SEARCH & UNDERSTANDING WHITEPAPER

Fireworks AI

Unlocking Enterprise Search, Long-Context Summarization, and RAG

Reliable, Context-Aware Knowledge Access for Enterprise Workflows

Executive Summary

01

Enterprises face an overwhelming volume of multi-modal data, including documents, chats, code, and transcripts, yet knowledge workers spend up to 30% of their time searching for answers. Legacy search tools and generic LLMs often return incomplete or context-blind results, slowing decisions, frustrating users, and increasing operational costs.

Fireworks AI delivers a production-grade platform for enterprise knowledge workflows, enabling 5X faster insight generation and up to 80% reduction in manual review effort, with sub-second response times for multi-source queries. Key capabilities include:

- **Enterprise Search**: Retrieve relevant information across documents, chats, code, and knowledge bases within milliseconds (sub-300ms).
- **Long-Context Summarization**: Consolidate multiple sources, calls, and threads into concise, structured summaries for quick decision-making.
- **Enterprise RAG**: Reason across repositories to deliver real-time, context-aware guidance.

Ownership, orchestration, and fine-tuning ensure models align with enterprise domains and workflows, letting organizations confidently extract actionable knowledge from multi-source data at scale. The result: faster decisions, reduced manual effort, and scalable, auditable knowledge processes. Enterprises like **Cresta** achieve up to 80% reduction in manual review effort and 5X faster insight generation, while reducing operational cost and improving customer experience.

Fireworks AI enables organizations to scale from early experimentation to fully operational knowledge workflows, converting unstructured information into operational insights across every workflow.

The Case for Search & Understanding in Enterprises

MACRO TREND — Enterprises generate massive volumes of data across documents, support tickets, chat logs, emails, and reports. Knowledge workers can spend up to 30% of their time searching for answers, slowing decision-making and reducing productivity (Forrester via CDP Institute).

CHALLENGE — Legacy search tools and generic LLMs often fall short in enterprise settings:

- Queries may return incomplete or context-blind results
- Summaries often require manual reconciliation and lack structure
- Limited observability, compliance, and integration capabilities make scaling difficult

SOLUTION — Fireworks AI delivers a unified platform for search, long-context summarization, and RAG. Enterprises gain:

- Real-time, accurate knowledge retrieval across multiple domains
- Consolidated and structured summaries across multi-modal sources
- Scalable, secure, and auditable knowledge workflows

With enterprise-grade ownership, orchestration, and fine-tuning, Fireworks AI provides the compliance, observability, and integration required to scale pilots into production-ready workflows, while sustaining efficiency and customer trust.

Fireworks AI: The Platform for Enterprise Knowledge Workflows

Fireworks AI provides a complete, enterprise-ready platform for search, summarization, and RAG.

Feature	Enterprise Value	How Fireworks Enables It	Metrics/Benchmarks
Model choice & customization	Domain-aligned knowledge retrieval	FireOptimizer supports open and proprietary LLMs fine-tuned on enterprise support logs, product catalogs, and knowledge bases	5X faster insight generation, 80% reduction in manual review effort
Enterprise-ready infrastructure	Low-latency, high-throughput query processing	GPU autoscaling, caching, and optimized pipelines for real-time search and summarization	Sub-second responses for multi-source queries; consistent throughput across enterprise workloads
Cost Efficiency	Reduce operational cost of manual review	Streamlined inference and model orchestration to minimize redundant processing	80% reduction in manual review effort
Control & Flexibility	Full governance, observability and compliance	RBAC, audit logging, secure pipelines, BYOC deployment	Enterprise-grade auditability
AI Orchestration/ Gateway	Centralized query and model management	Unified API, model catalog, monitoring, and feedback loops	Enables consistent operations across multi-source workflows
Model Lifecycle Management	Continuous performance and cost optimization	Build → Tune → Scale cycles with feedback-driven fine-tuning	Maintains consistent throughput and relevance across long-context summarization and RAG tasks

This combination of **speed, precision, cost efficiency, and enterprise control** helps knowledge workers access, summarize, and act on multi-modal data more efficiently across the enterprise.

SPEED

300ms p90 latency across hundreds of millions of records for long-context, multi-source queries

ACCURACY

5X faster insight generation, 80% reduction in manual review

SCALE

Millions of concurrent queries handled reliably

By orchestrating queries through the Fireworks AI Gateway, enterprises achieve fast, accurate, and scalable search and understanding across documents, chats, and knowledge bases, with full control and auditability.

Real-World Proof Points

Enterprises adopting Fireworks AI can expect faster, more accurate knowledge retrieval, reduced manual effort, and scalable, cost-efficient inference infrastructure compared with legacy or DIY solutions.

CRESTA

Cresta, a contact center platform, needed real-time guidance for agents and reliable decision workflows. Fireworks AI enabled:

- **5X faster call reviews** with automated next-step actions
- **80% reduction in manual review effort** for agents
- **Seamless fusion of long, multi-modal inputs** including chat, call transcripts, and knowledge base for accurate, actionable guidance

Read the Cresta case study.

These proof points show how Fireworks AI’s enterprise-grade search, long-context summarization, and RAG capabilities drive measurable gains in accuracy, efficiency, and business outcomes.

Patterns & Practices for Enterprise AI

03

People & Process

- Build and maintain expertise to fine-tune LLMs using enterprise data sources such as support logs, product catalogs, transcripts, and multi-modal inputs.
- Integrate AI-assisted search and summarization into knowledge workflows and support pipelines.
- Foster collaboration between ML, product, and operations teams to ensure models reflect real-world use cases.

Technology Implementation

- Select models, whether open, proprietary, or hybrid, tailored for enterprise search, long-context summarization, RAG, and multi-source retrieval.
- Deploy low-latency, high-throughput inference to power real-time query parsing, summarization, and retrieval.
- Use AI Gateway to decouple applications from underlying models, allowing seamless upgrades and flexibility.

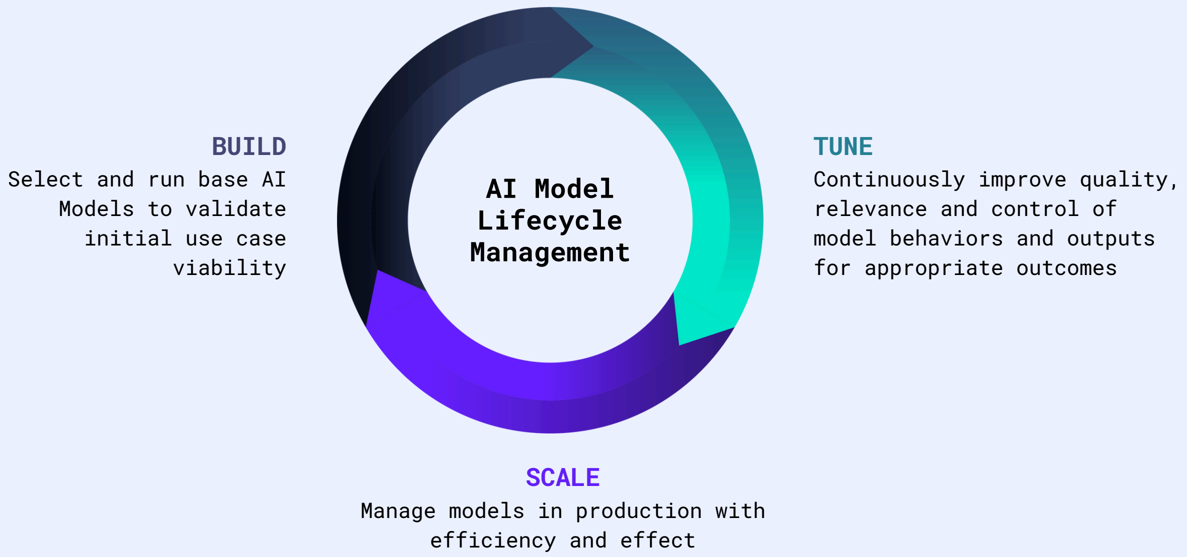
Risk Management

- Protect sensitive data across chat logs, documents, and transcripts with enterprise-grade security and governance.
- Ensure accuracy and reliability of search results and summaries through continuous evaluation and human-in-the-loop feedback.
- Minimize operational and vendor lock-in by owning fine-tuned models and deploying on Fireworks' flexible infrastructure.

Operational Patterns

- **AI Gateway**: Centralized orchestration for secure, scalable deployment of models for query understanding, summarization, and classification.
- **Model Lifecycle Management (Build, Tune, Scale)**: Streamlined process for model selection, fine-tuning, deployment, monitoring, and cost management.
- **Product-Model Co-Design**: Tight integration of search and summarization capabilities into enterprise products to accelerate time-to-insight and improve customer experience.

Model Lifecycle Management



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

AI Model Lifecycle Management

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

Search, summarization, and RAG are not just features, they are the backbone of customer experience and internal productivity. When AI systems fall short, the impact compounds across teams:

- **Customer Experience**: 75% of customers expect real-time, personalized support but most enterprises cannot deliver it at scale.
- **Productivity**: Knowledge workers lose up to 30% of their time searching for information or reworking low-quality summaries.
- **Cost Efficiency**: External APIs and untuned models lead to unpredictable spend and latency. Fine-tuned, in-house inference reduces cost by up to 100x.
- **Competitive Edge**: Enterprises that understand and act on data in real time move faster and deliver better experiences than those relying on legacy search and static reports.

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

Support search, knowledge assistants, enterprise RAG, or multi-modal summarization.

2

Run a pilot

Deploy Fireworks AI with preferred models to validate retrieval accuracy, summarization quality, and query throughput.

3

Scale and customize

Customize models with enterprise data, feedback loops, and domain-specific tuning.

4

Implement operational patterns

Implement AI Gateway, Model Lifecycle Management, and product-model co-design to ensure sustainable, enterprise-grade adoption.

Conclusion

05

The future of enterprise knowledge and insights is AI-assisted. Fireworks AI enables organizations to turn unstructured data into actionable intelligence, accelerating decisions, reducing manual effort, and delivering superior customer experiences. With Fireworks AI, your organization can deploy fine-tuned models optimized for your domain, scale to global traffic while meeting latency SLAs, own your AI stack to avoid vendor lock-in, and leverage unstructured enterprise data to drive measurable business outcomes.

Next Steps

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

Assess workflows

Identify high-value search and understanding use cases, such as support search, knowledge summarization, or long-context fusion.

Run a pilot

Deploy Fireworks AI with fine-tuned 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.