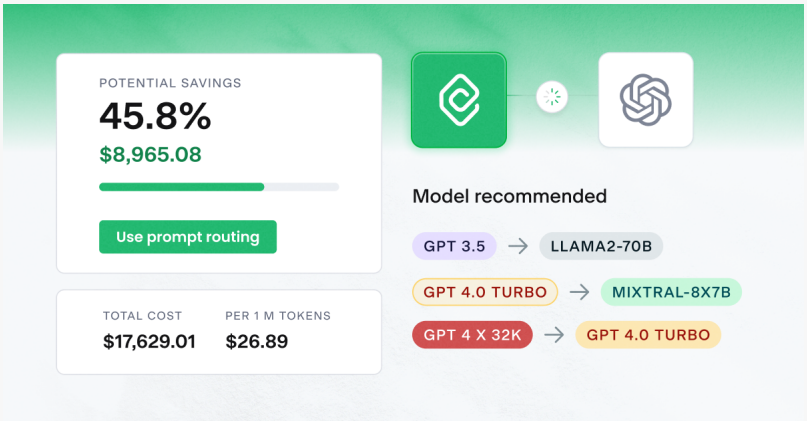
CAST AI Automatically Reduces the Cost of Deploying Generative AI

LAS VEGAS, NV, UNITED STATES, April 9, 2024 /EINPresswire.com/ -- Today at Google Cloud Next '24, CAST AI, the leading Kubernetes automation platform, announced AI Optimizer - a new service that automatically reduces the cost of deploying Large Language Models (LLMs). It integrates with any OpenAI-compatible API endpoint and automatically identifies the LLM across commercial vendors and open source that offers the most optimal performance and the lowest inference costs. It then deploys the LLM on CAST AI optimized Kubernetes clusters, unlocking unprecedented generative AI savings. AI Optimizer comes with deep insights into model usage, fine-tuning costs, and explainability in optimization decisions – including model selection.



CAST AI's new AI Optimizer service automatically reduces the cost of deploying Large Language Models.

Generative AI and LLM adoption are growing rapidly, but the diversity of model choice, compute availability, and cost of running models at scale cause sticker shock among organizations that are early adopters of the technology. Most commercially available LLMs leverage usage-based pricing, which means that spend increases at the same rate as an organization's adoption.

"According to Futurum Intelligence research, the worldwide adoption of AI used in development tools will exceed $3.6 billion in 2024," said Paul Nashawaty, Practice Lead for Application Development and Modernization at The Futurum Group. "However, key barriers to widespread adoption persist, including computational resource requirements and the costs they generate. Addressing this challenge will be critical in harnessing the full potential of Generative AI for diverse industries."

"Not all large language models are created equal. Some may be more efficient than others in terms of cost, performance, and accuracy across numerous use cases. But organizations haven't had a way to identify and deploy the most optimal model in terms of performance and cost,"

said CAST AI Co-Founder and CTO Leon Kuperman. “We’re addressing that gap by extending our expertise in automation to LLM optimization. What makes AI Optimizer so compelling is that it significantly reduces costs without requiring organizations to swap their existing technology stacks or even change a line of application code, which will help democratize generative AI.”

To identify the LLM that offers the most optimal performance and the lowest inference costs, AI Optimizer analyzes the cost associated with specific users and API keys, overall usage patterns, the balance of input versus output tokens, and the potential cost benefits of model fine-tuning. It ensures the automated selection of the most available GPUs, including Spot instances. AI Optimizer also creates budgets and alerts for customers based on these dimensions. Combined with the most efficient autoscaler, organizations can expect substantial cost reductions on AWS, Azure, and GCP.

“The combination of our large language model orchestration framework and the deployment of models on Kubernetes clusters ultra-optimized by CAST AI will unlock the most efficient and scalable deployment of LLM today,” said Kuperman.

The ability to automatically identify the LLM that offers the most optimal performance and the lowest inference costs is generally available today. Automated deployment of the LLM on CAST AI optimized Kubernetes clusters will be generally available in late Q2. Visit this page to request to join the private beta and use automated deployment, or visit CAST AI’s booth (1450) at Google Cloud Next ‘24.

About CAST AI
CAST AI is the leading Kubernetes automation platform that cuts AWS, Azure, and GCP customers’ cloud costs by over 50%. CAST AI goes beyond monitoring clusters and making recommendations. The platform utilizes advanced machine learning algorithms to analyze and automatically optimize clusters in real time, reducing customers’ cloud costs, and improving performance and reliability to bolster DevOps and engineering productivity.

Learn more: https://cast.ai/

Erika Rosenstein
CAST AI
erika@cast.ai
Visit us on social media:
Twitter
LinkedIn