Real-time insights into AI latency with voice-enabled observability

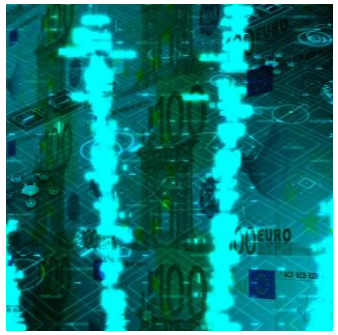
OPSVOICE ASSISTANT: UNMASKING THE BLACK BOX OF GENERATIVE AI LATENCY – INVESTOR PITCH DECK

AGENDA OVERVIEW

- The Problem: Lack of Transparency in AI-Powered Applications
- The Solution: Introducing OpsVoice – A Voice-Activated Observability Assistant
- Technical Workflow: How OpsVoice Works Seamlessly
- Demonstrated Value: Real World Proof of Effectiveness
- Modern Tech Stack: Designed for Scale and Efficiency
- Market Opportunity: Bridging the Gap in GenAI Observability
- Business Model: Scalable SaaS for Enterprise Observability
- The Team: Experts in Distributed Systems and Cloud Operations
- Call to Action: Join Us in Transforming AI Observability

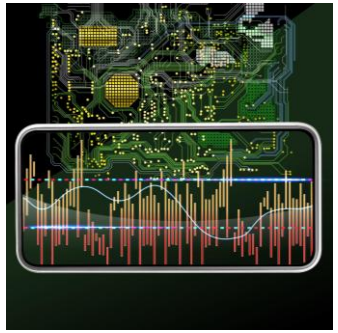
THE PROBLEM: LACK OF TRANSPARENCY IN AI-POWERED APPLICATIONS

GROWING ADOPTION OF LLMS INTRODUCES OBSERVABILITY CHALLENGES



Complex Latency Sources

LLMs introduce multiple latency sources making performance monitoring more complex than before.



Limitations of Traditional Tools

Conventional monitoring tools struggle to fully capture performance metrics of modern LLM architectures.

Blind Spots in Diagnostics

Incomplete observability leads to blind spots, impacting the ability to diagnose performance issues accurately.

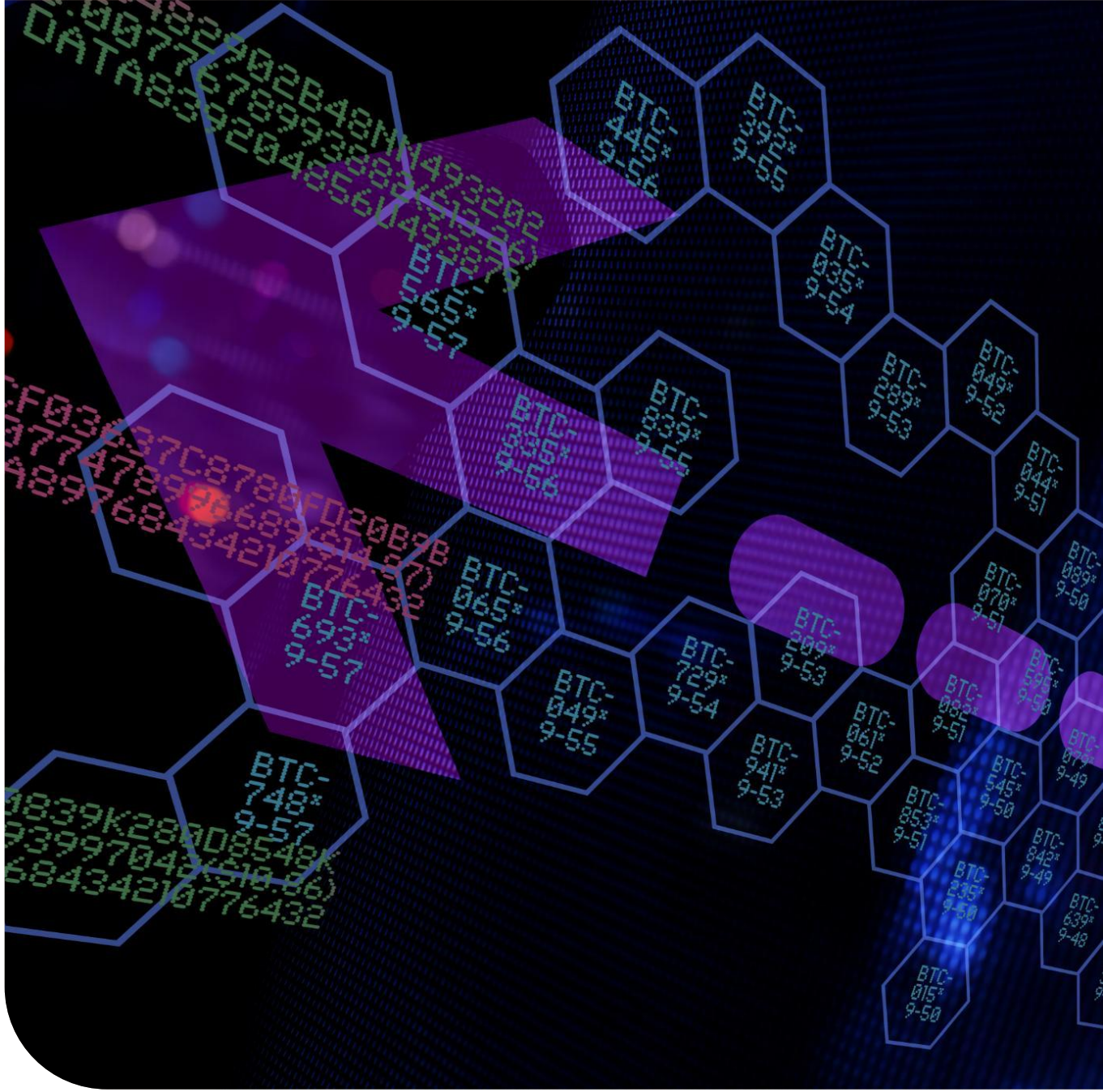
SRES STRUGGLE TO DIAGNOSE IF LATENCY IS FROM NETWORK OR MODEL

Challenges in Latency Diagnosis

SREs face complex challenges distinguishing latency caused by network issues from delays in AI model processing.

Impact on Incident Response

Difficulty in diagnosing latency origins slows incident response and prolongs system downtime.



HIGH LATENCY LEADS TO POOR USER RETENTION AND LOST REVENUE

Impact of Latency on Experience

High latency causes slow AI responses, which deteriorates the overall user experience and satisfaction.

User Retention Decline

Frustration from delays leads to reduced user retention as customers abandon the service.

Revenue Loss

Declining retention caused by latency issues results in negative impacts on revenue streams.

THE SOLUTION: INTRODUCING OPSVOICE – A VOICE-ACTIVATED OBSERVABILITY ASSISTANT

VOICE INTERFACE ENABLES NATURAL, REAL-TIME QUERIES

Hands-Free Interaction

Voice commands allow users to interact without using their hands, improving convenience and efficiency.

Immediate Data Access

Real-time voice queries provide instant access to observability data for faster troubleshooting.

VISUALIZES 'AI LAYER' PERFORMANCE VERSUS 'APP LAYER'

AI Model Latency

Shows the delay caused specifically by the AI model processing to identify performance bottlenecks.

Application Layer Delays

Highlights delays occurring in the application layer separate from AI processing for clearer diagnosis.

Actionable Insights

Empowers teams to optimize system performance based on clear distinctions between latency sources.

PINPOINTS LATENCY SOURCES ON DEMAND FOR FASTER INCIDENT RESPONSE

Rapid Bottleneck Identification

Pinpointing latency sources on demand enables immediate focus on problematic areas for faster troubleshooting.

Reduced Resolution Time

Quick identification of issues significantly reduces the mean time to resolution during incidents.

Minimized Downtime Impact

Faster incident response minimizes downtime, improving system reliability and user experience.

TECHNICAL WORKFLOW: HOW OPSVOICE WORKS SEAMLESSLY

USER SPEAKS QUERY TO OPSVOICE ASSISTANT

Voice-Activated Diagnostics

Users can start diagnostics by speaking directly to the voice assistant, making the process faster and more intuitive.

Simplified User Interaction

Speaking queries simplifies the user interface by reducing need for manual input and enhancing user experience.

GOOGLE SPEECH-TO-TEXT RAPIDLY CONVERTS AUDIO

Fast Audio Transcription

The API quickly converts spoken words into written text with minimal delay, improving user experience.

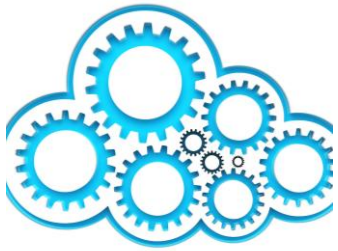
Accurate Voice Recognition

High accuracy in recognizing speech ensures reliable conversion of voice input to text queries.

PYTHON BACKEND ROUTES TO VERTEX AI GEMINI AND CAPTURES DATADOG TRACES

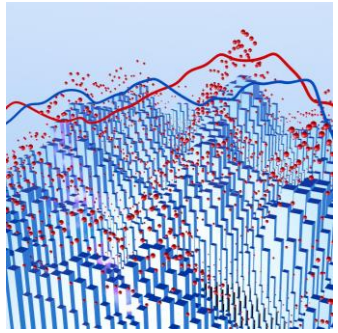
Python Backend Processing

The Python backend handles incoming queries and routes them to appropriate AI services for processing.



AI Integration with Vertex AI Gemini

Integration with Vertex AI Gemini enables advanced AI insights and intelligent response generation.



Datadog Trace Capture

Datadog APM captures distributed tracing data, providing detailed monitoring and performance insights.

DEMONSTRATED VALUE: REAL WORLD PROOF OF EFFECTIVENESS

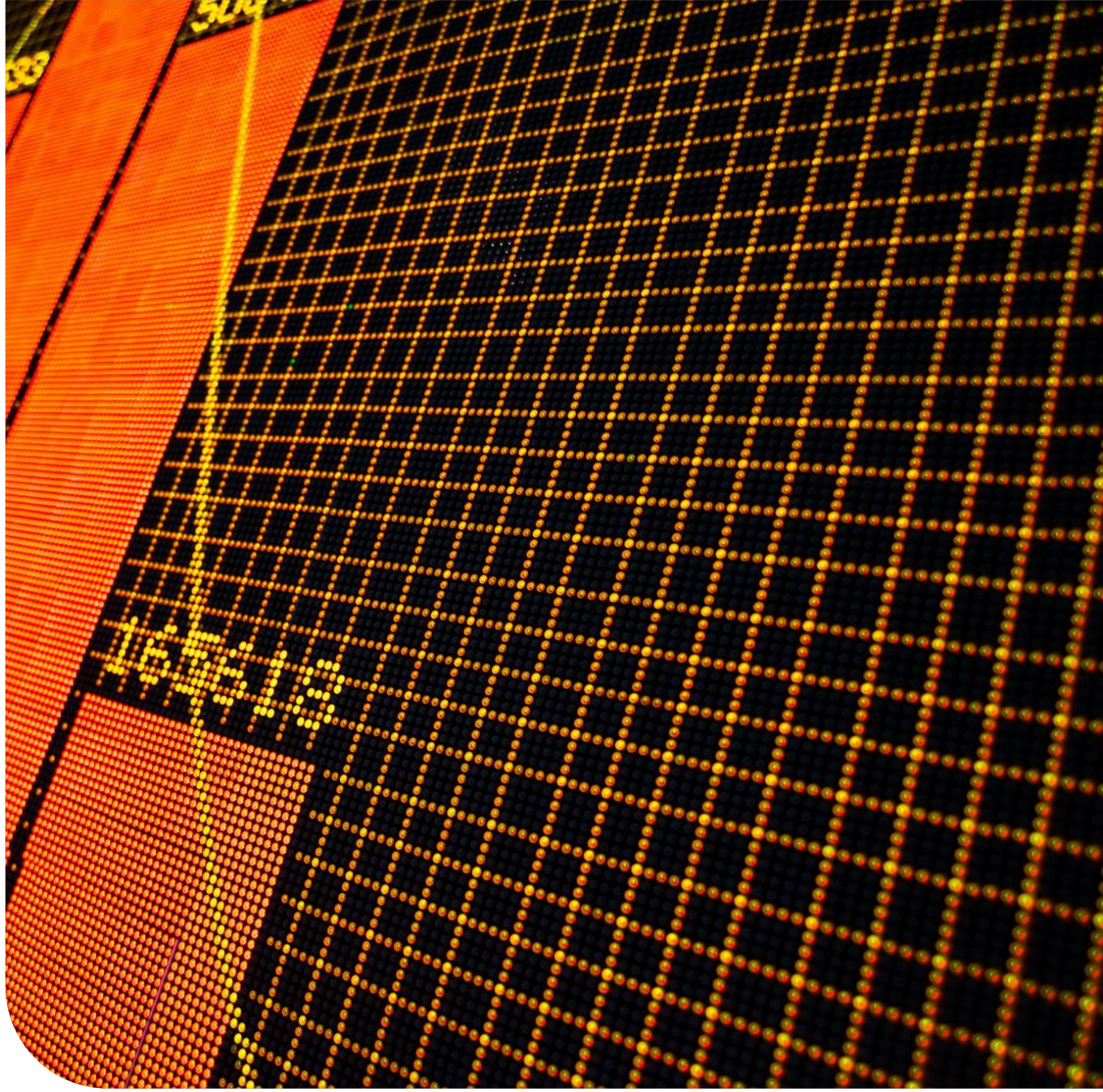
LIVE DEMO REVEALED 33-SECOND LATENCY ANOMALY

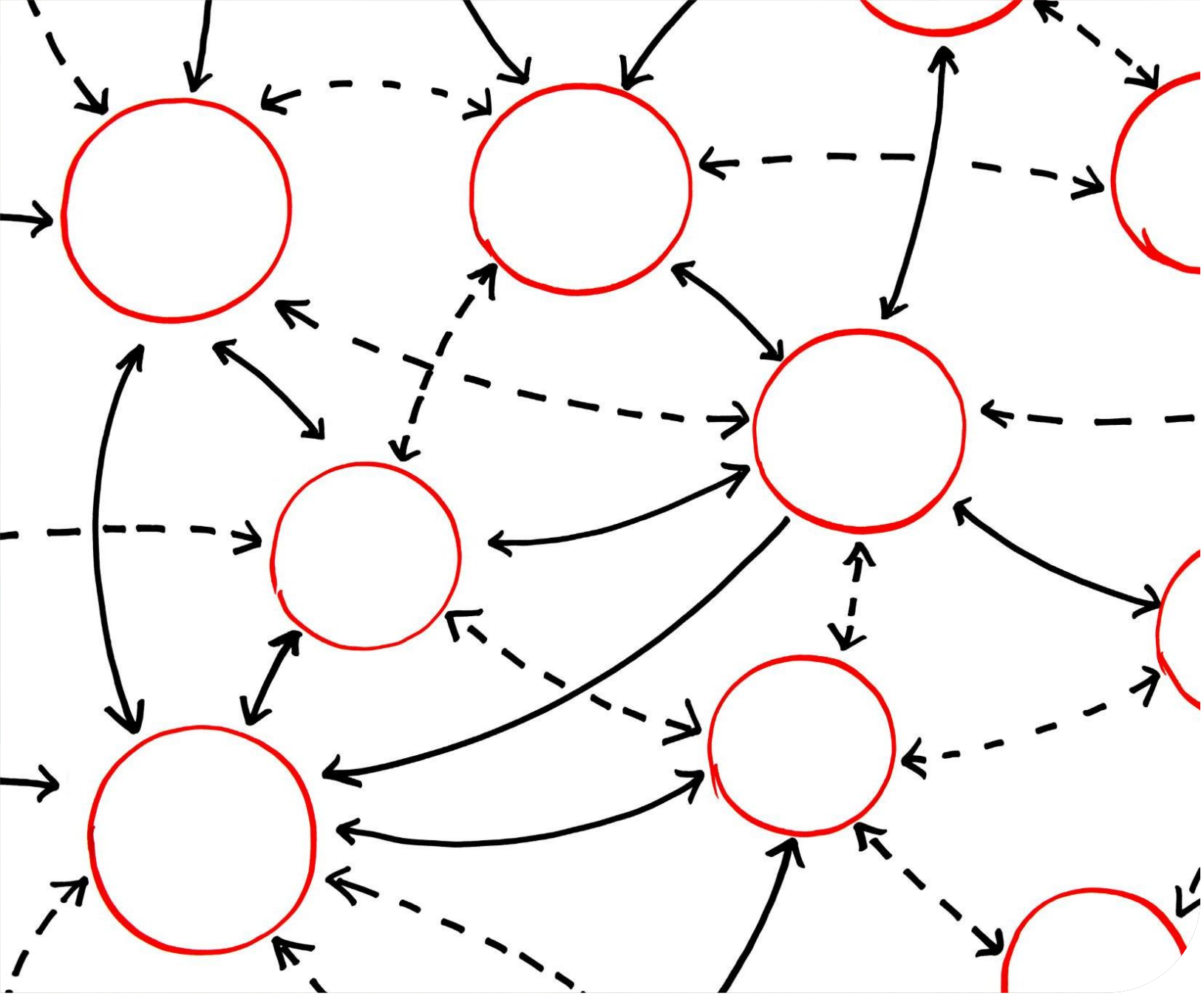
Latency Detection

OpsVoice detected a critical 33-second latency during system testing that could impact performance.

Importance of Monitoring

Continuous monitoring helps identify delays that might otherwise be overlooked or wrongly attributed.





OPSVOICE ACCURATELY TRACED ROOT CAUSE TO GEMINI MODEL

Accurate Root Cause Analysis

The system precisely pinpointed the source of latency within the AI model, ensuring targeted troubleshooting.

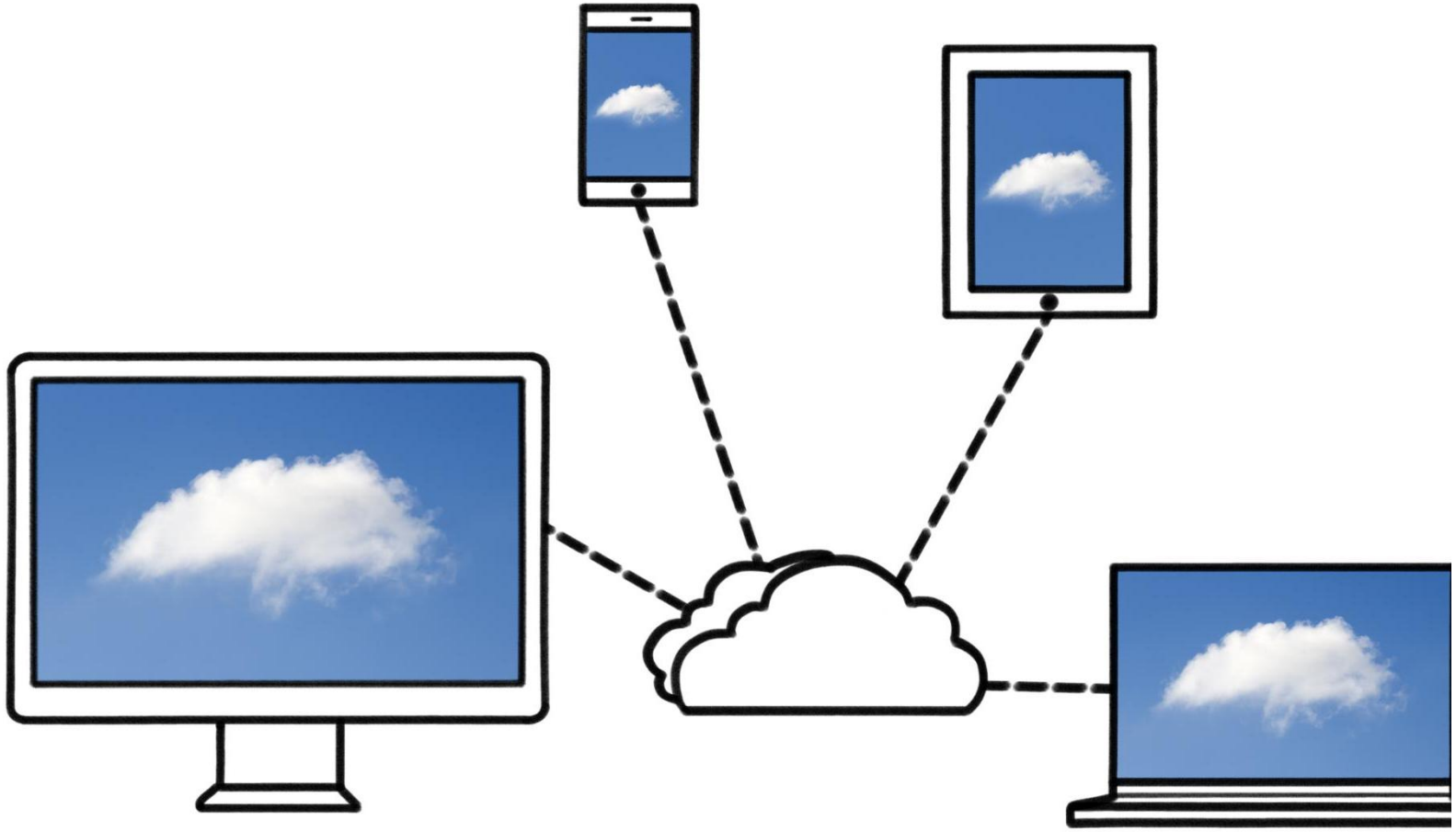
Latency Origin in AI Model

Latency was found specifically within the AI model, excluding other infrastructure elements from the issue.

Infrastructure Component Differentiation

The system differentiated between AI model issues and other infrastructure components to isolate the problem efficiently.

MODERN TECH STACK: DESIGNED FOR SCALE AND EFFICIENCY



GOOGLE CLOUD RUN FOR SCALABLE, SERVERLESS INFRASTRUCTURE

Serverless Architecture

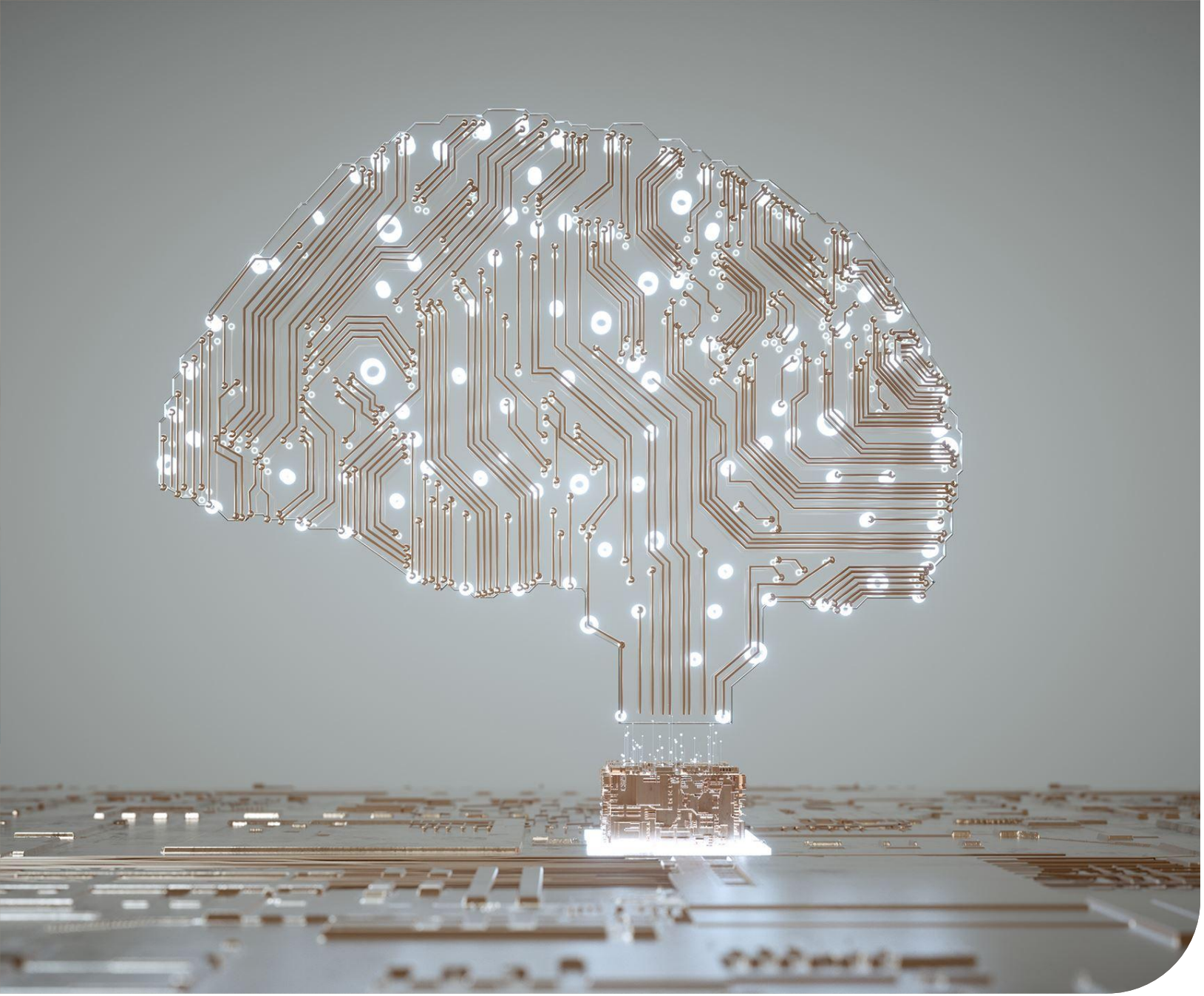
Google Cloud Run provides a serverless platform that abstracts infrastructure management for developers.

Auto-Scaling Features

Cloud Run automatically scales backend services based on demand, ensuring efficient resource use and performance.

Cost-Effective Management

Pay only for the compute time used, optimizing costs with Cloud Run's efficient serverless model.



GOOGLE VERTEX AI (GEMINI PRO) POWERS LLM FEATURES

Advanced AI Platform

Google’s Vertex AI Gemini Pro offers a powerful platform for building and deploying advanced AI models.

Natural Language Processing

The system enhances natural language understanding and processing capabilities for complex text analysis.

Large Language Models

Vertex AI Gemini Pro powers large language models enabling sophisticated language generation and comprehension.

INTEGRATED DATADOG APM FOR COMPREHENSIVE DISTRIBUTED TRACING

Detailed Distributed Tracing

Datadog APM provides detailed tracing capabilities to monitor distributed systems effectively.

Latency Source Identification

The integration helps pinpoint latency sources, enabling faster issue resolution in complex systems.

MARKET OPPORTUNITY: BRIDGING THE GAP IN GENAI OBSERVABILITY

TRADITIONAL DEVOPS TOOLS LACK AI-LAYER INSIGHT

Limitations of DevOps Tools

Traditional DevOps tools lack specialized metrics to monitor AI-specific latency effectively, causing blind spots in performance analysis.

Critical Visibility Gap

The absence of AI-layer latency insights creates a critical monitoring gap, impacting timely detection and resolution of AI system issues.

EXPLOSION OF GENAI ADOPTION FUELS NEW MONITORING NEEDS

Rapid GenAI Adoption

The rapid increase in generative AI adoption is transforming enterprise operations and driving new challenges.

Monitoring Challenges

Enterprises must develop effective monitoring strategies to ensure optimal AI performance and reliability.

Optimization Imperatives

Optimizing generative AI models is critical to maximizing value and minimizing operational risks in businesses.

UNIQUE METRIC: 'TOKENS PER SECOND' FOR AI PERFORMANCE TRACKING

Definition of Tokens per Second

'Tokens per Second' measures the speed at which an AI model processes language tokens, indicating throughput.

Importance for AI Efficiency

This metric helps assess efficiency and performance of AI models in real-time applications and workloads.

BUSINESS MODEL: SCALABLE SAAS FOR ENTERPRISE OBSERVABILITY

SUBSCRIPTION-BASED SAAS TARGETING ENTERPRISE SRE TEAMS

Subscription Service Model

The product operates on a subscription model to provide continuous access and updates for enterprise users.

Focus on SRE Teams

Designed specifically for Site Reliability Engineering teams to enhance system reliability and performance monitoring.

AI Observability at Scale

Utilizes AI-powered observability tools to manage complex systems and large-scale environments effectively.

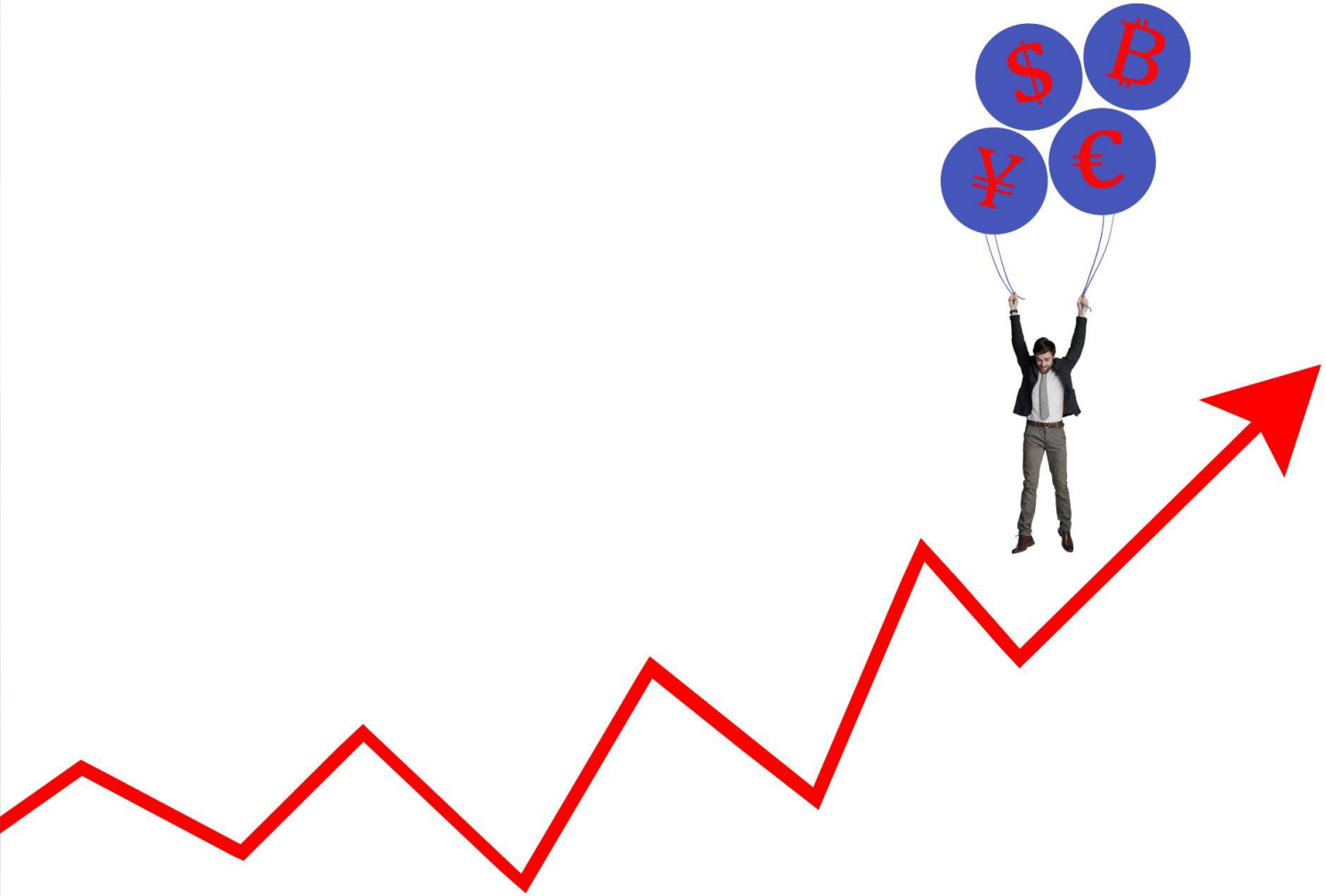
FUTURE FEATURES: COST-PER-QUERY ANALYTICS FOR BUDGET CONTROL

Detailed Cost Analytics

Future tools will provide detailed breakdowns of costs incurred per AI query for better transparency.

Budget Management

Organizations can use cost-per-query data to effectively manage and optimize their AI budgets.



EXPANDABLE USAGE TIERS FOR GROWING ORGANIZATIONS

Tiered Usage Options

OpsVoice offers tiered plans designed to meet the needs of organizations at different growth stages.

Scalability for Growth

Plans are scalable to support startups evolving into large enterprises with increasing observability demands.

Observability Needs

Different tiers cater to various observability requirements ensuring optimal resource allocation.

THE TEAM: EXPERTS IN DISTRIBUTED SYSTEMS AND CLOUD OPERATIONS

SENIOR BACKEND ENGINEERS WITH CLOUD-NATIVE EXPERTISE

Experienced Backend Engineers

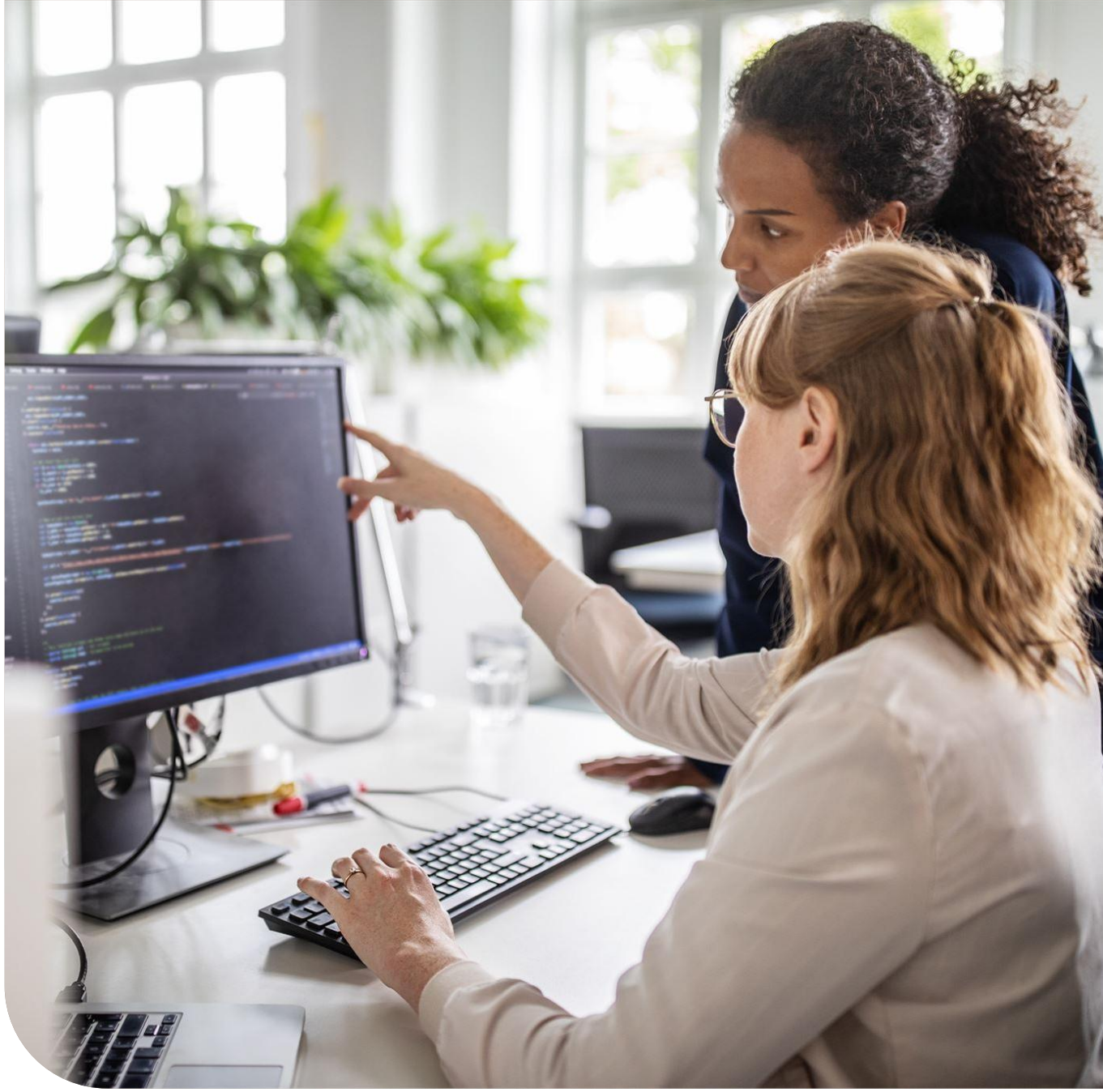
Senior engineers possess deep knowledge in designing and implementing backend systems.

Cloud-Native Optimization

Systems are built and optimized specifically for cloud environments to enhance performance.

Scalable Backend Systems

Backend systems are designed to scale efficiently as demand grows, ensuring reliability.



TRACK RECORD IN DISTRIBUTED SYSTEMS AND OBSERVABILITY

Distributed Systems Design

Expertise in designing scalable and reliable distributed systems enhances performance and fault tolerance.

Monitoring and Observability

Effective monitoring tools facilitate real-time observability for proactive incident management.

Incident Management

Efficient incident response minimizes downtime and improves system reliability in distributed environments.

AGILE AND HACKATHON-PROVEN DEVELOPMENT TEAM

Agile Methodology

The team follows agile practices to ensure flexibility and quick adaptation during development cycles.

Hackathon Innovation

Hackathons provide a platform for rapid idea generation and innovative solutions under time constraints.

Iterative Development

Iterative processes allow continuous improvement and faster delivery of product features.

CALL TO ACTION: JOIN US IN TRANSFORMING AI OBSERVABILITY

STOP GUESSING – START OBSERVING AI PERFORMANCE IN REAL TIME

Eliminate Guesswork

OpsVoice helps teams remove uncertainty in diagnosing AI latency issues by delivering precise insights immediately.

Real-Time Observability

Provides instant monitoring of AI performance to quickly identify and resolve bottlenecks and latency problems.

Accurate AI Insights

Enables accurate analysis of AI system behavior to improve operational efficiency and user experience.

PLACEHOLDER LINKS: GITHUB FOR SOURCE CODE ACCESS

Source Code Access

Access the complete OpsVoice source code through the GitHub repository for transparency and collaboration.

Contribution Invitation

Encouraging developers to contribute improvements and fixes via pull requests on GitHub.

PLACEHOLDER LINKS: DEVPOST FOR PROJECT DETAILS AND DEMO

Project Documentation Access

Detailed project documentation is available on the Devpost platform for in-depth understanding.

Live Demo Availability

Live demos showcase project functionality and key features interactively on Devpost.

Centralized Project Hub

Devpost serves as a centralized hub for both project details and demo presentations.

CONCLUSION

Unique Market Position

OpsVoice offers an innovative and scalable approach to AI performance observability, meeting growing market demands.

Future Collaboration

Excited to partner and collaborate on advancing AI observability solutions for a transformative journey ahead.