Evaluating the Role of Verifiers in Test-Time Scaling for Legal Reasoning Tasks

Davide Romano, Jonathan Richard Schwarz, Daniele Giofrè

Thomson Reuters

{Davide.Romano2, Jonathan.Schwarz, Daniele.Giofre}@thomsonreuters.com

Abstract

Test-time scaling (TTS) techniques can improve the performance of large language models (LLMs) at the expense of additional computation and latency. While TTS has proven effective in formal domains such as mathematics and programming (Snell et al., 2024; Chen et al., 2024), its value in argumentative domains such as law remains underexplored. We present an empirical study of verifier-based TTS methods for legal multiple-choice QA (MCQA) across five benchmarks. Using a family of 7 reward models, we evaluate both outcome-level (Best-of-N) and process-level (tree search) verification under realistic low-N budgets. Our analysis systematically investigates how verifier utility is affected by key properties such as domain specialization, model size, and supervision type (process-supervised PRMs vs. outcome-only ORMs), even when applied across different roles.

1 Introduction

Test-Time Scaling (TTS) methods aim to enhance Large Language Model (LLM) performance by trading additional compute for improved accuracy at inference time. The broad spectrum of these techniques range from single-path approaches like generating longer Chains-of-Thought (CoT) (Wei et al., 2022; Guo et al., 2025; Jaech et al., 2024) to more complex parallel and verifier-guided methods such as Best-of-N (BoN) selection and tree search. Systematic investigations of these verifier-guided methods in formal domains like math and programming have demonstrated substantial accuracy improvements on multiple choice QA (MCQA) tasks (Brown et al., 2024; Wu et al., 2024; Snell et al., 2024). However, the legal domain presents distinct challenges; its reasoning is often defeasible and accommodates multiple valid analytical paths. While prior work has explored single-path inference for legal reasoning (Yu et al., 2025), investigations into
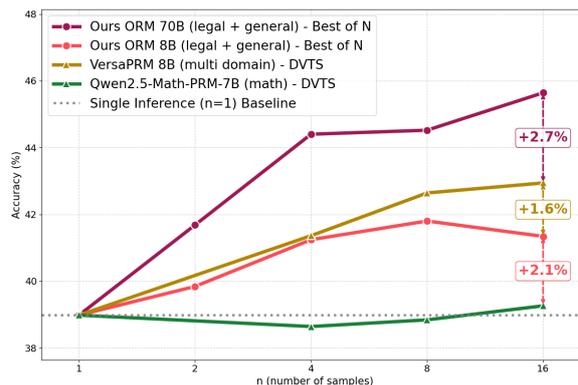

Figure 1: TTS with Llama-3.1-8B-Instruct with four different verifiers from N=4 to N=16, average over 5 legal MCQA benchmarks

parallel and verifier-based TTS are notably absent from the literature. This gap is critical because the verifiers underpinning these methods are often trained on general-purpose or formal-domain data. It remains an open question whether such models can reliably evaluate legal reasoning, or if domain-specific verifiers are required to achieve meaningful gains. These verifiers are broadly categorized into two types: Outcome Reward Models (ORMs), which assign a single score to a complete output, and Process Reward Models (PRMs), which provide fine-grained, step-by-step feedback (Uesato et al., 2022; Lightman et al., 2023). This paper addresses the aforementioned gap by empirically investigating whether the performance enhancements observed in formal domains translate to legal MCQA. Through an extensive comparison of reward models, we analyze how to optimize these verification strategies by evaluating the impact of verifier domain specialization, model size, and supervision type (PRM vs. ORM).

We examine verifier-based TTS for legal reasoning to answer the following research questions:

- **RQ1 (Value of verification under matched compute).** Under equal compute budgets, do

outcome-verified Best-of-N (BoN) and Diverse Verifier Tree Search (DVTS) outperform simple Majority Vote (MV) on legal MCQA benchmarks?

- **RQ2 (Importance of domain specialization and verifier size).** With method and compute held constant, does a legal-specialized verifier outperform a general-domain verifier, and how large is the additional effect of scaling the verifier size?

- **RQ3 (Role transfer between verifiers)**: Under matched size and compute, could these reward models be used also out-of-role (for PRMs in outcome verification and for ORMs in process-verification) in legal MCQA tasks?

Contributions This paper makes three key contributions to the understanding of verifier-based test-time scaling for legal reasoning. First, we conduct a comprehensive comparison between MV, BoN, and process-verified DVTS using open-source models, revealing that verifier-based methods rarely outperform simple voting baselines by significant margins in legal reasoning. Second, through systematic ablation studies, we show that both verifier model size and domain specialization are crucial for improving performance, with legal-domain training providing a distinct advantage that becomes most apparent at larger scales. Notably, we find that the utility of all methods diminishes as generator model capability increases, with even sophisticated verification providing minimal gains. Finally, our analysis of supervision type shows that PRMs deliver superior performance compared to ORMs of similar size, even when PRMs are deployed outside their intended role for outcome verification tasks. These findings provide valuable guidance for practitioners seeking to optimize computational resources in legal NLP applications.

2 Experimental Setup

We test three generators: **Llama-3.2-3B-Instruct**, **Llama-3.1-8B-Instruct**, and **Llama-3.1-70B-Instruct** (Dubey et al., 2024). We ran our tests with CoT prompting (Wei et al., 2022) and temperature $T = 0.8$ and the system prompts in appendix B. Our evaluation compares three methods:

Majority Vote (MV): Sample k CoT responses and select the most frequent answer.

Best-of-N (BoN): Sample N CoT responses, score each with an Outcome Reward Model (ORM), and

select the one with highest reward.

DVTS (Beeching et al.): A tree search guided by a Process Reward Model (PRM) scoring partial steps. For optimal hyperparameter choice, such as expansion width or aggregation strategy, we ran an ablation study on MBE exam that can be found in appendix C.

While generating N trajectories of length T dominates computational cost at $\mathcal{O}(T^2)$, the reward-model scoring methods (Best-of-N and DVTS) add only modest linear $\mathcal{O}(T)$ overhead, making all three approaches comparable in runtime (see more details in Appendix A) .

We use the verifiers detailed in Table 1 and evaluate on five legal benchmarks: binary **COLIEE Task 4** (Goebel et al., 2025); four-option **MBE** BAR Exam and **LEXam** (Fan et al., 2025); eight-option **SuperGPQA (Law subset)** (Du et al., 2025); and thirty-two-option **LEXam-32**(Fan et al., 2025). **MBE** BAR Exam is the only restricted-access benchmark.

Type	Verifier Model	Size	Training data
ORM	Our RMs	8B, 70B	General + Legal[†]
	Skywork-Reward (Liu et al., 2024)	8B, 27B	General
PRM	VersaPRM (Zeng et al., 2025)	8B	Multi-domain*
	Qwen2.5-Math-PRM (Zhang et al., 2025)	7B, 72B	Math

Table 1: Verifiers used in our study, grouped by supervision type (ORM/PRM). *Multi-domain includes Law, Philosophy, Biology and others. [†]Both our models were trained on identical datasets, comprising general knowledge sources such as UltraFeedback (Cui et al., 2023) and restricted-access legal data from US and UK jurisdictions. This training corpus encompasses various task types: legal reasoning, legal information retrieval, legal summarization, and basic instruction following.

2.1 Study Design

RQ1 (Value of Verification): We compare MV against both BoN with Skywork-RM-27B, and DVTS (with VersaPRM-8B). For BoN, we use Skywork-RM-27B due to its performance on the RewardBench (Lambert et al., 2024) benchmark. For DVTS, we selected VersaPRM-8B as it is the first open-source, multi-domain PRM available for research. **RQ2 (Impact of Domain & Size):** We compare legal (Ours RMs) vs. general (Skyworks) ORMs using BoN, and the multi-domain (VersaPRM) vs. out-of-domain
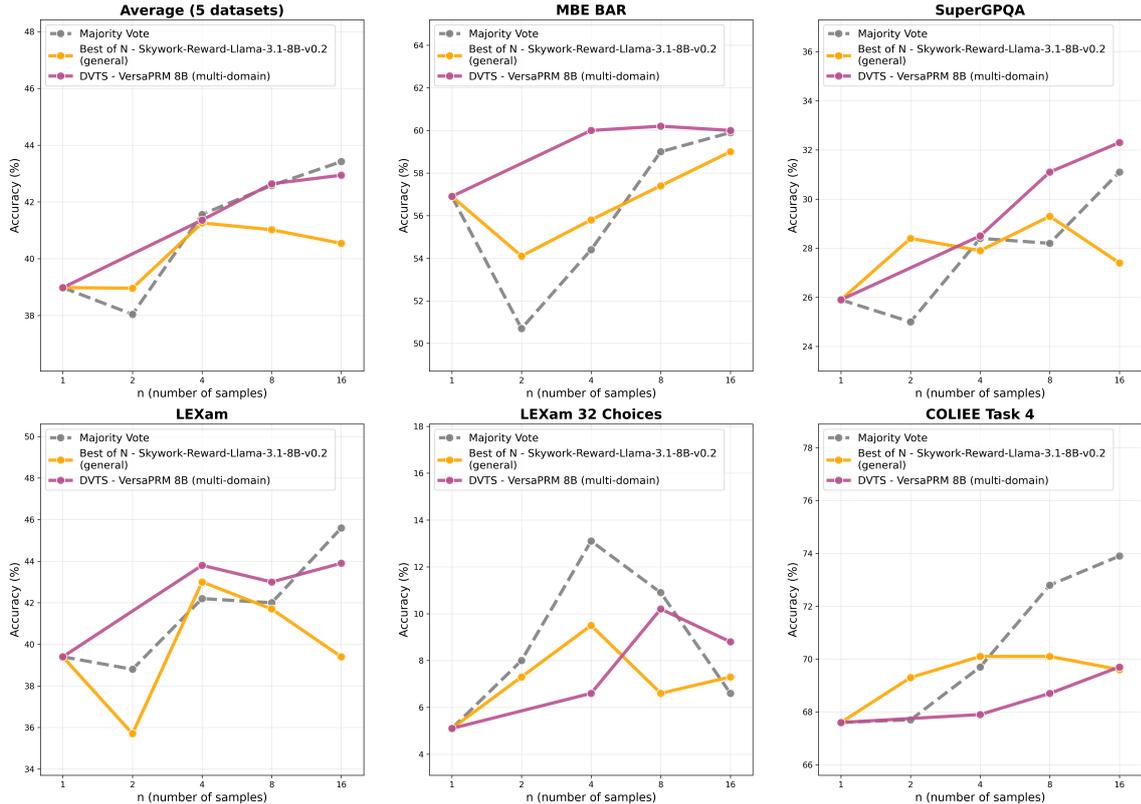
Figure 2: RQ1 results across all benchmarks with `Llama 8B` as the generator

(`Qwen-Math-PRM`) PRMs using DVTS. **RQ3 (Supervision Type):** We test our legal ORMs (`Ours`) against the legal PRM (`VersaPRM`) on both BoN and DVTS.

3 Results

3.1 RQ1: Value of verification under matched compute

In Figure 2 and figures 5-6 in appendix D we can see that MV remains a strong baseline across benchmarks and generator scales. The only model where BoN and DVTS surpass on average across the 5 benchmarks the MV baseline is `Llama-3.2-3B-Instruct`, where they achieve around 1.4% average improvement. For larger models, however, verifications provides no performance benefit or even show a decrease compared to MV. These limited benefits of verification prompted us to further explore the potential of domain specialization and size as two variables to obtain better verification in 3.2.

Performance Variation Across Benchmarks A closer analysis of the per-benchmark performance reveals that the utility of verification is heavily influenced by the task's complexity, specifically the

Table 2: Relative performance Gains on Llama 70B Generator at N=16 against Majority Vote baseline. Best-of-N uses Legal ORM 70B and DVTS uses QwenPRM 72B.

Benchmark	Method	Rel. Gain
MBE BAR	Best-of-N	+0.6
	DVTS	-4.8
SuperGPQA	Best-of-N	-2.1
	DVTS	-4.7
LEXam	Best-of-N	-1.3
	DVTS	-0.5
LEXam 32	Best-of-N	+10.2
	DVTS	**+12.4**
COLIEE Task 4	Best-of-N	-1.4
	DVTS	-2.7

cardinality of the answer space. While this is more limited with smaller generators, with 70B generator as shown in Table 2 the difference is considerable. On benchmarks with a small number of answer choices, such as COLIEE Task 4, MBE, and LEXam, verifier-based methods offer marginal or even negative gains over the highly effective MV
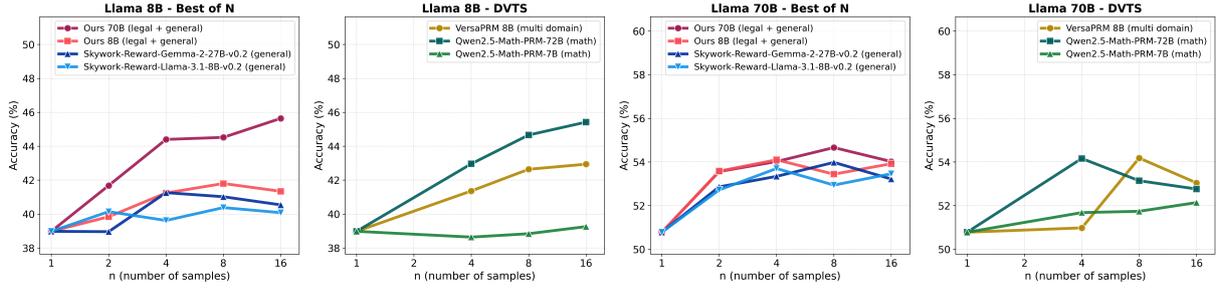
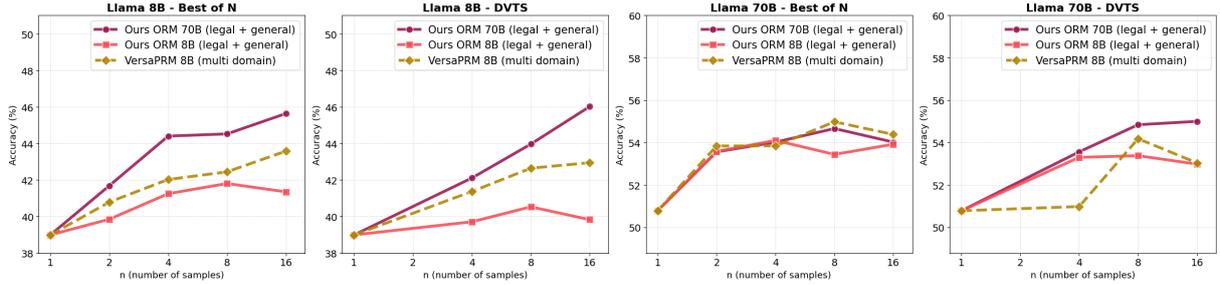Figure 3: RQ2 average results with both `Llama 8B` and `Llama 70B`

Figure 4: RQ3 average results with both `Llama 8B` and `Llama 70B`

baseline. However, this trend is starkly reversed in the high-complexity **LEXam-32** task. With 32 possible answers, the output space becomes significantly noisier, causing the MV baseline to struggle. In this scenario, DVTS achieves a substantial relative gain of **+12.4%**.

3.2 RQ2: Domain specialization vs. verifier size

Figure 3 and figures 7-8 in appendix D demonstrate that BoN with *our* reward models (both 8B and 70B variants) match or outperform general-domain verifiers across evaluations. While the 8B model shows minimal performance advantages over general verifiers, the 70B model consistently delivers superior results across numerous benchmarks. Regarding the performance of PRMs with DVTS, `QwenPRM 72B` produces the most significant enhancement when coupled with smaller generator models. Direct comparison between similarly sized `VersaPRM 8B` and `QwenPRM 7B` reveals that `VersaPRM` consistently delivers superior performance.

3.3 RQ3: Supervision type and transfer (PRM vs. ORM)

Figure 4 and figures 12-16 in Appendix D show that PRMs provide consistent benefits: as BoN scorers they yield stronger reranking than size-matched ORMs, and within DVTS they offer more effective guidance. Improvements are concentrated on

the smaller generators, similar to the other results. `Ours 70B` can still perform better than `VersaPRM` also in process supervision even though it has received no process training.

3.4 Discussion

Table 3: Relative improvement values across Llama models at N=16 over the Majority Vote baseline at N=16.

Method + Reward Model	Llama 3B	Llama 8B	Llama 70B
BoN + VersaPRM 8B	+2.94	+0.16	**+1.56**
BoN + Legal ORM 70B	**+4.46**	**+2.22**	+1.20
DVTS + QwenPRM 72B	+4.00	+2.00	-0.06

Diminishing Returns of Verification The performance gains from verifier-based TTS decrease as the capability of the generator model improves. At same value of N when using the 70B generator, even well-configured verifiers provide only small improvements over the MV baseline (Table 3), which proves to be a very competitive method.

Task Complexity as a Key Differentiator The strong performance on the LEXam-32 benchmark provides a crucial insight into the practical limits of MV method. While MV is a robust baseline for tasks with a small set of discrete answers, its utility appears to degrade significantly as the solution space expands. It is in this high-cardinality

environment that verifier-guided methods demonstrate their value. This suggests that for problems with a constrained output space, the simplicity of MV may be sufficient, but as task complexity and the number of possible outcomes grow, the computational overhead of verification becomes a more justifiable investment. This relationship warrants further investigation, including in Open QA settings.

The Dual Impact of Scale and Specialization Our findings highlight two key drivers of verifier performance: model scale and domain specialization. Scaling a verifier from an 8B to a 70B model consistently yields substantial performance gains. Similarly, models trained on specialized legal data regularly outperform their general-domain counterparts. However, these two factors are linked. The advantage from specialization is most pronounced at the 70B scale, while at the 8B scale, specialized models like VersaPRM offer modest, yet clear, improvements. This indicates that while larger models are inherently more capable, targeted training provides a distinct, scale-dependent advantage.

The Generalization of Process Supervision Finally, we find that PRMs seem to work well also in both outcome and process verification for legal tasks. This may indicate that the step-by-step feedback used to train PRMs helps them develop a more robust measure of reasoning quality.

4 Conclusions

In this work, we presented a systematic evaluation of verifier-based TTS for legal multiple-choice question answering.

A consistent observation is the diminishing return of verification as the generator model's power increases; while gains are evident for smaller generators, they shrink significantly for more powerful ones, where a simple Majority Vote often remains competitive. However, the utility of Majority Vote is challenged by task complexity, and we find that verifier-based methods provide substantial gains in high-cardinality benchmarks where the answer space is large.

Crucially, we find that effective verification relies on the dual impact of model scale and domain specialization. The advantage of in-domain training is most pronounced at larger verifier scales. This is complemented by the notable generalization of process supervision: VersaPRM proved highly

versatile, outperforming size-matched Outcome Reward Models even when used out-of-role for outcome reranking.

For practitioners, these findings suggest that investing in high-quality, in-domain reward models is a promising direction for improving inference-time legal reasoning.

5 Limitations and Scope

The scope of this study is limited to legal reasoning MCQA, and our findings may not generalize to other legal tasks such as summarization or open-ended QA where verification is arguably more complex. Additionally, our experiments primarily focused on a single model family (i.e. `Llama 3.1` and `Llama 3.2`), and other model architectures might exhibit different improvements from verification. Future work should explore additional legal domains and open QA, expand the verifier pool with more recent reward models such as `Skywork-v2` Reward Models, and evaluate newer generators like `Qwen3` models. It should be noted that some verifiers used in this study are restricted-access data (*ours* RMs), which limits the full reproducibility of certain results.

References

Edward Beeching, Lewis Tunstall, and Sasha Rush. Scaling test-time compute with open models.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yu Fan, Jingwei Ni, Jakob Merane, Etienne Salimbeni, Ya ng Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Florian Geering, Oliver Dreyer, and 1 others. 2025. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*.

Randy Goebel, Yoshinobu Kano, Japan Calum Kawn, Mi-Young Kim, and Masaharu Yoshioka. 2025. International competition on legal information extraction and entailment (coliee 2025).

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

Yaoyao Yu, Leilei Gan, Yinghao Hu, Bin Wei, Kun Kuang, and Fei Wu. 2025. Evaluating test-time scaling llms for legal reasoning: Openai o1, deepseek-r1, and beyond. *arXiv preprint arXiv:2503.16040*.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, and 1 others. 2025. Versaprm: Multi-domain process reward model via synthetic reasoning data. *arXiv preprint arXiv:2502.06737*.

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

A TTS methods compute cost

For N trajectories with average length T, generation dominates total cost. Without KV caching, the generator cost scales as $\Theta(P_M N T^2)$. In contrast, reward-model scoring is linear in (T): Best-of-N adds one verifier forward per trajectory ($\Theta(P_R N T)$), and DVTS scores multiple partial paths across s reasoning steps, about $\frac{s+1}{2}$ times the BoN cost, $\Theta(P_R N T \frac{s+1}{2})$. With typical CoT lengths of hundreds of tokens (average is 1000 for our CoTs) and $s \approx 10$, these $\mathcal{O}(T)$ verifier terms remain much smaller than the $\mathcal{O}(T^2)$ generation term, even when the verifier is larger than the generator, so for a fixed sample count N, MV, BoN, and DVTS have comparable runtime, with only modest linear overheads for BoN and DVTS.

B Generator prompt templates

This section details the prompt templates used for the generator models.

B.1 Majority Vote & Best-of-N system prompt

For the Best-of-N (BoN) generation method, we use a custom system template for MCQA and the classic CoT preprompt.

Generator	Reward Model	Dataset	N	Expansion Width	Independent Subtrees	Accuracy
Llama-3.1-8B-Instruct	VersaPRM	BAR	16	8	2	55.6%
		BAR	16	4	4	59.5%
		BAR	16	**2**	**8**	**61.8%**
Llama-3.1-70B-Instruct	Ours 70B	LEXam 32	16	4	4	19.7%
		LEXam 32	16	**2**	**8**	**22.6%**
	VersaPRM	LEXam 32	16	4	4	**19.0%**
				2	8	21.2%
	Qwen2.5-Math-PRM-72B	LEXam 32	16	**4**	**4**	**20.4%**
				2	8	19.0%

Table 4: Expansion width tuning tests

Please complete the following user request.

When answering questions, first reflect on the problem step by step. At the end ALWAYS conclude with this phrase:

Therefore, the final answer is: \boxed{answer}. I hope it is correct.

Where answer CAN BE ONLY [answer_options].

The {answer_options} are specific for each dataset, and it represents the list of accepted final answers. For example for BAR:

Where answer CAN BE ONLY ONE OF THE FOLLOWING: "A", "B", "C", "D"'

For parsing we accepted both formats "\boxed{answer}" and "Therefore, the final answer is: {answer}" as Llama3.1 family didn't output the \boxed{} very often. When we do the selection of the final answer we filter for the ones that passed successfully the parsing.

B.2 DVTS system prompt

Please complete the following user request.

Use this step-by-step format:

Step 1
[Reasoning step description]

Step 2
[Reasoning step description]

...

Regardless of the approach, ALWAYS conclude with this phrase:

Therefore, the final answer is: \boxed{answer}. I hope it is correct.

Where answer CAN BE ONLY [answer_options].

C RMs hyperparameter tuning

C.1 Expansion width tuning

Diverse Verifier Tree Search (DVTS) (Beeching et al.) requires to set a hyperparameter called "expansion width" W which corresponds to the number of next steps expansions for each tree. Together with it, we have the number of initial subtrees T (or

also called "beams") at the start of the algorithm. The number N used in our paper is the corresponding of $W \cdot T$. To study the best parameter for W given a fixed budget N we performed the experiments in Table 4.

The results indicate that with smaller models such as `Llama-3.1-8B-Instruct` having a smaller expansion width (and therefore higher number of subtrees T) leads to better results. This is caused by the fact that diversifying more the generations at the start will lead to less formatting errors to parse the final answer. Therefore, in all our experiments we used $W = 2$ and the number T is N/W.

C.2 Score aggregation method tuning

From (Beeching et al.; Zeng et al., 2025) there are four common options to the aggregation strategy choice for the PRM scores:

Min-Aggregation

$$\text{Aggr}_{\min}(S) = \min_{i \in [k]} \text{PRM}(S)_i.$$

Last-Aggregation

$$\text{Aggr}_{\text{last}}(S) = \text{PRM}(S)_k.$$

Average-Aggregation

$$\text{Aggr}_{\text{avg}}(S) = \frac{1}{k} \sum_{i \in [k]} \text{PRM}(S)_i.$$

Prod-Aggregation

$$\text{Aggr}_{\text{prod}}(S) = \prod_{i \in [k]} \text{PRM}(S)_i.$$

To select our selection strategy we ran tests for VersaPRM and `Qwen2.5-Math-PRM-72B` on the BAR exam with `Llama-3.1-8B-Instruct` as generator with $N = 16$. The results are in Table 5 These results brought us to use for `VersaPRM` the **Mean** aggregation strategy. While for the Qwen-PRMs (both 72B and 7B) to use **Last**.

D   Full results

Full results are added, I just commented them for faster compilation

PRM	Aggregation Strategy	BAR accuracy
VersaPRM	**Mean**	**62.7%**
	Min	62.2%
	Last	59.5%
Qwen2.5-Math-PRM-72B	Mean	57.7%
	Prod	59.2%
	Min	57.6%
	Last	**60.5%**

Table 5: Aggregation strategy ablation tests
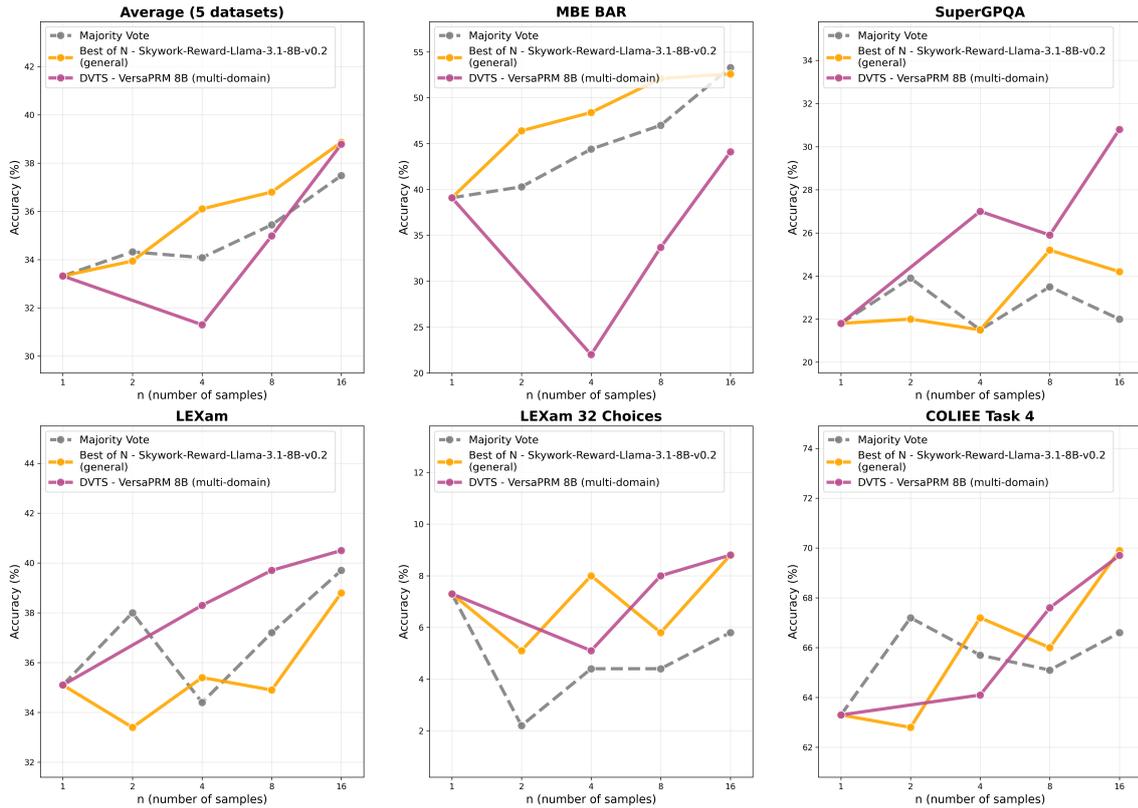
Figure 5: RQ1 average and individual benchmarks results using `Llama-3.2-3B-Instruct`
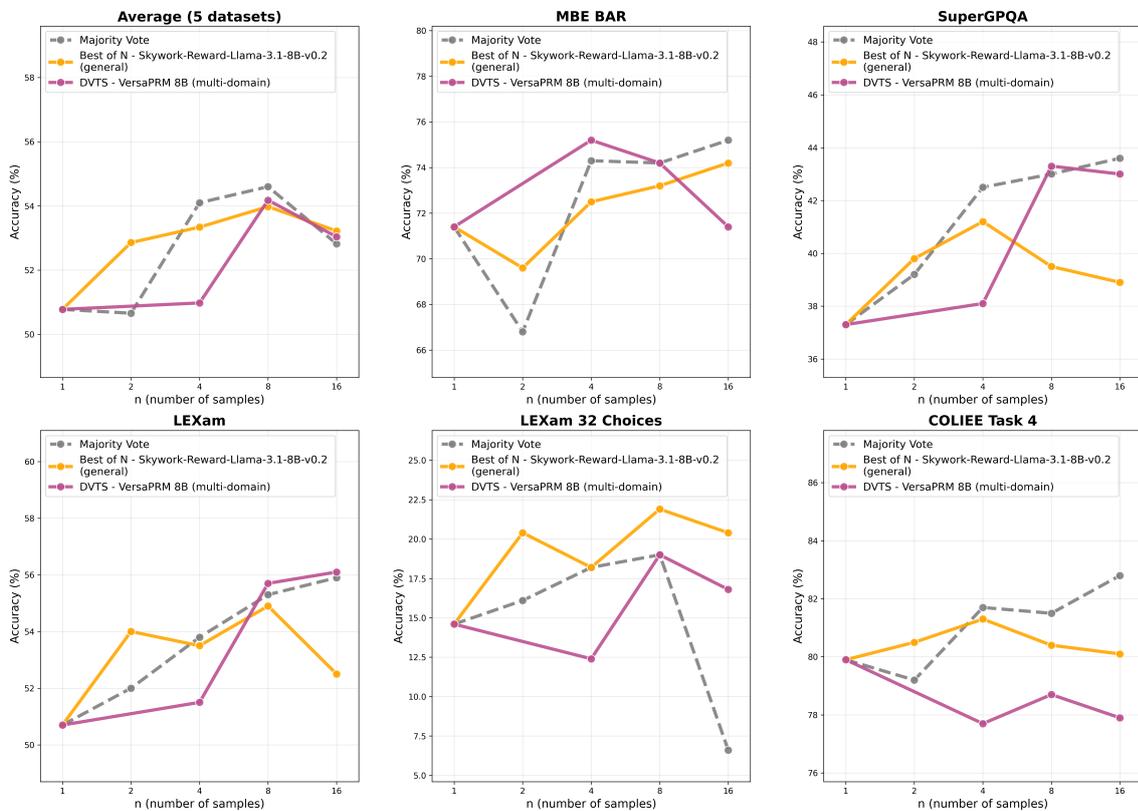
Figure 6: RQ1 average and individual benchmarks results using `Llama-3.1-70B-Instruct`
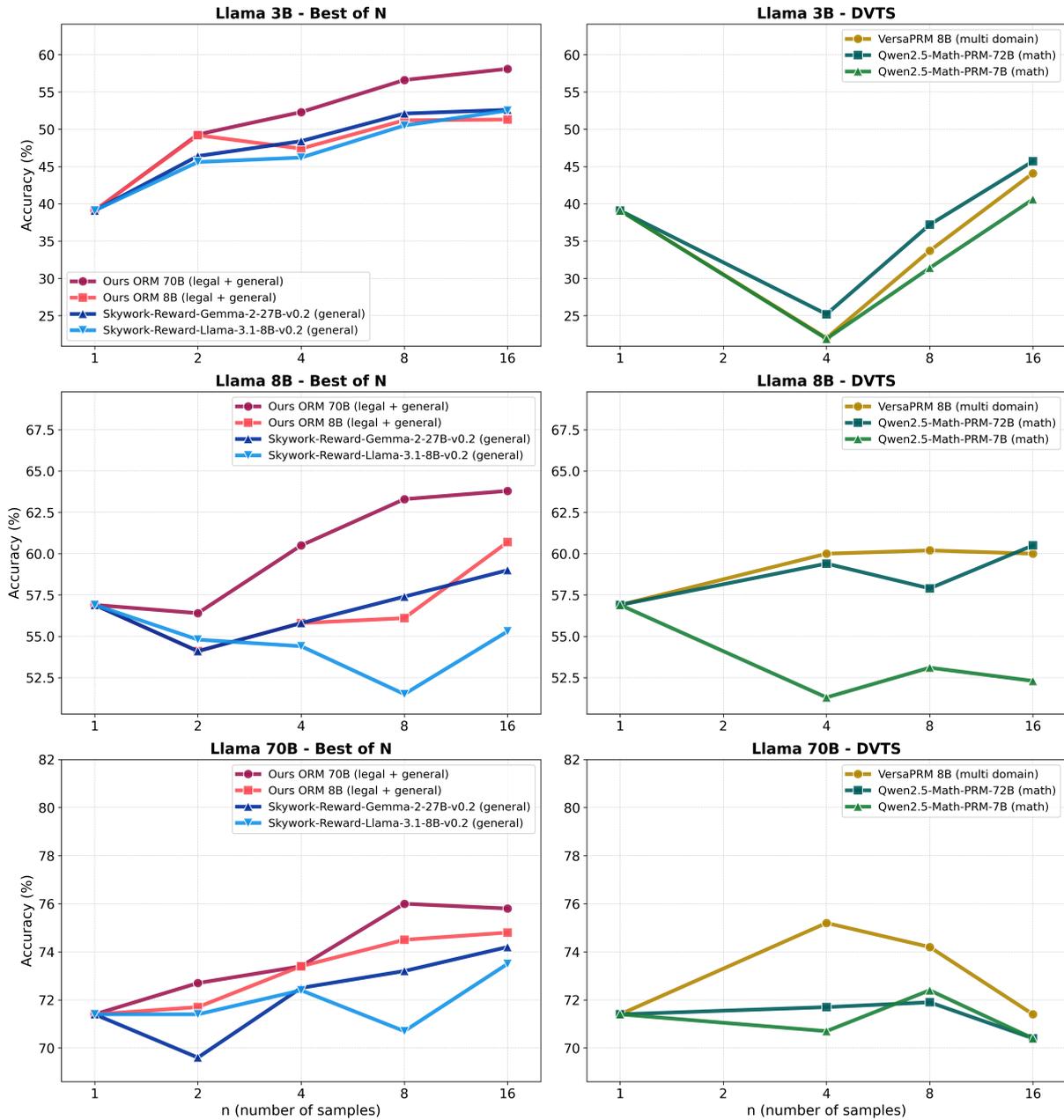
Figure 7: MBE bar exam RQ2 results with Best-of-N and DVTS
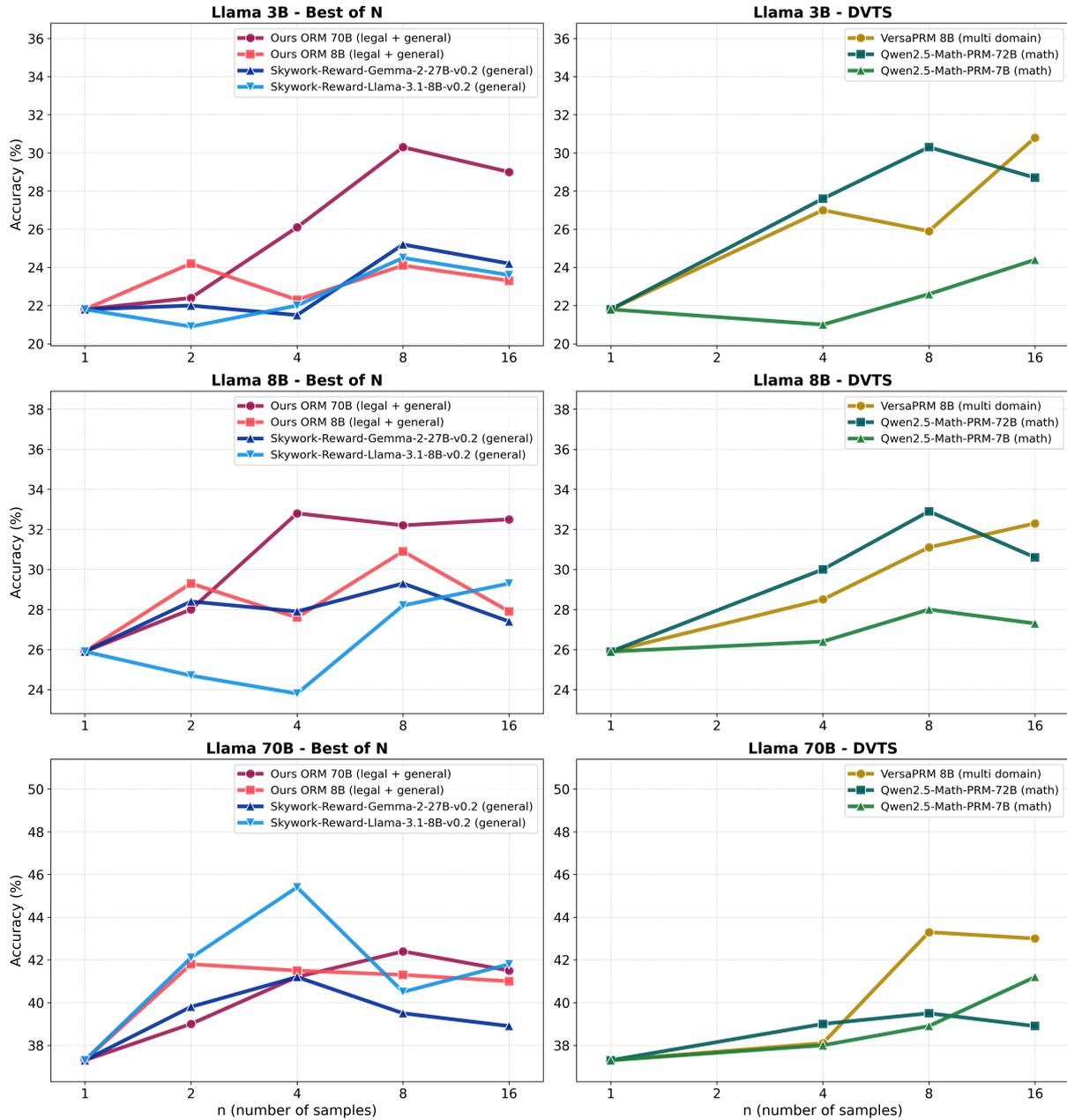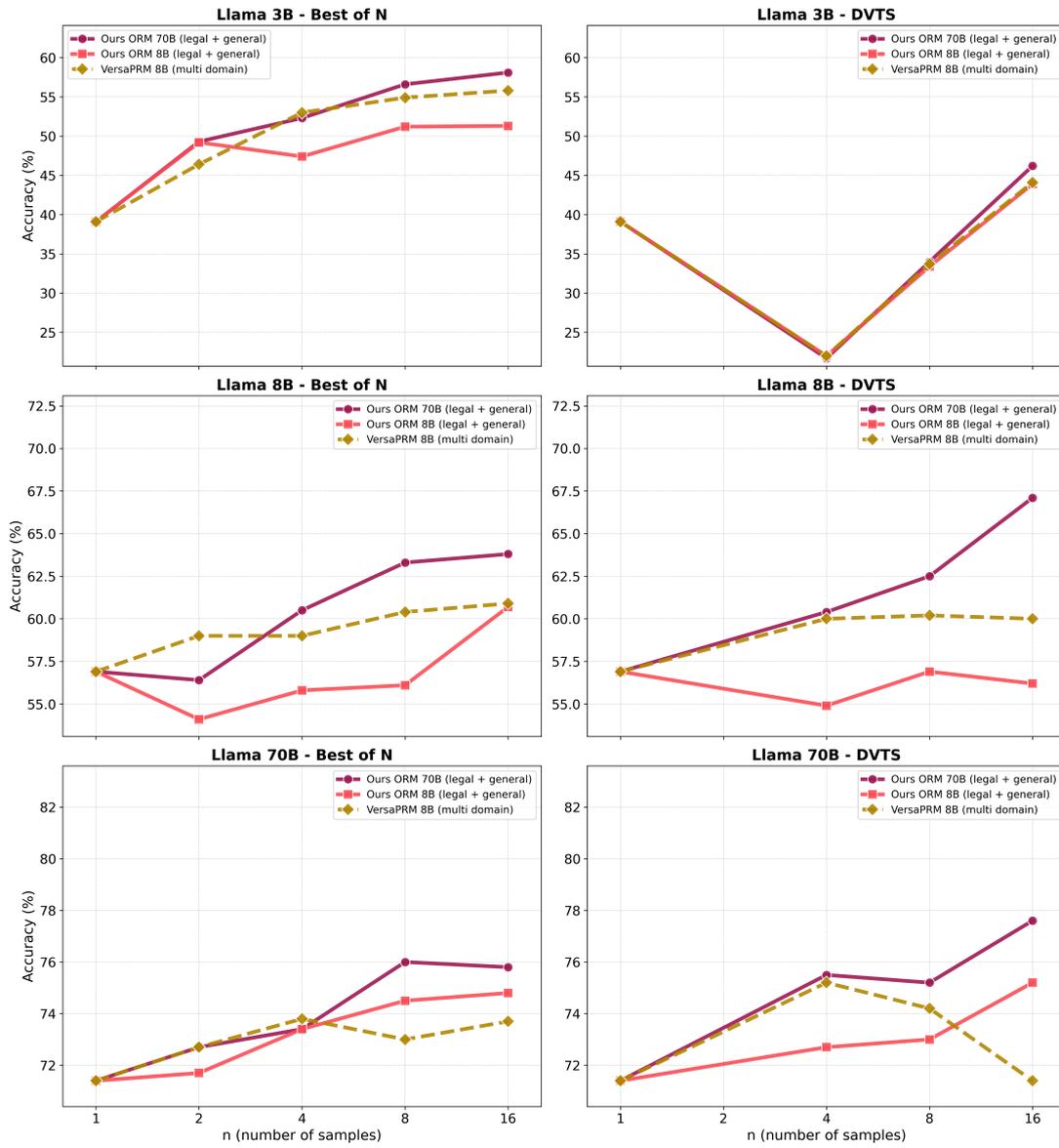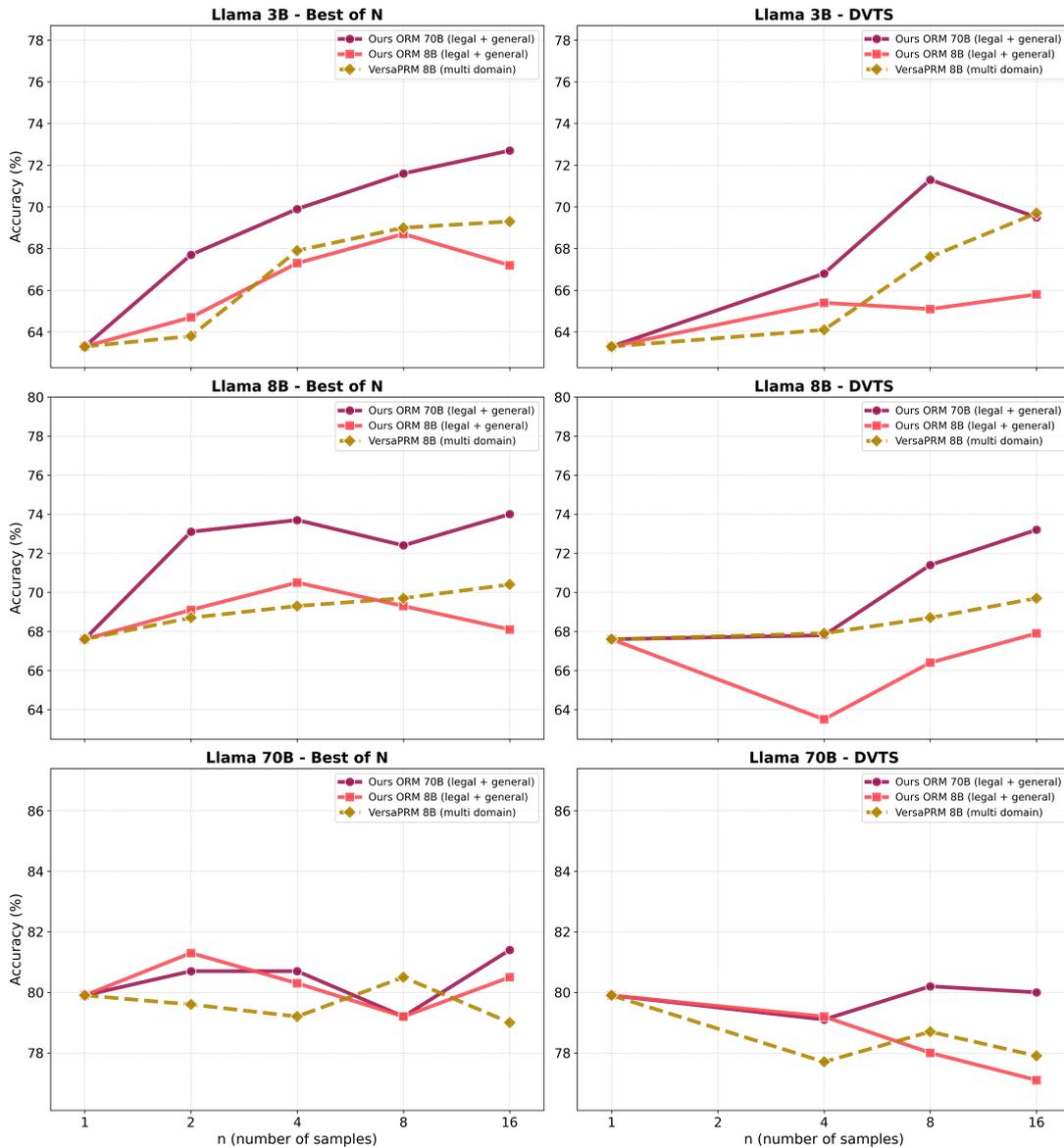
Figure 8: Coliee Task 4 RQ2 results with Best-of-N and DVTS
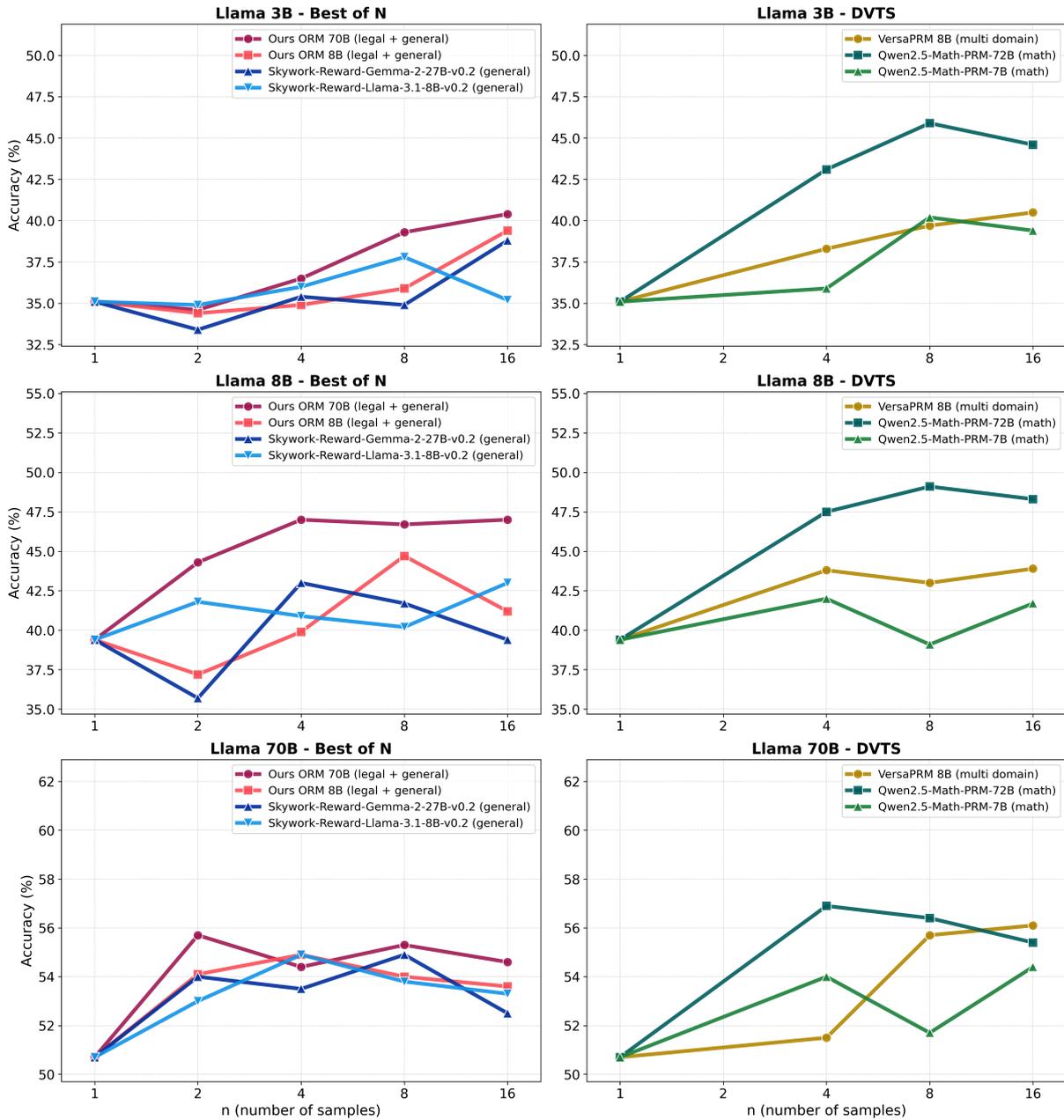
Figure 9: LEXam RQ2 results with Best-of-N and DVTS
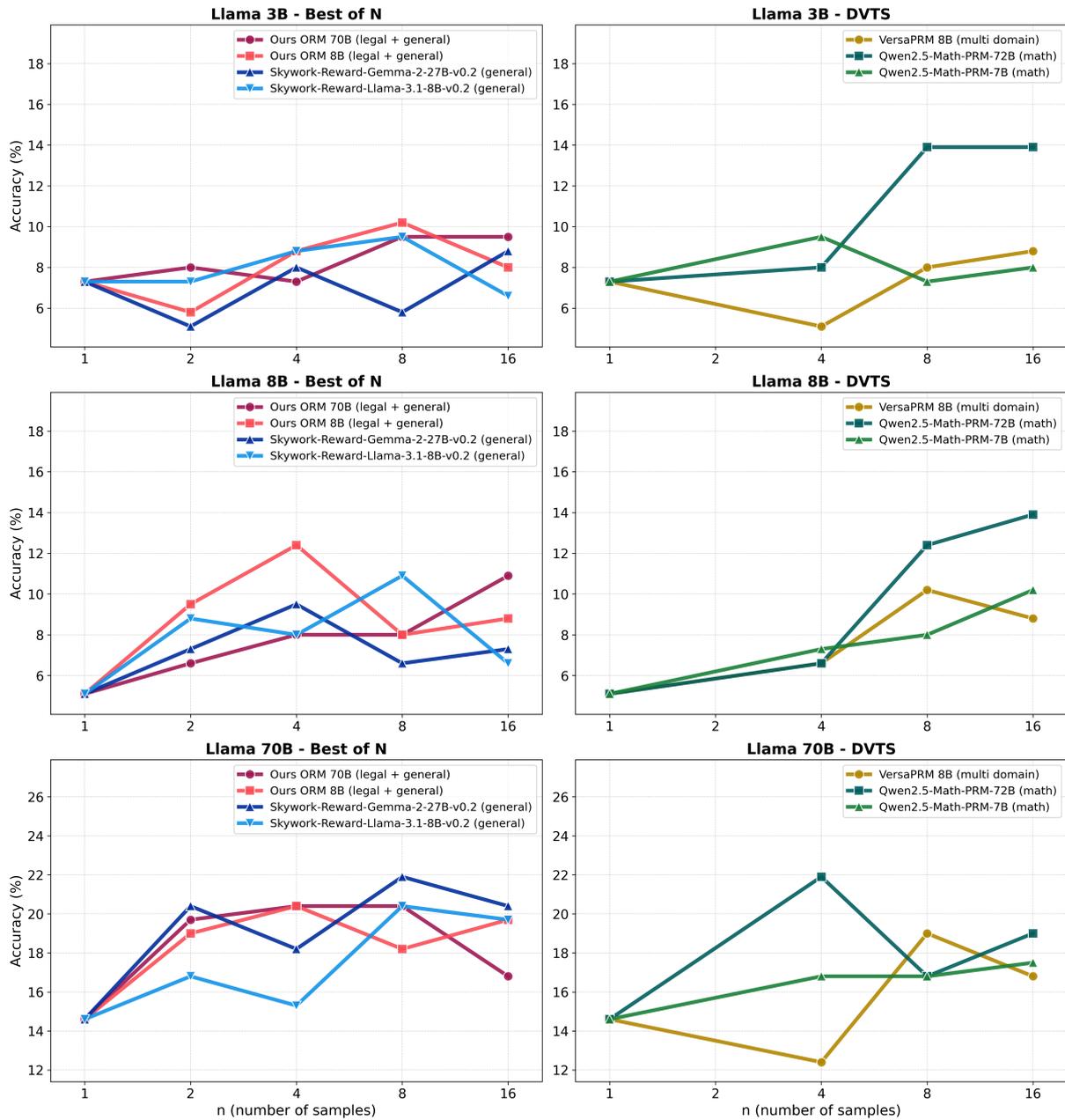
Figure 10: LEXam (32 options) RQ2 results with Best-of-N and DVTS
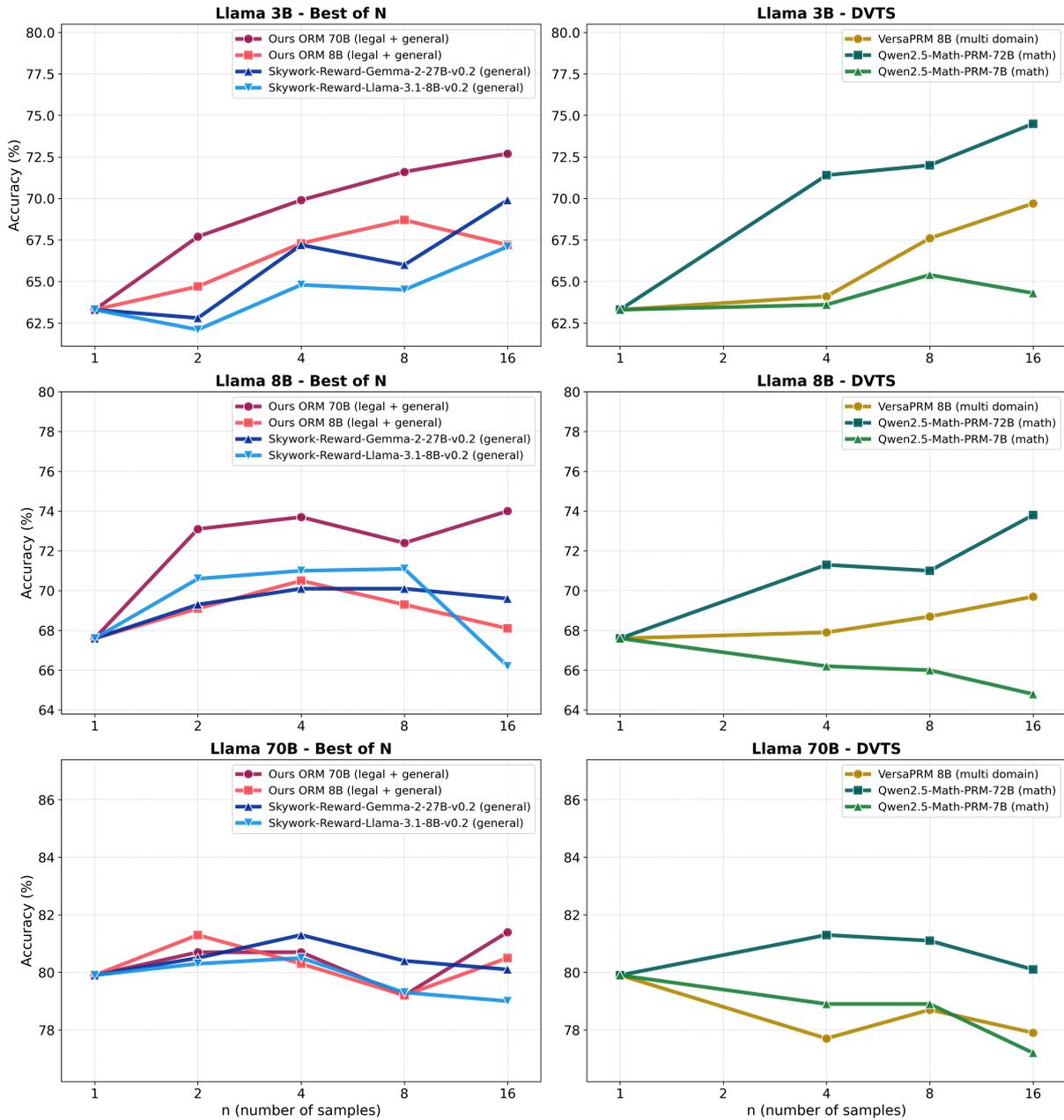
Figure 11: SuperGPQA RQ2 results with Best-of-N and DVTS

Figure 12: MBE bar exam RQ3 results with Best-of-N and DVTS

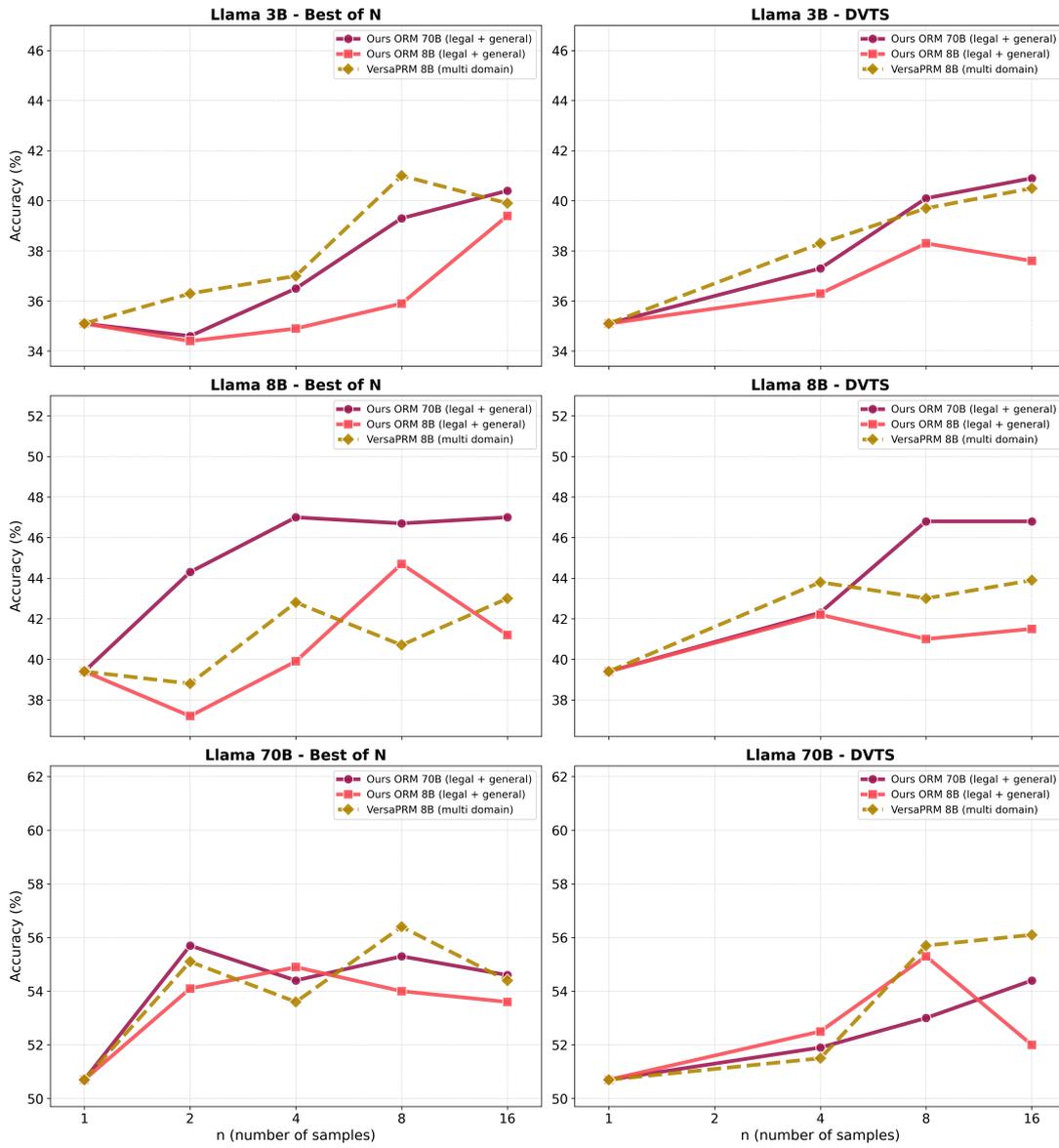Figure 13: Coliee Task 4 RQ3 results with Best-of-N and DVTS

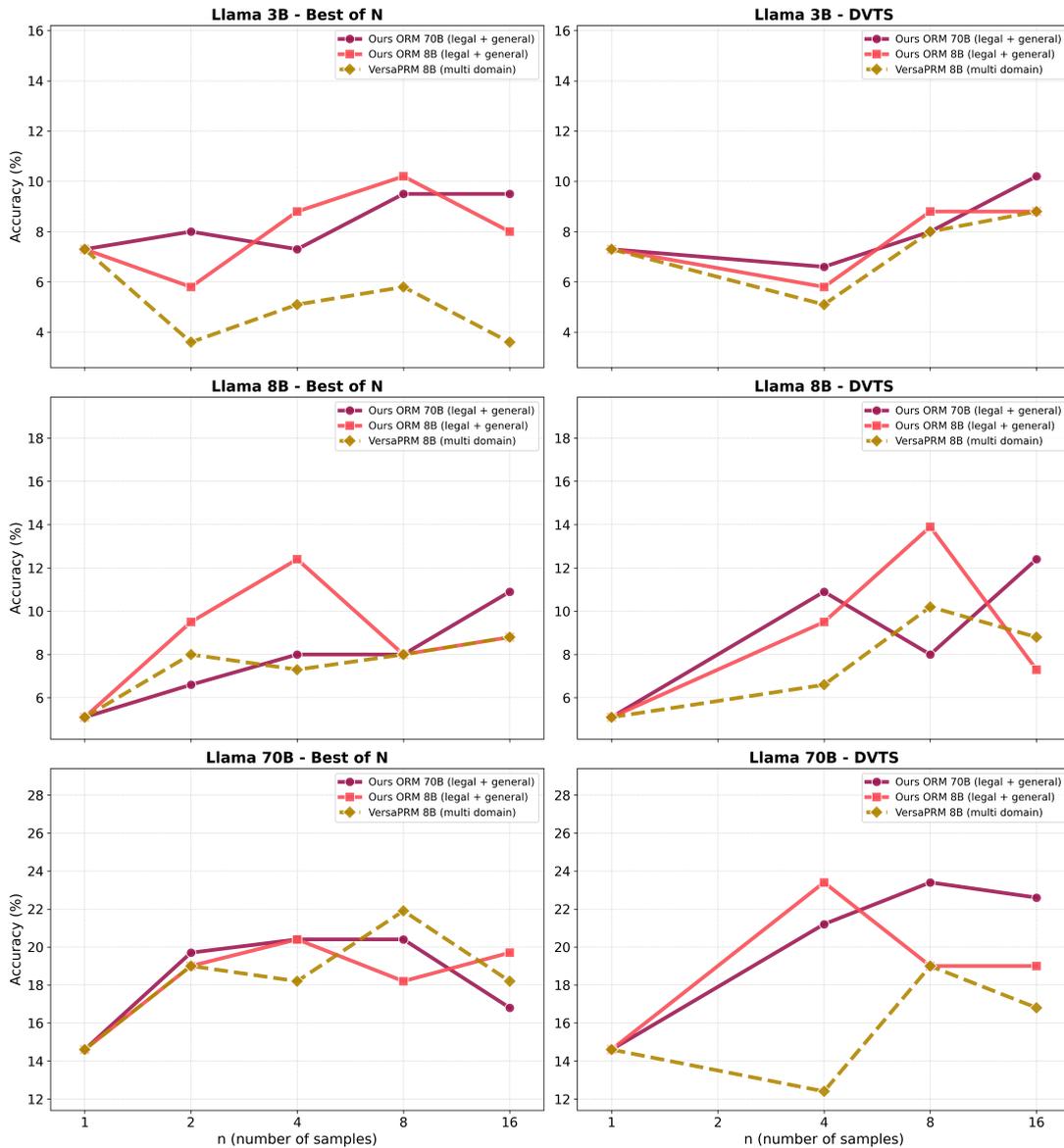Figure 14: LEXam RQ3 results with Best-of-N and DVTS
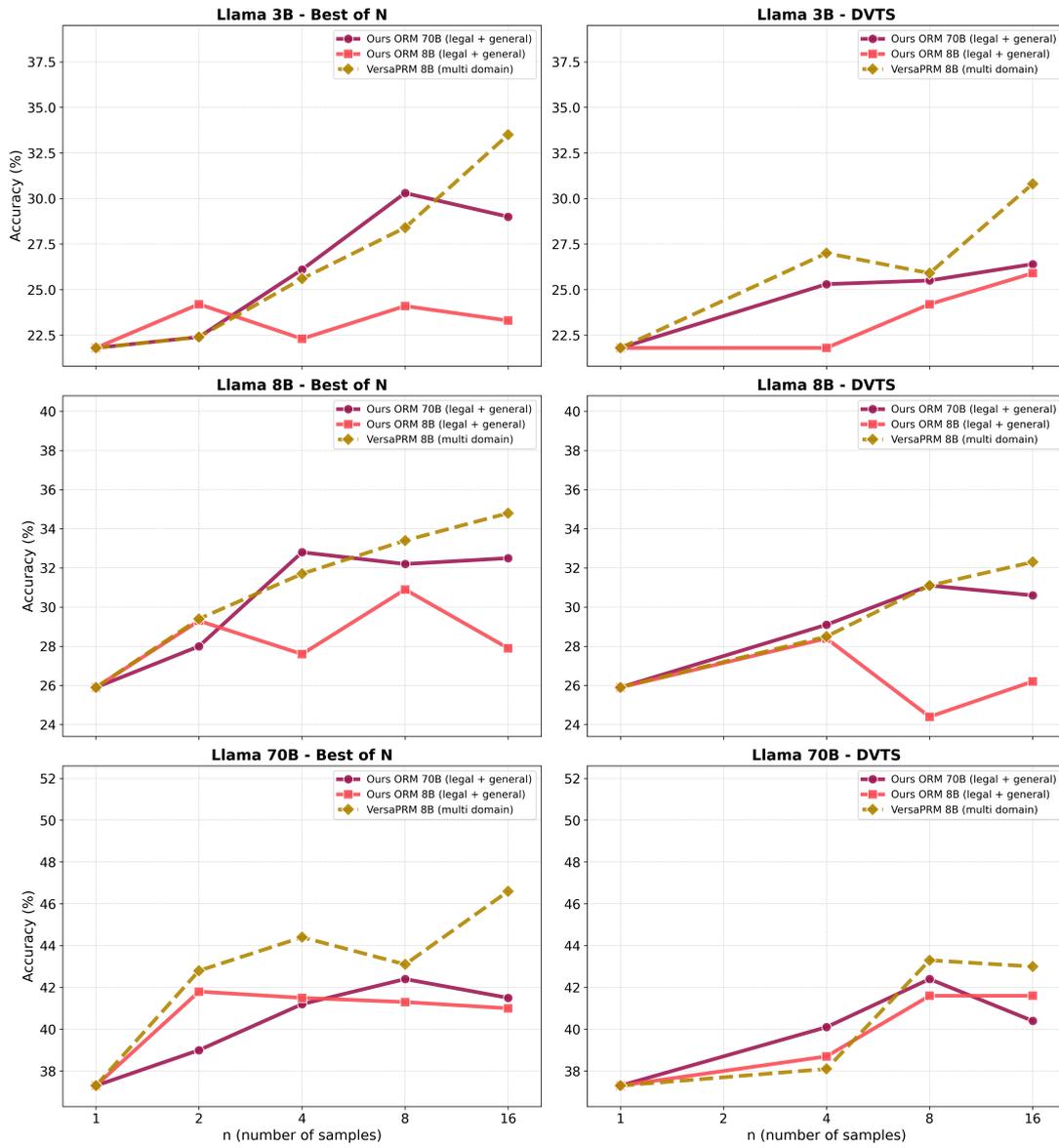
Figure 15: LEXam (32 options) RQ3 results with Best-of-N and DVTS

Figure 16: SuperGPQA RQ3 results with Best-of-N and DVTS