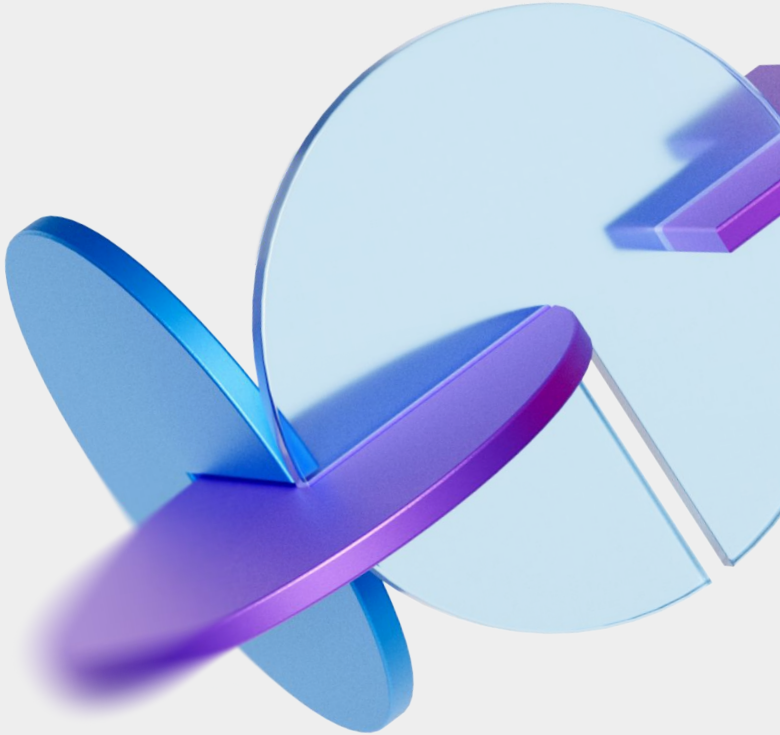
together.ai

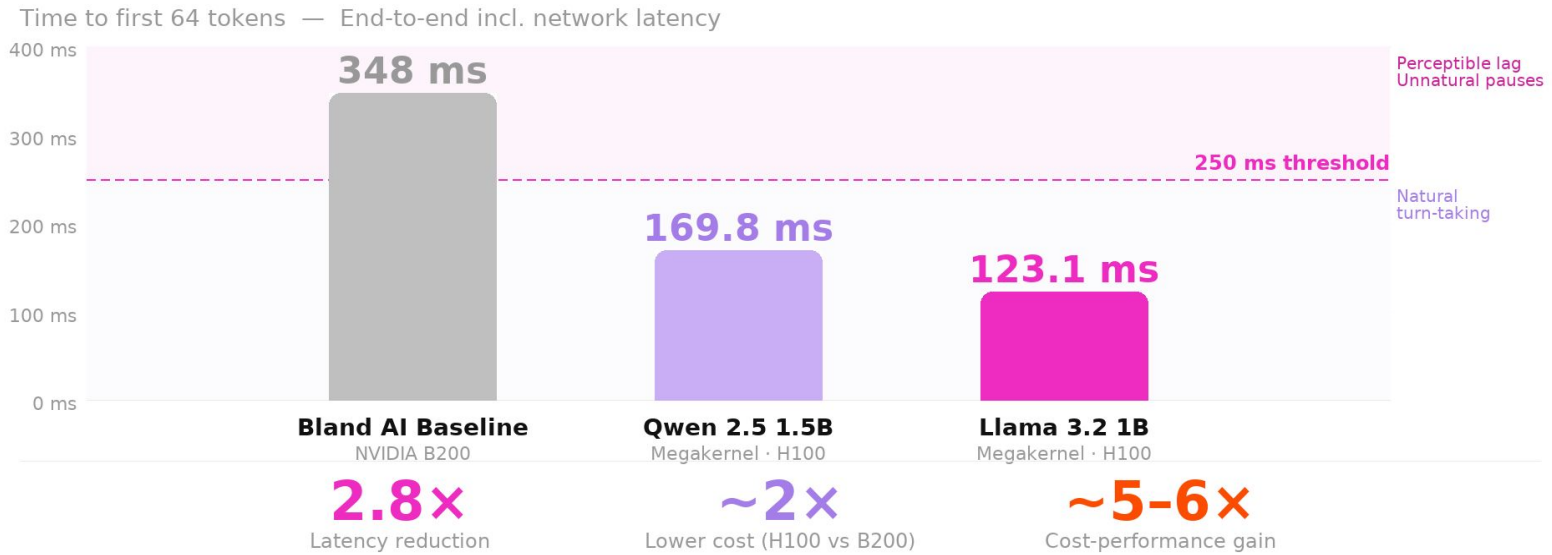


The AI Native Cloud

Train, fine-tune, and run your own enterprise AI models, faster ⚡

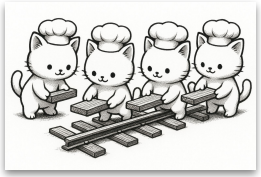
Megakernels in Production: Bland

Built on a research agenda started at Stanford and Together!



ThunderKittens DSL: Primitives and building blocks to write a fast kernel.

ThunderKittens: Simple, Fast, and Adorable AI Kernels
ParallelKittens: Simple and Fast Multi-GPU AI Kernels



MegaKernels system: Scheduling many sequential kernels in end-to-end workloads.

ThunderMLA: FlashMLA, Faster and Fused-er!
Look Ma, No Bubbles! Designing a Low-Latency Megakernel for Llama-1B
How Many Llamas Can Dance in the Span of a Kernel?

Achieving Peak Performance at Scale

Challenges with today's systems

MegaKernels

- Hand-tuned instructions for each operator
- Written with custom DSLs

Torch Compile

- Limited kernel fusion opportunities
- High warm start penalties

Together Compile

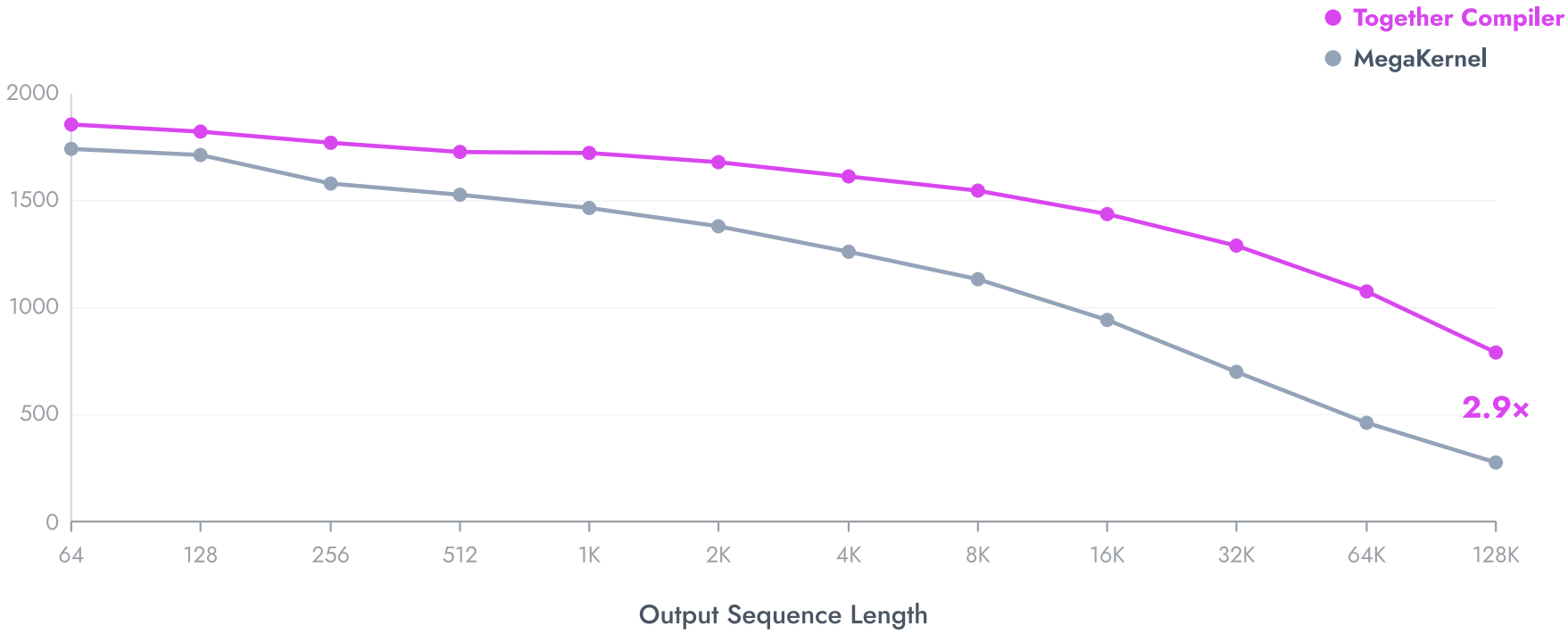
✔️ Interoperable with Torch frontend
✔️ Generalizes to any model and codebase
✔️ No runtime overhead

Only need to change **one** line of code!
@torch.compile → @together.compile

Together Compiler Demo: Llama 1B

B200

Tokens per second — higher is better



Join Us!

Build cutting-edge, efficient AI systems with our world-class team

Systems Research Scientist, GPU Programming

- Develop high-performance GPU kernels
- Contribute to open-source AI innovation, such as ThunderKittens
- Frontier ML systems and compilers research

Machine Learning Engineer, Inference

- Build high-performance, scalable AI inference systems
- Apply deep systems knowledge for performance gains
- Translate frontier inference research into production

AI Researcher, Core ML

- Advance SOTA models through novel research
- Contribute to open-source and publish impactful findings
- Push the frontier of efficient inference and RL-driven training

Chat with our team at MLSys!

Simran Arora
PRINCIPAL SCIENTIST

Reyna Abhyankar
SR SYSTEMS RESEARCH ENGINEER

Dan Fu
VP KERNELS

Drew Wadsworth
SR SYSTEMS RESEARCH ENGINEER