hard"), and "Domain" and "Dataset" balance for diversity sampling.** Difficulty filtering consistently yields the largest gains across both 1K and 23K training scales, while domain and dataset balancing provide complementary improvements. Notably, scaling up from 1K to 23K substantially boosts accuracy, underscoring the importance of data scale. The same header abbreviations as Table 1.

Larger model capacity helps. When scaling model size from 7B to 32B parameters, we observe a more pronounced benefit from test-time scaling as larger models inherently possess richer medical knowledge. In Table 1, `m1-32B-1K` consistently outperforms or matches even larger (70B+) specialized medical LLMs, demonstrating that pairing a larger base model with simple supervised thinking traces and inference-time scaling yields strong results. This trend is also apparent in Figure 1, where the 32B model's accuracy curve leads across most datasets as the thinking budget increases.

Budget forcing does not help. Unlike mathematical tasks, where prompting the model to repeatedly refine its chain-of-thought can yield further gains, our experiments show diminishing returns from forced re-thinking (Figure 3). Although the model will generate additional intermediate tokens when repeatedly prompted to "keep thinking", we see minimal improvement, suggesting that medical reasoning may differ from math domains in how additional iterative reasoning is best leveraged. We analyze such failure cases in the following sections.

SFT data ablation. We ablate two key data curation steps used in our SFT process: difficulty filtering and diversity sampling (comprising domain and dataset balancing). As shown in Table 2, at the 1K training scale, models fine-tuned on difficulty-filtered data outperform those trained on randomly sampled data by +0.27% percentage points on average. Adding domain balance further improves performance by 0.43%, and incorporating both domain and dataset balance yields additional gains, reaching up to 56.55% average accuracy. At the 23K scale, overall performance improves substantially, and difficulty filtering alone provides a +0.65% accuracy boost on average. These results highlight the critical role of both data quality and scale in enhancing model performance.

Question: Which of the following are not a branch of external carotid Aery in Kiesselbach's plexus.

Options: (A) Sphenopalatine aery; (B) Anterior ethmoidal aery; (C) Greater palatine aery; (D) Septal branch of superior labial aery.

Ground-truth Answer: Anterior ethmoidal aery

Sample from [MedMCQA]

m1-7B-1K

Reasoning Tokens: 3405

Okay, let's tackle this question about Kiesselbach's plexus and its branches... in Kiesselbach's plexus, the main contributors are the anterior ethmoidal (from ECA), sphenopalatine (from maxillary), greater palatine (from maxillary), and the septal branch of the superior labial (which is a branch of the facial artery, which comes off the ECA). So, the plexus includes the anterior ethmoidal artery (direct from ECA) .... The anterior ethmoidal is definitely a branch of the ECA. **So the answer would be A. Let me check. Wait**, the Kiesselbach's plexus is formed by the anastomosis of the following arteries:\n\n1. Anterior ethmoidal artery (direct from ECA)\n ...Therefore, Sphenopalatine artery is not a direct branch of the ECA. **Therefore, the answer is A.**

m1-7B-23K

Reasoning Tokens: 1562

Okay, let's tackle this question. Hmm, the question is asking which of the options are not branches of the external carotid artery in Kiesselbach's plexus. ... Now looking at the options:\n\n A. Sphenopalatine artery. Yes, that's one of the main branches, so it is part of the plexus. B. Anterior ethmoidal artery That's from the ophthalmic, which is internal carotid. So it does contribute to the plexus, so this is part of it. ... But let me confirm once again. If the anterior ethmoidal is part of Kiesselbach's plexus but comes from the internal carotid, then yes, it's the correct answer. The others are all from the external. **Therefore, the answer is B**.

m1-32B-1K

Reasoning Tokens: 995

Okay, let's see. The question is about Kiesselbach's plexus and which of the listed arteries are not branches of the external carotid artery in that plexus. ... Wait, the external carotid artery gives rise to the facial artery, and the facial artery then branches into several arteries, including the inferior labial artery and the angular artery. . The greater palatine artery is also a branch of the maxillary artery. The anterior ethmoidal artery comes from the ophthalmic artery, which is a branch of the internal carotid artery, not the external... So B is the one that's not from external carotid. **Therefore, the answer is B**

Figure 4: **A failure case of test-time scaling** with the Qwen2.5-7B using 1K reasoning data. Although the m1-7B-1K conducts the longest reasoning, its deficiency in essential medical knowledge prevents it from producing the right answer. On the other hand, both m1-7B-23K and m1-32B-1K effectively resolve the question with a relatively brief reasoning procedure.

Failure cases. In this section, we discuss several fail cases where the models fail to perform test-time scaling, underscoring the *critical role of accurate knowledge in medical reasoning models*. Specifically, our investigation can be distilled into the following key points:

• **The extent of knowledge is crucial for effective medical reasoning.** As illustrated in Figure 4, m1-7B-1K is unable to generate accurate reasoning because it lacks crucial knowledge: 'Anterior ethmoidal artery belongs to the internal carotid'. In contrast, both m1-7B-23K and m1-32B-1K possess this essential knowledge. Consequently, having a substantial amount of accurate knowledge, whether from fine-tuning data or the pre-training model, enhances the model's capability for medical reasoning. This is similarly evidenced in Table 1, where m1-7B-23K and m1-32B-1K exhibit a significantly better performance.

• **Incorrect knowledge obstructs the reasoning.** As demonstrated in Figure 5, even when the model generates the correct answer at first, forcing it to re-think cause it to retrieve faulty information, which results in an incorrect response. Therefore, such erroneous knowledge may lead to unstable reasoning process, highlighting the importance of verifying the accuracy of the training data for medical reasoning models.

• **Test-time scaling fails to rectify incorrect knowledge.** In fields like math and coding, scaling up thinking processes can enhance a model's reasoning by allowing it to conduct self-reflection and identify errors in its previous logic. However, in the medical domain, errors largely stem from misconceptions in knowledge. These are difficult to correct merely by increasing the reasoning budget. As illustrated in Figure 4, despite m1-7B-1K executing the most extended reasoning, its lack of crucial medical knowledge hinders it from arriving at the correct answer. Additionally, in Appendix Figure 7, even when the model is forced to re-think multiple times, it is unable to rectify the inaccurate knowledge.

5 Conclusions

We introduced `m1`, demonstrating that test-time scaling significantly improves medical reasoning in large language models without requiring extensive fine-tuning. Performance across diverse

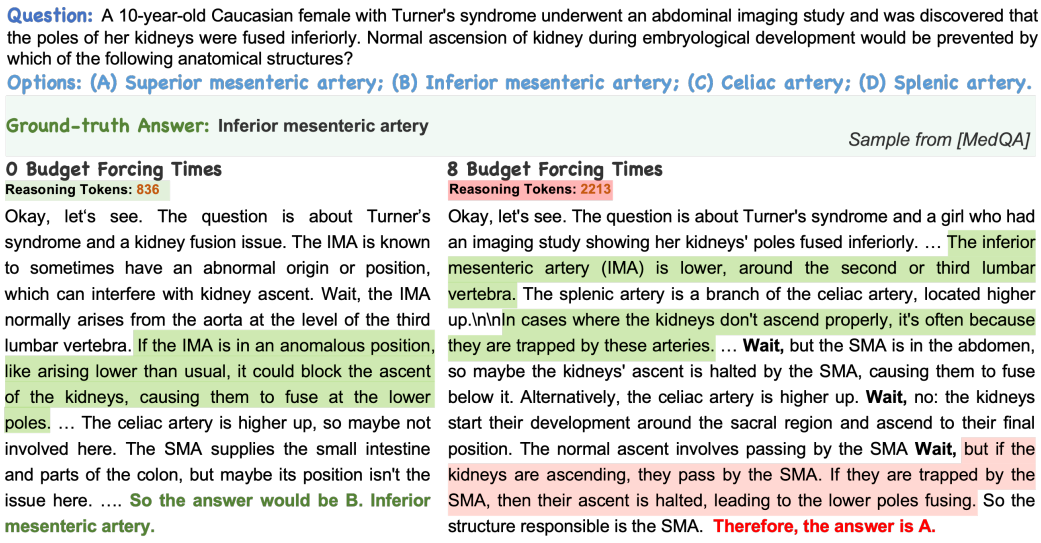


Figure 5: **A failure case of budget forcing.** Initially, the model produces the correct answer, but forcing it to re-think causes the model to retrieve incorrect knowledge, ultimately resulting in an erroneous answer.

medical QA benchmarks consistently improved with increased inference-time reasoning budgets. Crucially, we found that test-time scaling alone cannot remedy fundamental deficiencies in medical knowledge, emphasizing the necessity for high-quality medical data and model scale expansion. m1 achieves strong performance, outperforming more expensive approaches such as HuatuoGPT-o1 and UltraMedical on various benchmarks. Our 7B model trained on 23K data establishes a new state-of-the-art in the $\leq 10B$ parameter category, and our 32B model rivals models $2\times$ in size. We will release a full-stack open-source package including the curated dataset (m1K), fine-tuned model weights, and inference code with budget control, to encourage future exploration in optimizing inference strategies in clinical AI applications.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

[3] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. *arXiv preprint arXiv:2402.18060*, 2024.

[4] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

[5] 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.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[7] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, and Heung-Yeung Shum Xiangyu Zhang. Open-reasoner-zero: An open source approach to scaling reinforcement learning on the base model. `https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero`, 2025.

[8] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[9] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[10] Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. MedS^3: Towards medical small language models with self-evolved slow thinking. *arXiv preprint arXiv:2501.12051*, 2025.

[11] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[12] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[13] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[14] Mary M Lucas, Justin Yang, Jon K Pomeroy, and Christopher C Yang. Reasoning with large language models for medical question answering. *Journal of the American Medical Informatics Association*, 31(9):1964–1975, 2024.

[15] Chunyang Meng, Shijie Song, Haogang Tong, Maolin Pan, and Yang Yu. Deepscaler: Holistic autoscaling for microservices based on spatiotemporal gnn with adaptive graph learning. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 53–65. IEEE, 2023.

[16] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[17] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

[18] Malaikannan Sankarasubbu Ankit Pal and Malaikannan Sankarasubbu. Openbiollms: Advancing open-source large language models for healthcare and life sciences. *Hugging Face repository*, 2024.

[19] Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384, 2024.

[20] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[21] 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. *arXiv preprint arXiv:2212.13138*, 2022.

[22] 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*, pages 1–8, 2025.

[23] David Vilares and Carlos Gómez-Rodríguez. Head-qa: A healthcare dataset for complex reasoning. *arXiv preprint arXiv:1906.04701*, 2019.

[24] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[25] 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.

[26] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

[27] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint*, 2025.

[28] Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*, 2024.

[29] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific, 2024.

[30] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[31] 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.

[32] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081, 2024.

[33] 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.

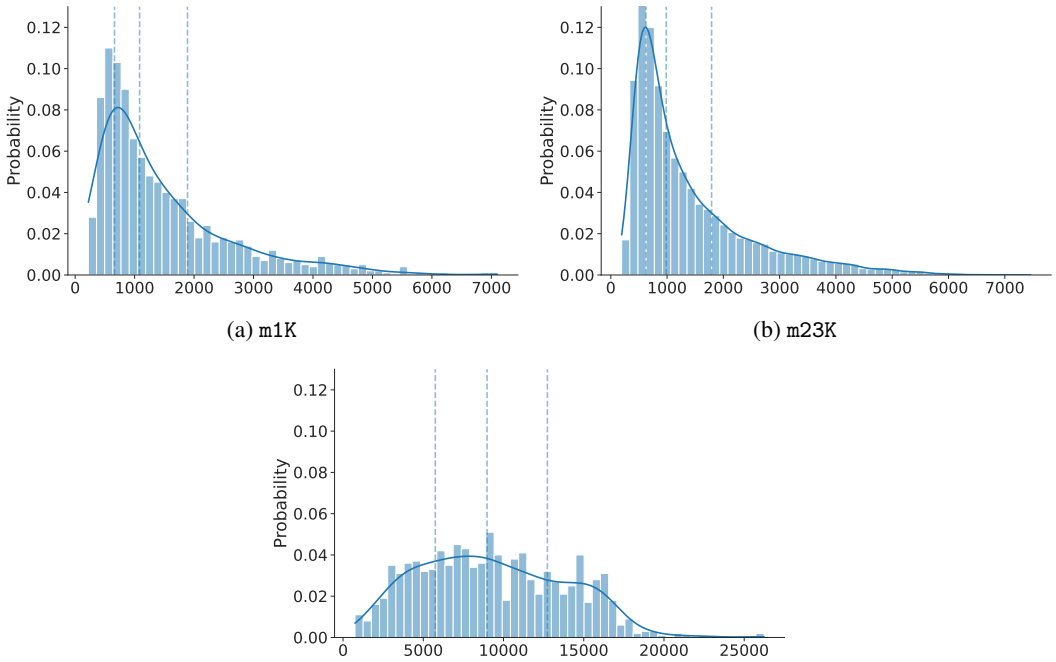
A Data Statistics

Dataset statistics. Table 3 shows the statistics of all datasets used in the paper. Note that the sample counts across datasets are highly imbalanced.

Table 3: The statistics of all datasets used in the paper.

Dataset	MedQA	HeadQA	MedMCQA	PubMedQA	Summation
Initial collection	10,178	2,657	182,822	500	196,157
+After difficulty filtering	2,099	331	35,270	116	37,816
+Generating thinking data	1,628	209	21,628	39	23,504
+Decontamination & deduplication (m23K)	1,628	209	21,628	28	23,493
Random 23K	1,316	317	21,831	29	23,493
Random 1K	61	8	929	2	1,000
Hard random 1K	78	10	909	3	1,000
Hard Domain Balanced 1K	52	20	924	4	1,000
Hard Domain Dataset Balanced 1K (m1K)	274	123	575	28	1,000

Token length statistics. As illustrated in Figure 6, the distributions of training token lengths between the m1K/m23K and s1K [16] datasets exhibit clear differences. The m1K/m23K dataset shows a strongly right-skewed distribution, with most samples having token lengths clustered around 1,000 tokens, quickly diminishing toward lengths beyond 3,000 tokens. In contrast, the s1K dataset displays a more uniform and broader distribution, spanning widely from about 2,500 to over 15,000 tokens, with peaks around 5,000 to 10,000 tokens. These contrasting distributions reflect differing data preparation strategies: m1K/m23K focuses on concise medical knowledge without longer thinking steps, whereas s1K includes longer, more detailed reasoning traces suitable for complex multi-step inference tasks.



(a) `m1K` (b) `m23K`

(c) s1K

Figure 6: The token length distributions of *m1K* and s1K [16]. The 25%/50%/75% quantile is marked in transparent vertical dotted lines.

Sample domain statistics. We list the statistics of sample domains for `m1K` and `m23K` in Table 4. We use the domain label from MeSH Qualifiers with Scope Notes [4].

Table 4: The statistics of sample domain for `m1K` and `m23k`.

Domain	m23K					m1K				
	MedQA	PubMedQA	HeadQA	MedMCQA	Summation	MedQA	PubMedQA	HeadQA	MedMCQA	Summation
Abnormalities	24	1	0	646	671	10	1	0	10	21
Administration & Dosage	9	0	3	590	602	6	0	3	9	18
Adverse Effects	116	2	2	936	1,056	9	2	2	7	20
Agonists	2	0	1	20	23	2	0	1	13	16
Analogs & Derivatives	1	0	0	10	11	1	0	0	10	11
Analysis	2	2	14	222	240	2	2	2	6	12
Anatomy & Histology	17	0	3	2,328	2,348	10	0	3	11	24
Antagonists & Inhibitors	14	0	2	113	129	7	0	2	6	15
Biosynthesis	4	0	3	73	80	4	0	3	6	13
Blood	45	1	0	360	406	7	1	0	12	20
Blood Supply	19	0	0	228	247	6	0	0	8	14
Cerebrospinal Fluid	7	0	0	34	41	4	0	0	10	14
Chemical Synthesis	0	0	5	1	6	0	0	5	1	6
Chemically Induced	16	0	1	82	99	5	0	1	8	14
Chemistry	0	0	21	342	363	0	0	7	8	15
Classification	0	0	5	341	346	0	0	5	8	13
Complications	73	1	2	625	701	6	1	2	3	12
Congenital	59	0	0	382	441	7	0	0	4	11
Cytology	3	0	1	90	94	3	0	1	18	22
Deficiency	58	1	2	245	306	6	1	2	3	12
Diagnosis	229	1	17	1,681	1,928	2	1	3	4	10
Diagnostic Imaging	25	4	2	720	751	7	4	2	11	24
Diet Therapy	3	1	2	41	47	2	1	2	3	8
Drug Effects	12	0	0	107	119	11	0	0	9	20
Drug Therapy	164	0	6	509	679	1	0	4	5	10
Economics	1	1	1	17	20	1	1	1	11	14
Education	0	1	1	20	22	0	1	1	12	14
Embryology	12	0	1	245	258	7	0	1	6	14
Enzymology	5	0	3	95	103	4	0	1	2	7
Epidemiology	8	0	3	265	276	4	0	3	4	11
Ethics	8	0	2	20	30	5	0	2	2	9
Ethnology	0	0	0	5	5	0	0	0	5	5
Etiology	133	1	5	657	796	4	1	5	3	13
Genetics	66	0	5	293	364	4	0	2	3	9
Growth & Development	8	1	3	245	257	0	1	1	5	7
History	0	0	2	72	74	0	0	2	5	7
Immunology	17	0	2	199	218	10	0	2	7	19
Injuries	12	0	1	430	443	12	0	1	10	23
Innervation	10	0	1	157	168	3	0	1	9	13
Instrumentation	0	0	0	151	151	0	0	0	17	17
Isolation & Purification	0	0	0	7	7	0	0	0	7	7
Legislation & Jurisprudence	0	0	1	204	205	0	0	1	15	16
Metabolism	15	0	5	158	178	1	0	4	5	10
Methods	1	1	3	260	265	1	1	3	9	14
Microbiology	27	0	0	358	385	5	0	0	8	13
Mortality	3	0	0	30	33	3	0	0	5	8
Nursing	0	0	5	6	11	0	0	5	6	11
Organization & Administration	0	1	3	111	115	0	1	3	10	14
Parasitology	5	0	1	209	215	5	0	1	9	15
Pathogenicity	10	0	0	71	81	5	0	0	4	9
Pathology	68	0	2	1,532	1,602	11	0	2	7	20
Pharmacokinetics	4	0	6	99	109	3	0	6	1	10
Pharmacology	31	0	1	315	347	5	0	1	6	12
Physiology	45	1	17	1,168	1,231	5	1	5	13	24
Physiopathology	92	1	3	1,093	1,189	5	1	3	6	15
Poisoning	17	0	0	171	188	5	0	0	7	12
Prevention & Control	17	0	1	131	149	6	0	1	10	17
Psychology	20	1	21	165	207	0	1	1	4	6
Radiation Effects	1	0	0	56	57	1	0	0	23	24
Radiotherapy	0	0	0	28	28	0	0	0	17	17
Rehabilitation	0	0	1	10	11	0	0	1	10	11
Secondary	2	0	0	44	46	2	0	0	6	8
Standards	2	0	0	67	69	2	0	0	10	12
Statistics & Numerical Data	5	1	2	53	61	5	1	2	4	12
Supply & Distribution	0	0	0	10	10	0	0	0	10	10
Surgery	5	2	1	749	757	5	2	1	7	15
Therapeutic Use	5	0	1	230	236	5	0	1	6	12
Therapy	45	2	7	373	427	5	2	3	3	13
Toxicity	2	0	0	29	31	2	0	0	7	9
Transmission	1	0	1	86	88	1	0	1	15	17
Transplantation	0	0	0	31	31	0	0	0	11	11
Trends	0	0	0	1	1	0	0	0	1	1
Ultrastructure	0	0	0	7	7	0	0	0	7	7
Urine	11	0	0	77	88	8	0	0	6	14
Veterinary	0	0	0	2	2	0	0	0	2	2
Virology	12	0	5	90	107	6	0	5	4	15

[4] `https://www.nlm.nih.gov/mesh/qualifiers_scopenotes.html`

B Implementation Details

B.1 Data Generation Details

We generate the reasoning traces and answers using the API of `deepseek-ai/DeepSeek-R1` model on the *SiliconFlow* platform [5]. We call the API with its default sampling parameters. The API calls are scheduled with `curator` [6].

According to the initial observation on the length of the outputs, we set the limit to 8K as no samples have outputs with length larger than 8K. The prompt is formatted as `"Return your final response within \\boxed{{ }}.\n{Question}\n{Options}"`, thus the answers are enclosed and are easy to be extracted and verified.

We perform data decontamination and dedupliation following OpenThoughts project [7].

B.2 SFT Details

All fine-tuning runs use standard language modeling training with 'trl' library: we optimize the model to minimize the output cross-entropy on the `reasoning \to answer` sequences (teacher-forcing the entire sequence). We use the same training hyperparameters as the s1 paper for consistency [16]: 5 epochs of training, a batch size equals 16, a low learning rate of 1e-4 with warmup and cosine decay, a modest weight decay of 1e-4, Adam betas of 0.9 and 0.95. The thinking part is enclosed with `<|im_start|>think` and `<|im_start|>answer`. SFT is performed with *trl* and `transformers` libraries.

Training `m1-7B-1K` and `m1-7B-23K` is extremely fast on the order of minutes on 8 H100 GPUs, and m1-32B-1K can be trained in a few hours with 16 H100 GPUs. This underscores the efficiency of our approach: unlike massive instruction tuning efforts that require many days on tens of GPU nodes, our models reach convergence with modest compute. We did not apply any reward modeling or RL in training: the model purely learns to imitate the given chain-of-thought format.

B.3 Evaluation Details

Datasets. To thoroughly assess both in-domain performance and generalization, we evaluate on eight medical QA benchmarks, grouped as follows:

In-Distribution Tests:

1. MedMCQA [17] – a collection of 3.5K multiple-choice questions from Indian medical entrance exams, testing general medical knowledge.
2. MedQA-USMLE [11] – the USMLE question dataset (NYU MedQA) containing US medical licensing exam MCQs; we use the standard test split.
3. PubMedQA [12] – a dataset of biomedical research questions (factoid Q paired with abstracts) where the task is to answer yes/no/maybe or short answer. These three were part of our training data pool (though we filtered and sampled from them), so they represent in-domain evaluations. We report accuracy (for MCQ, percentage of correct choices; for PubMedQA, percentage of correct yes/no/maybe).

Out-of-Distribution Tests:

1. MMLU-Pro [24] (Medical) – the medical category subset of the Massive Multitask Language Understanding benchmark, which includes professional medicine questions and related subjects. We specifically evaluate on the Professional Medicine section (and report accuracy). We follow the split from [4].
2. GPQA (Medical) [20] – the biomedical portion of the Graduate-Level Physics/Chemistry/Biology QA dataset GPQA. This dataset contains extremely

[5] `https://siliconflow.cn/`

[6] `https://github.com/bespokelabsai/curator/`

[7] `https://github.com/open-thoughts/open-thoughts/tree/main/open_thoughts`

challenging, Google-proof multiple-choice questions created by experts, requiring high reasoning (we use the biology/medical questions, 150 in total). We follow the split from [4].

3. Lancet & NEJM – we compiled two small sets of QA pairs from The Lancet [8] and New England Journal of Medicine [9] (NEJM) clinical case reports (answers verified from the text). These assess how models handle medical literature style questions.
4. MedBullets [3] – a collection of practice questions from the MedBullets medical education platform. We specifically take subsets of difficulty level 4 and 5 (on a 1–5 scale, 5 being hardest), denoted MedBullets Op4 and MedBullets Op5, about 100 questions each, to serve as challenging test sets.
5. MedXpertQA [33] – a custom set of 50 expert-written multi-step medical reasoning questions we created for qualitative evaluation (free-form answers). For MedXpertQA we report the percentage of questions answered correctly.

These out-of-distribution (OOD) sets were not used in training and often differ in style from our training data (e.g. long clinical vignettes, or extremely tricky edge cases). They allow us to test how well `m1`'s reasoning generalizes.

Methods. We compare our models against a broad range of baselines, including both general LLMs and specialized medical LLMs:

1. Qwen2.5 Instruct (7B, 32B, 72B) – The base instruct models (no medical fine-tuning). We include these to show the starting performance of the underlying models before our fine-tuning. We also test Qwen2.5 with a chain-of-thought prompting (+CoT), where we simply prompt it to "think step by step" at inference, to see if prompting alone can elicit similar reasoning (this baseline uses no additional training).
2. MedLlama3 (8B)[10] – An 8B instruction-tuned model released by M42 (Johns Hopkins/APL), one of the early open medical LLMs. We list two versions from their releases.
3. OpenBioLLM (8B) [18] – The 8B model from Saama AI, fine-tuned with expert-curated medical data.
4. MMed-Llama (8B) [19] – A multilingual medical model from MedS3 work, which underwent additional pre-training (denoted MMedS or MMed in results).
5. Med42 (8B) [5] – The 8B model from the Med42-v2 suite, instruction and preference-tuned on clinical data.
6. UltraMedical (8B) [32] – The 8B model from Tsinghua's UltraMedical project (we test both the v3.0 and v3.1 versions if available).
7. HuatuoGPT-o1 (7B & 8B) [4] – The smaller versions of HuatuoGPT-o1 (the 70B model's distilled or intermediate checkpoints) as reported in their paper.
8. Larger models (>10B): We also compare to state-of-the-art open models in the larger size class: Med42-70B [5], OpenBioLLM-70B [18], UltraMedical-70B [32], and HuatuoGPT-o1-70B/72B [4] (if available). These represent the current best open medical LLMs (some claim parity with GPT-4).

It is worth noting that some of these baselines (e.g. HuatuoGPT-o1 [4], UltraMedical [32]) involve complex training regimes (RL or extensive preference tuning), and in cases like Med42 and OpenBioLLM, they incorporate expert feedback. Our approach does not, so beating or matching them would be a strong indication of the power of test-time scaling.

Inference. We use `SGLang` as our inference engine. We use `bfloat16` precision and greedy sampling (i.e., temperature=0) for inference. A fixed seed of 42 is used during inference. The prompt format is: `"{Question}\n{Options}\n{Instruction}"`. We format options as: `"A. yes\nB. no\nC. maybe"`. The default instruction is:

[8] `https://www.thelancet.com/`

[9] `https://www.nejm.org/`

[10] `https://huggingface.co/johnsnowlabs/`

"Return your final response within \\boxed {{}}." For chain-of-thought inference with baseline LLMs Qwen2.5 7B/32B/72B Instruct, we update the instruction to: "Let's think step by step. Return your final response within \\boxed{{}}.".

Answer matching. We try to directly extract the answers from "\\boxed{{}}". If the extraction fails, we follow [4] to match answers via regex. If multiple answers are matched, we only choose the first one.

C Failure Case of Budget Forcing

We illustrate a failure case of budget forcing in Figure 7. Initially, the model arrives at the correct answer with concise and accurate reasoning. However, when forced to continue thinking for longer, the extended reasoning introduces confusion and incorporates incorrect anatomical associations, ultimately leading to the wrong answer. This highlights a key limitation of budget forcing in medical QA: more reasoning does not always equate to better reasoning.

Question: The lateral spread of dental caries is facilitated mostly by the:
Options: (A) Enamel spindles; (B) Dentinoenamel junction; (C) Enamel lamellae; (D) Striae of Retzius.

Ground-truth Answer: Dentinoenamel junction

Sample from [MedMCQA]

0 Budget Forcing Times
Reasoning Tokens: 741

Okay, so I need to figure out which of the options A to D is correct for the question about what facilitates the lateral spread of dental caries… Oh, right! The striae of Retzius are the incremental lines of enamel… I remember that these striae create microscopic grooves in the enamel, which can act as pathways for cariogenic bacteria to spread laterally. Since they are connected horizontally, bacteria can move along these lines without needing to penetrate deep into the enamel, allowing the caries to spread sideways. ... **So the answer should be D, Striae of Retzius**

8 Budget Forcing Times
Reasoning Tokens: 2303

Okay, I need to figure out which of the options A to D is the correct answer … Because the striae run across the enamel, so maybe they provide a pathway for the bacteria or acid to move sideways, parallel to the tooth surface. That would make the striae a possible answer for lateral spread. … **Wait**, let me check my notes. Oh, right, the striae of Retzius are indeed pathways for caries to spread laterally. Because they are horizontal lines, so if a carious lesion starts in one area, the acid or the bacteria can move along the striae to other parts of the enamel, causing a lateral spread. That makes sense. … **Wait**, but another mnemonic I remember: DEJ is the site for initial penetration into dentin, but lateral spread in enamel is via striae. … Yes, I think **the answer is D. Striae of Retzius.**

Figure 7: A failure case of budget forcing.