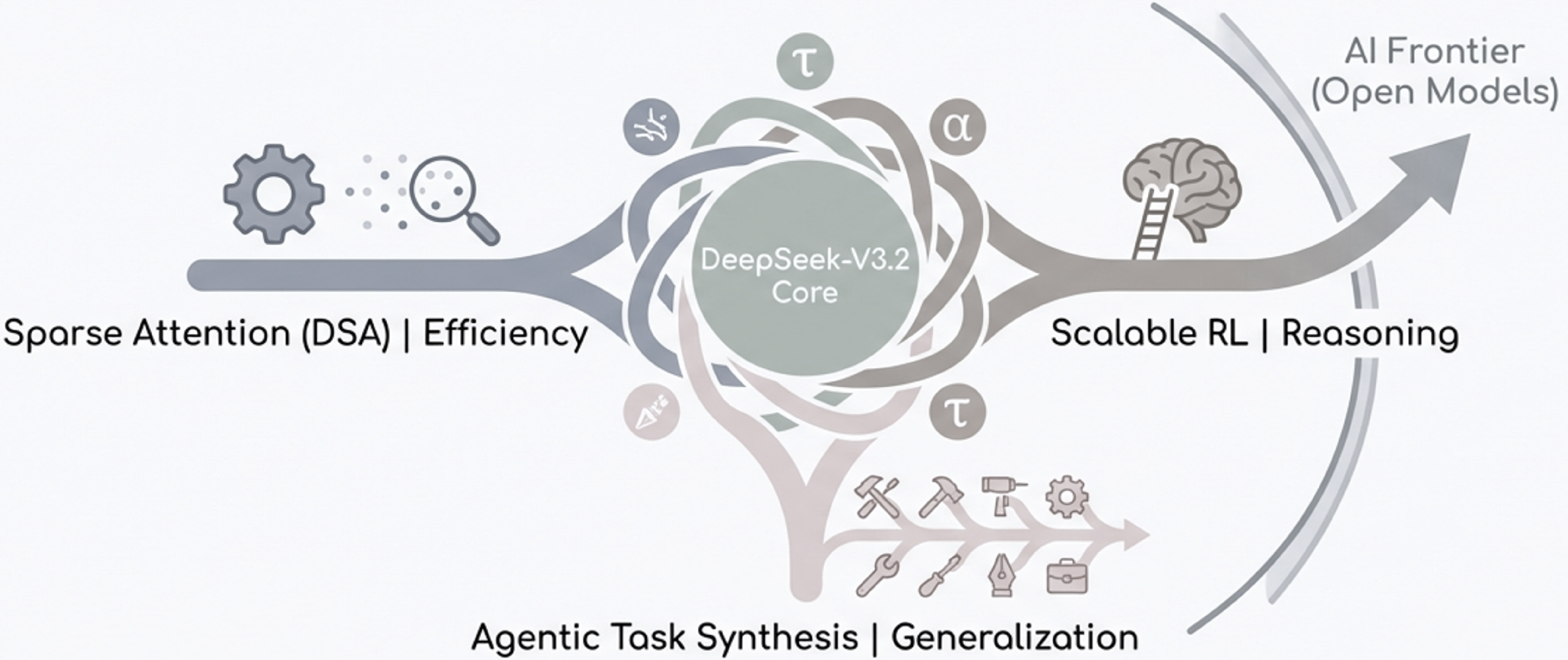
DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models



AI Frontier (Open Models)

DeepSeek-V3.2 Core

Sparse Attention (DSA) | Efficiency

Scalable RL | Reasoning

Agentic Task Synthesis | Generalization

Authors: DeepSeek-AI

Background and Problem Statement

Efficiency vs. Reasoning

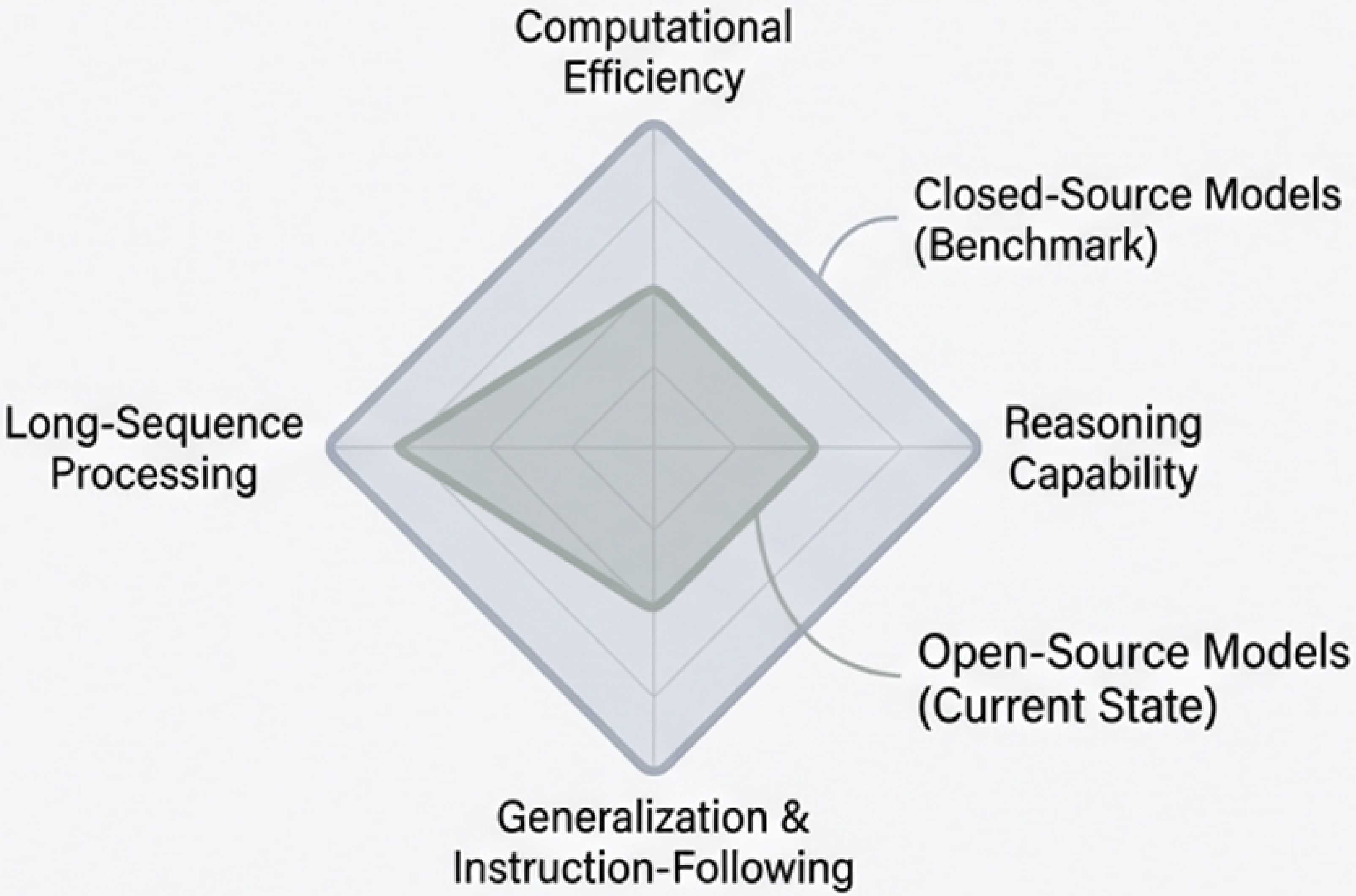
The primary research problem of DeepSeek-V3.2 is to address the gap between high computational efficiency and superior reasoning and agentic performance in AI models.

Existing Limitations

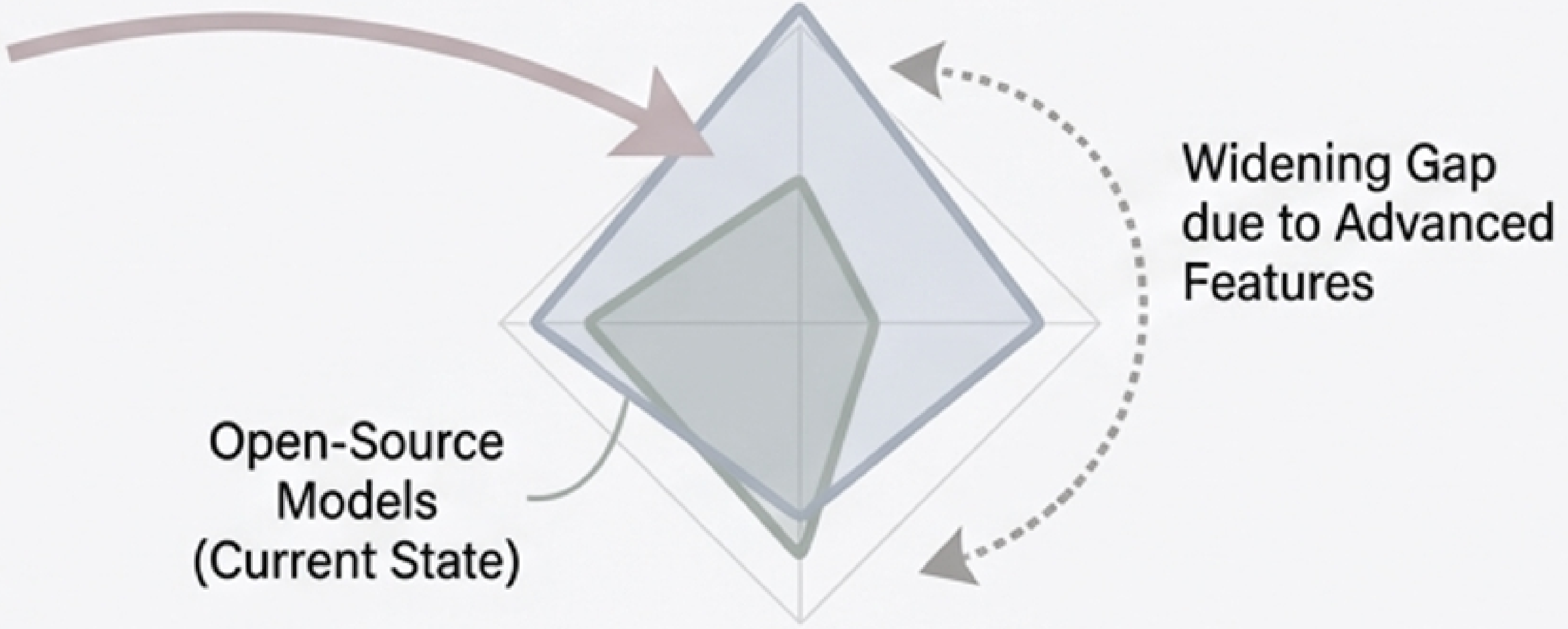
Existing methods struggle with long-sequence processing due to their reliance on vanilla attention mechanisms, resulting in inefficiencies and scalability issues.

Open-Source Gap

Open-source models suffer from insufficient computational resources and demonstrate poor generalization and instruction-following capabilities, creating a performance gap compared to closed-source models.


Computational Efficiency
Closed-Source Models (Benchmark)
Long-Sequence Processing
Reasoning Capability
Open-Source Models (Current State)
Generalization & Instruction-Following

Performance Gap and Current Challenges


Widening Gap due to Advanced Features
Open-Source Models (Current State)

→ This gap is further widened by a declining performance trajectory and integration challenges with advanced features.

Methodology: Framework Overview

DeepSeek Sparse Attention (DSA):

Reduces computational complexity while maintaining performance for long sequences. The lightning indexer computes relevance scores.

Scalable RL Framework:

Expands computational resources during post-training, enhancing generalization capabilities.

Agentic Task Synthesis Pipeline:
Enhances instruction-following by integrated reasoning and tool-use systematically.

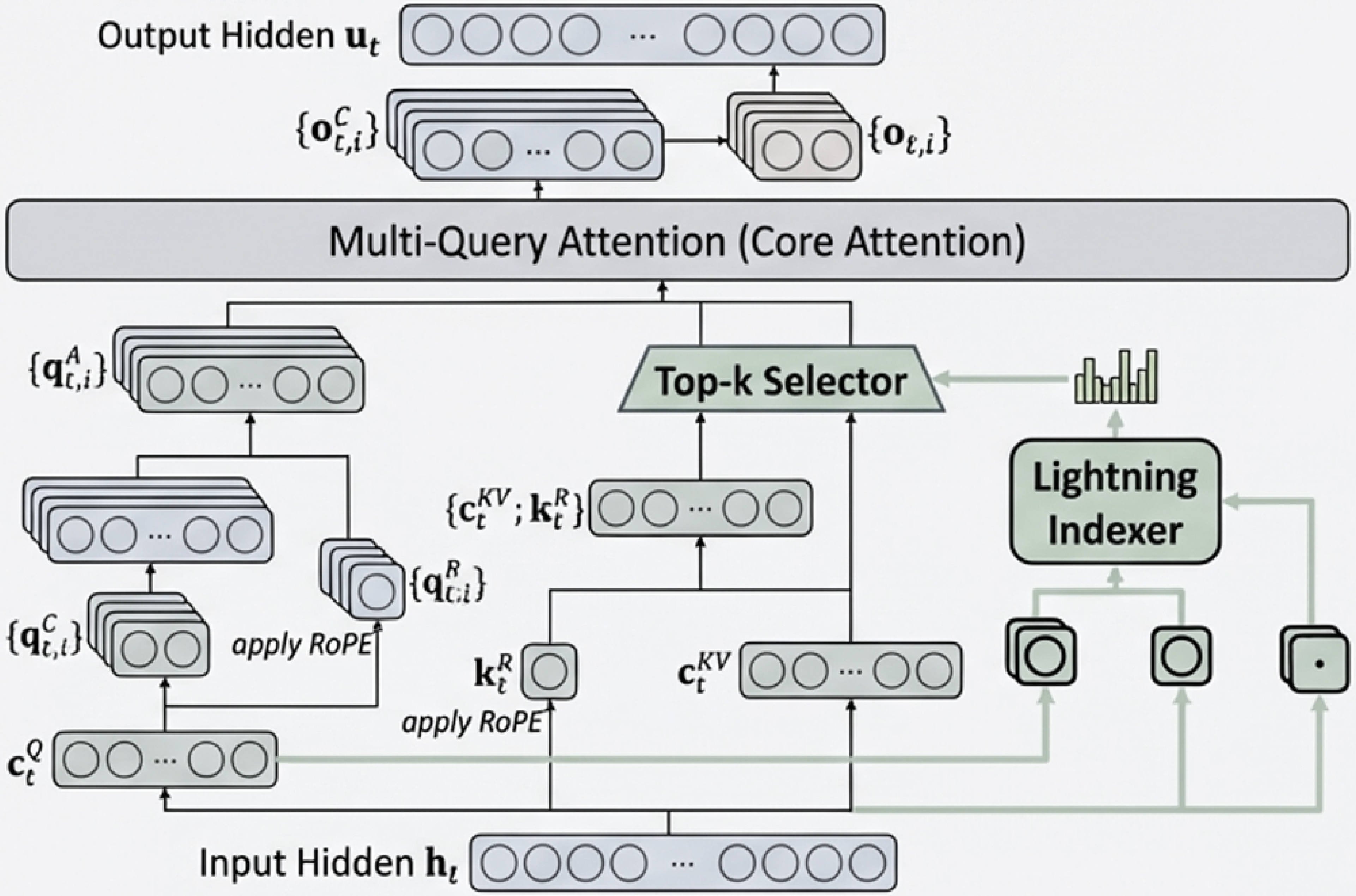

Output Hidden $\mathbf{u}_t$
$\{\mathbf{o}_{t,i}^{C}\}$
$\{\mathbf{o}_{t,i}\}$
Multi-Query Attention (Core Attention)
$\{\mathbf{q}_{t,i}^{A}\}$
Top-k Selector
Lightning Indexer
$\{\mathbf{c}_t^{KV};\mathbf{k}_t^{R}\}$
$\{\mathbf{q}_{t,i}^{R}\}$
$\{\mathbf{q}_{t,i}^{C}\}$
apply RoPE
$\mathbf{k}_t^{R}$
$\mathbf{c}_t^{KV}$
apply RoPE
$\mathbf{c}_t^{Q}$
Input Hidden $\mathbf{h}_t$

Figure 2: Attention architecture of DeepSeek-V3.2, where DSA is instantiated under MLA. The green part illustrates how DSA selects the top-k key-value entries according to the indexer.

Mathematical Formulations and Key Components

DeepSeek-V3.2 employs several mathematical formulations to illustrate its attention mechanisms. One key equation, the Sparse Attention Output Equation, is:

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

Here, $\mathbf{u}_t$ represents the attention output vector for the t^{th} query token, $\mathbf{h}_t$ is the query token embedding, and $\mathbf{c}_s$ the context vectors for the top k tokens. Additionally, an index score $I_{t,s}$ is calculated by:

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

These equations are crucial for processing complex reasoning tasks efficiently.

Experimental Results

DeepSeek-V3.2 was evaluated across various benchmarks and achieved competitive performance compared to both open and closed-source models. It was tested on tasks including code and mathematical competitions, showing strong results in metrics such as Pass@1 and rating benchmarks. For instance, in LiveCodeBench and Codeforces, DeepSeek-V3.2 achieved Pass@1 rates and competitive ratings compared to proprietary models, demonstrating its robust reasoning and agentic task capabilities.

Table 2: DeepSeek-V3.2 Performance Metrics

Benchmark (Metric)	Pass@1 / Rating
LiveCodeBench	83.3
Codeforces	2386

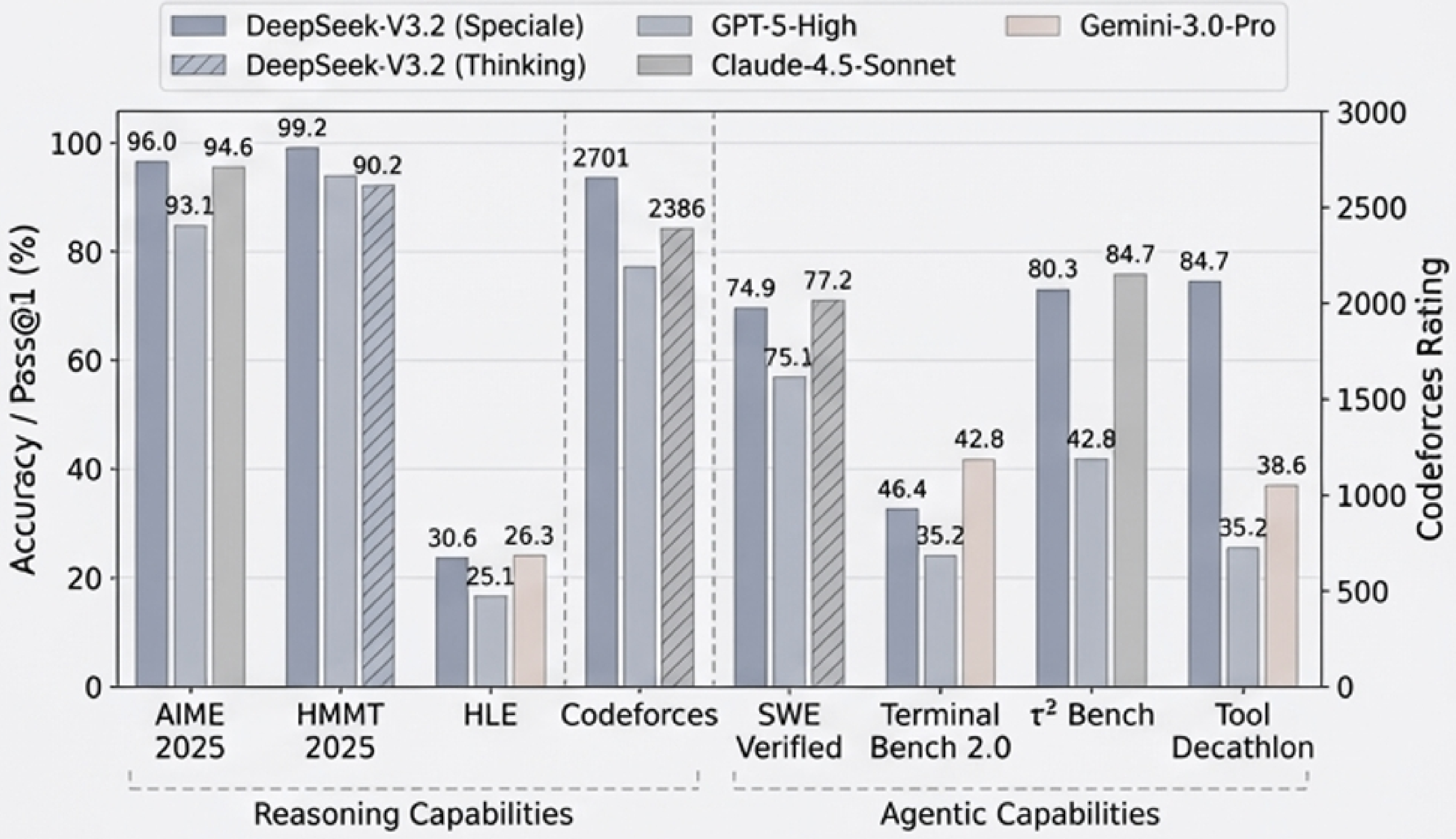


Figure 1: Benchmark Comparison

Conclusion and Key Contributions

Efficient DSA Mechanism

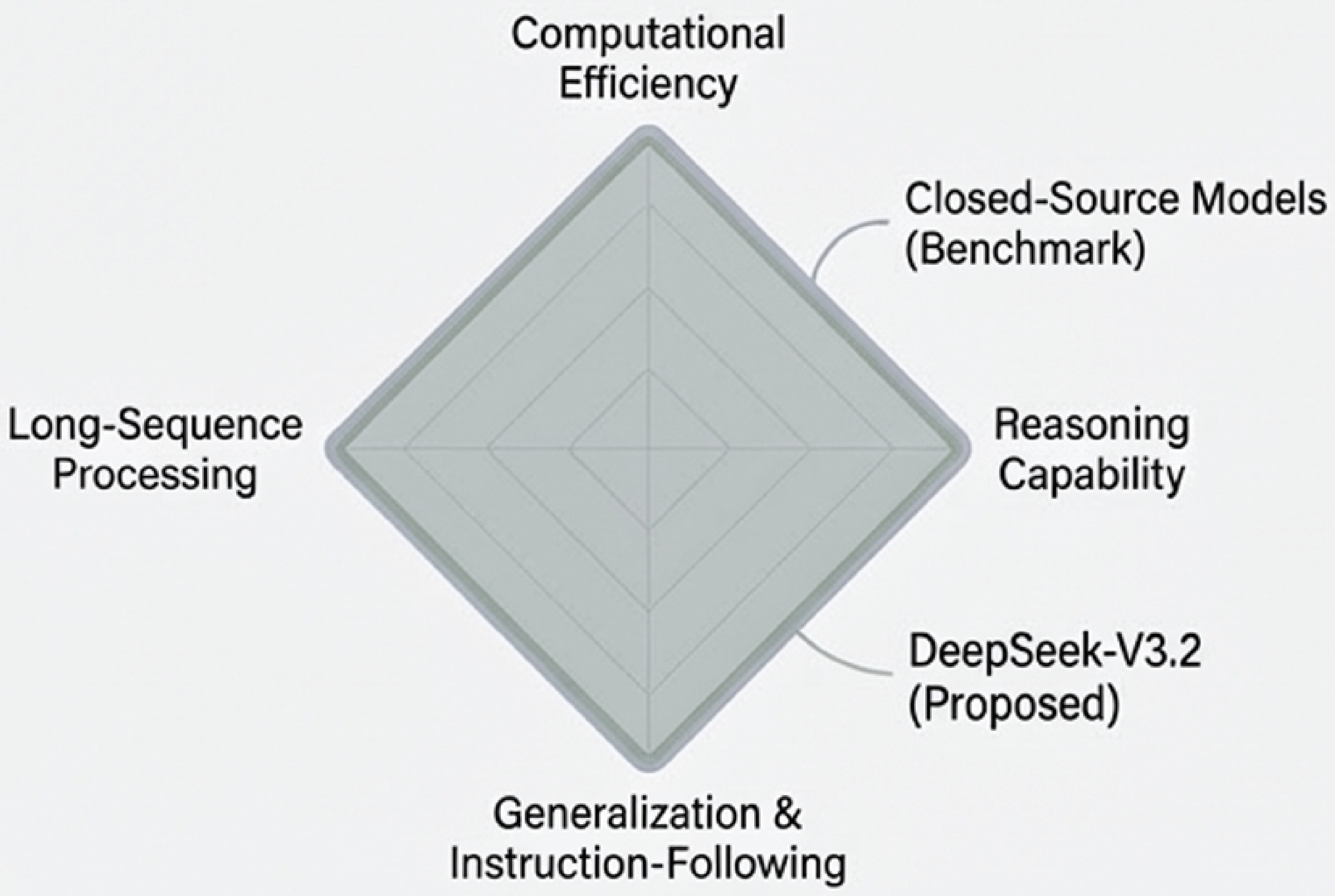
- Innovations in attention mechanisms, reducing complexity.

Robust RL Framework

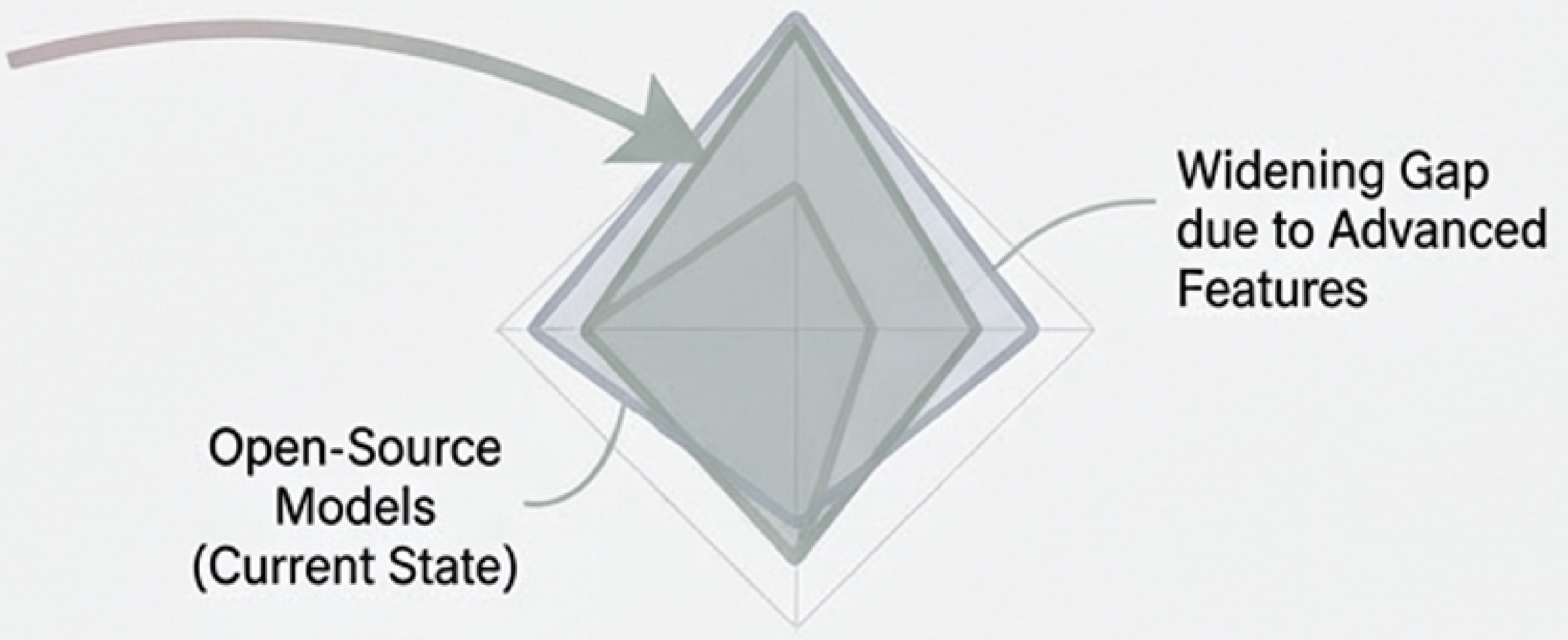
- Facilitating powerful computing resources and scalability.

Large-Scale Task Synthesis Pipeline

- Improving generalization across agentic tasks.



Performance Gap Bridged & Enhanced Capabilities


Widening Gap due to Advanced Features
Open-Source Models (Current State)

→ Effective bridging of gaps; positioning DeepSeek-V3.2 as a competitive and efficient open-source model.