

DeepSeek-V3 Technical Report

DeepSeek-AI

research@deepseek.com

Abstract

We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. We pre-train DeepSeek-V3 on 14.8 trillion diverse and high-quality tokens, followed by Supervised Fine-Tuning and Reinforcement Learning stages to fully harness its capabilities. Comprehensive evaluations reveal that DeepSeek-V3 outperforms other open-source models and achieves performance comparable to leading closed-source models. Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for its full training. In addition, its training process is remarkably stable. Throughout the entire training process, we did not experience any irrecoverable loss spikes or perform any rollbacks. The model checkpoints are available at <https://github.com/deepseek-ai/DeepSeek-V3>.

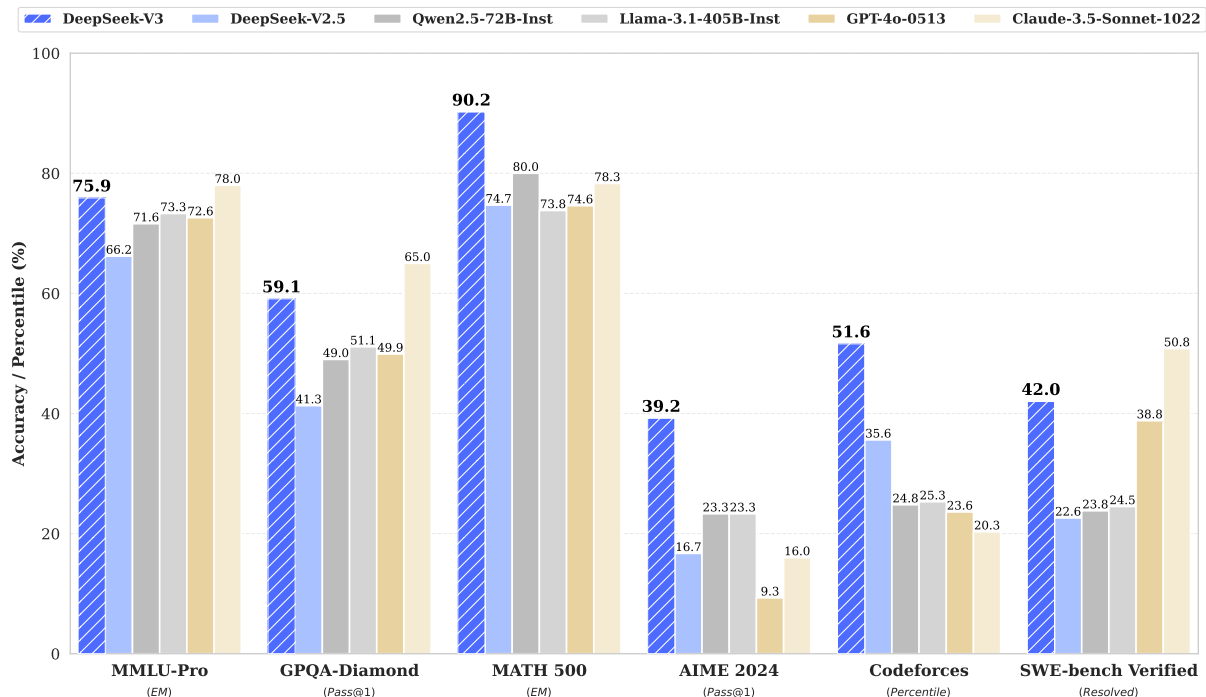


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

Contents

1	Introduction	4
2	Architecture	6
2.1	Basic Architecture	6
2.1.1	Multi-Head Latent Attention	7
2.1.2	DeepSeekMoE with Auxiliary-Loss-Free Load Balancing	8
2.2	Multi-Token Prediction	10
3	Infrastructures	11
3.1	Compute Clusters	11
3.2	Training Framework	12
3.2.1	DualPipe and Computation-Communication Overlap	12
3.2.2	Efficient Implementation of Cross-Node All-to-All Communication	13
3.2.3	Extremely Memory Saving with Minimal Overhead	14
3.3	FP8 Training	14
3.3.1	Mixed Precision Framework	15
3.3.2	Improved Precision from Quantization and Multiplication	16
3.3.3	Low-Precision Storage and Communication	18
3.4	Inference and Deployment	18
3.4.1	Prefilling	19
3.4.2	Decoding	19
3.5	Suggestions on Hardware Design	20
3.5.1	Communication Hardware	20
3.5.2	Compute Hardware	20
4	Pre-Training	21
4.1	Data Construction	21
4.2	Hyper-Parameters	22
4.3	Long Context Extension	23
4.4	Evaluations	24
4.4.1	Evaluation Benchmarks	24
4.4.2	Evaluation Results	24
4.5	Discussion	26
4.5.1	Ablation Studies for Multi-Token Prediction	26
4.5.2	Ablation Studies for the Auxiliary-Loss-Free Balancing Strategy	26

4.5.3	Batch-Wise Load Balance VS. Sequence-Wise Load Balance	27
5	Post-Training	28
5.1	Supervised Fine-Tuning	28
5.2	Reinforcement Learning	29
5.2.1	Reward Model	29
5.2.2	Group Relative Policy Optimization	30
5.3	Evaluations	30
5.3.1	Evaluation Settings	30
5.3.2	Standard Evaluation	31
5.3.3	Open-Ended Evaluation	33
5.3.4	DeepSeek-V3 as a Generative Reward Model	33
5.4	Discussion	34
5.4.1	Distillation from DeepSeek-R1	34
5.4.2	Self-Rewarding	34
5.4.3	Multi-Token Prediction Evaluation	35
6	Conclusion, Limitations, and Future Directions	35
A	Contributions and Acknowledgments	45
B	Ablation Studies for Low-Precision Training	47
B.1	FP8 v.s. BF16 Training	47
B.2	Discussion About Block-Wise Quantization	47
C	Expert Specialization Patterns of the 16B Aux-Loss-Based and Aux-Loss-Free Models	48