

DEEPSEEK-V3.2: PUSHING THE FRONTIER OF OPEN LARGE LANGUAGE MODELS



AUTHORS: DEEPSEEK-AI

Motivation and Research Problem

Bridging Efficiency & Reasoning Gap

DeepSeek-V3.2 aims to bridge the gap between high computational efficiency and superior reasoning performance in AI.



Long-Sequence Processing Inefficiencies

Tackles inefficiencies in long-sequence processing from vanilla attention, constraining scalability and post-training effectiveness.



Computational Resource Constraints

Addresses limited resources hindering open-source models compared to better-funded closed-source counterparts.



Generalization & Instruction Deficits

Generalization and instruction-following deficits limit open-source models' real-world applicability.



Performance Trajectory Decline

Competitive disadvantage highlighted by the decline in performance trajectory vs. proprietary models.



Integration & Knowledge Limitations

Challenges with advanced features and knowledge limitations require better knowledge representation and retrieval. DeepSeek-V3.2 focuses on efficiency and reasoning to mitigate these issues.



Framework Overview: DeepSeek-V3.2



DeepSeek Sparse Attention (DSA)

Reduces computational complexity, maintains long-sequence performance. Uses lightning indexer and fine-grained token selection.



Lightning Indexer

Computes relevance between tokens, optimizing retention with top-k selections.



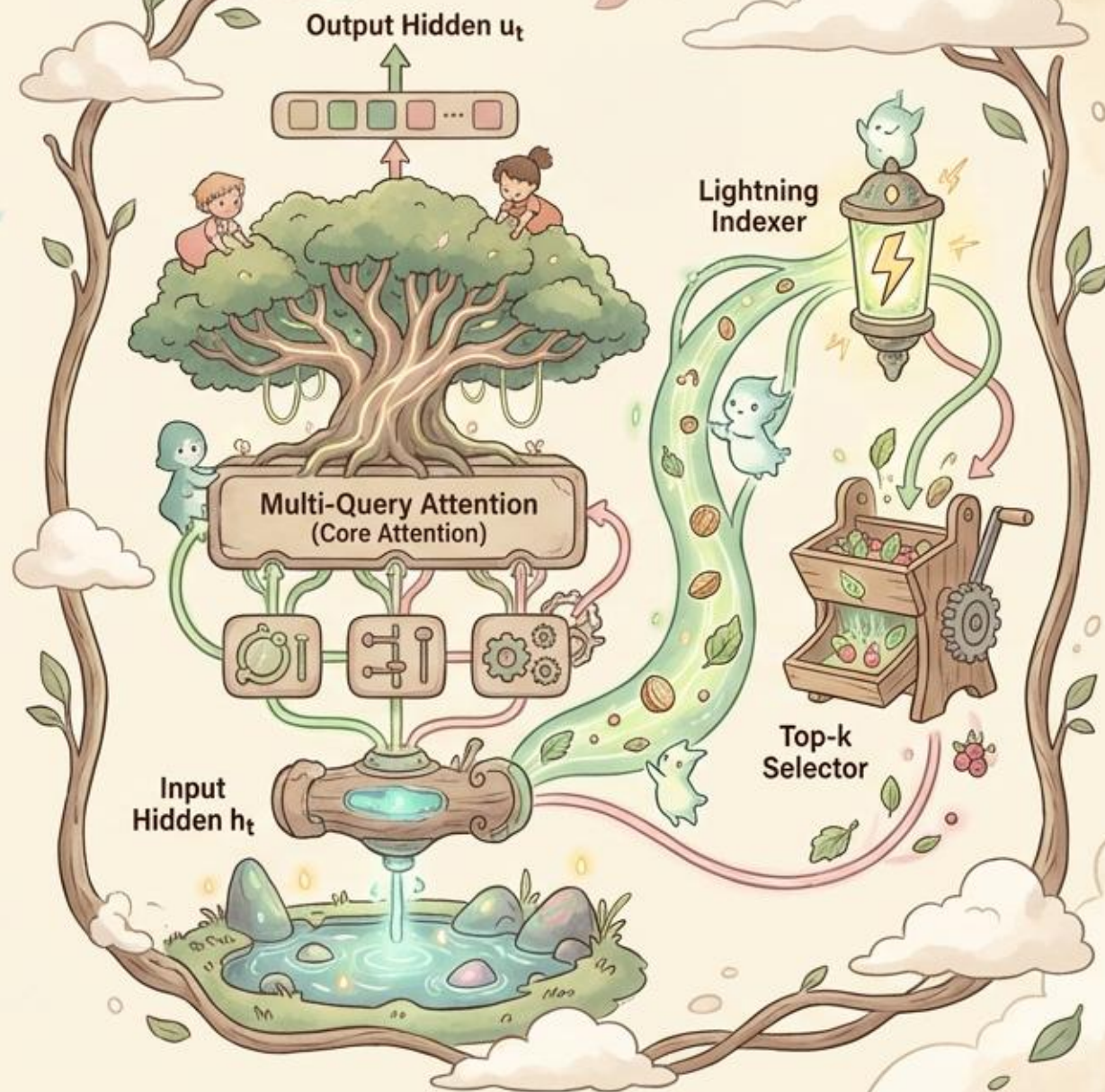
Scalable Reinforcement Learning

Enhances post-training performance with additional resources & integrated training strategies for efficiency.



Large-Scale Task Synthesis

Generates diverse environments for better task generalization. Cognitive integration and context management enable advanced reasoning and long interaction history.



Mathematical Foundations

The DeepSeek-V3.2 framework is grounded in advanced mathematical formulations that enhance its capability in long-sequence processing.

Sparse Attention Output

$$\mathbf{u}_t = \text{Attn}(\mathbf{h}_t, \{\mathbf{c}_s \mid I_{t,s} \in \text{Top-}k(I_{t,:})\})$$



These formulations underpin the DSA mechanism, ensuring reduced computational load while preserving essential information.

Index Score Calculation

$$I_{t,s} = \sum_{j=1}^{H^I} w_{t,j}^I \cdot \text{ReLU}(\mathbf{q}_{t,j}^I \cdot \mathbf{k}_s^I)$$



Experimental Results and Performance

DeepSeek-V3.2 demonstrates significant improvements in problem-solving tasks and performance metrics across various benchmarks. Evaluated against closed and open-source models, DeepSeek-V3.2 offers competitive performance in reasoning capabilities in math and coding tasks, achieving tasks, achieving scores of 96.0 in AIME 2025 (Pass@1), 99.2 in HMMT Feb 2025 (Pass@1), and a remarkable improvement in tool-assisted scenarios.

The scalable RL framework shows efficacy with synthetic data, as depicted in Figure 5.

Performance in coding and mathematical competitions, such as ICPC and IOI 2025, further underline the model's advanced capabilities. These metrics showcase DeepSeek-V3.2's ability to narrow the gap with proprietary models in multiple domains.

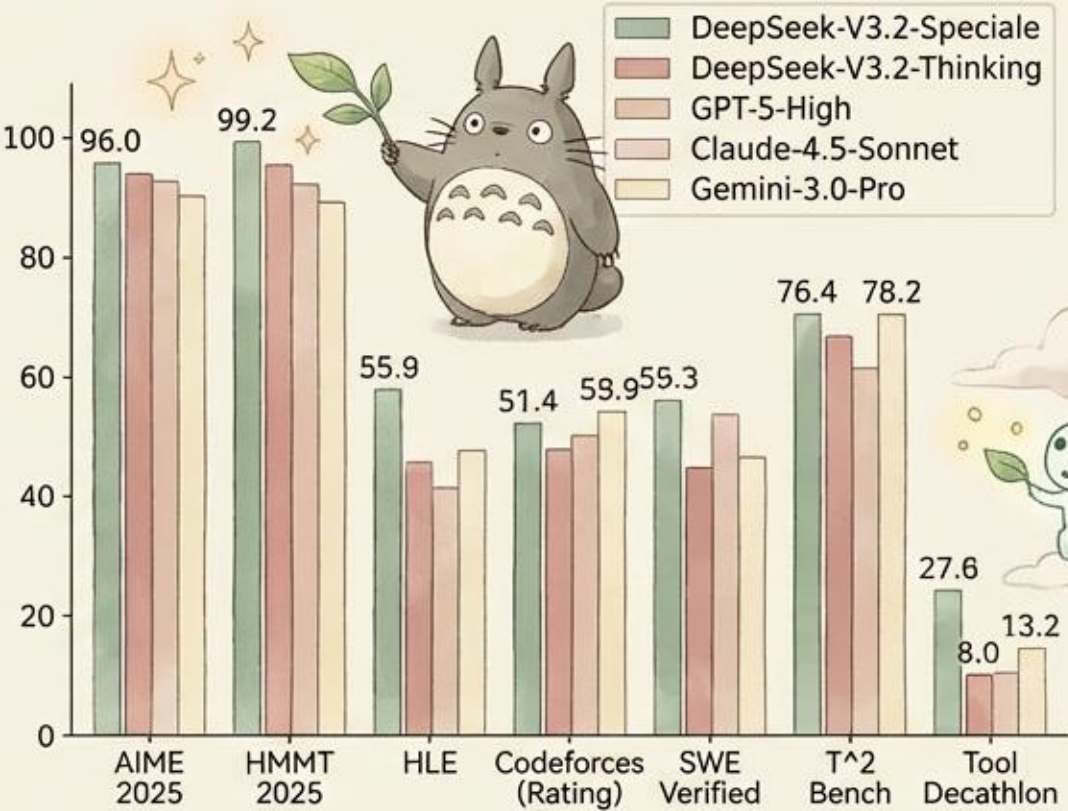


Figure 1. Compared with bar di...proem.veriving, ununreadng non+entivst; and eniðoiedbimt.

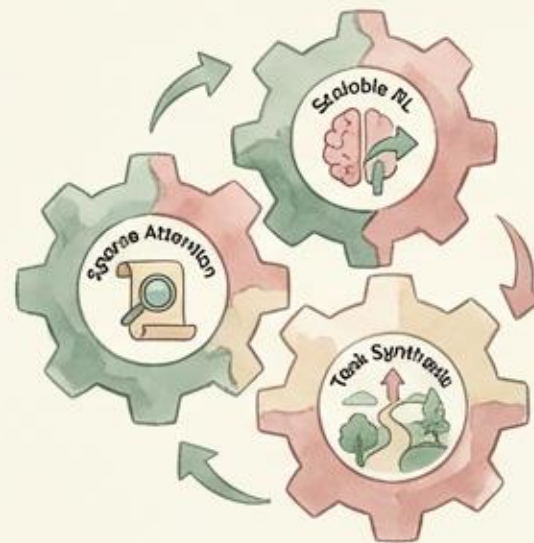
Table 2 (focus: performance comparison):

Benchmark	Ours	Baseline
AIME 2025	96.0	95.0

Conclusion

New Paradigm in Open-Source AI

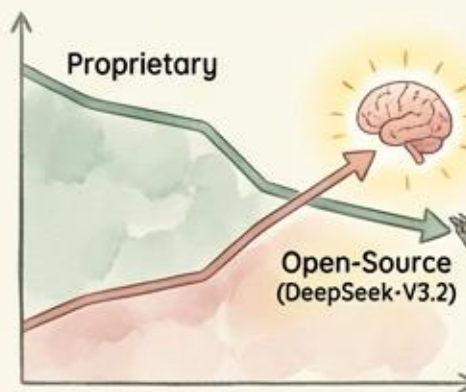
DeepSeek-V3.2 introduces a new paradigm, demonstrating significant computational efficiency and reasoned task execution.



Key Contributions (DSA, Scalable RL, Task Synthesis)

Integrates DeepSeek Sparse Attention, scalable RL frameworks, and large-scale task synthesis to collectively bolster performance.

Narrowed Performance Gaps & Enhanced Reasoning



Significantly narrows performance gaps with proprietary systems and enhances reasoning in diverse fields.

Establishing a New Standard

Sets a new standard for open-source LLMs, overcoming previous limitations with superior metrics in top-tier mathematical and coding competitions.

