LLMs: Taking the World By Storm

Large Language Models (LLMs) are revolutionizing various sectors, including cybersecurity, government, research, finance, and enterprise communications. These advanced language processing tools can generate, classify, and summarize text with remarkable coherence and accuracy with the ability to predict the next word in a sentence by analyzing vast amounts of data. LLMs are transforming the way we work and communicate and their potential applications are vast, with their impact on various markets only beginning to be realized.

Fully unlocking the potential of LLMs requires:

1. **Real-time interactions**
2. **Differentiated performance from your competition**

Introducing the LPU™ Inference Engine

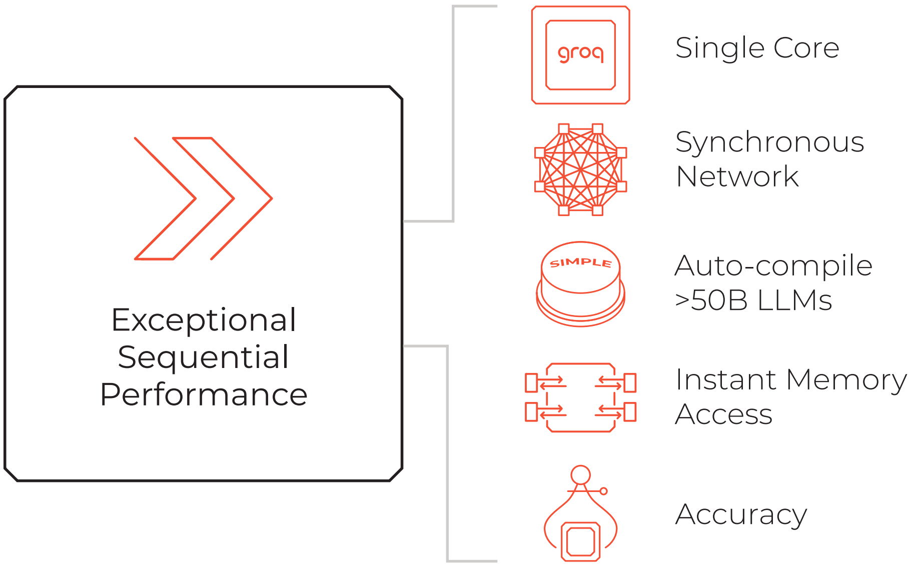
Groq is a real-time AI inference company and the creator of the LPU™ Inference Engine, the fastest language processing accelerator on the market. It is architected from the ground up to achieve low latency, energy-efficient, and repeatable inference performance at scale.

- We offer superior performance for LLMs at scale, delivering real-time outcomes, higher throughput, greater accuracy, and more efficient scalability.
- We accelerate the pace of AI workload development through rapid, kernel-less compilation, improving time to market, reducing resource requirements, and keeping up with the pace of innovation.
- We deliver predictability, providing accurate data on workload performance and costs at compile time so that developers can optimize software design with full understanding of how it will run when deployed.
- We enable higher accuracy, as lower latency allows for software techniques to deliver better predictions and improved business results in real-time.

We envision an AI solutions ecosystem that moves at the pace of software, unconstrained by the slow pace and high costs of big chip makers’ hardware development cycles.

KEY ATTRIBUTES OF AN LPU INFERENCE ENGINE

- **10X performance advantage**
- **Lower cost & lower power**
- **No supply constraints**



Exceptional Sequential Performance

Single Core

Synchronous Network

Auto-compile >50B LLMs

Instant Memory Access

Accuracy

LPU Inference Engine Performance

Leveraging our software and hardware ecosystem, Groq can get multi-billion parameter LLMs and other AI models models up and **running in less than five days.** Today, we're running Llama-2 70B at over 300 tokens per second per user, record-breaking performance that has continued to 2024. Groq offers the fastest inference performance, in tokens per second per user, in the industry. Customers rely on the LPU Inference Engine as an end-to-end solution for running Large Language Models and other generative AI applications at 10x the speed. Access to the LPU Inference Engine is available through GroqCloud™ via an API with public and private cloud options.

July 18th	July 24th	July 29th	August 3rd	August 31	**Today**
Model released	Model compiling five days after first download	Performance five days after first compile	Performance 10 days after first compile	Performance 38 days after first compile	**30x performance increase since first compile**
Llama-2 70B released	10 T/s per user initial performance	65 T/s per user	100 T/s per user	240 T/s per user	**>300 T/s per user**

Go to groq.com to experience real-time LLM performance, or reach out to us at groq.com/contact.