Cross Domain Test Time Scaling: Scale Knowledge and Reasoning on Cross Domains

Minxi Yan[1], **Yihua Shao**[1,2], **Yanling Pan**[3], **Siyu Chen**[2], **Hongjuan Pei**[4], **Hao Tang**[5], **Fei Ma**[6], **Jingcai Guo**[1*] and **Nicu Sebe**[8]
[1]The Hong Kong Polytechnic University, Hong Kong SAR
[2]Institute of Automation, Chinese Academy of Sciences, China
[3]Zhejiang University, China
[4]University of Chinese Academy of Sciences, China
[5]Peking University, China
[6]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China
[8]University of Trento, Italy
yanmx8@gmail.com, jc-jingcai.guo@polyu.edu.hk

Abstract

Test-time scaling (TTS) has demonstrated remarkable potential in enhancing the reasoning capabilities of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs). However, its application has primarily been limited to domains such as mathematics and programming, owing to their reasoning-intensive nature and the ease of result verification. Its utility in other knowledge-intensive fields, such as medicine and general scientific research, remains underexplored. To bridge this gap and unlock the potential of TTS in broader domains, we propose Cross-Domain TTS, a novel framework that enables task-tailored scaling. This framework consists of two key components: a conformal prediction-based cold-start strategy and an information-gain-based dynamic reasoning adjustment. The CP-based cold-start strategy guides the model's initialization during test-time scaling based on conformal prediction theory, while the information-gain-based dynamic reasoning adjustment guides the model's reasoning progress through a progress vector according to the information gain of reasoning steps. We conducted experiments using LLMs and LVLMs on cross-domain benchmarks. Our results demonstrate that the proposed framework consistently improves performance across various domain-specific datasets. For instance, in the medical domain, it achieves an improvement of up to 17% in pass@1 accuracy while reducing inference latency and saving up to 30% in token consumption. Code is available at https://github.com/Yan0613/Cross-Domain-TTS.

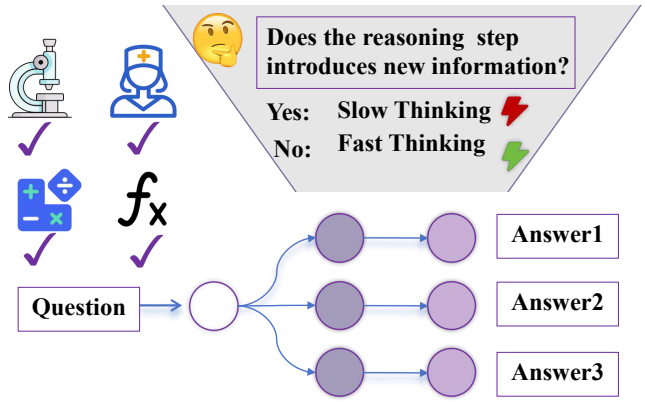

Figure 1: We propose a task-tailored TTS approach based on the characteristics of tasks, such as reasoning-heavy problems in mathematics or knowledge-driven queries in medicine.

1 Introduction

Recently, test-time scaling (TTS) has demonstrated its power in large language models (LLMs) [Snell *et al.*, 2025]. If pre-trained model size can be traded off for additional inference compute, this would enable the deployment of smaller on-device models in place of datacenter-scale LLMs. The most mainstream TTS methods generally fall into two categories: **sequential scaling** and **parallel scaling**. Sequential scaling iteratively updates intermediate states, iterative refinement, and revises the answer. The most widely used sequential scaling approach includes budget forcing, self-refinement, and CoT [Wei *et al.*, 2022; Kamoi *et al.*, 2024]. Parallel scaling enhances the reasoning abilities of LLMs by generating multiple answers and selecting the best one. For parallel scaling, the most widely used approach is repeated sampling, including techniques like best-of-n sampling, majority voting, and self-consistency [Inoue *et al.*, 2026; Li *et al.*, 2025]. Recently, some researchers have also tried to combine sequential scaling and parallel scaling. This method, called hybrid scaling,

*Corresponding author: *Jingcai Guo*.

exploits the complementary benefits of both parallel and sequential approaches.[Inoue *et al.*, 2026]

However, these existing TTS technologies primarily rely on the mathematical and coding domains due to their inherent ease of validation and reasoning-intensive nature. The TTS capabilities of LRMs in other domains, such as medical and general scientific fields, remain largely underexplored. Recent research has shown that s1[Muennighoff *et al.*, 2025] is ineffective in medical domains[Huang *et al.*, 2025]. Furthermore, the general reasoning abilities acquired by R1-distilled models do not transfer effectively to the medical domain through either Supervised Fine-Tuning (SFT) or Reinforcement Learning [Oh *et al.*, 2025]. Other studies also indicate that SFT on mathematical data can even lead to a performance decrease in non-reasoning domains[Huan *et al.*, 2025].

Inspired by [Wu *et al.*, 2025b], we can divided a reasoning task into two parts, reasoning depth and reasoning width, the former represent the reasoning, and the later represents the inner knowledge of LRMs. Furthermore, during the thinking process, if a path stops providing new information, we should accelerate the pace to reach the conclusion more quickly.

Our contributions are threefold:

- We introduce a novel unified test-time scaling (TTS) method that demonstrates robust effectiveness across both reasoning-intensive and knowledge-intensive domains.
- We present a method to successfully mitigate the previously observed ineffectiveness of TTS in knowledge-intensive tasks, thereby broadening the applicability of test-time scaling for large reasoning models.
- We propose a cross-domain reasoning method based on test-time scaling. Unlike most training-based approaches, ours is training-free and significantly more resource-efficient.

2 Related Works

2.1 Large Reasoning Models

The field of Large Reasoning Models (LRMs) has witnessed significant progress in recent years. Notably, models such as the DeepSeek-R1 series [Guo *et al.*, 2025] and the OpenAI-O1 series [Jaech *et al.*, 2024] have demonstrated strong performance in domains like mathematics and coding. However, the development of LRMs in other critical areas, such as medicine and finance, remains relatively underexplored. In the medical domain, recent advancements have shown promise. For instance, Huatuo-GPT-O1 [Chen *et al.*, 2024a], MedReason [Wu *et al.*, 2025a], and Med-R1 [Lai *et al.*, 2025] have made significant strides in enhancing medical reasoning capabilities. Also, in the financial domain, models like Fin-R1 [Liu *et al.*, 2025] have been developed to address the unique challenges of financial reasoning. Fin-R1 incorporates domain-specific knowledge graphs and reinforcement learning techniques to improve its ability to analyze financial data and make accurate predictions. However, despite these advancements, LRMs in domains other than mathematics and coding still face significant challenges and opportunities for further development.

2.2 Cross Domain Reasoning

Large Reasoning Models have shown their promising potential. However, open research has been narrowly focused on math and code domains. Recently, the academic community has noticed this lack. Cross Domain Reasoning has get advancement. Reasoning360 [Cheng *et al.*, 2026] introduces GURU-7B and GURU-32B models, achieving state-of-the-art performance for open models trained with publicly available data. Similarly, Open-Reasoner-Zero [Hu *et al.*, 2026] scale reinforcement learning directly on base large language models for reasoning tasks. Also, General-Reasoner [Ma *et al.*, 2026] developed a generative model-based verifier for robust answer assessment within a Zero RL training framework based on a verifiable all-domain reasoning dataset. CrossThink [Akter *et al.*, 2026] systematically incorporates multi-domain and multi-format data with rule-based rewards, leading to LLMs that outperform the base model by 8.55% to 13.36% across diverse reasoning benchmarks and generate more efficient, concise, correct answers. Crossing the Reward Bridge [Su *et al.*, 2025] employs a generative, model-based soft reward method to extend RLVR to more diverse, broader domains. However, research on cross-domain reasoning is still in its nascent stages, and most methods remain confined to reinforcement learning.

2.3 Test Time Scaling for LLMs

Test-time scaling (TTS) refers to techniques that adjust the computational effort exerted by a model during inference to improve performance. This paradigm is particularly relevant for LRMs, where a trade-off between model size and inference compute can enable more flexible deployment. TTS methods aim to enhance the quality of the output by allocating additional computational budget at inference time, rather than solely relying on a fixed, pre-trained model. Test Time Scaling can be generally divided into three categories: sequential scaling, parallel scaling, and hybrid scaling.

Sequential scaling methods [Muennighoff *et al.*, 2025] enhance reasoning depth. These methods iteratively update intermediate states, often by prompting the model to engage in more extensive thinking. For instance, s1 [Muennighoff *et al.*, 2025] introduces test-time scaling through *budget forcing*, which controls computational effort during inference by appending "Wait" tokens to extend thinking or by forcefully terminating a response. Similarly, alphaone [Zhang *et al.*, 2025a] employs a hyperparameter α to adjust the balance between slow thinking and fast reasoning processes, thereby achieving strong performance.

In contrast, parallel scaling can expand the reasoning width. Unlike sequential scaling, parallel scaling will generate multiple outputs for one sample, and then, by techniques such as best of N [Snell *et al.*, 2025], aggregating them into a final answer[Zhang *et al.*, 2025b; Zeng *et al.*, 2025].

Recently, hybrid scaling has also emerged as a promising approach, combining the strengths of sequential scaling and parallel scaling[Bi *et al.*, 2025; Inoue *et al.*, 2026]. Some methods employ multi-agent frameworks to further enhance the robustness of scaling [Wang *et al.*, 2025; Chen *et al.*, 2024b]. However, most of these TTS methods currently focus

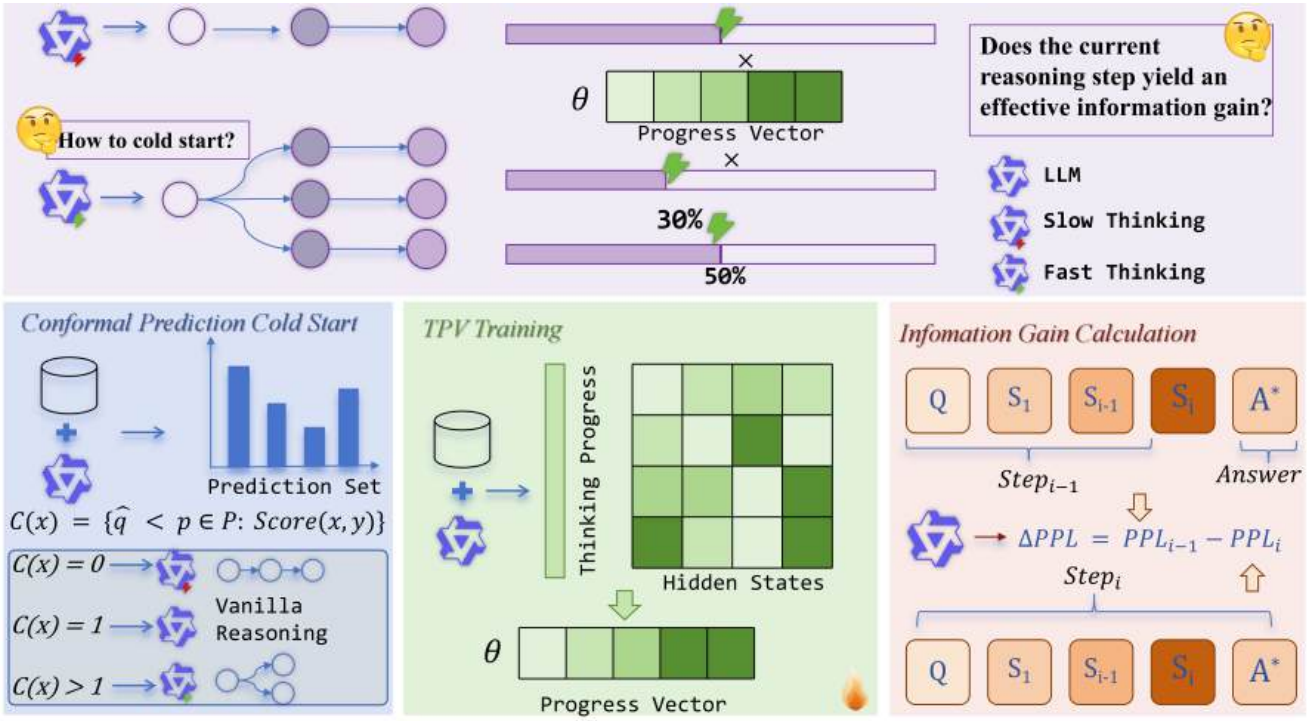

Figure 2: Overview of Cross-Domain Test-Time Scaling. Our framework comprises three components: First, a Conformal Prediction-based cold-start strategy determines whether a problem requires parallel scaling at processing onset (indicated in blue). In this stage, we calculate a quantile $\hat{q}$ for the dataset, where $C(x)$ represents the prediction candidate set. Second, a Task Progress Vector (TPV) training phase (indicated in green). The TPV is trained on a task-specific calibration dataset to establish a mapping between the thinking process and LLM hidden states. Third, an information-gain-based dynamic reasoning adjustment mechanism. This component modulates reasoning pace by evaluating the incremental information gain of each reasoning step.

primarily on the mathematical and coding domains, neglecting their potential applications in other domains.

3 Method

Our framework is composed of three primary parts. First, it utilizes a Conformal Prediction-based cold-start strategy to determine the necessity of parallel scaling at the start of LLM processing. Second, a Thinking Progress Vector (TPV) is trained on a calibration dataset to identify a progress vector that defines the acceleration direction within the reasoning trajectory. Third, we implement a dynamic speed adjustment approach based on information gain, which adaptively modulates the reasoning pace by quantifying the marginal utility of each incremental step in the model's output.

3.1 Conformal Prediction Cold Start

To establish the cold-start strategy for reasoning, we employ **Conformal Prediction (CP)** [Angelopoulos and Bates, 2023], a rigorous statistical framework designed to quantify the uncertainty of model predictions without requiring specific distributional assumptions. During the calibration phase, we utilize a hold-out calibration dataset $\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i=1}^{n}$. For each instance, the LLM generates a probability distribution over the label space $\mathcal{Y} = \{A, B, C, D\}$. Let $\hat{f}(x_i)_y$ denote the softmax probability assigned to label y. We define the conformity score s_i as the probability of the true label:

$$s_i = \hat{f}(x_i)y_i, \tag{1}$$

Given a pre-defined error budget $\alpha \in (0, 1)$, we calculate the threshold $\hat{q}$ as the α-quantile of the scores in $\mathcal{D}_{cal}$:

$$\hat{q} = \text{Quantile}\left(s_i i = 1^n; \frac{\lceil (n+1)\alpha \rceil}{n}\right), \tag{2}$$

The theoretical strength of CP lies in its *marginal coverage guarantee*, which ensures that for a new test point x_{n+1}, the true label y_{n+1} is contained in the prediction set $C(x)$ with a probability of at least $1 - \alpha$:

$$P(y_{n+1} \in C(x_{n+1})) \geq 1 - \alpha, \tag{3}$$

During inference, for any input query X, the prediction set $C(X)$ is constructed by collecting all options whose softmax probabilities exceed the calibrated threshold $\hat{q}$:

$$C(X) = y \in \mathcal{Y} : \hat{f}(X)_y \geq \hat{q}. \tag{4}$$

The cardinality of $C(X)$ serves as a proxy for the model's epistemic uncertainty: When $|C(X)| = 0$, it indicates that the questions in the dataset are of high difficulty or logical complexity for the model. Such data is considered reasoning-intensive and necessitates deep deliberation. If $|C(X)| > 1$, it suggests that the model exhibits significant uncertainty regarding the answers, often due to knowledge ambiguity or

internal conflicts. In this scenario, we employ parallel expansion strategy. Finally, when $|C(X)| = 1$, it signifies that the answers are generally deterministic and well-represented within the model's parameters. Under these conditions, the model operates without adjustments to its reasoning progression.

3.2 TPV Training

To capture the internal progression of reasoning, we first construct a calibration training set $\mathcal{Q}$. For each problem $q \in \mathcal{Q}$, we prompt the base reasoning model to generate multiple independent trajectories. For a reasoning chain containing N tokens, we extract the sequence of hidden states $H = \{h_1, h_2, \ldots, h_N\}$ from the final layer. We define the relative reasoning progress p_j for each token as its normalized position within the sequence:

$$p_j = \frac{j}{N}, \quad j \in 1, 2, \ldots, N, \tag{5}$$

This results in a dataset of pairs $\mathcal{D} = \{(h_j, p_j)\}_{j=1}^{\sum N}$ mapping internal states to progress values. Empirical evidence suggests that the reasoning progress of LLMs is linearly encoded within the hidden state space.[Eisenstadt *et al.*, 2025] Consequently, we define the TPV θ as the weight of a linear predictor, optimized by minimizing the Mean Squared Error (MSE):

$$\min_{\theta,b} \mathcal{L} = \sum_{(h_j,p_j)\in\mathcal{D}} |\theta^T h_j + b - p_j|^2, \tag{6}$$

where $\theta \in R^d$ represents the TPV. Once the Thinking Progress Vector (TPV) is obtained, we can manually modify the model's hidden states during the inference stage to control the reasoning progress. The formula is:

$$h_\alpha = h + \alpha\theta. \tag{7}$$

where α is the intervention strength, h_α represents the adjusted hidden states, h represents the original hidden states of the model, and θ is the Thinking Progress Vector.

3.3 Dynamical Thinking Path Adjusting

Building upon the TPV framework, we propose a dynamic intervention mechanism that adjusts the intervention intensity α based on the Information Gain (ΔI) of different reasoning stages. Following the *InfoGain* framework, we quantify the extent to which each reasoning step s_i within a trajectory $\mathcal{S} = \{s_1, s_2, \ldots, s_t\}$ reduces the uncertainty regarding the final answer A^*. The Information Gain ΔI_i is defined as the reduction in the Perplexity (PPL) of the correct answer:

$$\Delta I_i = PPL(A^*|Q, s_{1:i-1}) - PPL(A^*|Q, s_{1:i}), \tag{8}$$

To establish a robust mapping, we first partition the normalized progress $p \in [0, 1]$ into multiple discrete intervals(accroding to the previous reasoning steps). Using a calibration set, we generate reasoning trajectories and calculate the average Information Gain ΔI_{avg} for each interval. This allowed us to identify low-gain stages where the model is prone to overthinking or redundancy. Based on these calibration results, we pre-define a mapping function $\alpha(p)$ that assigns an intervention strength to each progress stage:

$$\alpha(p) = \text{Mapping}(\Delta I_{avg}[p]). \tag{9}$$

During the inference stage, the model only needs to compute the current progress value $p = \theta^T h + b$ at each step. The corresponding intervention strength α is then retrieved from the pre-defined mapping based on the current interval.

4 Experiment

To evaluate our method's effectiveness, we conducted several experiments across different domains, including medical, mathematical, and scientific. For the purpose of a straightforward evaluation, we specifically examined queries with definite answers, such as those found in medical question-answering or multiple-choice problems, and fill-in-the-blank mathematical problems.

4.1 Settings

Datasets

We selected three domains for our evaluation: medical, as a typical knowledge-intensive domain; mathematics, as a reasoning-intensive domain; and science, chosen as a compromise option that have both knowledge-intensive and reasoning-intensive features. For the medical domain, we evaluated on five medical question-answering benchmarks. Specifically, we used MedQA-USMLE (MedQA) [Jin *et al.*, 2021], the 4-Options (MedBullet (4)) and 5-Options (MedBullet (5)) splits from the MedBullets platform (MedBullet) [Chen *et al.*, 2025], Humanity's Last Exam (Med) [Kalinin and others, 2026], and MedXpertQA [Zuo *et al.*, 2025]. All of these are single or multiple-choice QA problems, chosen for their verifiability and high quality. For multimodal tasks, we performed evaluations using medical datasets that feature easily verifiable answers. Specifically, we utilized MedXpertQA(MM) for our experiments. For the mathematics domain, we evaluated on four math datasets: AIME 2024 (AIME24) [Codeforces,], AMC23 [AIMO, 2024], and MATH500[Hendrycks *et al.*, 2021]. For the science domain, we employed GPQA-diamond [Rein *et al.*, 2024].

Models

We conducted experiments using six large reasoning models to validate the effectiveness of our method. For knowledge-intensive problems like those in medicine, large reasoning models need to possess a certain level of medical knowledge. Therefore, for medical problems, we selected a medical-domain-specific large model, HuatuoGPT-O1 [Chen *et al.*, 2024a], a cross-domain large model, GURU [Cheng *et al.*, 2026], and two general-purpose large models, Deepseek Distill Qwen [Guo *et al.*, 2025] and Qwen3 [Yang *et al.*, 2025]. For mathematics and science problems, general-purpose large models typically have the requisite knowledge, especially for mathematical problems. Thus, for these two domains, we employed a cross-domain large model, GURU, two sizes of the Deepseek Distill Qwen model, and a general-purpose large model, Qwen3. For multimodal benchmarks, we test our method on Qwen2.5-VL-7B-Instruct [Bai *et al.*, 2025].

Method	MedQA		MedBullet		MedBullet(5)		HLE (Med)		MedXpertQA	
	Pass@1	TK	Pass@1	TK	Pass@1	TK	Pass@1	TK	Pass@1	TK
DeepSeek-r1 Distill Qwen 7B										
Original	53.04	344	43.77	350	36.62	387	15.53	417	10.92	435
s1*	52.13	607	40.19	615	35.65	627	17.82	883	11.95	879
AlphaOne	51.53	378	42.68	417	35.90	442	15.84	679	11.20	768
SMV	**54.01**	344	**47.01**	332	36.95	407	15.53	352	11.11	420
Ours	53.85	300	**47.01**	321	**40.27**	367	**19.42**	307	**16.14**	370
Qwen3 8B Thinking										
Original	68.48	504	57.47	574	49.03	532	**16.67**	513	10.27	709
s1*	70.92	858	60.91	1035	54.07	1265	10.78	980	17.20	1378
AlphaOne	78.24	816	67.53	876	63.63	903	10.68	769	19.88	910
SMV	75.86	477	66.23	416	61.24	507	12.87	480	17.74	598
Ours	**79.03**	429	**69.48**	322	**64.72**	401	11.13	431	**21.33**	570
GURU 7B										
Original	57.63	670	**50.00**	770	42.21	497	17.65	706	**13.67**	673
s1*	57.27	1238	49.35	1571	41.88	1283	17.65	988	13.32	1096
AlphaOne	57.58	1037	**50.00**	1227	44.48	890	21.35	695	13.10	934
SMV	60.17	670	49.67	703	44.48	497	21.35	685	13.10	597
Ours	**60.80**	481	48.53	603	**44.86**	452	**22.75**	697	13.25	591
HuatuoGPT-O1 7B										
Original	68.16	568	56.17	521	49.68	476	15.69	572	15.96	608
s1*	67.64	907	56.49	970	47.73	1049	**16.83**	1539	14.60	1481
AlphaOne	64.27	830	57.31	573	52.22	896	**16.83**	798	14.56	673
SMV	69.52	562	60.06	521	54.22	476	16.32	572	15.67	552
Ours	**69.99**	387	**62.34**	411	**55.32**	443	**16.83**	523	**15.97**	479
M1 23K 7B										
Original	71.01	2161	58.77	2451	51.95	2339	17.32	3135	16.29	3265
s1*	70.36	2762	60.38	3104	55.19	3292	**19.41**	4716	19.04	3789
AlphaOne	70.46	2439	66.88	2156	60.06	1976	18.45	3217	**20.43**	2981
SMV	73.13	2161	65.58	2451	57.79	2339	16.50	3135	18.01	3265
Ours	**74.23**	1876	**75.58**	1917	**60.71**	2002	18.45	2731	**20.43**	2456

Table 1: We compare the pass@1(%) and TK(number of average generated tokens) of our cross-domain TTS method against the original model and other methods (SMV represents the shortest majority vote). The highest and second-best scores are indicated in **bold** and underlined, respectively.

Baselines

For language tasks, to validate the effectiveness of our method, we compared it with four baseline approaches. The first, original output, represents the model's Pass@1 output without employing any test-time scaling methods. The second, S1 [Muennighoff *et al.*, 2025], forces the model to continue thinking, refining, and correcting its reasoning by inserting a "wait" token. The third, alphaone [Zhang *et al.*, 2025a], introduces a unified adjustable α-moment to regulate the model's processing speed. The fourth is a parallel sampling method, shortest majority vote strategy(SMV) [Zeng *et al.*, 2025], which addresses the model's overthinking problem by selecting answers with shorter reasoning paths but the highest vote count. For multimodal tasks, most of the TTS methods cannot be transferred directly to align with the multimodal inference process, so we chose zero-shot and CoT as baselines.

4.2 Main Result

Medical Domain

Table 1 presents the results of our method compared to baseline approaches, demonstrating that our proposed cross-domain test-time scaling method outperforms most baselines.

For example, in experiments with the M1 7B model on the more reasoning-intensive HLE (Med) task, s1 and AlphaOne perform better than the short majority vote (SMV). Our method is adaptive and opts for more efficiency and precision for these reasoning-intensive tasks, outperforming the short majority vote by 2.43%. As shown in 3, in multimodal tasks, our method is still effective.

Furthermore, we found that our method is more effective on models that tend to generate extensive reasoning steps,

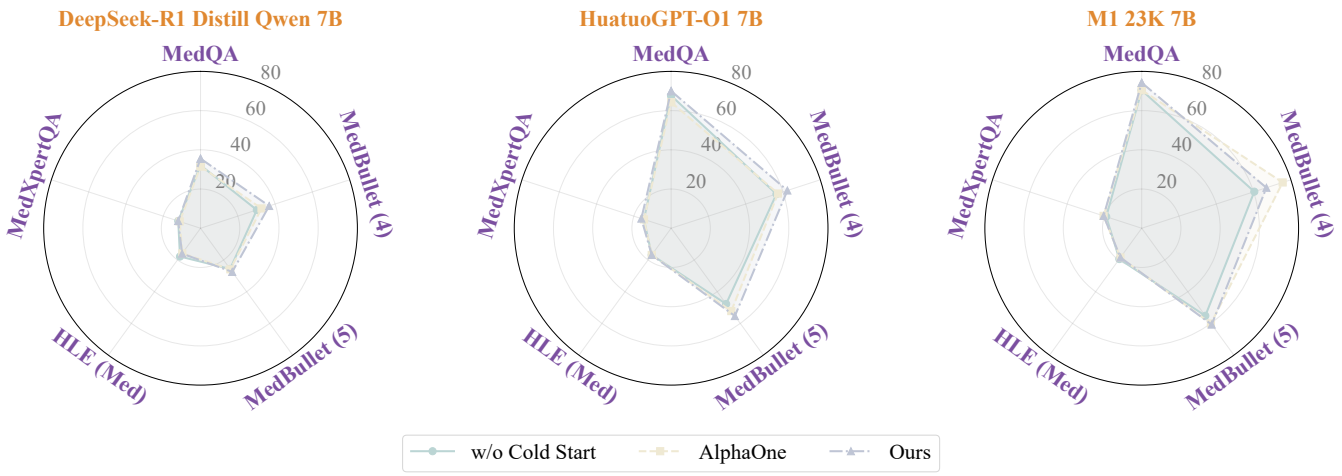

Figure 3: Comparison of performance with and without cold start applied. From left to right, the results are for DeepSeek-R1-Distill-Qwen-7B, HuatuoGPT-O1, and M1-23K-7B in the medical domain.

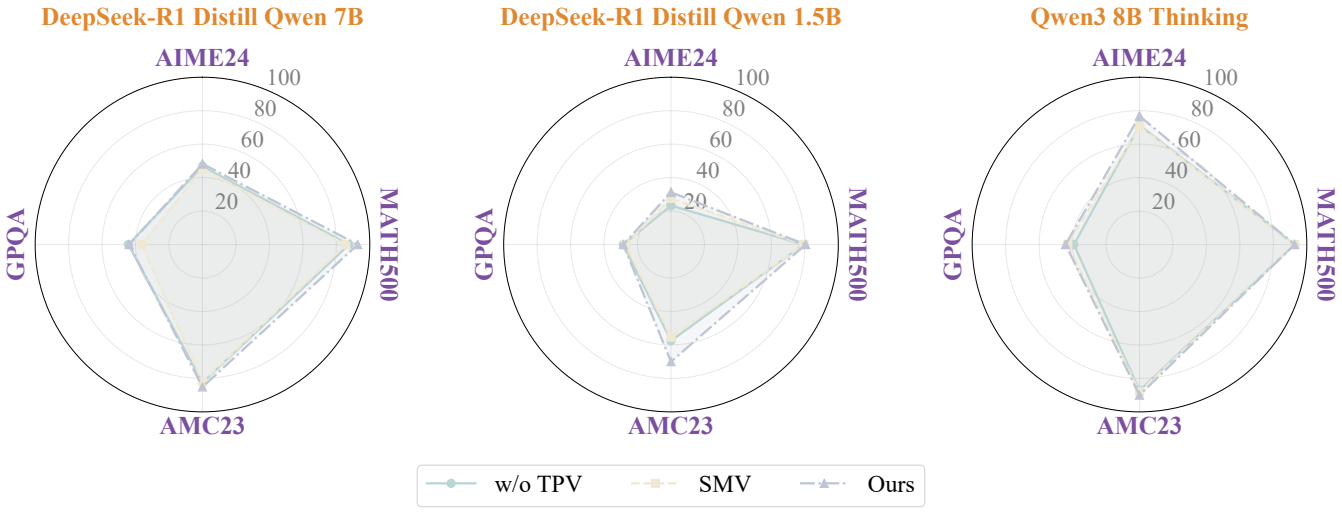

Figure 4: Comparison of performance with and without cold start applied. From left to right, the results are for DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, and Qwen3 8B Thinking in the medical domain.

such as M1. When the model fails to obtain useful information during reasoning, our approach can accelerate the reasoning process and enable the model to reach a conclusion more quickly. This prevents the model from repeatedly overturning its own conclusions, reduces token consumption by approximately 50%, and achieves higher pass@1 accuracy.

Math and Science Domain

For the math domain, which is highly reasoning-intensive, as shown in Table 2, our method effectively improves accuracy while reducing token consumption. Experimental results demonstrate that deep reasoning approaches, such as S1 and AlphaOne, remain effective for mathematics-related problems that prioritize logical inference; they guide the model into deeper cognitive processing. For example, on capacity-limited models such as DeepSeek-R1 distilled Qwen-1.5B, the performance gains from test-time scaling are particularly obvious, with our method achieving up to about 12% improvement. Compared with S1 and AlphaOne, our method not only maintains comparable accuracy but also significantly reduces the number of tokens required. Specifically, our results show an average reduction of approximately 30% in token consumption. On benchmarks like scientific problems, where reasoning and knowledge components are more balanced, the advantages of our method are even more pronounced. Our approach achieved the highest accuracy across all three models while saving a substantial number of inference tokens.

4.3 Ablation Studies

Study of the effectiveness of cold start strategy

To validate the effectiveness of the parallel expansion component in our method, we conducted ablation experiments on medical domain, which is a typical knowledge-intensive do-

Method	AIME24		MATH500		AMC23		GPQA	
	#P@1	#TK	#P@1	#TK	#P@1	#TK	#P@1	#TK
DeepSeek-R1 Distill Qwen 7B								
Original	46.17	6648	87.46	3638	82.15	4624	44.23	8944
s1*	46.27	7295	**92.78**	4513	80.0	5673	42.11	9444
AlphaOne	**50.0**	6078	91.21	3079	**90.0**	4397	43.28	9279
SMV	45.17	6648	85.64	3638	82.12	4624	44.23	8944
Ours	48.32	5726	**92.18**	2832	85.61	4315	**45.73**	8162
Qwen3 8B Thinking								
Original	71.31	7932	92.86	5732	87.25	7876	39.23	7451
s1*	72.91	8645	**93.37**	6499	88.95	8328	40.52	7999
AlphaOne	73.12	6498	**93.38**	4332	88.72	7287	**44.63**	6829
SMV	70.62	7732	92.81	5732	88.0	7876	42.92	7451
Ours	**76.84**	6538	92.77	4675	**90.0**	6945	44.25	6847
GURU 7B								
Original	**17.54**	8932	77.35	7352	52.28	7832	40.82	7892
s1*	15.31	9577	78.24	8975	53.25	8644	42.25	8537
AlphaOne	**17.55**	9653	79.82	9054	55.22	8307	44.33	7598
SMV	16.77	8932	77.83	7352	53.56	7832	44.82	7892
Ours	17.23	8387	**80.0**	8321	**55.81**	7306	**47.42**	6399
DeepSeek-R1 Distill Qwen 1.5B								
Original	23.33	7280	79.22	3637	57.48	5339	28.50	8943
s1*	26.72	7798	78.24	4374	57.45	6418	29.17	9554
AlphaOne	30.0	5916	**81.0**	3782	**70.0**	4952	29.75	8790
SMV	27.62	7280	79.45	3637	55.33	5339	27.24	8943
Ours	**31.25**	6832	80.52	3245	**70.0**	4655	**30.50**	8211

Table 2: Performance comparison across different backbones. We report P@1 (%) and TK for each benchmark, #P@1: Pass@1 (%); #Tk: number of generated tokens. The highest and second-best scores are indicated in **bold** and underlined, respectively.

main on three models: M1, HuatuoGPT-O1, and Deepseek-R1 Distill Qwen 7B. As shown in Figure 3, without the cold start mechanism, the Pass@1 of all three models decreased on most benchmarks, particularly on knowledge-intensive tasks such as MedQA, MedBullet(4), and MedBullet(5), which demonstrates the necessity of parallel scaling, especially for knowledge-intensive tasks such as medicine.

To validate the effectiveness of the slow thinking adjustment component of our method, we conducted ablation experiments on three models: DeepSeek-R1 Distill Qwen 1.5B, DeepSeek-R1 Distill Qwen 7B, and Qwen3 8B, within the mathematics domain. As shown in Figure 4, without the slow thinking adjustment, the Pass@1 of all three models decreased across all benchmarks. These results have shown the necessity of the effectiveness of our cold start strategy.

Study of the effect of information gain-based thinking adjusting

We conducted ablation experiments on the Information Gain-based thinking path, adjusting part of our framework. The ablation experiment was done using the DeepSeek-R1-Distill-Qwen-7B model across benchmark datasets in three domains: MedQA (medicine domain), AIME24 (mathematics domain), and GPQA (science domain). As shown in Table 4, the results demonstrate the effectiveness of information gain-based thinking. Adjusting this mechanism leads to a performance decline in both pass@1 accuracy and token consumption.

Method	#P@1 (%)	#TK
Original	19.15	7839
CoT	17.25	8475
Ours	**21.25**	**6732**

Table 3: Multimodal results comparison on Qwen-VL 2.5 7B Instruct on MedxpertQA benchmark. #P@1: Pass@1 (%); #TK: number of generated tokens. The highest and second-best scores are indicated in **bold** and underlined, respectively.

Method	MedXpertQA		AIME24		GPQA	
	#P@1	#TK	#P@1	#TK	#P@1	#TK
Original	10.92	435	46.17	6648	44.23	8944
w/o IG	11.75	432	45.17	6648	44.23	8944
w IG	16.14	370	48.32	5726	45.73	8162

Table 4: Ablation study on Information gain-based thinking path adjusting. #P@1: Pass@1 (%); #Tk: number of generated tokens.

We further observe that the Information Gain for knowledge-oriented tasks is lower than that for tasks requiring complex reasoning. Taking evaluations in the medical domain as an example, we find that the Information Gain measured on MedQA is approximately 0.15, while on MedXpertQA it can reach 0.3. This result suggests that for knowledge-intensive tasks, reasoning steps do not bring about significant Information Gain to help the model arrive at the correct answer. This may also explain why the reasoning process does not substantially improve the answer accuracy of large language models on knowledge-based tasks.

5 Conclusion

In this paper, we propose a cross-domain test-time scaling (TTS) framework comprising a conformal prediction-based cold-start strategy and information gain-based reasoning path adjustment. To be more specific, we use conformal prediction to set the starting point for test-time scaling, then leverage internal progress vectors to dynamically adjust the model's reasoning speed based on the effectiveness of each step in the inference process. We conducted experiments in three representative domains, and both experimental results and ablation studies demonstrate our method's overall effectiveness and the contribution of its key components. We hope our work inspires future research on cross-domain reasoning in LLMs and LVLMs, moving past the limitations of traditional reasoning-intensive and knowledge-intensive tasks.

Acknowledgements

This research was supported by funding from the Hong Kong RGC General Research Fund (Grant Nos. 15221123, 15216424, and 15211525), and the Hong Kong PolyU Internal Research Fund (Grant Nos. P0058468 and P0056171). This research was also supported by funding from the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515010184).

References

[AI-MO, 2024] AI-MO. Ai-mo. AIMO Validation Dataset - AMC, 2024. Accessed: 2025-05-19.

[Akter *et al.*, 2026] Syeda Nahida Akter, Shrimai Prabhumoye, Matvei Novikov, Seungju Han, Ying Lin, Evelina Bakhturina, Eric Nyberg, Yejin Choi, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-CrossThink: Scaling self-learning beyond math reasoning. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 984–1002, Rabat, Morocco, March 2026. Association for Computational Linguistics.

[Angelopoulos and Bates, 2023] Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Bi *et al.*, 2025] Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing LLM reasoning. In *Forty-second International Conference on Machine Learning*, 2025.

[Chen *et al.*, 2024a] 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.

[Chen *et al.*, 2024b] Shuhao Chen, Weisen Jiang, Baijiong Lin, James Kwok, and Yu Zhang. Routerdc: Query-based router by dual contrastive learning for assembling large language models. *Advances in Neural Information Processing Systems*, 37:66305–66328, 2024.

[Chen *et al.*, 2025] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, 2025.

[Cheng *et al.*, 2026] Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Taylor W. Killian, Haonan Li, Mikhail Yurochkin, Eric P. Xing, Zhengzhong Liu, and Zhiting Hu. Revisiting reinforcement learning for LLM reasoning from a cross-domain perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.

[Codeforces,] MAA Codeforces. American invitational mathematics examination-aime 2024, 2024.

[Eisenstadt *et al.*, 2025] Roy Eisenstadt, Itamar Zimerman, and Lior Wolf. Overclocking llm reasoning: Monitoring and controlling thinking path lengths in llms. *arXiv preprint arXiv:2506.07240*, 2025.

[Guo *et al.*, 2025] 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.

[Hendrycks *et al.*, 2021] 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. *NeurIPS*, 2021.

[Hu *et al.*, 2026] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[Huan *et al.*, 2025] Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*, 2025.

[Huang *et al.*, 2025] 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.

[Inoue *et al.*, 2026] Yuichi Inoue, Kou Misaki, Yuki Imajuku, So Kuroki, Taishi Nakamura, and Takuya Akiba. Wider or deeper? scaling LLM inference-time compute with adaptive branching tree search. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[Jaech *et al.*, 2024] 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.

[Jin *et al.*, 2021] 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.

[Kalinin and others, 2026] Mikhail Kalinin et al. A benchmark of expert-level academic questions to assess ai capabilities. *Nature*, 649(8099):1139–1146, 2026.

[Kamoi *et al.*, 2024] Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

[Lai *et al.*, 2025] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learn-

ing for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[Li *et al.*, 2025] Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E. Gonzalez, and Ion Stoica. S*: Test time scaling for code generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15964–15978, Suzhou, China, November 2025. Association for Computational Linguistics.

[Liu *et al.*, 2025] Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, et al. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*, 2025.

[Ma *et al.*, 2026] Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun MA, and Wenhu Chen. General-reasoner: Advancing LLM reasoning across all domains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[Muennighoff *et al.*, 2025] 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. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics.

[Oh *et al.*, 2025] Gyutaek Oh, Seoyeon Kim, Sangjoon Park, and Byung-Hoon Kim. Rethinking test-time scaling for medical ai: Model and task-aware strategies for llms and vlms. *arXiv preprint arXiv:2506.13102*, 2025.

[Rein *et al.*, 2024] 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.

[Snell *et al.*, 2025] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Su *et al.*, 2025] Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.

[Wang *et al.*, 2025] Junlin Wang, Jue WANG, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Wei *et al.*, 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.

[Wu *et al.*, 2025a] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

[Wu *et al.*, 2025b] Juncheng Wu, Sheng Liu, Haoqin Tu, Hang Yu, Xiaoke Huang, James Zou, Cihang Xie, and Yuyin Zhou. Knowledge or reasoning? a close look at how llms think across domains. *arXiv preprint arXiv:2506.02126*, 2025.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[Zeng *et al.*, 2025] Zhiyuan Zeng, Qinyuan Cheng, Zhangyue Yin, Yunhua Zhou, and Xipeng Qiu. Revisiting the test-time scaling of o1-like models: Do they truly possess test-time scaling capabilities? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4651–4665, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Zhang *et al.*, 2025a] Junyu Zhang, Runpei Dong, Han Wang, Xuying Ning, Haoran Geng, Peihao Li, Xialin He, Yutong Bai, Jitendra Malik, Saurabh Gupta, and Huan Zhang. AlphaOne: Reasoning models thinking slow and fast at test time. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11329–11354, Suzhou, China, November 2025. Association for Computational Linguistics.

[Zhang *et al.*, 2025b] Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. Crowd comparative reasoning: Unlocking comprehensive evaluations for LLM-as-a-judge. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5059–5074, Vienna, Austria, July 2025. Association for Computational Linguistics.

[Zuo *et al.*, 2025] Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedxpertQA: Benchmarking expert-level medical reasoning and understanding. In *Forty-second International Conference on Machine Learning*, 2025.