

## A Limitations

Despite the importance of our work in highlighting and addressing numerical precision’s effect on reproducibility, there are limitations that are worth noting. First, our experiments focus on a certain set of models and settings. Due to resource constraints, we do not conduct large-scale experiments on much larger models (>70B) or a wide range of GPU architectures or accelerators; these settings might exhibit different levels of numerical behaviors. Additionally, our study primarily addresses numerical precision issues in transformer-based LLMs and may not fully generalize to other architectures or modalities.

## B Broader Impacts

This research has several societal impacts. First, by raising the discussion of reproducibility issues in LLM inference, our work promotes greater scientific rigor in AI research, calling for a more reliable comparison of models and techniques. This is particularly important as LLMs are increasingly deployed in critical areas like healthcare, education, and public services, where consistency and reproducibility are essential. On the other hand, our findings reveal that true reproducibility may require additional computational resources, potentially exacerbating the already significant environmental effects of LLM development. We hope our work will encourage the AI community to establish better standards of evaluation—ultimately leading to more trustworthy and reliable AI systems.

## C Supplementary Results on Accuracy Variance under BFloat16

In Section 1, we demonstrated the variance of accuracy on BF16 with the same seed and greedy decoding but different batch sizes and number of GPUs in Figure 1. To provide a better sense of the challenges of reproducibility, in Figure 9, we further demonstrated this phenomenon across all datasets. We can see that the phenomenon is moderate in MATH500 and LiveCodeBench-Easy, but much more pronounced in LiveCodeBench-Medium and LiveCodeBench-Hard.

## D Supplementary Results on Std@Acc, Div\_Percent & Average Div\_Index

Following up on Figure 5 and Table 3 from Section 3.2. Here, we show full results on Std@Acc, Div\_Percent, and Average Div\_Index under all settings. In Table 5, we present the standard deviation of accuracy across different settings for two additional datasets. In Table 6, we report the average divergence indices for different models and datasets. In Table 7, we show the percentage of examples exhibiting divergent outputs across different settings. All of the results conform with the conclusion that during greedy decoding, rounding errors from floating point may cause serious reproducibility problems, and as precision increases from BF16 to FP32, the divergence phenomenon is largely reduced (with lower Std@Acc, lower Div\_Percent, and larger Div\_Index).

Table 5: Std@Acc across 12 different settings under LiveCodeBench datasets

	LiveCodeBench-Medium			LiveCodeBench-Hard		
	BF16	FP16	FP32	BF16	FP16	FP32
<b>DeepSeek-R1-Distill-Qwen-7B</b>	2.28%	2.21%	0.44%	1.19%	1.98%	0.42%
<b>DeepSeek-R1-Distill-Llama-8B</b>	2.08%	1.84%	0.78%	2.15%	1.90%	0.53%
<b>Qwen2.5-7B-Instruct</b>	1.44%	0.24%	0	0.40%	1.45e-17	1.45e-17
<b>Llama-3.1-8B-Instruct</b>	0.70%	0.48%	0.44%	0.65%	3.62e-18	3.62e-18

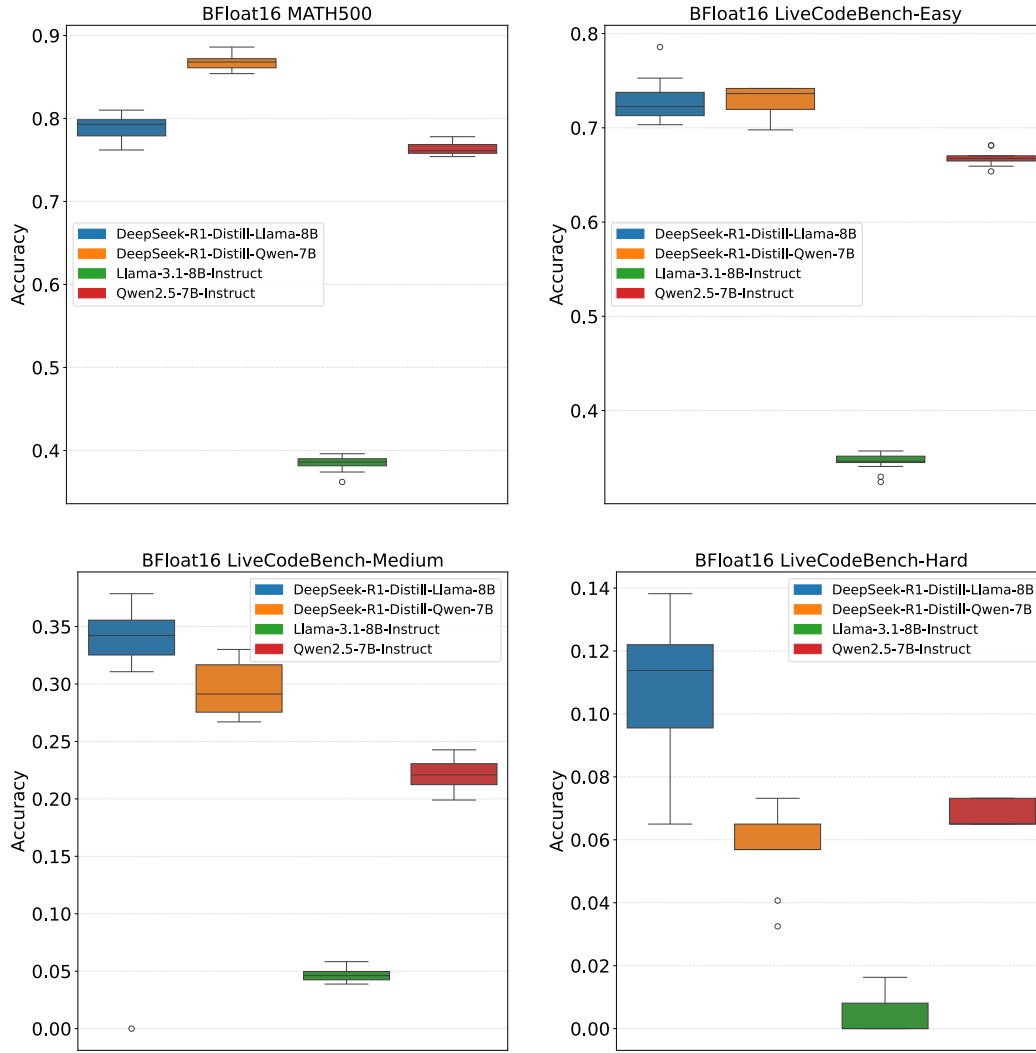


Figure 9: Accuracy varies significantly across different settings under BFloat16.

Table 6: Average Div\_Index across 12 different settings under BF16, FP16, and FP32 Numerical Precisions. In the table, a value of “-1” indicates that no divergence occurred for any example in the dataset under the given setting.

		DeepSeek-R1-Distill-Qwen-7B	DeepSeek-R1-Distill-Llama-8B	Qwen2.5-7B-Instruct	Llama-3.1-8B-Instruct
<b>AIME’24</b>	<b>BF16</b>	45.67	67.37	82.37	22.44
	<b>FP16</b>	430.13	430.47	457.41	635.43
	<b>FP32</b>	-1	13520	2000.00	1976.40
<b>MATH500</b>	<b>BF16</b>	65.33	76.85	103.58	38.07
	<b>FP16</b>	395.46	463.14	267.59	284.89
	<b>FP32</b>	1825.79	1855.17	2000.00	1509.61
<b>LCB-Easy</b>	<b>BF16</b>	35.11	47.43	46.98	53.99
	<b>FP16</b>	292.99	327.53	87.53	133.80
	<b>FP32</b>	1143.30	1036.04	-1	108.86
<b>LCB-Medium</b>	<b>BF16</b>	44.23	47.59	44.23	54.33
	<b>FP16</b>	267.81	287.83	107.07	172.73
	<b>FP32</b>	2291.83	2585.83	-1	125.00
<b>LCB-Hard</b>	<b>BF16</b>	43.85	31.57	43.64	54.95
	<b>FP16</b>	216.16	329.38	136.89	221.76
	<b>FP32</b>	5807.83	3832.54	-1	352.78

Table 7: Div\_Percent across 12 different settings under BF16, FP16, and FP32 numerical precisions

		DeepSeek-R1-Distill-Qwen-7B	DeepSeek-R1-Distill-Llama-8B	Qwen2.5-7B-Instruct	Llama-3.1-8B-Instruct
<b>AIME’24</b>	<b>BF16</b>	100%	100%	100%	53.33%
	<b>FP16</b>	100%	100%	73.33%	23.33%
	<b>FP32</b>	0	6.67%	6.67%	16.67%
<b>MATH500</b>	<b>BF16</b>	99.60%	100%	90.20%	77.60%
	<b>FP16</b>	86.00%	87.80%	37.60%	36%
	<b>FP32</b>	5.80%	6.00%	1.80%	9.20%
<b>LCB-Easy</b>	<b>BF16</b>	100%	100%	72.53%	96.15%
	<b>FP16</b>	100%	97.25%	16.48%	41.76%
	<b>FP32</b>	10.99%	13.19%	0	3.85%
<b>LCB-Medium</b>	<b>BF16</b>	100%	100%	89.32%	98.06%
	<b>FP16</b>	100%	100%	35.44%	48.06%
	<b>FP32</b>	19.90%	23.30%	0	3.40%
<b>LCB-Hard</b>	<b>BF16</b>	100%	100%	95.12%	100%
	<b>FP16</b>	100%	100%	50.41%	58.54%
	<b>FP32</b>	24.39%	30.08%	0	7.32%

## E Supplementary Results on Ablation Study

Continuing the discussion in Section 3.4, we provide more results on isolating other runtime configurations to show how they affect reproducibility. In *greedy decoding*, since FP32 precision can effectively mitigate numerical-precision-related errors, we report experimental results only under BF16 and FP16 precisions.

Tables 8 and 9 discuss the effect of different numbers of GPUs since, intuitively, the effect has a dependency on the GPU version, so we consider the situation on L40S and A100, separately. Table 8 reports the Avg\_Std@top1\_prob of LLM inference responses for the same question using 2 or 4 L40S GPUs under three different batch sizes (8, 16, and 32), while Table 9 presents the corresponding results using A100 GPUs. As shown in Table 8, increasing the number of GPUs from 2 to 4 generally leads to higher Avg\_Std@top1\_prob. **This observation suggests that increasing the number of GPUs may introduce greater inference nondeterminism.** However, this trend becomes less consistent in the A100 setting as shown in Table 9, where in many cases the 2 GPU results are even slightly higher than those of 4 GPUs. One reason behind it is that A100s do have higher instability in our experiments, which may have more influence beyond the number of GPUs (see Figure 6(c)).

Table 8: Instability (Avg\_Std@top1\_prob ( $\times 10^{-4}$ )) under different numbers of L40S GPUs

		DeepSeek-R1-Distill-Qwen-7B		DeepSeek-R1-Distill-Llama-8B		Qwen2.5-7B-Instruct		Llama-3.1-8B-Instruct	
		2GPU	4GPU	2GPU	4GPU	2GPU	4GPU	2GPU	4GPU
AIME'24	BF16	29.66	24.52	23.02	28.01	29.40	34.05	22.12	44.54
	FP16	5.83	5.19	3.18	3.95	2.94	3.91	2.89	7.30
MATH500	BF16	32.91	32.67	23.03	27.33	26.51	24.91	35.92	42.31
	FP16	4.48	4.72	3.19	3.55	3.79	3.75	4.80	5.75
LCB-Easy	BF16	42.80	46.82	28.46	34.80	28.48	30.31	28.05	33.93
	FP16	6.03	6.55	4.27	4.68	4.29	4.42	3.62	4.45
LCB-Medium	BF16	38.44	44.52	30.25	39.24	29.68	34.56	27.62	36.34
	FP16	7.04	7.68	4.52	5.00	4.84	4.74	3.45	4.05
LCB-Hard	BF16	38.63	48.74	29.76	36.10	34.31	29.60	28.90	34.60
	FP16	6.62	7.36	5.04	5.58	5.03	4.85	3.17	5.61

Table 9: Instability (Avg\_Std@top1\_prob ( $\times 10^{-4}$ )) under different numbers of A100 GPUs

		DeepSeek-R1-Distill-Qwen-7B		DeepSeek-R1-Distill-Llama-8B		Qwen2.5-7B-Instruct		Llama-3.1-8B-Instruct	
		2GPU	4GPU	2GPU	4GPU	2GPU	4GPU	2GPU	4GPU
AIME'24	BF16	32.91	33.90	31.23	25.86	30.64	41.65	53.72	35.56
	FP16	7.23	5.79	4.57	3.85	4.23	3.75	5.66	5.02
MATH500	BF16	34.59	30.88	31.69	25.63	26.59	21.88	43.03	44.33
	FP16	5.29	4.45	4.40	3.46	4.17	3.08	7.43	6.58
LCB-Easy	BF16	48.20	43.05	36.76	32.01	28.82	34.64	36.01	35.56
	FP16	7.55	6.38	6.14	4.52	4.75	4.70	5.36	4.26
LCB-Medium	BF16	44.86	44.26	42.58	37.60	31.02	35.01	39.96	33.40
	FP16	8.40	7.37	6.53	5.03	5.12	4.44	4.93	4.07
LCB-Hard	BF16	46.25	47.95	41.58	37.75	36.88	39.98	41.82	33.43
	FP16	7.95	6.79	6.63	5.41	5.11	4.76	5.29	5.88

Table 10 discuss the effect of different batch sizes. We report the Avg\_Std@top1\_prob of LLM inference responses for the same question under varying GPU counts and GPU types. From the table we can see that, **larger batch sizes consistently lead to lower Avg\_Std@top1\_prob, indicating**

**better reproducibility in model outputs.** That stands with our assumption that higher parallelism will have smaller error accumulations.

Finally, we compare the effect of the two GPU versions we used (L40S and A100). Table II reveals that under the same experimental settings, the Avg\_Std@top1\_prob on the A100 GPU is consistently slightly higher than that on the L40S GPU. **This conforms to the previous findings that, in our experiments, A100 exhibits slightly higher hardware-induced variability, which may contribute to less stable top-1 token predictions across different runtime configurations.**

Table 10: Instability (Avg\_Std@top1\_prob ( $\times 10^{-4}$ )) under different batch sizes

		DeepSeek-R1-Distill-Qwen-7B			DeepSeek-R1-Distill-Llama-8B			Qwen2.5-7B-Instruct			Llama-3.1-8B-Instruct		
		BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32
AIME'24	BF16	42.35	37.96	39.21	30.81	36.82	34.41	35.88	32.09	37.75	53.86	58.55	47.23
	FP16	6.66	6.17	6.16	4.40	4.14	4.10	4.03	4.49	3.91	3.79	4.96	3.41
MATH500	BF16	37.61	36.91	34.41	32.34	31.72	26.93	31.16	31.84	29.22	54.17	50.32	47.86
	FP16	5.26	5.23	4.79	4.28	4.19	3.68	3.87	3.96	3.57	6.69	5.59	6.14
LCB-Easy	BF16	56.00	58.99	51.39	41.93	41.56	36.80	39.14	38.27	36.21	44.63	45.88	39.60
	FP16	7.51	7.81	7.15	5.97	5.90	5.13	5.11	5.52	4.70	5.42	4.68	4.52
LCB-Medium	BF16	53.20	51.56	48.35	45.21	46.54	41.43	41.11	39.74	39.48	47.32	46.33	43.92
	FP16	8.55	8.36	8.11	6.21	6.25	5.46	5.98	5.94	5.95	4.92	4.68	4.52
LCB-Hard	BF16	57.69	59.04	54.60	47.51	47.89	41.43	49.21	47.05	46.57	46.12	44.22	39.27
	FP16	8.41	8.50	7.74	6.59	6.54	5.90	5.96	6.01	6.15	5.54	5.75	5.30

Table 11: Instability (Avg\_Std@top1\_prob ( $\times 10^{-4}$ )) under different GPU versions

		DeepSeek-R1-Distill-Qwen-7B		DeepSeek-R1-Distill-Llama-8B		Qwen2.5-7B-Instruct		Llama-3.1-8B-Instruct	
		A100	L40S	A100	L40S	A100	L40S	A100	L40S
AIME'24	BF16	42.43	37.74	33.89	31.48	39.08	34.86	55.99	42.36
	FP16	7.22	6.34	4.72	4.23	4.69	4.34	6.86	6.01
MATH500	BF16	39.12	36.11	33.94	29.32	30.96	30.19	53.43	47.06
	FP16	5.50	5.09	4.51	3.87	4.22	4.00	6.86	5.58
LCB-Easy	BF16	58.36	54.88	43.33	36.87	37.45	35.12	44.84	41.29
	FP16	7.96	7.15	6.28	5.32	5.23	4.82	5.45	4.49
LCB-Medium	BF16	54.76	48.67	47.79	43.59	38.53	37.90	46.17	42.22
	FP16	8.81	8.26	6.67	5.65	5.54	5.59	5.08	4.62
LCB-Hard	BF16	60.11	54.66	49.39	42.40	47.11	44.72	45.59	40.64
	FP16	8.38	8.10	7.04	6.18	5.91	5.91	5.59	5.13

## F Supplementary Results on LayerCast

We evaluate the LayerCast method proposed in Section 4 on the DeepSeek-R1-Distill-Qwen-7B model across five benchmarks: AIME'24, MATH500, LiveCodeBench-Easy, LiveCodeBench-Medium, and LiveCodeBench-Hard. These evaluations span different batch sizes (8, 16, 32), GPU counts (2 and 4), and numeric precision options (BF16, FP32, and LayerCast). All experiments for LayerCast have the same settings as the main experiments.

Table 12 reports both the accuracy and its standard deviation across different GPU count and batch size configurations. To facilitate comparison, results under FP32 and BF16 are also included. As expected, BF16 exhibits the most instability, with significantly higher standard deviations. LayerCast matches FP32 in terms of stability (often with zero or near-zero std) while preserving memory

484 efficiency. These results corroborate our main findings in Section 4, demonstrating that LayerCast  
485 delivers reproducibility on par with FP32 while operating within a lower memory footprint.

Table 12: Accuracy and standard deviation of DeepSeek-R1-Distill-Qwen-7B under different GPU counts, precisions, and batch sizes across 5 benchmarks.

Benchmark	2x A100			4x A100			Std@Acc
	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	
<b>AIME'24</b>							
LayerCast	0.4333	0.4333	0.4333	0.4333	0.4333	0.4333	<b>0</b>
FP32	0.4333	0.4333	0.4333	0.4333	0.4333	0.4333	<b>0</b>
BF16	0.4667	0.4667	0.3667	0.4333	0.5333	0.4667	<b>0.0544</b>
<b>MATH500</b>							
LayerCast	0.8680	0.8680	0.8680	0.8660	0.8660	0.8660	<b>0.0011</b>
FP32	0.8680	0.8680	0.8680	0.8680	0.8660	0.8700	<b>0.0013</b>
BF16	0.8700	0.8620	0.8580	0.8540	0.8860	0.8700	<b>0.0114</b>
<b>LCB-Easy</b>							
LayerCast	0.7308	0.7308	0.7363	0.7363	0.7418	0.7418	<b>0.0049</b>
FP32	0.7363	0.7363	0.7363	0.7308	0.7363	0.7363	<b>0.0022</b>
BF16	0.7198	0.7418	0.7253	0.7418	0.6978	0.7363	<b>0.0169</b>
<b>LCB-Medium</b>							
LayerCast	0.3058	0.3010	0.2961	0.2961	0.3010	0.3058	<b>0.0043</b>
FP32	0.3058	0.3010	0.3058	0.3058	0.3058	0.3010	<b>0.0025</b>
BF16	0.2816	0.2767	0.3204	0.2961	0.2864	0.2718	<b>0.0176</b>
<b>LCB-Hard</b>							
LayerCast	0.0488	0.0488	0.0569	0.0569	0.0488	0.0569	<b>0.0044</b>
FP32	0.0488	0.0488	0.0488	0.0569	0.0569	0.0569	<b>0.0044</b>
BF16	0.0569	0.0732	0.0569	0.0650	0.0569	0.0650	<b>0.0066</b>

## 486 G Supplementary Results on Random Sampling

487 In this section, we show more results related to the discussion in Section 3.3. In the random sampling  
488 setting, we report the complete set of Pass@1 results in AIME'24 and MATH500 to further evaluate  
489 reproducibility under different precision settings. These experiments are conducted across different  
490 batch sizes (8, 16, 32), and GPU counts (2 and 4). The standard deviations are calculated along  
491 numeric precision types (BF16, FP16, FP32).

492 Tables 13, 14, 15, and 16 summarize the Pass@1 accuracies and their standard deviation on DeepSeek-  
493 R1-Distill-Qwen-7B, Qwen2.5-7B-Instruct, DeepSeek-R1-Distill-Llama-8B, Llama-3.1-8B-Instruct,  
494 respectively. In most cases, FP32 shows the most stable variance (with the lowest standard deviations).  
495 However, it is worth noticing that there are (3 out of 8) cases where FP16 results are more stable.  
496 This suggests that during random sampling, the two sources of randomness can interleave. We still  
497 urge researchers to sample more extensively to obtain more reproducible results.

Table 13: Pass@1 accuracies (%) and their standard deviations, on DeepSeek-R1-Distill-Qwen-7B.

Dtype	AIME'24						Stddev	MATH500						Stddev
	2 A100			4 A100				2 A100			4 A100			
	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32		BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	
FP32	53.33	56.25	55.21	53.13	54.17	54.17	1.1784	90.60	90.80	90.80	90.75	90.90	90.70	0.1021
FP16	52.71	55.00	52.92	53.33	53.13	53.13	0.8273	90.25	90.20	90.20	90.45	90.50	90.15	0.1463
BF16	53.33	56.88	55.42	56.67	53.54	53.13	1.7151	90.65	91.15	90.40	90.95	90.85	90.35	0.3158

Table 14: Pass@1 accuracies (%) and their standard deviations, on Qwen2.5-7B-Instruct

Dtype	AIME'24						Stdev	MATH500						Stdev
	2 A100			4 A100				2 A100			4 A100			
	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32		BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	
FP32	11.46	11.46	11.46	11.46	11.46	11.46	0	74.85	74.85	74.85	74.80	74.80	74.80	0.0274
FP16	11.04	11.67	11.46	11.25	11.67	11.25	0.2523	74.50	74.90	74.80	75.00	74.80	74.75	0.1686
BF16	11.25	10.21	11.04	9.38	11.04	10.42	0.7056	74.05	74.00	74.60	74.15	74.80	74.15	0.4663

Table 15: Pass@1 accuracies (%) and their standard deviations, on DeepSeek-R1-Distill-Llama-8B.

Dtype	AIME'24							Stdev	MATH500							Stdev
	2 A100			4 A100			2 A100			4 A100						
	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	BS=8		BS=16	BS=32	BS=8	BS=16	BS=32			
FP32	43.33	43.75	41.88	42.92	43.96	42.08	0.8606	83.85	83.95	83.75	83.65	83.95	83.75	0.1211		
FP16	43.96	45.63	43.75	41.04	40.83	42.08	1.8792	84.55	83.95	83.95	84.15	84.25	83.55	0.3371		
BF16	46.46	45.42	42.71	43.13	43.54	43.13	1.5124	84.50	84.80	84.35	84.60	85.20	85.20	0.3602		

Table 16: Pass@1 accuracies (%) and their standard deviations, on Llama-3.1-8B-Instruct.

Dtype	AIME'24						Stdev	MATH500						Stdev
	2 A100			4 A100				2 A100			4 A100			
	BS=8	BS=16	BS=32	BS=8	BS=16	BS=32		BS=8	BS=16	BS=32	BS=8	BS=16	BS=32	
FP32	5.00	5.42	3.75	5.00	5.42	3.75	0.7759	37.00	36.50	36.30	36.80	36.30	36.15	0.3293
FP16	5.00	5.00	4.58	5.21	5.00	5.21	0.2282	37.10	36.60	37.00	36.90	36.95	37.00	0.1725
BF16	3.96	4.38	4.79	5.42	5.21	5.42	0.5992	36.25	36.95	36.65	35.35	35.90	36.75	0.6020