Product Summary

AI Accelerator

The AI Accelerator enhances large language model (LLM) application performance by caching semantically similar responses globally, reducing API calls to LLM providers.

Faster AI starts with semantic caching

- ✔ Boost Performance for User Engagement
- ✔ Cut Costs by Optimizing API Requests
- ✔ Enable Seamless Developer Integration

9x

Faster AI APIs in just 5 minutes

The AI Accelerator enhances large language model (LLM) application performance by caching semantically similar responses globally, reducing API calls to LLM providers. It empowers developers to save costs, improve latency, and innovate faster with minimal effort.

Boost Performance for User Engagement

- Increase application responsiveness by improving LLM performance with faster, smarter caching.

Cut Costs by Optimizing API Requests

- Save on AI computation costs by reducing API token usage by 20% through intelligent semantic caching at the edge.

Enable Seamless Developer Integration

- Implement semantic caching with as little as one line of code, maximizing developer productivity and efficiency without additional engineering investments.

Learn more at www.fastly.com