

AI Cloud for Agent Development

3-dimensional optimization

Lin Qiao |

CEO & Cofounder, Fireworks AI

2025 is the year of Agents

Coding Agents

Document
Agents

Sales/Marketing
Agents

Hiring Agents

Customer
Service Agents

Retail

Insurance

Finance

Education

Health/Medical

Manufacturing

Security





CURSOR

 Fireworks AI



CURSOR **Uber**

 Fireworks AI



CURSOR

Uber

 **UiPath**

Principle 1:

Model is Your Product

with product-model codesign



Principle 2:

Model is Your IP

through data flywheel





Build



Customize

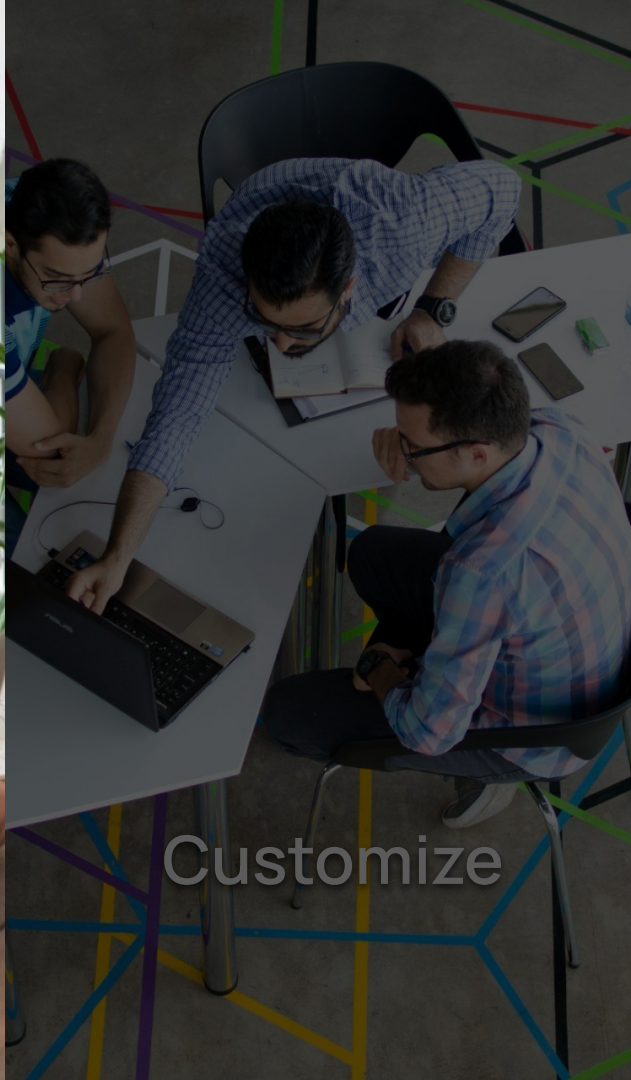


Scale





Build



Customize



Scale



Fireworks AI @FireworksAI_HQ · Dec 13, 2023

Latest @MistralAI's Mixtral MoE 8x7B model has been on Fireworks since Saturday (2 days before it was officially released)

🏆 Quality: beating GPT-3.5 on most benchmarks

🚀 Speed: fastest inference engine reaching 1000 tokens/sec

[Show more](#)

Gautam Goswami @gautam5669

Fireworks always first in decoding @MistralAI's model 🤖

11:56 PM · Apr 11, 2024 · 731 Views

Artificial Analysis @ArtificialAnlys

It is bizarre that third-party providers are hosting Gemma 2 before Google is.

We are now seeing third-party API providers host Gemma 2 with @FireworksAI_HQ launching their Gemma 2 9B API.

Fireworks AI @FireworksAI_HQ · Apr 28

Qwen 3 is now live on Fireworks AI!

Qwen 3 is a SOTA open-weights model for building AI agents:

🔧 SOTA agentic tool use: Highest ranked open model on BFCL (tool-calling) benchmark, below gpt4-2024-11-20 and above gpt4.5. Tuned to support Model Context Protocol (MCP) tools with

[Show more](#)

Qwen3

LATEST RELEASE - APRIL 2025

Fireworks AI

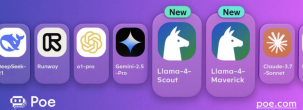
🔧 SOTA agentic tool use

Highest ranked open model on BFCL (tool-calling) benchmark, tuned to support MCP tools with the Qwen agent

🔧 Top-tier coding and reasoning capabilities

Outperforms every open and models on usefulness and consistency

Llama 4 Scout & Maverick are now available



AI at Meta @AIatMeta · Apr 5

Today is the start of a new era of natively multimodal AI innovation.

Today, we're introducing the first Llama 4 models: Llama 4 Scout and Llama 4 Maverick — our most...

Fireworks AI @FireworksAI_HQ · Mar 24

We're excited to announce that DeepSeek AI's latest model, DeepSeek V3-0324, released today, is officially available in the Fireworks AI Model Library.

We are also the official inference partner for DeepSeek V3-0324 on Hugging Face, supporting the community in bringing advanced

[Show more](#)

Fireworks AI deepseek

DeepSeek v3-0324

Now available on fireworks.ai!

Fireworks AI @FireworksAI_HQ · Jan 20

Playground: [fireworks.ai/models/firewor...](#)

Get started with API: [docs.fireworks.ai/getting-starte...](#)

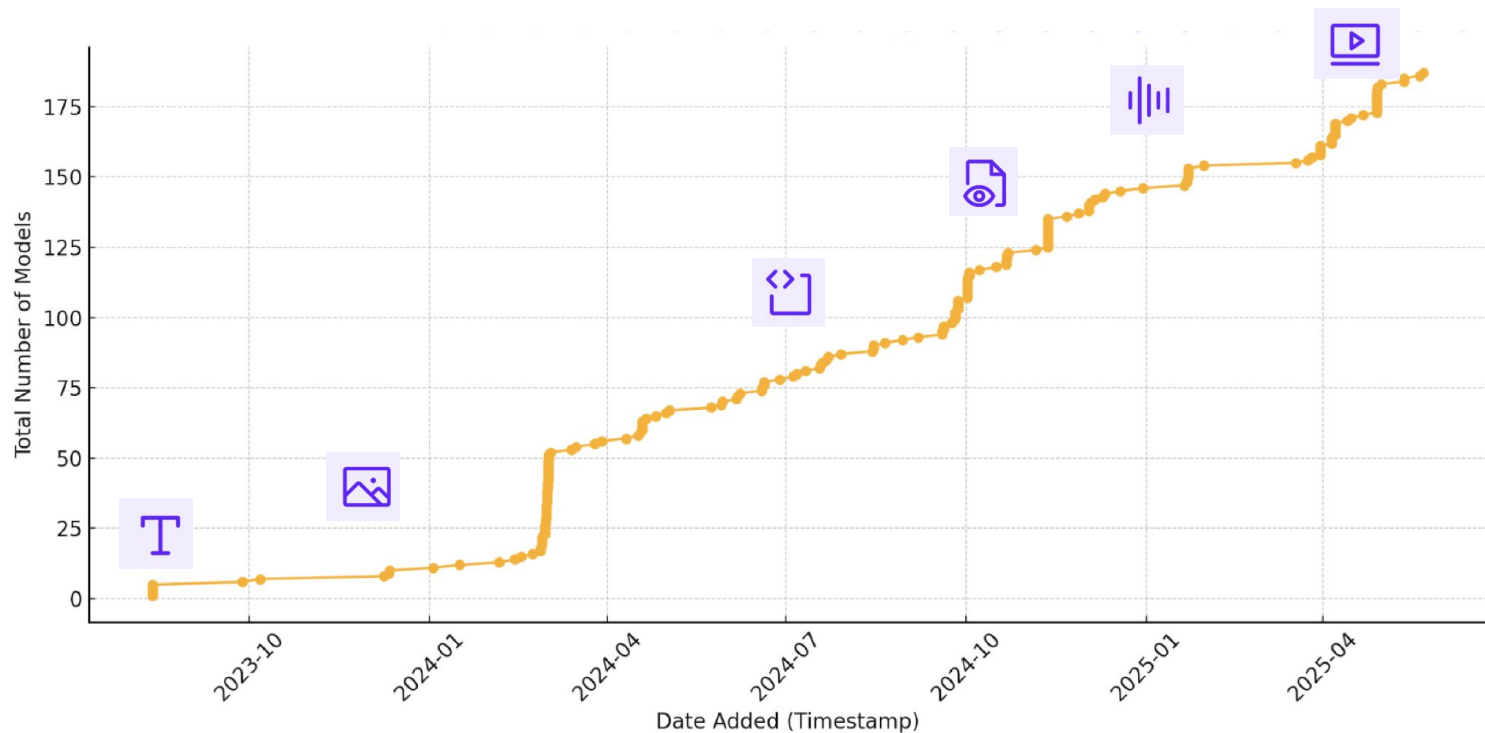
Contact us for enterprise deployments: [fireworks.ai/company/contac...](#)

Fireworks AI deepseek

DeepSeek R1 model is LIVE on Fireworks AI

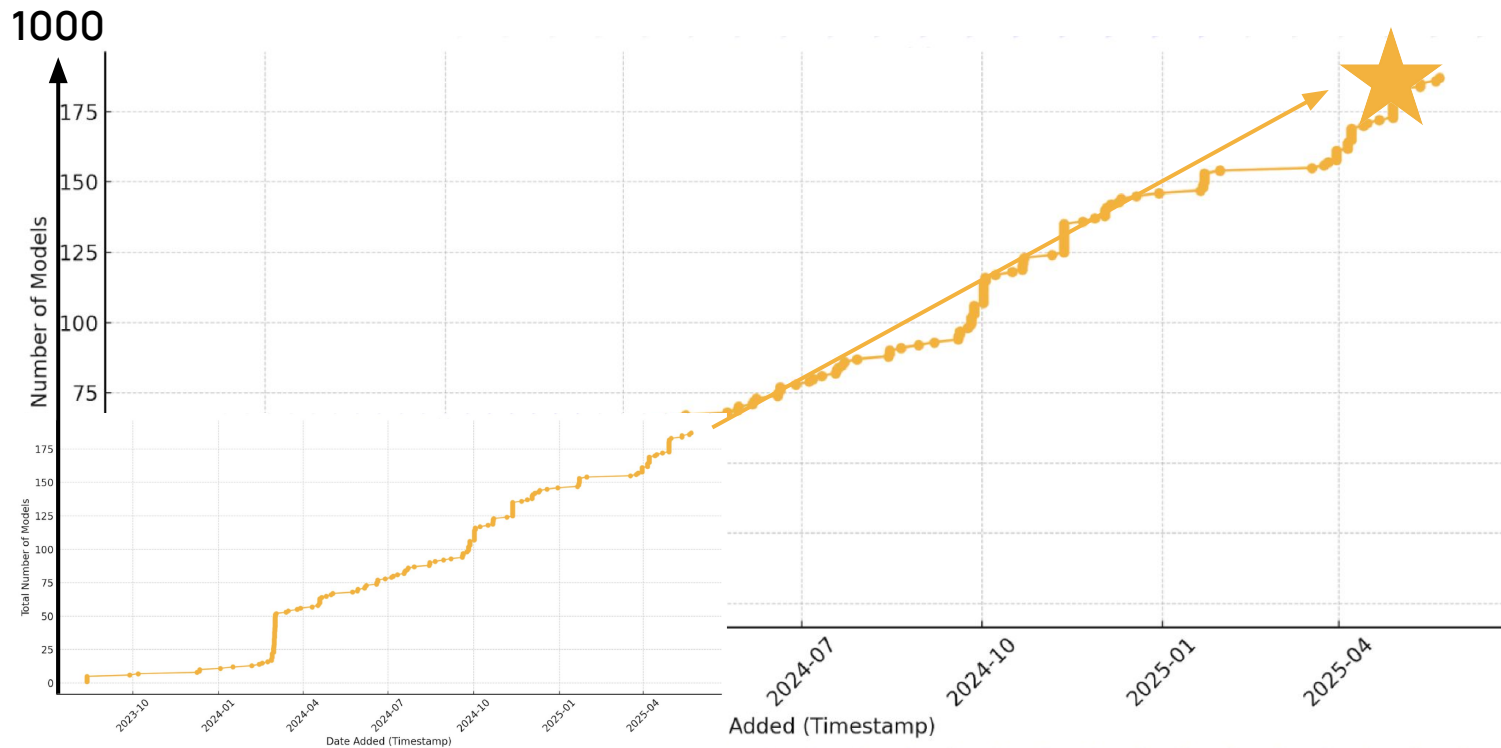


SOTA Open Models Enabled





Access to 1000+ models today!



What a summer of open models!

Model	Total Params, B	Active Params, B	KV cache entry, KB	Max Context	Layer s	Hidden Dim	Experts (MoE)			
							Total	Active	Sparsity	Routing
gpt-oss-120b	120	5.1	36	131,072	36	2880	128	4	32	Simple
gpt-oss-20b	20	3.6	24	131,072	24	2880	32	4	8	Simple
DeepSeek V3/R1	671	37	69	163,840	61	7168	256	8	32	Grouped
Kimi K2	1000	32	69	131,072	61	7168	384	8	48	Simple
Qwen3-235B	235	22	188	262,144	94	4096	128	8	16	Simple
Qwen3-30B	30.5	3.3	96	262,144	48	2048	128	8	16	Simple
GLM-4.5	355	32	368	131,072	92	5120	160	8	20	Simple
GLM-4.5-Air	106	12	184	131,072	46	4096	128	8	16	Simple

Experiment Platform (GA)

 Fireworks AI

Platform ▾ Model Library Customers Docs Pricing Blog Company ▾

LOG IN GET STARTED



OpenAI gpt-oss-120b

Welcome to the gpt-oss series, OpenAI's open-weight models designed for powerful reasoning, agentic tasks, and versatile developer use cases. gpt-oss-120b is used for production, general purpose, high reasoning use-cases that fits into a single H100 GPU.

[TRY MODEL](#)

Fireworks Features

**Fine-tuning**

OpenAI gpt-oss-120b can be customized with your data to improve responses. Fireworks uses LoRA to efficiently train and deploy your personalized model

[LEARN MORE](#)

**Serverless**

Immediately run model on pre-configured GPUs and pay-per-token

[LEARN MORE](#)

**On-demand Deployment**

On-demand deployments give you dedicated GPUs for OpenAI gpt-oss-120b using Fireworks' reliable, high-performance system with no rate limits.

[LEARN MORE](#)

Info & Pricing

	Provider	OpenAI
	Model Type	LLM
	Context Length	131072
	Serverless	Available
	Fine-Tuning	Available
	Pricing Per 1M Tokens Input/Output	\$0.15 / \$0.6

- SOTA Open models
- Fast, Powerful Tuning with Zero Setup
- Flexible Capacity & 1-click Deploy
- Private and Secure
- Build SDK – experiment as code

Evaluation is a big pain

Model selection

CI/CD

Reinforcement tuning



eval protocol

evalprotocol.io

github.com/eval-protocol



The open-source toolkit for building your internal model leaderboard.

pypi v0.2.22 license MIT

[Documentation](#)



What is Eval Protocol?

When you have multiple AI models to choose from—different versions, providers, or configurations—how do you know which is best for your use case?



Organization Repositories

Core Projects

Repository	Description
eval-protocol	Main specification, documentation, and examples
python-sdk	Python implementation

Build with us

100x

Faster iteration

10K+

Companies

18x

Growth in a year





Build



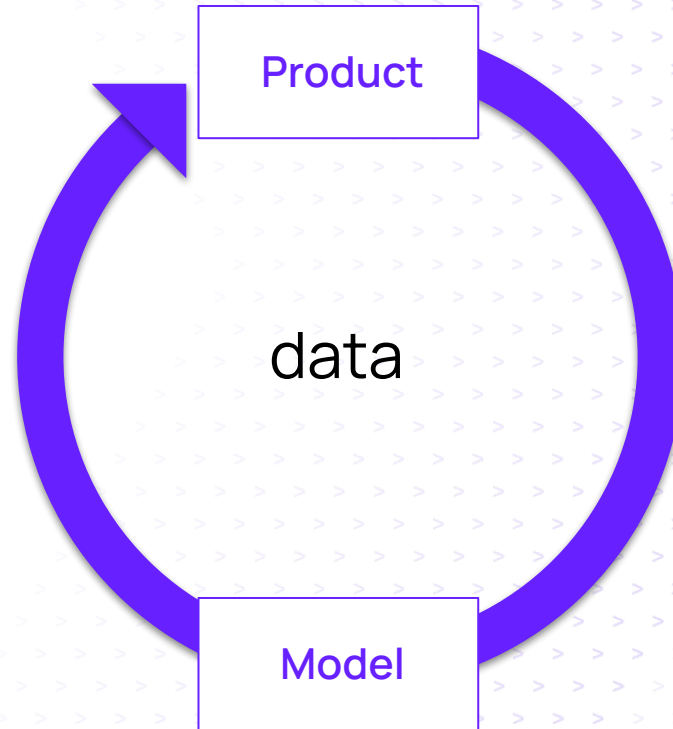
Customize



Scale



Customize with data flywheel



Supervised Fine Tuning update

- Big SOTA models
- Longer context
- Better quality for quantization
- Faster training



Announcing

Reinforcement Fine Tuning (Beta)

- RFT SOTA open models
- Composing SFT and RL
- Flexible evaluation
- Ease of use with reward-kit SDK and web IDE



Supervised Fine Tuning V2

Better Quality

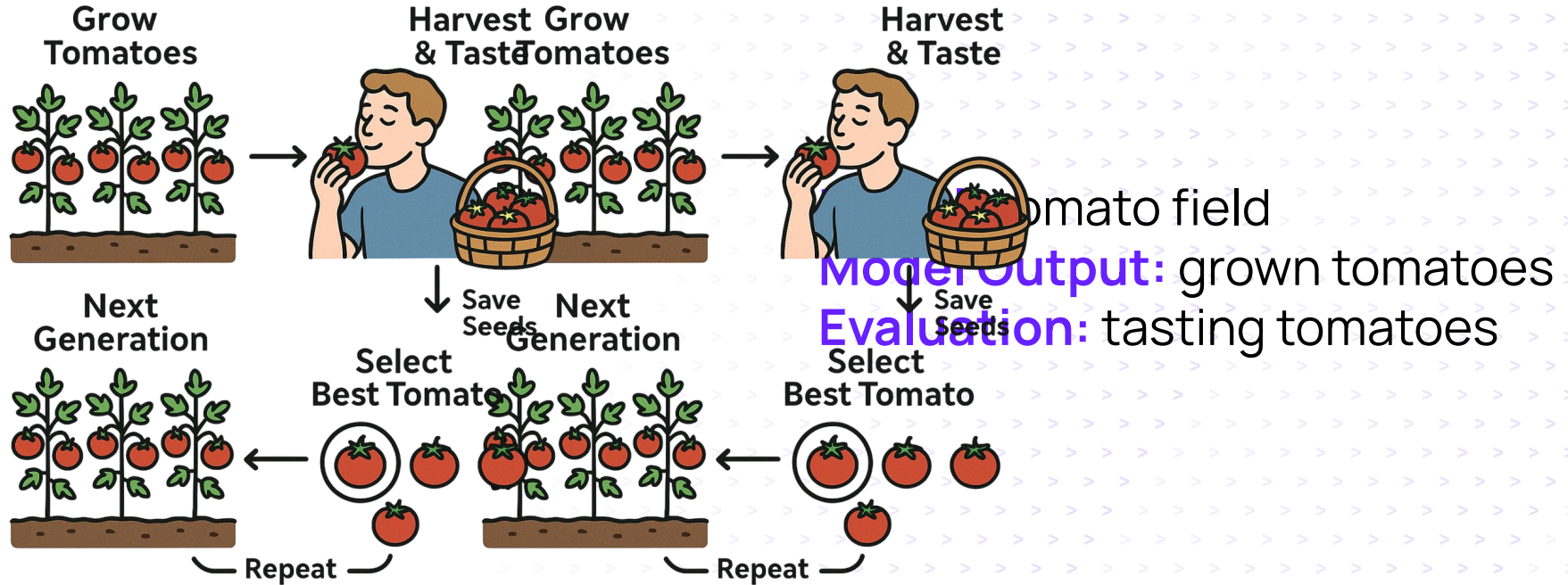
- Fine tune DeepSeek R1, V3, Llama and other SOTA models
- Access full 131K context window for training
- Preserve model quality with Quantization Aware Training

Faster Speed

- 2X faster training speed
- Better speed, quality with Multi-token Prediction Tuning



Reinforcement Fine Tuning



Reinforcement Fine Tuning Workflow

```
Evaluator

def evaluate(messages: list[dict], ground_truth: int, **kwargs) -> dict:

    answer = messages[-1]['content']
    last_line = answer.split('\n')[-1]
    mg = re.match(r"Answer: (\d+)", last_line)
    score = 0.0
    if mg is not None:
        extracted_answer = int(mg.group(1))
        score = 1.0 if extracted_answer == ground_truth else 0.0
    return {
        "score": score,
        "reason": str(ground_truth),
        "is_score_valid": True,
    }
```

Create Evaluator
(Reward Function)

Select Model & Dataset

Run Training Job

Review Results
and Deploy

Reinforcement Fine Tuning DEMO



Better results than proprietary models

Candidate selection

9x

Faster than
o4-mini

20%

Better accuracy than
o4-mini

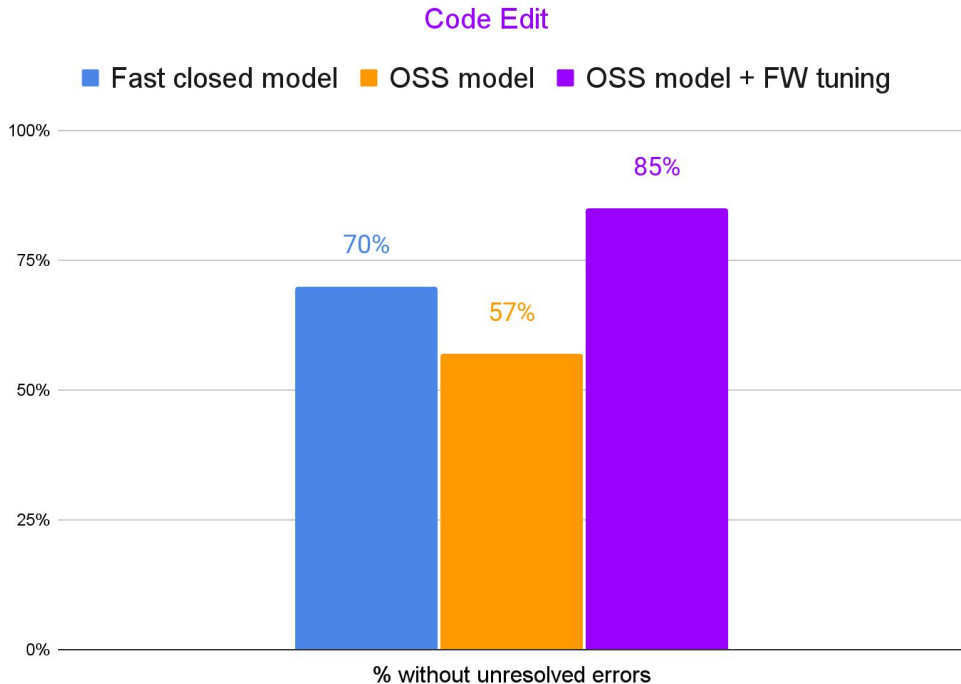


Code Edit

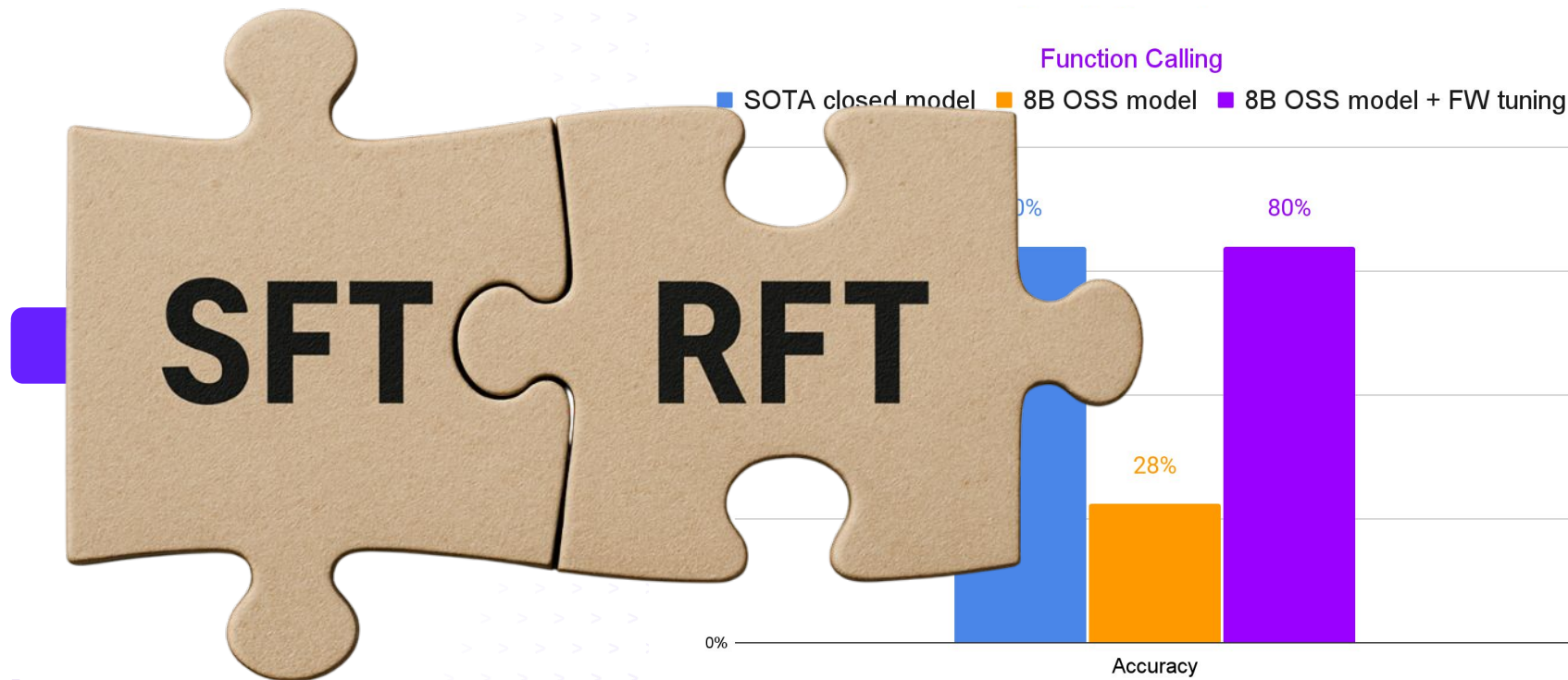
“The awesome news is our current model takes 2 passes to fix it, while using Fireworks this new model took 1. On a 800 LOC file, that is huge!”

Better results than proprietary models

Code Rewrite for very large files

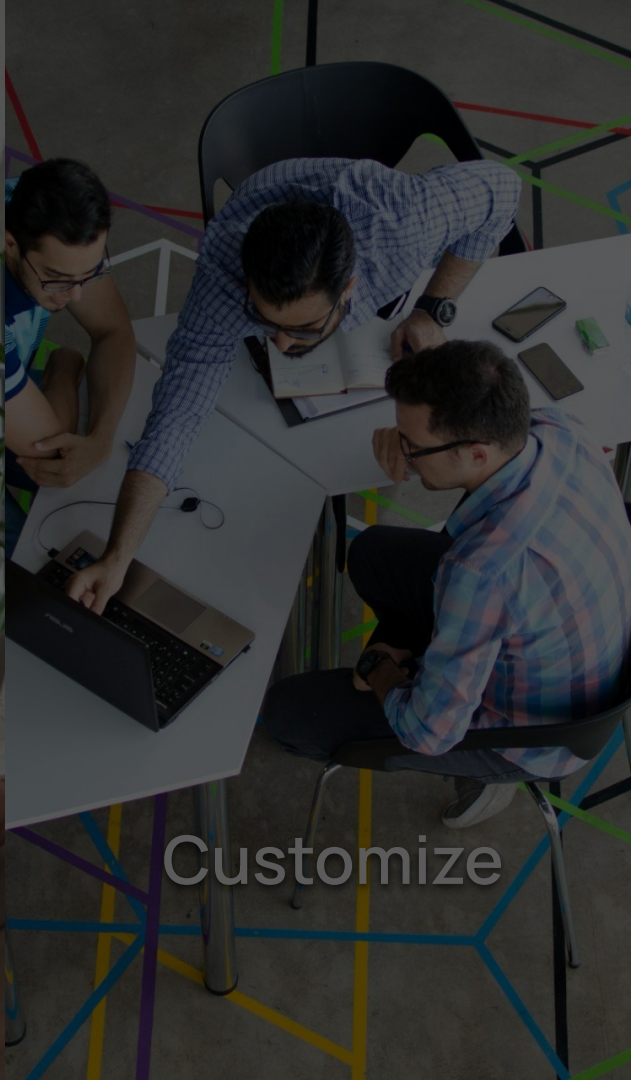


SOTA results with Fireworks Tuning





Build



Customize



Scale



Scale

Reliably with high performance



Which hardware?



Where is capacity?



System failures with a thousand cuts



How to optimize for production



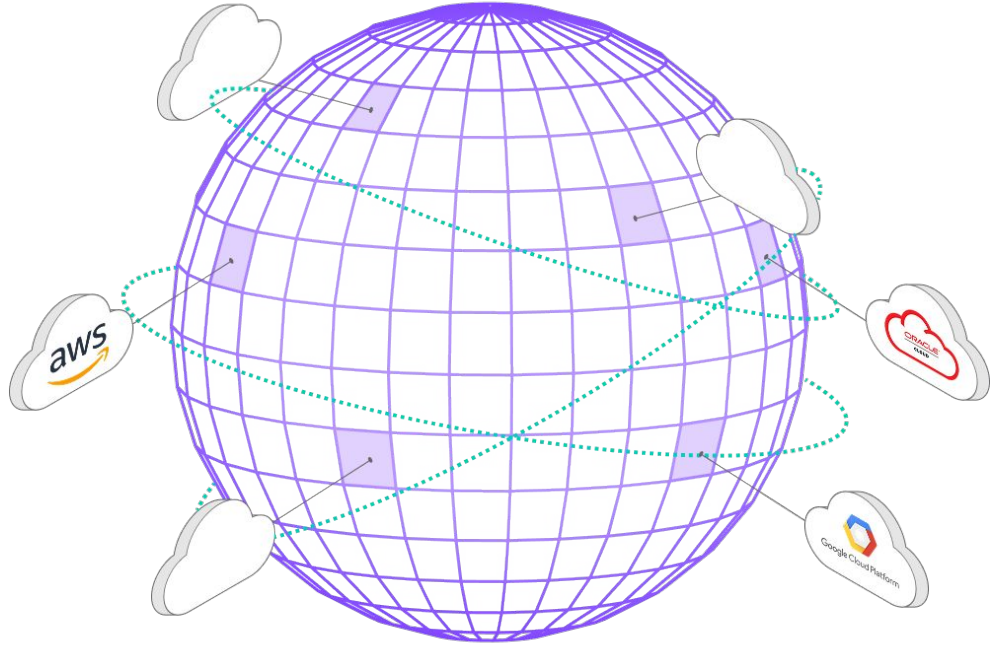
Global distribution

8

clouds

18

regions



+ bring your own cloud



Announcing

Virtual Cloud

- Across 8 CSPs and SOTA hardware
- Disaggregated engine for max efficiency
- High reliability and global scheduling
- Enterprise grade privacy and compliance
- Bring your own cloud or bucket



3-D Optimizer v2 for Quality, Speed & Concurrency

5 speculative decoding modes

8 quantization modes

7 hardware SKUs

4 sharding schemes

5 multi-host setups

10+ kernel options

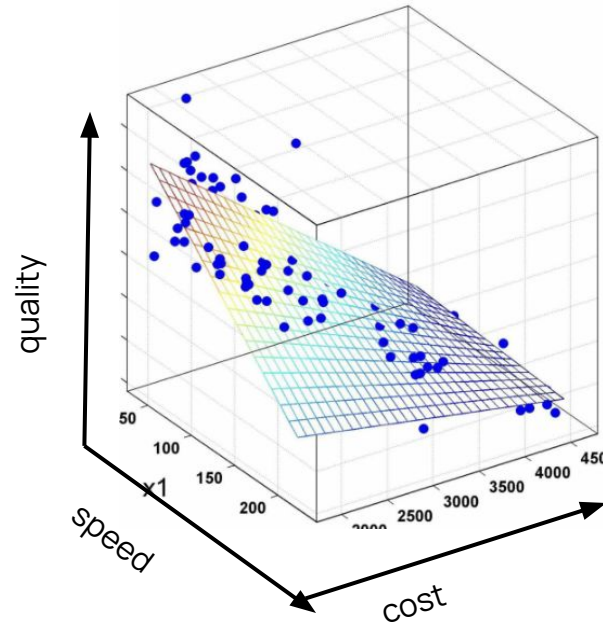
4 quality tuning approaches

100,000+
possible
combinations



3-D Optimizer for Quality, Speed & Concurrency

All in One



10,000,000,000,000+

tokens / day

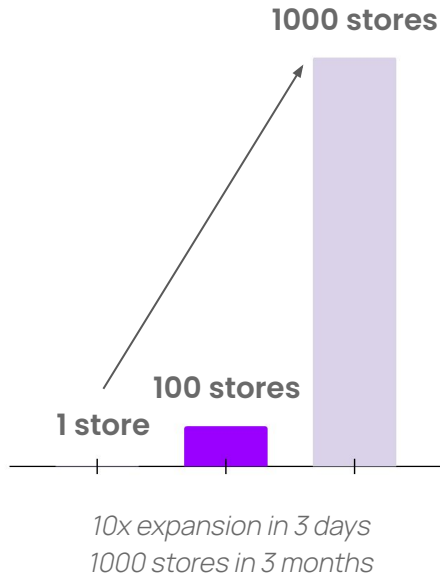
180,000+

requests / second

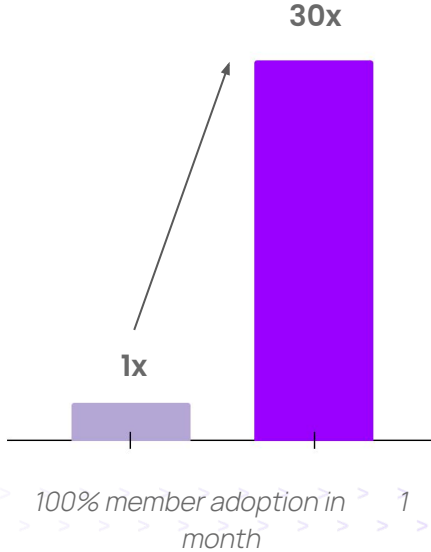


Enterprise customers scale fast

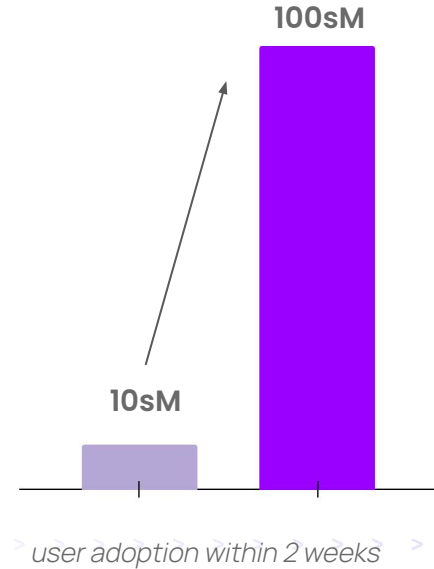
Food chain



Digitally Native



Software Productivity



Starting scale



Current scale



Start building today!

- \$5k credits
- New product links

