202502 Open-Source Week

We're a tiny team @deepseek-ai pushing our limits in AGI exploration.

Starting this week , Feb 24, 2025 we'll open-source 5 repos – one daily drop – not because we've made grand claims, but simply as developers sharing our small-but-sincere progress with full transparency.

These are humble building blocks of our online service: documented, deployed and battle-tested in production. No vaporware, just sincere code that moved our tiny yet ambitious dream forward.

Why? Because every line shared becomes collective momentum that accelerates the journey. Daily unlocks begin soon. No ivory towers - just pure garage-energy and community-driven innovation 🔧

Stay tuned – let's geek out in the open together.

Day 1 - FlashMLA

Efficient MLA Decoding Kernel for Hopper GPUs

Optimized for variable-length sequences, battle-tested in production

FlashMLA GitHub Repo

BF16 support

Paged KV cache (block size 64)

Performance: 3000 GB/s memory-bound | BF16 580 TFLOPS compute-bound on H800

Day 2 - DeepEP

Excited to introduce DeepEP - the first open-source EP communication library for MoE model training and inference.

DeepEP GitHub Repo

Efficient and optimized all-to-all communication

Both intranode and internode support with NVLink and RDMA

High-throughput kernels for training and inference prefilling

Low-latency kernels for inference decoding

Native FP8 dispatch support

Flexible GPU resource control for computation-communication overlapping

Day 3 - DeepGEMM

Introducing DeepGEMM - an FP8 GEMM library that supports both dense and MoE GEMMs, powering V3/R1 training and inference.

DeepGEMM GitHub Repo

Up to 1350+ FP8 TFLOPS on Hopper GPUs

No heavy dependency, as clean as a tutorial

Fully Just-In-Time compiled

Core logic at ~300 lines - yet outperforms expert-tuned kernels across most matrix sizes

Supports dense layout and two MoE layouts

Day 4 - Optimized Parallelism Strategies

DualPipe - a bidirectional pipeline parallelism algorithm for computation-communication overlap in V3/R1 training.

GitHub Repo

EPLB - an expert-parallel load balancer for V3/R1.

GitHub Repo

Analyze computation-communication overlap in V3/R1.

GitHub Repo

Day 5 - 3FS, Thruster for All DeepSeek Data Access

Fire-Flyer File System (3FS) - a parallel file system that utilizes the full bandwidth of modern SSDs and RDMA networks.

6.6 TiB/s aggregate read throughput in a 180-node cluster

3.66 TiB/min throughput on GraySort benchmark in a 25-node cluster

40+ GiB/s peak throughput per client node for KVCache lookup

Disaggregated architecture with strong consistency semantics

Training data preprocessing, dataset loading, checkpoint saving/reloading, embedding vector search & KVCache lookups for inference in V3/R1

3FS → https://github.com/deepseek-ai/3FS

Smallpond - data processing framework on 3FS → https://github.com/deepseek-ai/smallpond

Day 6 - One More Thing: DeepSeek-V3/R1 Inference System Overview

Optimized throughput and latency via:

Cross-node EP-powered batch scaling

Computation-communication overlap

Load balancing

Production data of V3/R1 online services:

73.7k/14.8k input/output tokens per second per H800 node

Cost profit margin 545%