AGENTIC SYSTEMS WHITEPAPER

Fireworks AI

Unlocking Function Calling & AI Agents at Enterprise Scale

Real-Time, Action-Oriented Agents that Execute Workflows Reliably and Securely

Executive Summary

01

Enterprises face growing pressure to automate complex workflows reliably and at scale. Passive AI assistants can answer questions, but they cannot take action across systems, APIs, and applications. Enterprises need reliable, context-aware function execution, scalable multi-agent orchestration, and fine-tuned models that integrate seamlessly into workflows.

This whitepaper shows how Fireworks AI enables enterprises to harness function calling and AI agents for real-world agentic workflows, including:

- **Enterprise AI Assistants**: In-product assistants that take action across internal or customer-facing systems.
- **Vertical Agents**: Tool-using agents tailored to domains such as e-commerce, logistics, and productivity.
- **Voice AI Agents**: Real-time speech-to-action pipelines for in-flow automation.

Customers like **Notion** use Fireworks AI to automate complex workflows reliably, accelerate task completion, and scale operations with full control and enterprise-grade governance.

The Case for Agentic AI in the Enterprise

MACRO TREND — AI is moving from answering questions to taking action. Gartner projects that by 2028, one-third of GenAI interactions will involve autonomous agents completing tasks.

CHALLENGE — Legacy models often fail due to context-blind function calls, unreliable multi-step task execution, and limited observability, monitoring, and compliance controls.

SOLUTION — Fireworks AI delivers structured, production-grade function calling and agentic reasoning, enabling enterprises to orchestrate complex, multi-step workflows reliably, securely, and at scale.

Fireworks AI: The Platform for Enterprise Agentic Workflows

Fireworks AI provides a complete, enterprise-ready platform for deploying tool-using, action-oriented AI agents.

Feature	Enterprise Value	How Fireworks Enables It	Metrics/Benchmarks
Model choice & customization	Align agentic workflows to enterprise APIs, internal tools, and domain-specific tasks	FireOptimizer fine-tunes open and proprietary LLMs; RL-optimized models for multi-step workflows	72% win rate in production tasks; multi-step task reliability
Enterprise-ready infrastructure	Support high concurrency and global deployment of agentic workflows	GPU autoscaling, real-time STT pipelines, and resilient multi-region architecture	Millions of concurrent agentic tasks handled reliably; <500ms function execution
Cost Efficiency	Optimize resource usage and minimize compute spend	Predictable GPU scaling, off-the-shelf model baseline with selective fine-tuning	2X throughput vs GPT-4omini; 2X faster task completion
Control & Flexibility	Full observability, traceability, and governance across workflows	RBAC, audit logging, monitoring, SOC2/GDPR-ready deployments	Enterprise-grade compliance and end-to-end workflow traceability
AI Orchestration/ Gateway	Coordinate multi-agent, multi-step workflows in real-time	Central AI Gateway for function execution, monitoring, and feedback loops	Low-latency, deterministic execution; millions of calls per day
Model Lifecycle Management	Continuously improve agent performance and reliability	Build → Tune → Scale workflow for fine-tuning, evaluation, and deployment	Multi-region deployment with high concurrency; continuous RL fine-tuning improves task accuracy

By combining real-time execution, multi-step workflow orchestration, fine-tuned models, and enterprise governance, Fireworks AI solves the key challenges of scaling agentic AI reliably and securely.

SPEED

300ms latency for real-time, in-flow workflow execution

ACCURACY

4X faster AI-powered responses with fewer errors in multi-step tasks

SCALE

100M+ users and millions of concurrent agentic tasks

By centralizing multi-agent orchestration, real-time function execution, and workflow monitoring through the Fireworks AI Gateway, enterprises achieve faster, more reliable, and fully traceable agentic workflows at scale.

Real-World Proof Points

Enterprises adopting Fireworks AI can expect faster feature delivery, reduced manual debugging time, and cost savings on inference infrastructure compared with DIY or closed APIs.

Notion faced challenges scaling AI beyond chat to enterprise-grade agentic workflows across 100+ million users and complex enterprise tools. Fireworks AI enabled:

- AI-powered responses 4X faster than standard open-source models
- Reliable, low-latency function execution integrated with Slack, Jira, and GitHub
- Fine-tuned, RL-optimized models for multi-step agentic workflows
- Scalable infrastructure to support rapid iteration and enterprise adoption

Read the Notion case study.

By centralizing multi-agent orchestration, real-time function execution, and workflow monitoring through the Fireworks AI Gateway, enterprises achieve faster, more reliable, and fully traceable agentic workflows at scale.

Patterns & Practices for Enterprise AI

03

People & Process

- Build expertise in agent orchestration and function calling
- Integrate agent workflows into existing processes
- Foster collaboration across platform, product, and AI teams

Technology Implementation

- Select and fine-tune models for domain-specific tasks
- Deploy high-throughput inference with continuous evaluation
- Use Fireworks AI Gateway to decouple agents from underlying models
- Enforce secure, auditable pipelines with RBAC, monitoring, and encryption

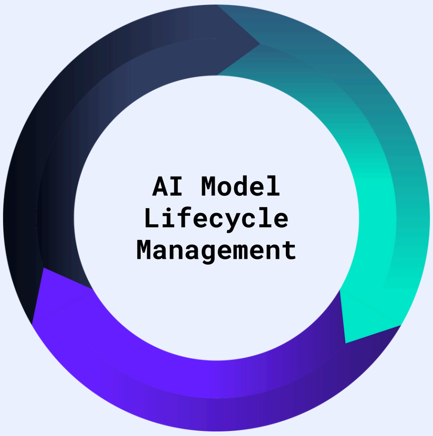
Risk Management

- Ensure secure, auditable, and traceable execution of agentic workflows
- Minimize operational and vendor lock-in risks
- Continuously monitor and fine-tune for accuracy, compliance, and governance requirements.

Operational Patterns

- **AI Gateway**: Central service for orchestration, monitoring, and deployment
- **Model Lifecycle Management**: Run, Tune, Scale loop for agentic workflows
- **Product-Model Co-Design**: Align agent capabilities tightly with business processes while ensuring compliance

Model Lifecycle Management



AI Model Lifecycle Management

BUILD
Select and run base AI Models to validate initial use case viability

TUNE
Continuously improve quality, relevance and control of model behaviors and outputs for appropriate outcomes

SCALE
Manage models in production with efficiency and effect

Why This Matters for Enterprises

Agentic AI can:

- Automate complex tasks across domains and apps
- Improve workflow reliability and speed
- Reduce operational risk with traceable, governed execution
- Scale seamlessly across enterprise traffic

With metrics, proof points, and operational patterns in place, enterprises can confidently move from experimentation to production while retaining control over their AI stack.

Getting Started

1

Identify high-value workflows

Start with multi-step, tool-using, or voice-driven tasks.

2

Run a pilot

Deploy Fireworks AI with preferred models to validate performance and reliability.

3

Scale and customize

Align agents with enterprise APIs, internal tools, and domain workflows.

4

Implement operational patterns

Adopt AI Gateway, Model Lifecycle Management, and product-model co-design.

Conclusion

05

The future of enterprise AI is agentic. Fireworks AI empowers enterprises to deploy agents that act, not just respond, with speed, reliability, and scale. Real-world proof points, enterprise-grade infrastructure, and fine-tuned function calling make scalable agentic AI a practical reality.

Next Steps

Fireworks AI helps enterprises move from experimentation to production with confidence. To get started:

Assess workflows	Identify tool-using, multi-step, or voice-driven tasks.
Run a pilot	Deploy Fireworks AI with fine-tuned models for reliable performance.
Fine-tune and scale	Customize agents for domain-specific APIs and workflows
Adopt operational patterns	Implement AI Gateway, Model Lifecycle Management, and product-model co-design practices for sustainable, enterprise-grade AI adoption. For a broader perspective on enterprise AI architecture and operational maturity, see our enterprise page.

Fireworks AI

Contact Fireworks AI today to schedule a pilot, explore model customization, and unlock scalable, high-performance code intelligence for your development teams.