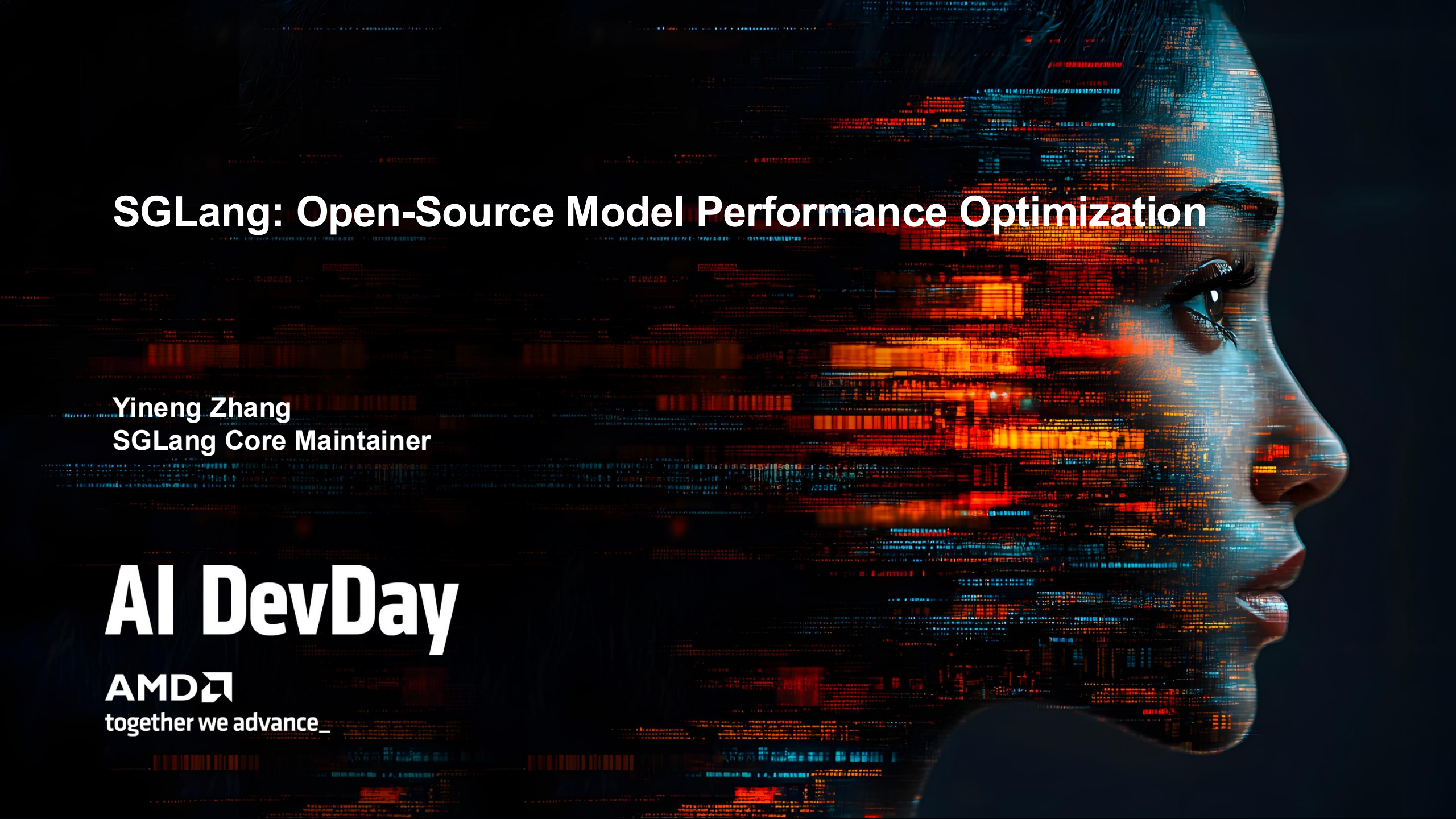
Yineng Zhang
SGLang Core Maintainer

- GitHub: https://github.com/zhyncs
- X: https://x.com/zhyncs42
- LinkedIn: https://www.linkedin.com/in/zhyncs



SGLang: Open-Source Model Performance Optimization

Yineng Zhang
SGLang Core Maintainer

AI DevDay

AMD
together we advance_

Outline

SGLang is a fast serving framework for large language models and vision language models.

- Review 2025 Highlights
- Outlook for the Rest of 2025
- Feedback (Q & A)

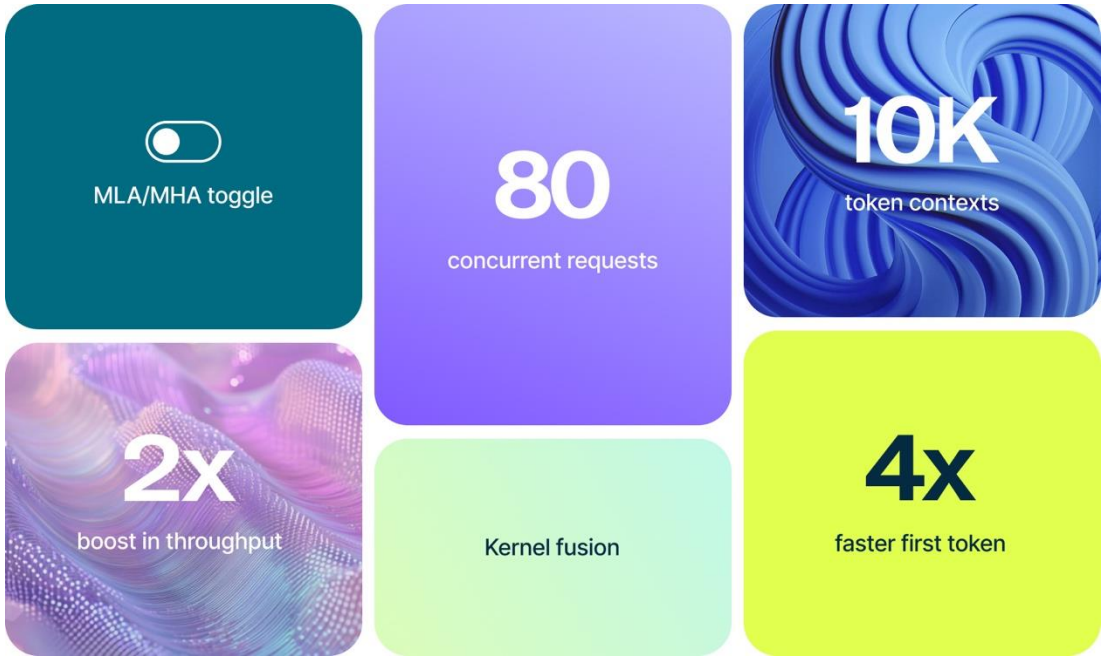
Links:

- https://github.com/sgl-project/sglang/issues/4042
- https://github.com/sgl-project/sglang/issues/7736
- https://lmsys.org/blog

Review 2025 Highlights

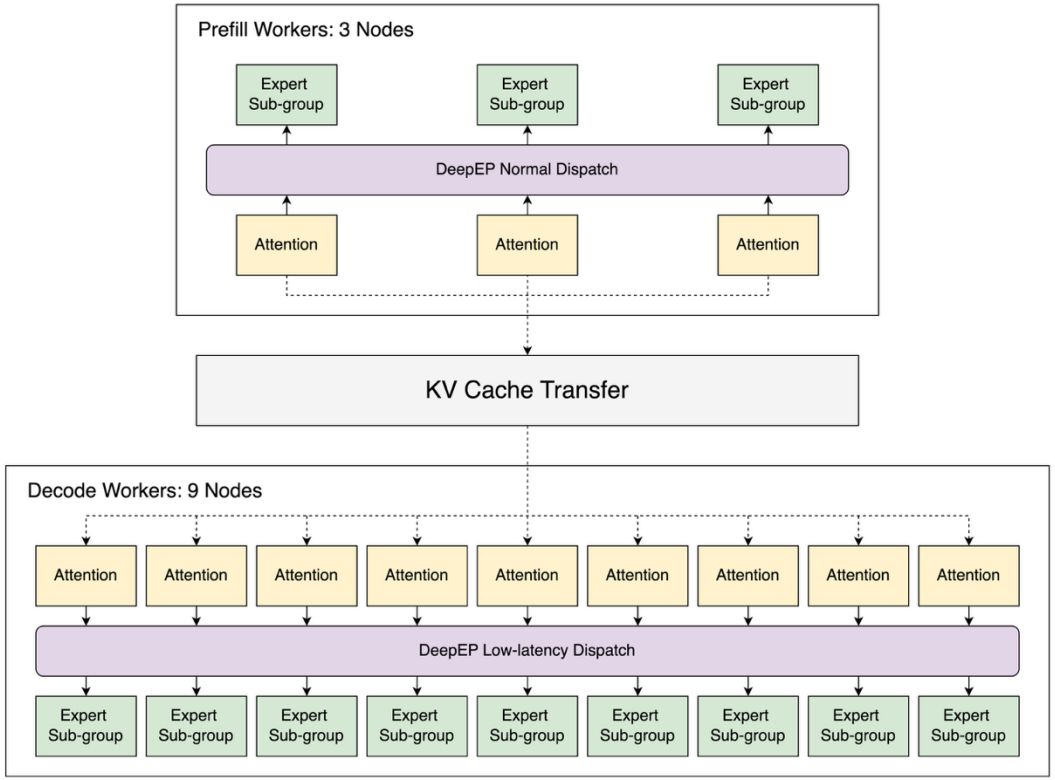
- DeepSeek V3 Optimization
- Large-Scale Deployment
- Reinforcement Learning Integration
- Speculative Decoding Training Acceleration
- Hierarchical KV Caching Integration
- Deterministic Inference
- New models day-0 support
- Model Deployment Orchestration
- Distributed Inference on AMD

2025 Highlight: DeepSeek V3 Optimization



- FlashAttention-3
- Dynamic MLA-to-MHA switching
- DeepGEMM (FP8 matrix multiplication)
- Kernel fusion
- Data-parallel attention computation

2025 Highlight: Large-Scale Deployment

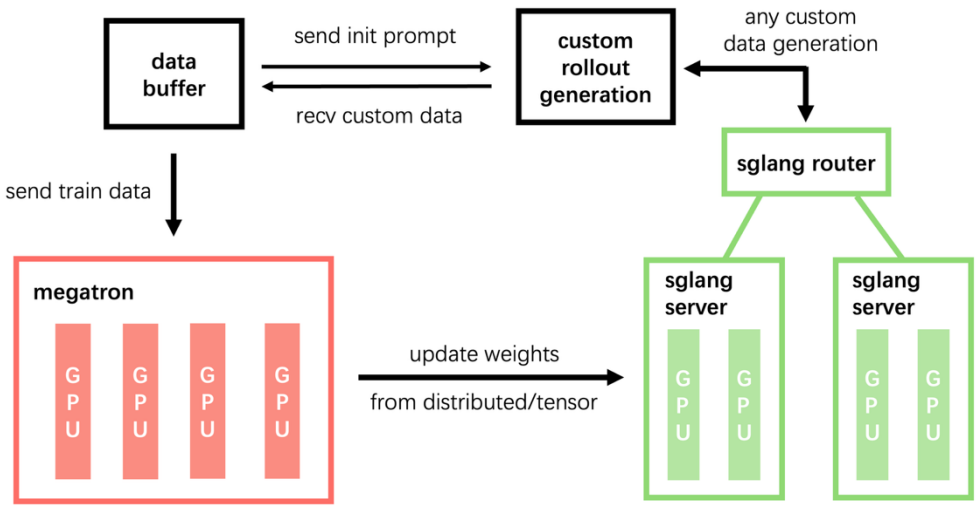


Performance at (May. 2025)

- 52.3k input token/s/node
- 22.3k output token/s/node
- 5x cheaper than DeepSeek API price

Reproduced by 10+ other teams

2025 Highlight: Reinforcement Learning Integration



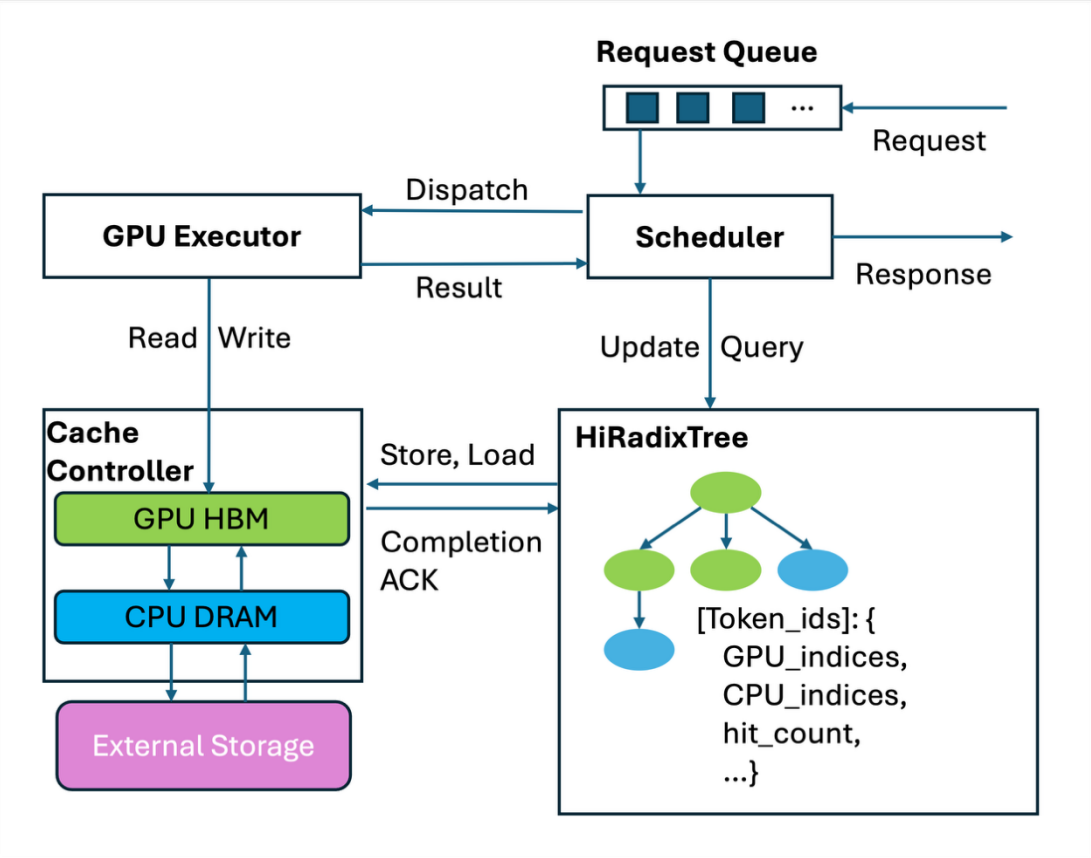
- SGLang-exclusive reinforcement learning framework: slime
  - Architecture: Built on Megatron-LM and SGLang for extreme scalability
  - Use case: Powers large-scale training for GLM 4.5 and GLM 4.6
- SGLang other general integration: veRL, AReaL

2025 Highlight: Speculative Decoding Training Acceleration

Train Eagle 3 for SGLang with SpecForge

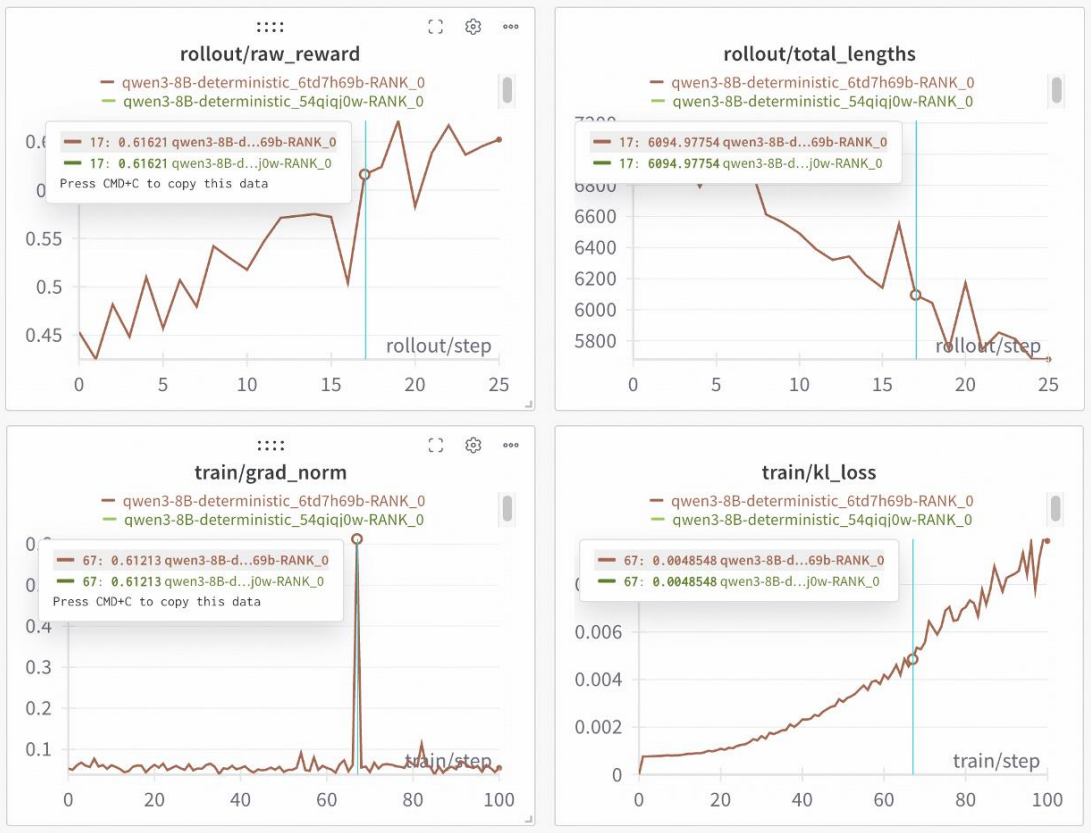
- Native Support for Advanced Architectures
- Scalable Distributed Training
- Memory-Efficient Training
- Collaborate with the official EAGLE team

2025 Highlight: Hierarchical KV Caching Integration



- Up to 6× throughput and 84% lower TTFT via hierarchical KV reuse
- HiRadixTree + GPU-assisted I/O → efficient cross-tier caching (GPU / CPU / storage)
- Smart prefetch & write policies hide transfer latency, boost cache hit rate
- Pluggable backends: Mooncake, 3FS — easy get/exist/set integration
- Adopted by Ant Group, Novita AI, Alibaba Cloud and others

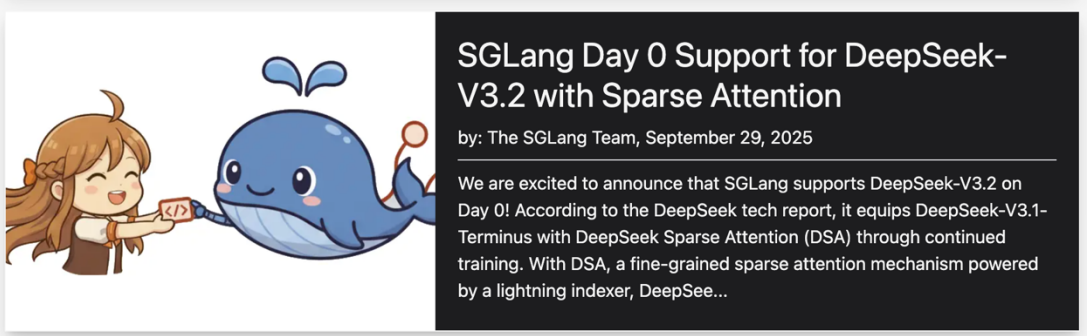
2025 Highlight: Deterministic Inference



- SGLang enables fully deterministic inference with batch-invariant kernels
- 2.8x faster via CUDA Graphs
- slime achieves 100% reproducible RL training on Qwen3-8B

2025 Highlight: New models day-0 support

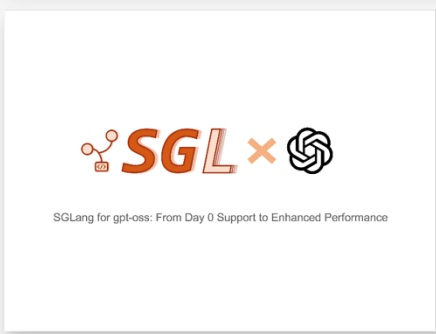
DeepSeek, GPT-OSS, Qwen, Kimi, GLM, LongCat, and so on



SGLang Day 0 Support for DeepSeek-V3.2 with Sparse Attention

by: The SGLang Team, September 29, 2025

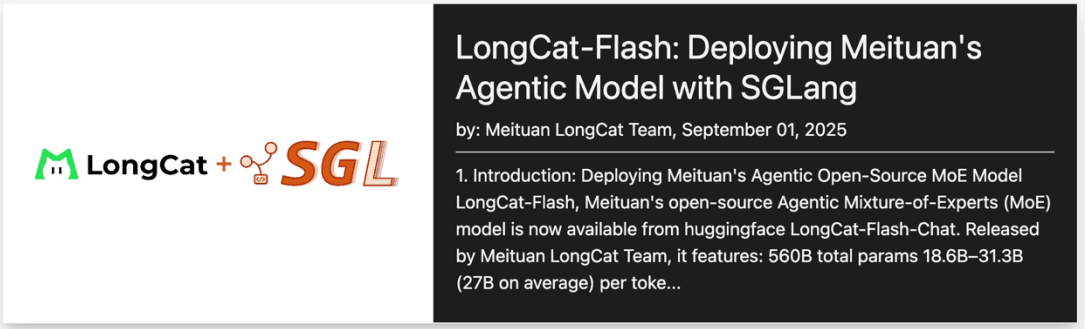
We are excited to announce that SGLang supports DeepSeek-V3.2 on Day 0! According to the DeepSeek tech report, it equips DeepSeek-V3.1-Terminus with DeepSeek Sparse Attention (DSA) through continued training. With DSA, a fine-grained sparse attention mechanism powered by a lightning indexer, DeepSee...



SGLang for gpt-oss: From Day 0 Support to Enhanced Performance

by: Liangsheng Yin, Ke Bao, August 27, 2025

We are excited to announce a major update for SGLang, focusing on deep performance optimizations and new features for the recently released openai/gpt-oss-120b model. While we had support from day zero, we took the last few weeks to enhance our engine to ensure you get the best possible performance....



LongCat-Flash: Deploying Meituan's Agentic Model with SGLang

by: Meituan LongCat Team, September 01, 2025

1. Introduction: Deploying Meituan's Agentic Open-Source MoE Model LongCat-Flash, Meituan's open-source Agentic Mixture-of-Experts (MoE) model is now available from huggingface LongCat-Flash-Chat. Released by Meituan LongCat Team, it features: 560B total params 18.6B–31.3B (27B on average) per toke...

GLM-4.5 Meets SGLang: Reasoning, Coding, and Agentic Abilities

by: GLM Team, July 31, 2025

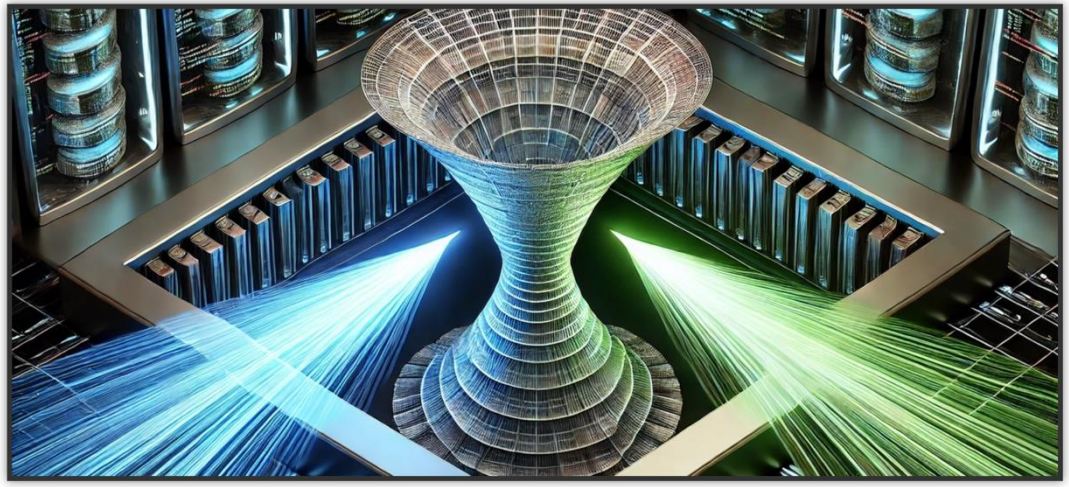
Today, we are excited to introduce our latest flagship models GLM-4.5 and GLM-4.5-Air, along with their FP8 variants. All models are now available with day-one support on SGLang. GLM-4.5 and GLM-4.5-Air are both powerful models designed to unify reasoning, coding, and agentic capabilities, with 355B...

2025 Highlight: Model Deployment Orchestration

- OME is a Kubernetes operator for enterprise-grade management and serving of Large Language Models (LLMs).
- Use case: launch a 128-GPU cluster for deploying kimi-k2 with PD and EP with one click

2025 Highlight: Distributed Inference on AMD

Unleashing AMD Instinct™ MI300X GPUs for LLM Serving: Disaggregating Prefill & Decode with SGLang



6 min read. | 1377 total words.
August 28, 2025 by Bill He, Andy Luo.

AI
AI/ML, Optimization, Performance
Software tools & optimizations

- Prefill / Decode Disaggregation — isolate compute vs memory workloads
- RDMA-based KV transfer (Mooncake) — fast, non-blocking communication
- Up to 7× latency & 6.9× Goodput improvement
- Result: lower TTFT / TPOT, higher GPU efficiency

Outlook for the Rest of 2025

- Speculative decoding refactor (compatible with CPU overlap scheduler)
- Memory pool refactor
- Multi platform abstraction refactor
- Optimize DeepSeek, GPT-OSS, Qwen (SOTA performance and reliability)
- Support performant combination of all major features
- SGLang Model Gateway (sgl-router successor)
- And so on

Question & Answer

sglang Public

SGLang is a fast serving framework for large language models and vision language models.

Python 18,886 Apache-2.0 3,036 549 (29 issues need help) 762

- GitHub: https://github.com/sgl-project/sglang
- Slack: https://slack.sglang.ai
- X: https://x.com/lmsysorg
- LinkedIn: https://www.linkedin.com/company/sgl-project

Contributors 770

+ 756 contributors


AMD