

Unlocking Code Generation and Code Fixing at Enterprise Scale

Reliable, Scalable, and Low-Latency Developer Workflows

Executive Summary

01

Software development is undergoing a major transformation. Large language models (LLMs) are becoming central to how enterprises and AI-native companies approach code generation, code fixing, and developer productivity. With the right infrastructure and operational patterns, these models drive measurable gains in developer productivity, accelerate feature delivery, and improve code quality across large codebases.

This whitepaper shows how Fireworks AI enables enterprises to harness state-of-the-art LLMs for real-world software development use cases, including:

- **Code generation:** Automating boilerplate, accelerating feature development, and generating context-aware code suggestions.
- **Code fixing and editing:** Detecting errors, refactoring, and improving code at scale.
- **Developer Productivity:** Reducing time spent on debugging, repetitive tasks, and integration challenges.

Customer examples from **Cursor** and **Sourcegraph** illustrate real-world impact, including performance, cost, and scalability gains.



The Case for AI in Software Development

MACRO TREND — Enterprises face growing pressure to deliver software faster while maintaining quality. [Gartner](#) reports that 47% of organizations prioritize development speed, yet 70% of developers struggle to scale AI tools effectively.

CHALLENGE — Balancing speed with quality while maintaining secure, compliant code is difficult. Debugging alone wastes [\\$61B annually](#) and consumes up to 75% of a developer's day ([Undo & Cambridge Judge, 2024](#)).

SOLUTION — LLMs integrated with an enterprise-grade platform like Fireworks AI streamline workflows from code suggestions to automated bug fixes, enabling faster, safer, and more scalable development.

Fireworks AI: The Platform for Enterprise Code Intelligence

02

Fireworks AI provides a complete, enterprise-ready platform for accelerating developer productivity, code generation, and code fixing.

Feature	Enterprise Value	How Fireworks Enables It	Metrics/Benchmarks
Model choice & customization	Tailored AI that fits your domain and workflow	Support for high-performing open and proprietary LLMs, fine-tuned for code	Up to 2X higher relevance for code suggestions with custom models
Enterprise-ready infrastructure	Sub-second latency for interactive coding, repository-wide edits	Secure, high-throughput inference at scale	1000+ tokens/sec.generation
Cost Efficiency	Reduce operational costs. Significant savings vs DIY or public API solutions	Optimized serving architecture	62% reduction in cost per request , 3X lower cost vs alternatives
Control & Flexibility	Full governance and data security	BYOC deployment, fine-tuning with FireOptimizer, audit/RBAC controls	Enterprise-grade compliance and traceability
AI Orchestration/ Gateway	Centralized access to models and services	Unified API, model catalog, monitoring, flexible deployment	5T tokens/day processing, multi-region deployment
Model Lifecycle Management	Continuous performance, reliability, and cost optimization	Evaluation, fine-tuning, monitoring, and scaling	Reduced downtime, consistent throughput, cost-optimized deployment

This combination of **speed, precision, cost efficiency, and enterprise control** addresses the most common pain points for software development teams.

SPEED

Sub-second responses, 1000+ tokens/sec for interactive coding

ACCURACY

2X more relevant code suggestions with custom models

COST

62% reduction in cost per request vs alternative

By centralizing model orchestration, monitoring, and scaling through the Fireworks AI Gateway, enterprises achieve faster, more accurate, and cost-efficient code generation and fixing across their development teams.

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.



Cursor, the AI-native IDE, relies on Fireworks AI for high-performance inference to power code completion and editing workflows. With Fireworks, Cursor has been able to:

- Reduce latency for interactive coding.
- Scale cost-effectively with growing user demand.
- Deliver advanced developer experiences without compromise.
- Achieve **1000+ tokens/sec generation** using Fireworks' Speculative Decoding API with a custom-trained 70B model

Read the [Cursor case study](#).



Sourcegraph uses Fireworks AI to enhance code generation and large-scale code editing. In production, they demonstrated:

- **62% reduction in cost per request** using Fireworks inference infrastructure.
- **2.7X higher throughput** compared to prior deployment.
- Efficient handling of complex, repository-wide code editing tasks.

Read the [Sourcegraph case study](#).

These proof points show how **AI Gateway and lifecycle management** enable real-world enterprise-scale outcomes.

People & Process

- Build and maintain expertise to customize and manage LLMs for code.
- Integrate AI-assisted workflows into existing development pipelines
- Foster collaboration between ML teams and broader engineering teams.

Technology Implementation

- Select models (open, proprietary, hybrid) for domain needs
- Deploy high-throughput inference with continuous evaluation
- Use AI Gateway to decouple product from model for seamless upgrades

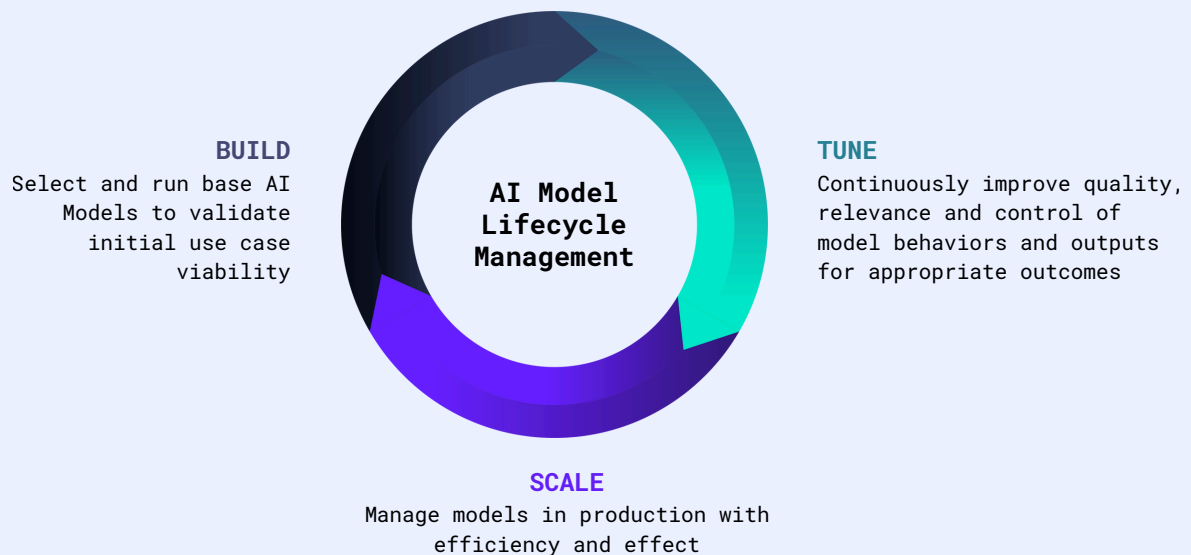
Risk Management

- Protect proprietary code and maintain compliance
- Ensure output quality and reliability through continuous evaluation
- Minimize operational and vendor lock-in risks

Operational Patterns

1. **AI Gateway:** Central service for model orchestration, customization, monitoring, and scalable deployment.
2. **Model Lifecycle Management:** Build → Tune → Scale selection, tuning, continuous deployment, monitoring, cost management.
3. **Product-Model Co-Design:** Tight integration of AI capabilities with product development for faster, higher-quality feature delivery.

Model Lifecycle Management



AI-driven code generation and fixing can:

- **Accelerate feature delivery** while reducing manual overhead.
- **Elevate developer productivity** by automating repetitive tasks.
- **Improve code quality** through consistent, model-driven suggestions and edits.
- **Scale securely** with enterprise-grade deployment and governance.

Deploying scalable, low-latency multimodal AI allows organizations to unlock faster, more accurate, and cost-effective insights across text, images, and audio, enabling reliable automation and informed business decisions at enterprise scale.

Getting Started

1

Identify high-value workflows

Start with product catalog enrichment, document parsing, real-time audio transcription.

2

Run a pilot

Deploy Fireworks AI with recommended models to validate performance, latency, and cost efficiency.

3

Scale and customize

Fine-tune models with enterprise-specific data and feedback loops to maximize ROI.

4





Implement operational patterns


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

Enterprises can unlock real-time, structured insights from unstructured data at scale with Fireworks AI. Fine-tuned multi-modal models, combined with enterprise-grade infrastructure and lifecycle management, transform image, document, and audio data into actionable intelligence faster, more accurately, and cost-effectively.

Next Steps

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

	Assess workflows	Identify high-value code workflows, such as code suggestions, bug fixing, or repository-wide edits
	Run a pilot	Deploy Fireworks AI with preferred models to validate performance, scalability, and cost efficiency.
	Fine-tune and scale	Customize models with enterprise-specific data and feedback loops to maximize ROI.
	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.