



Fireworks AI

Reviews, tips, and advice from real users



Powered by  **PeerSpot**

Contents

Product Recap..... 3 - 4

Valuable Features..... 5 - 12

Other Solutions Considered..... 13 - 14

ROI..... 15 - 17

Use Case..... 18 - 23

Setup..... 24 - 25

Customer Service and Support..... 26

Other Advice..... 27 - 30

Trends..... 31 - 32

About PeerSpot..... 33 - 34

Product Recap



Fireworks AI

Fireworks AI Recap

Fireworks AI uses advanced technologies to streamline operations and enhance user experience, catering to industry-specific requirements and driving innovation.

Fireworks AI integrates cutting-edge tools for data processing, offering seamless automation in managing complex workflows. It addresses industry needs through scalable solutions adaptable to personalized requirements. Fireworks AI ensures optimized performance, enhancing decision-making efficiency across businesses.

What are the crucial features of Fireworks AI?

- **Automated Workflow Management:** Streamlines complex processes efficiently.
- **Data Analytics:** Provides insightful analytics for informed strategy development.
- **Scalability:** Offers flexible scaling to match business growth trajectories.
- **Industry Integration:** Solutions tailored to the unique demands of different industries.

Which benefits and ROI should users evaluate?

- **Increased Efficiency:** Significant reduction in workload through automation.
- **Cost Savings:** Optimizes resources, leading to reduced operational costs.
- **Enhanced Decision-Making:** Supports data-driven decisions for better outcomes.
- **Scalability:** Facilitates growth without proportional cost increases.

Industries such as healthcare and finance benefit from Fireworks AI by streamlining data management, improving client interaction, and supporting compliance through automated document handling. Each deployment adjusts to specific sector demands, ensuring relevant application across diverse business environments.

Valuable Features

Excerpts from real customer reviews on PeerSpot:



“After introducing Fireworks AI's high-speed inference engine, I found that communication speed between agents was about twice as fast as before.”



Mdpman 김

Ai스페셜리스트매니저 at a tech vendor with 501-1,000 employees



“Fireworks AI has positively impacted our organization by making our AI features feel more production-ready instead of experimental.”



Hussain Gagan

FullStack Developer at EnactOn Technologies



“Fireworks AI has impacted us positively as it helps in offering us access to the open-source models by advancing fine-tuning options, a massive library where we can get information from the database that we can use in line with our company policy.”



Ike Christian

Assistant Banking Officer at Zenith-bank

- ✓ “Since using Fireworks AI, being part of their startup program has resulted in significant cost savings and has helped accelerate our development timeline.”



Vito Palermo

Co-Founder and CEO at Seismora

- ✓ “The inference itself works fine, with reasonable model breadth and speed, and my problem is entirely with billing, not the underlying technology.”



Verified user

Principal Software Engineer at a consultancy with 1-10 employees

- ✓ “Fireworks AI has positively impacted our organization by increasing our AI response time by twenty to fifty percent, as we now have AI agents and AI features that return answers twenty to fifty percent faster.”



Verified user

Product Manager at a tech vendor with 11-50 employees

- ✓ “Fireworks AI has positively impacted my organization by helping my productivity go up.”



Chaheti Jha

full stack developer intern at a tech services company with 1-10 employees

What users had to say about valuable features:

“The best features Fireworks AI offers are that the fine-tuning is really flexible and suitable. The customization is really good. I think it is providing so many models, which is the best part, and I do not need any GPU setup for using them. The number of models, which is very high, makes it very good.

The customization options in Fireworks AI are very good. I can adjust temperature there, or I can set a token limit there. That all helps me to customize my AI model and how I can use it better. That is really good.

Regarding the features of Fireworks AI, the integration in the back end and all is really good. Since I use it in my organization, the integration is pretty smooth.

Fireworks AI has positively impacted my organization by helping my productivity go up. It has saved me time. It has helped me to achieve more deadlines faster..”

Chaheti Jha

full stack developer intern at a tech services company with 1-10 employees

[Read full review](#) 

“The best features Fireworks AI offers for us are ease of use and connecting to the platform, along with the breadth of the models that they have available.

The ease of use was very straightforward to connect to Fireworks AI. I simply selected the model that I wanted and provided the API endpoint. We are currently working with Fireworks and are in discussions with them to begin moving from an API endpoint model to selecting specific individual points of presence that we can utilize across our mesh, particularly in North America and Asia.

Fireworks AI has positively impacted our organization as we are a member of their startup program. Being an early-stage startup, having access to their resources at this stage through their startup program was instrumental in allowing us to continue moving forward. We are also members of the NVIDIA Inception program and the AWS Activate program, and having access to these resources has enabled us to accelerate during this stage of our development.

Since using Fireworks AI, being part of their startup program has resulted in significant cost savings and has helped accelerate our development timeline..”

Vito Palermo

Co-Founder and CEO at Seismora

[Read full review](#) 

“The best feature Fireworks AI offers is speed. The speed of Fireworks AI stands out to me, as it is both the response time and scalability. The speed is very fast, so the inferencing happens very fast and we do not have to worry about the GPU running cost. Fireworks AI handles the scalability as well, so we have a few clients doing the inferencing at any point, and it is Fireworks AI's responsibility to scale up our GPU.

Fireworks AI has positively impacted our organization by increasing our AI response time by twenty to fifty percent, as we now have AI agents and AI features that return answers twenty to fifty percent faster. The engineering effort from the infrastructure side has been reduced, with our engineers not having to worry about hosting these trained models, resulting in a twenty to thirty percent reduction in engineering effort. The cost of hosting these models has gone down by fifteen to thirty-five percent.

We measure those improvements with Fireworks AI internally. Previously we used to host this model on our GPU on AWS cloud and knew the latency and inferencing time. After switching to Fireworks AI, we compared the response time and found the reduction in speed..”

Verified user

[Read full review](#) 

Product Manager at a tech vendor with 11-50 employees

“The most satisfying feature of Fireworks AI was the combination of efficient inference speed and stable function calling. The core of an autonomous agent system is the model's ability to interact with external tools in real time. Fireworks AI is not just fast in plain text generation; it is innovative in that it reduces the latency that occurs in the process where agents perform complex tasks and, through that, choose and call tools.

“Current LLM models have evolved from traditional foundation models into hybrid models, and while inference speed has improved, response time has become slower because they use things like Chain-of-Thought (CoT). To gain that inference speed, optimization of external function calling and similar aspects must be perfect; otherwise, the final answer will not come out quickly. By optimizing that through Fireworks, we were able to speed up the response time, which is a weakness of existing LLM models. A major advantage is that customers or business users can obtain answers quickly through text.

“In terms of metrics, in the case of health checkup data, it is at least two to three pages of PDF files or scans, so when a human reads it, it takes at least about one to three minutes. Using LLMs and Fireworks, we built an integrated system that can make a determination in about thirty seconds to one minute and then pass that result on to other systems based on that..”

Mdpman 김

Ai스페셜리스트매니저 at a tech vendor with 501-1,000 employees

[Read full review](#) 

“Fireworks AI has a very fast inference speed by providing minimal delay for our real-time applications.

“The best feature Fireworks AI offers is the multiple model tiers; it has very vast model applications where it's more about grasping the infrastructure component quickly, and I think it helps our team balance quality and cost.

“Having access to multiple model tiers helps our team balance quality and cost by giving us leverage where we can make options and look at what best suits our company and what we could use, which is beneficial because when you have multiple choices, you can tailor your approach and get what you actually need, so our options are not limited.

“Fireworks AI has impacted us positively as it helps in offering us access to the open-source models by advancing fine-tuning options, a massive library where we can get information from the database that we can use in line with our company policy.

“Fireworks AI helped us reduce costs, and it helps our team balance quality and improve customer satisfaction because interacting with us at that moment could provide them with easier access and quick answers and responses..”

Ike Christian

Assistant Banking Officer at Zenith-bank

[Read full review](#) 

“The best features Fireworks AI offers are speed and control over models. You can pick different open-source models and switch fairly easily. Additionally, the API layer feels developer-friendly.

“The API layer in Fireworks AI is developer-friendly because its consistency is a major factor. It follows standard OpenAI-compatible endpoints, which meant we could swap out models or integrate new ones without rewriting our entire service layer. For example, when we wanted to test a new Llama 3 variant against our existing deployment, it was literally just a one-line change in our configuration.

“The fine-tuning and customization options in Fireworks AI are useful, even though we didn't go very deep into them. The ability to experiment with multiple models in one setup is underrated. It saves time when comparing outputs. Fireworks AI has positively impacted our organization by making our AI features feel more production-ready instead of experimental. Teams became more confident shipping AI-based features, which also reduced dependency on a single vendor.

“Since we started using Fireworks AI, we've seen around a 20 to 30% improvement in latency for some endpoints. Cost-wise, we've achieved approximately 15 to 25% savings depending on the model we use. Nothing extraordinary, but definitely meaningful..”

Hussain Gagan

FullStack Developer at EnactOn Technologies

[Read full review](#) 

Other Solutions Considered

“We didn't switch from our primary inference, which runs on Amazon Bedrock. We were trialing Fireworks alongside it for its open-source model selection..”

Verified user

Principal Software Engineer at a consultancy with 1-10 employees

[Read full review](#) 

“We were mostly using hosted APIs from bigger providers before using Fireworks AI. We switched mainly for cost control and flexibility with models. I also wanted better performance for certain use cases..”

Hussain Gagan

FullStack Developer at EnactOn Technologies

[Read full review](#) 

“Before choosing Fireworks AI, we evaluated other options, including Claude and Groq AI, but then we had to look at the options available to us, considering the cost-effectiveness and the license model..”

Ike Christian

Assistant Banking Officer at Zenith-bank

[Read full review](#) 

“Before choosing Fireworks AI, we looked at things such as Together AI and some direct cloud GPU setups. We also briefly considered sticking with OpenAI APIs. Fireworks AI felt like a good middle ground..”

Hussain Gagan

FullStack Developer at EnactOn Technologies

[Read full review](#) 

“We used Amazon Bedrock as our main platform and other hosted open-source inference providers. We were evaluating Fireworks as an additional option for OSS models..”

Verified user

Principal Software Engineer at a consultancy with 1-10 employees

[Read full review](#) 

ROI

Real user quotes about their ROI:

“There is a return on investment as Fireworks AI's accuracy helps us with our turnaround time, and I think that's a return on investment for us. It saves us cost as well..”

Ike Christian

Assistant Banking Officer at Zenith-bank

[Read full review](#) 

“I have seen a return on investment with Fireworks AI as we have saved thousands of dollars. We do not need any additional employees, as we have been utilizing AI to avoid hiring at this stage, and time to market has been accelerated by six months..”

Vito Palermo

Co-Founder and CEO at Seismora

[Read full review](#) 

“Fireworks AI has saved my team around half of what we used to take because initially, we had to manually research all the models. Now, we can just use it, which saves the time of searching and using each one and then deciding which one to go with..”

Chaheti Jha

[Read full review](#) 

full stack developer intern at a tech services company with 1-10 employees

“The return on investment with Fireworks AI has been decent. We've experienced faster iteration and slightly lower costs, as well as reduced engineering time spent managing infrastructure ourselves. The savings are not huge, but definitely worth it..”

Hussain Gagan

[Read full review](#) 

FullStack Developer at EnactOn Technologies

“I have seen a return on investment with Fireworks AI. The speed of the response time has improved, and on the ROI side, we do not have to worry about engineering effort, leading to a twenty to thirty percent reduction in the engineering time for data engineers working on infrastructure..”

Verified user

[Read full review](#) 

Product Manager at a tech vendor with 11-50 employees

“The companies we usually work with are enterprise-level companies in Korea, so we cannot really provide actual company names or detailed data. However, in order to make decisions, when customers have certain requirements, we can quickly create agents for them, and in connecting those agents through connections like A2A and MCP, Fireworks has helped a lot. As a result, we experienced an innovative situation where time spent on simple repetitive tasks was reduced by over sixty percent. Additionally, task processing speed improved by about thirty percent. This naturally led to cost reduction or cost optimization. To be specific, if one person used to complete one unit of work before, it is now optimized so that they can do one point five or more units of work..”

Mdpman 김

Ai스페셜리스트매니저 at a tech vendor with 501-1,000 employees

[Read full review](#) 

Use Case

“Our main use case for Fireworks AI is running LLM-based APIs for things like summarization and internal search. We didn't want to rely fully on a closed model, so Fireworks AI helped us run an open-source model with decent performance. It fits well for production APIs where latency matters.

“We also experimented with embeddings and some lightweight fine-tuning in Fireworks AI. Not everything made it to production, but it was useful for testing different models quickly. It's good for teams that want flexibility rather than a fixed model..”

Hussain Gagan

FullStack Developer at EnactOn Technologies

[Read full review](#) 

“My main use case for Fireworks AI is evaluating it as our inference substrate for distributed inference networking.

I am using Fireworks AI for distributed inference networking by taking various AI workloads from very small to very large workloads, orchestrating those workloads across the edge of the network and utilizing various Fireworks AI API endpoints to provide the inference for each level of that GPU workload.

I believe we are fairly unique in that we are building a control plane for the agentic internet and utilizing Fireworks AI as the substrate, regardless of where the original workload request was initiated..”

Vito Palermo

Co-Founder and CEO at Seismora

[Read full review](#) 

“My main use case for Fireworks AI is typically for fine-tuning or choosing what models I want to use for my project. It is good for letting me use all the models, and it acts as a playground so I can even test them.

Recently, I used Fireworks AI to choose between models for a project. It was an assignment for one of the YC startups, and I wanted to see which model I should use for the audio transcript. I tested all of them using Fireworks AI, and I ended up choosing GPT 120B because of the help of Fireworks AI.

About my main use case for Fireworks AI, it is interesting that it lets me choose. There are so many models available. I think there were 300-plus models, which is really impressive..”

Chaheti Jha

full stack developer intern at a tech services company with 1-10 employees

[Read full review](#) 

“We use Fireworks AI as a powerful tool that helps us in building and scaling our customized AI application model for our business.

“We wanted to create a customer base where our customers could interact with us through our chatbot, and Fireworks AI helped us in scaling through that by customizing the AI application model for our business to suit our customer's taste.

“Fireworks AI helped us customize the application for our customers by creating strong platform leverage in the ecosystem around it, and that's what we leveraged by it providing us multiple model tiers, which we use in creating that customized AI application for our teams..”

Ike Christian

Assistant Banking Officer at Zenith-bank

[Read full review](#) 

“Fireworks AI hosts the large language model that we have trained, which is a large language model on behavior science and human capital data. We have a culture operating system, so whenever we need to do some kind of inferencing that goes via our large language model that we have trained, Fireworks AI is hosting the LLM that we have trained. Whenever we need AI capabilities in our product, we fire a query or API call to Fireworks AI and then we get a response, with the inferencing happening on Fireworks AI model.

Building AI capabilities on the culture operating system data with Fireworks AI allows our managers to query the LLM for insights. For example, if a manager wants to know what their team trust score is right now, it will query the LLM and then it will get the answer. If a manager wants to deep dive into how they can improve, the inferencing will happen on Fireworks AI and generate an answer to improve the trust score or any vital sign score that is being generated by our LLM that is running on Fireworks AI..”

Verified user

Product Manager at a tech vendor with 11-50 employees

[Read full review](#) 

“Fireworks AI was my choice because, with keywords like low latency, enterprise automation, and corporate automation, it was able to provide expert-level insights. I built a system for memory, autonomous collaboration, and controlled generation in the projects where I applied Fireworks AI. After introducing Fireworks AI's high-speed inference engine, I found that communication speed between agents became about twice as fast compared to before. The function calling capability for agents to invoke external tools was very stable, and I confirmed that it was possible to perfectly implement complex workflows that query and reflect enterprise data in real time. This was the most decisive differentiator that allowed us to practically apply AI automation in enterprise environments.

“A concrete example of how Fireworks AI helped would be a system with an access control system. By developing an agent, we built a system where, when people enter a factory inside a company, we receive their medical examination documents and, based on the information in those documents, we can determine whether to approve or deny their access registration or entry. By automating that, we quickly verify employees' health conditions and can impose entry restrictions..”

Mdpman 김

Ai스페셜리스트매니저 at a tech vendor with 501-1,000 employees

[Read full review](#) 

Setup

The setup process involves configuring and preparing the product or service for use, which may include tasks such as installation, account creation, initial configuration, and troubleshooting any issues that may arise. Below you can find real user quotes about the setup process.

“The initial setup was fairly easy. It took about eight to ten days, including integrating it into our solution, testing, and moving from scratch to production..”

Verified user

[Read full review](#) 

Senior Software Development Engineer at a tech services company with 1-10 employees

“My experience with pricing, setup cost, and licensing was a bit difficult, but the pricing was cost-effective for us, so we were able to get it done. I think it is renewable every year, so that's not a challenge for us..”

Ike Christian

[Read full review](#) 

Assistant Banking Officer at Zenith-bank

“My experience with Fireworks AI regarding pricing, setup cost, and licensing is good, as it is pretty easy and the UI was simple. Our engineer was able to deploy it easily with no support needed from Fireworks—it was straightforward..”

Verified user

[Read full review](#) 

Product Manager at a tech vendor with 11-50 employees

“My experience with pricing, setup cost, and licensing for Fireworks AI was fine. It was easy, and currently, because of the startup program, we are operating off of credits that were provided by Fireworks AI and AWS..”

Vito Palermo

[Read full review](#) 

Co-Founder and CEO at Seismora

Customer Service and Support

“I didn't interact heavily with Fireworks AI's customer support, but when we did, responses were decent. Responses were not super fast, but helpful enough..”

Hussain Gagan

FullStack Developer at EnactOn Technologies

[Read full review](#) 

“I think the customer support is good, but we did not have any chance to connect with the support team. The documentation was thorough and complete, so it is straightforward and you will find all the answers there..”

Verified user

Product Manager at a tech vendor with 11-50 employees

[Read full review](#) 

“The customer support for Fireworks AI is average.

I would rate the customer support with answers being a ten and timeliness a seven on a scale of one to ten..”

Vito Palermo

Co-Founder and CEO at Seismora

[Read full review](#) 

Other Advice

“The inference technology itself is fine, but the billing is a serious risk. We were billed \$40,000 on a pay-as-you-go plan for usage we never knowingly incurred and which our own usage dashboard does not reflect. Until there are enforceable spend limits, real-time alerts, and transparent metering, evaluate the billing exposure very carefully before relying on this in production..”

Verified user

Principal Software Engineer at a consultancy with 1-10 employees

[Read full review](#) 

“Based on my experience, I give Fireworks AI a rating of seven out of ten. Due to the fact that various connections are still somewhat lacking, I deducted about three points for this rating. Since we are basically a CSP partner, we use a public cloud as our base. However, depending on customer needs, enterprise-level customers want to apply it via their own in-house LLM or local LLM, so the hybrid concept is also under consideration. Our company fundamentally aims for a multi-cloud approach, so we use GCP, AWS, and Azure all together. Currently, I am mainly focused on the Azure side, so we deal only with Azure-based systems..”

Mdpman

이스페셜리스트매니저 at a tech vendor with 501-1,000 employees

[Read full review](#) 

“My advice regarding using Fireworks AI would be to go in with a clear use case instead of just experimenting randomly. Additionally, spend time understanding model selection, as that makes a big difference. Don't expect everything to work

perfectly out of the box.

“Fireworks AI is a good option if you want more control over your AI stack without managing everything yourself. Fireworks AI is not perfect, but definitely practical for real-world use. I found Fireworks AI to be a valuable tool in streamlining our workflows. I would definitely recommend exploring its capabilities for businesses looking to enhance their operations. I rated this review an eight overall..”

Hussain Gagan

FullStack Developer at EnactOn Technologies

[Read full review](#) 

“I advise others looking into using Fireworks AI to use it because the ecosystem around Fireworks creates strong platform leverage and provides multiple model tiers that can let their team balance quality and cost.

“Regarding Fireworks AI's AI capabilities, its governance and security policy is deeply rooted, following global standards, and I think that's a fair offering from them.

“Regarding Fireworks AI's AI capabilities and the reliability of the output, this has not posed any challenge for us. It's good and satisfactory. I rated this review eight out of ten overall..”

Ike Christian

Assistant Banking Officer at Zenith-bank

[Read full review](#) 

“Regarding Fireworks AI's AI capabilities, its accuracy and reliability are pretty accurate, as the quality of output depends on the LLM that we are hosting on this platform. We have trained our LLM and tested it, and speed is something that has

improved by hosting our model on Fireworks AI.

Fireworks AI's governance and security are pretty secure, as we have all the compliance certificates, including SOC 1 and SOC 2.

For others looking into using Fireworks AI, I advise you to know your costs if you are hosting. If you have one customer for in-house deployment, you do not have to worry about hosting. If you have few customers who want to use privately developed LLMs, then Fireworks AI is a very good place. I would rate my overall experience with Fireworks AI a ten out of ten..”

Verified user

Product Manager at a tech vendor with 11-50 employees

[Read full review](#) 

“My advice for others looking into using Fireworks AI is that if they are initially trying an AI model, I think it is a good option, but they can do more research and be better at security and all the other things we discussed. They have a huge library of all the open-source models. As I have already said, their fine-tuning features are very good. It is really good for a developer, but it is not that good for a businessman or someone who is non-technical.

Before we wrap up, I think Fireworks AI should have good build and integration so that a developer does not have to do a setup. I think it is similar to tools such as [Zendesk](#).

I think Fireworks AI handles security and data privacy in my organization pretty well, but security can be a concern. It does have unusual traffic patterns, and it would be better if the vulnerabilities are properly monitored.

The performance of Fireworks AI in terms of speed and reliability was good. It is pretty reliable, and it makes me work faster.

My overall rating for this product is an eight out of ten..”

Chaheti Jha

full stack developer intern at a tech services company with 1-10 employees

[Read full review](#) 

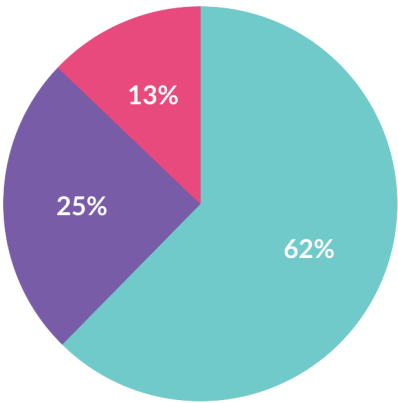
Top Industries

by visitors reading reviews

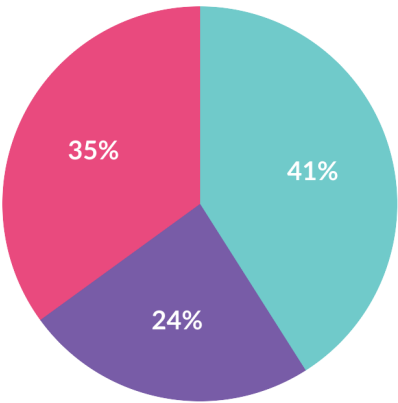


Company Size

by reviewers



by visitors reading reviews



Large Enterprise Midsized Enterprise Small Business

About this buyer's guide

Thanks for downloading this PeerSpot report.

The summaries, overviews and recaps in this report are all based on real user feedback and reviews collected by PeerSpot's team. Every reviewer on PeerSpot has been authenticated with our triple authentication process. This is done to ensure that every review provided is an unbiased review from a real user.

Get a custom version of this report... Personalized for you!

Please note that this is a generic report based on reviews and opinions from the collective PeerSpot community. We offer a [customized report](#) of solutions recommended for you based on:

- Your industry
- Company size
- Which solutions you're already considering

The customized report will include recommendations for you based on what other people like you are using and researching.

Answer a few questions in our short wizard to get your customized report.

[Get your personalized report here](#)

About PeerSpot

PeerSpot is the leading review site for cloud, AI, and business software. We created PeerSpot to provide a trusted platform to share information about software, applications, and services. Since 2012, over 22 million people have used PeerSpot to choose the right software for their business.

PeerSpot helps tech professionals by providing:

- A list of products recommended by real users
- In-depth reviews, including pros and cons
- Specific information to help you choose the best vendor for your needs

Use PeerSpot to:

- Read and post reviews of products
- Access over 30,000 buyer's guides and comparison reports
- Request or share information about functionality, quality, and pricing

Join PeerSpot to connect with peers to help you:

- Get immediate answers to questions
- Validate vendor claims
- Exchange tips for getting the best deals with vendor

Visit PeerSpot: www.peerspot.com

PeerSpot

244 5th Avenue, Suite R-230 • New York, NY 10001

reports@peerspot.com

+1 646.328.1944