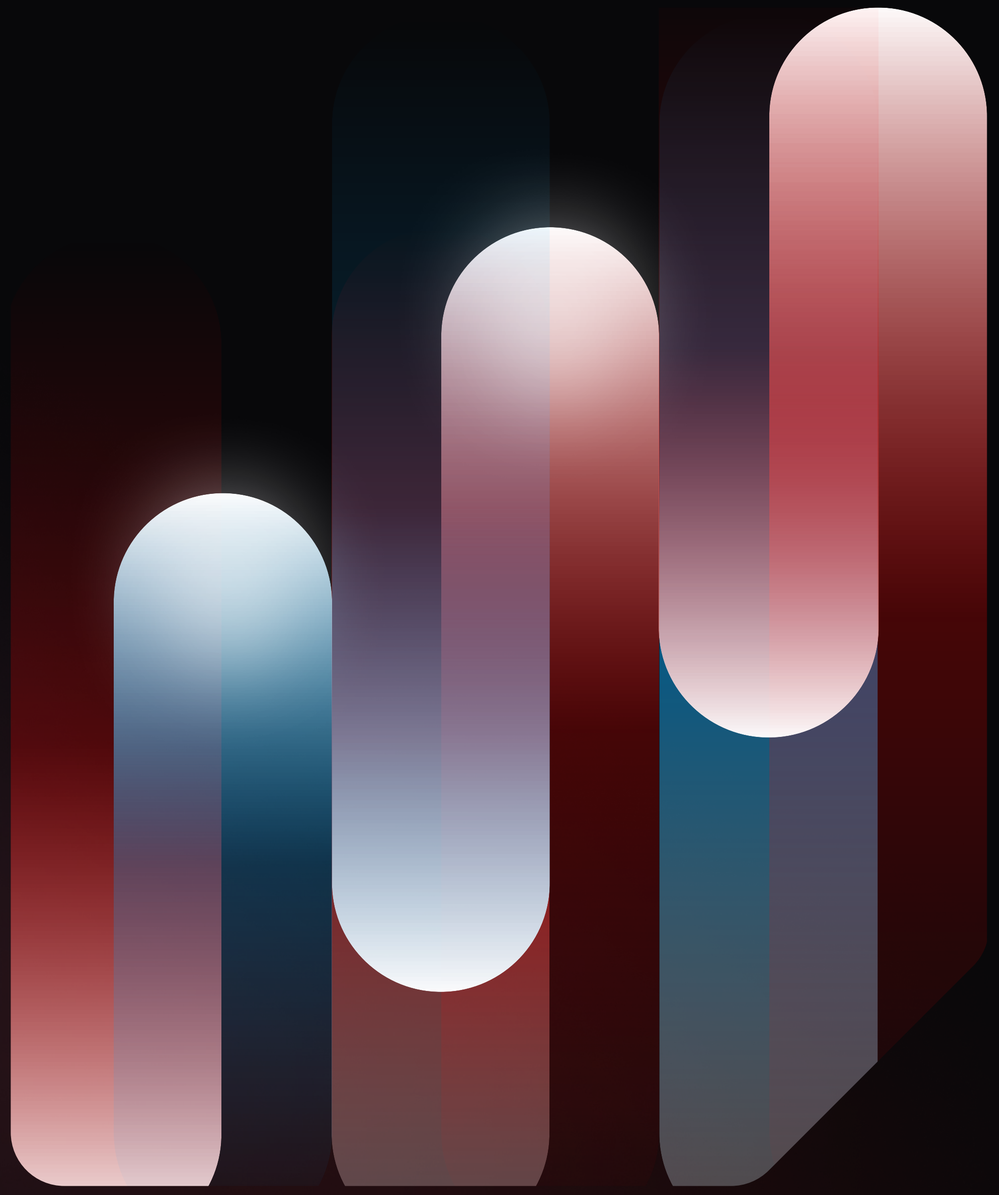
portkey

LLMs in Prod'25

Real-world insights from analyzing **2 trillion tokens across 90+ regions**



Booking.com BCG ByteDance clearcover

Entrepreneur POSTMAN Qoala SONY

About the report

Portkey has the front-row seats to AI's rapid evolution, witnessing firsthand how organizations build and scale their AI operations. Through our analysis of **2 trillion tokens** processed across **1600+ models** in **90+ regions**, this report offers deep insights into how enterprises are advancing their AI initiatives into production-ready systems.

What Sets This Report Apart:

1. **Comprehensive Scope:**
   - Captures the nuances of multi-provider strategies as adoption surges in less than a year.
   - Analyzes regional and organizational dynamics, highlighting the rise of mid-market players and emerging markets.
2. **Performance Insights:**
   - Explores critical metrics like error rates, latency patterns, and the impact of rate limits on production systems.
   - Benchmarks top providers, showcasing how OpenAI and Azure maintains leadership while challengers like Anthropic and Google gain momentum.
3. **Actionable Trends:**
   - Reveals the growing complexity of AI workflows, as organizations shift from simple queries to multi-step orchestrations.
   - Highlights caching's transformative role, in cost saving.

This report is an essential resource for leaders, developers, and data scientists looking to navigate the challenges and opportunities of enterprise AI deployment.

Executive Summary

2024 AI Infrastructure Reality Check

The way teams deploy LLMs in production has fundamentally changed. We went beyond the hype and analyzed **2T+ Tokens** Processed using Portkey's AI Gateway to understand the real patterns in enterprise LLM infrastructure - from provider dynamics to performance metrics.

Key Findings:

- Multi-provider adoption has jumped from **23% to 40% in just 10 months**
- Average token count per request has increased from **500 to 2,000**
- OpenAI maintains **53.8%** market share while specialized providers gain traction in specific use cases
- Caching strategies reduce costs by **38%** while improving response times by **30x**

As enterprises scale from proof-of-concept to production-grade AI applications, infrastructure decisions have become mission-critical. Our analysis shows that successful organizations are moving beyond basic single-provider dependencies to build more complex, resilient, cost-effective AI infrastructure.

Index

Introduction 01

1. AI Adoption by Region
2. Distribuition by Organization Sie
3. Global & Organization Analysis

LLM Adoption Patterns & Insights 04

1. LLM Provider Market Share
2. LLM Adoption Rates
3. LLM Request Growth Rates
4. Cloud LLM Adoption Rates
5. Cloud LLM Request Growth Rates
6. LLM Token Usage Patterns Peak Usage Hours

Performance: The Reality of LLMs 13

1. Server Error Rate Across Providers
2. Rate Limit Errors Across Providers
3. OpenAI vs Azure Error Rates
4. Error Rates for Claude Sonnet-3.5
5. Multi-Provider Adoption
6. Production-Grade LLM Apps with Portkey

Latency Patterns in Production 20

1. OpenAI vs Azure Latency Patterns
2. Claude 3.5 Sonnet Latency Analysis
3. Caching

Conclusion 25

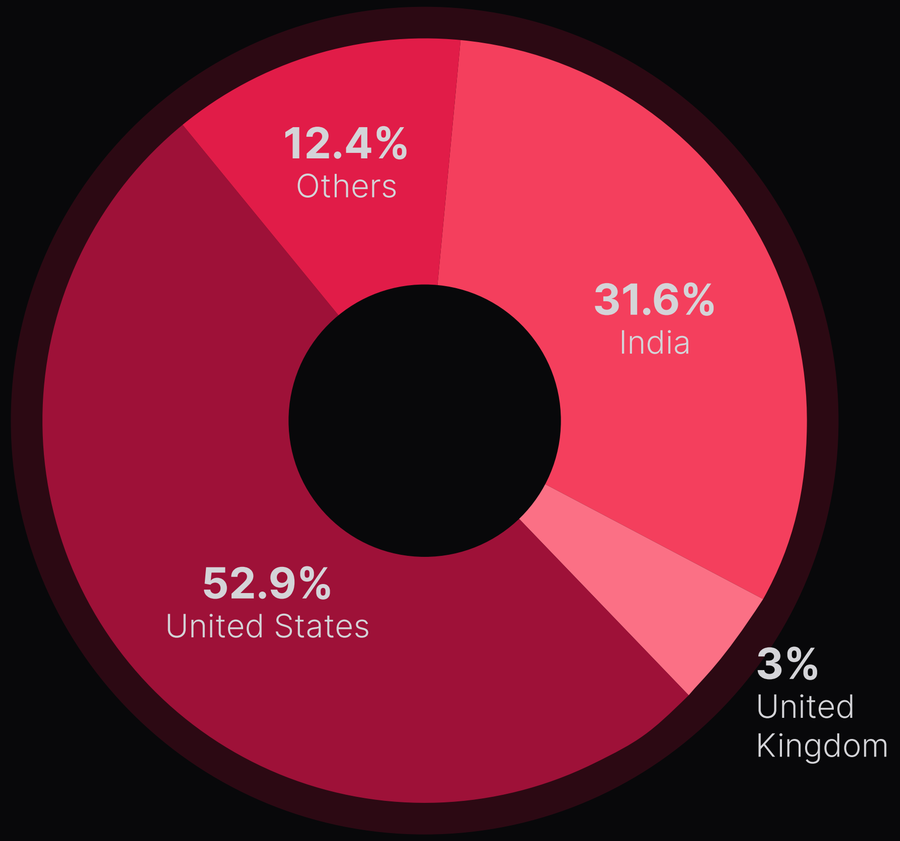
A. Introduction

The AI infrastructure landscape is entering a new era as organizations shift from experimental implementations to mission-critical applications. This transformation brings unprecedented challenges in scaling, reliability, and cost management, demanding fresh insights into how AI systems perform in production environments. Our analysis captures how organizations are solving the fundamental challenges of running AI at scale.

> Portkey stood out among AI Gateways we evaluated for several reasons: excellent, dedicated support even during the proof of concept phase, easy-to-use APIs that reduce time spent adapting code for different models, and detailed observability features that give deep insights into traces, errors, and caching

AI Leader
Fortune 500 Pharma Company

1. Organization Distribution by Region



While traditional markets lead adoption, emerging regions are rapidly closing the gap, reshaping the global AI implementation landscape. Key Metrics:

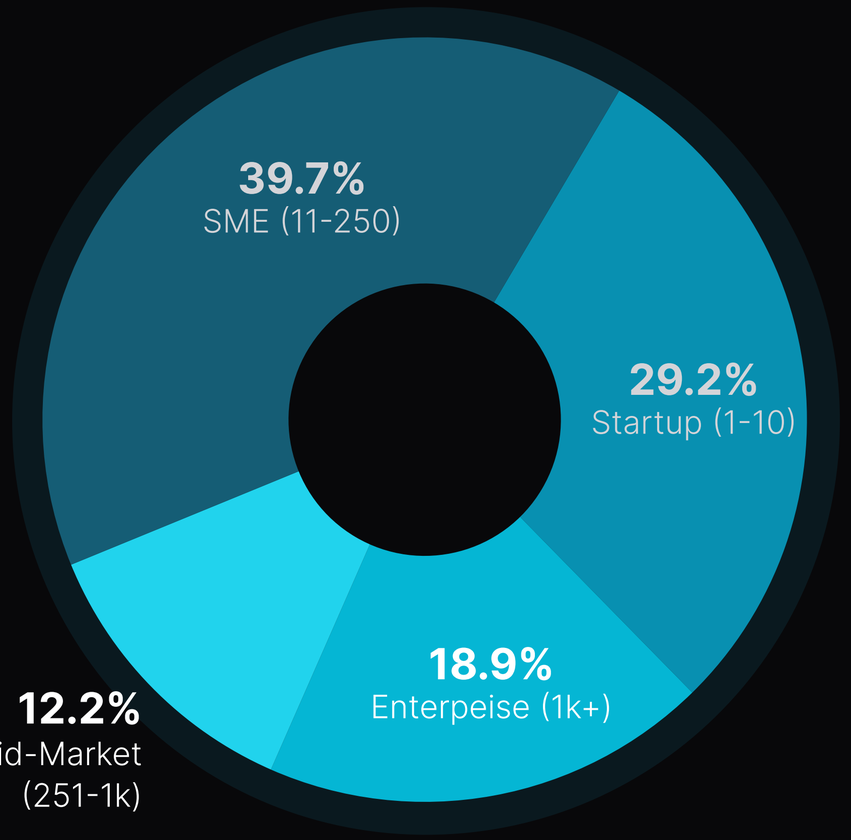
🇺🇸 **United State**
52.9%

🇮🇳 **India**
31.6%

🇬🇧 **United Kingdom**
3.0%

2. Distribution By Organization Size

The AI infrastructure landscape shows surprising diversity, with SMEs leading adoption at 39.7% while established enterprises and emerging startups maintain balanced representation. This distribution suggests a maturing market.



Enterprises (1k+)
18.9%

Mid Market (250-1k)
12.2%

SME (11-250)
39.7%

Startup (1-10)
29.2%

LLM Adoption
Patterns & Insights

B. **LLM Adoption Patterns & Insights**

AI companies are moving from experimenting with LLMs to deploying them at scale. Looking at hundreds of companies running LLMs in prod tells us both who's adopting them and how they're building more sophisticated applications.

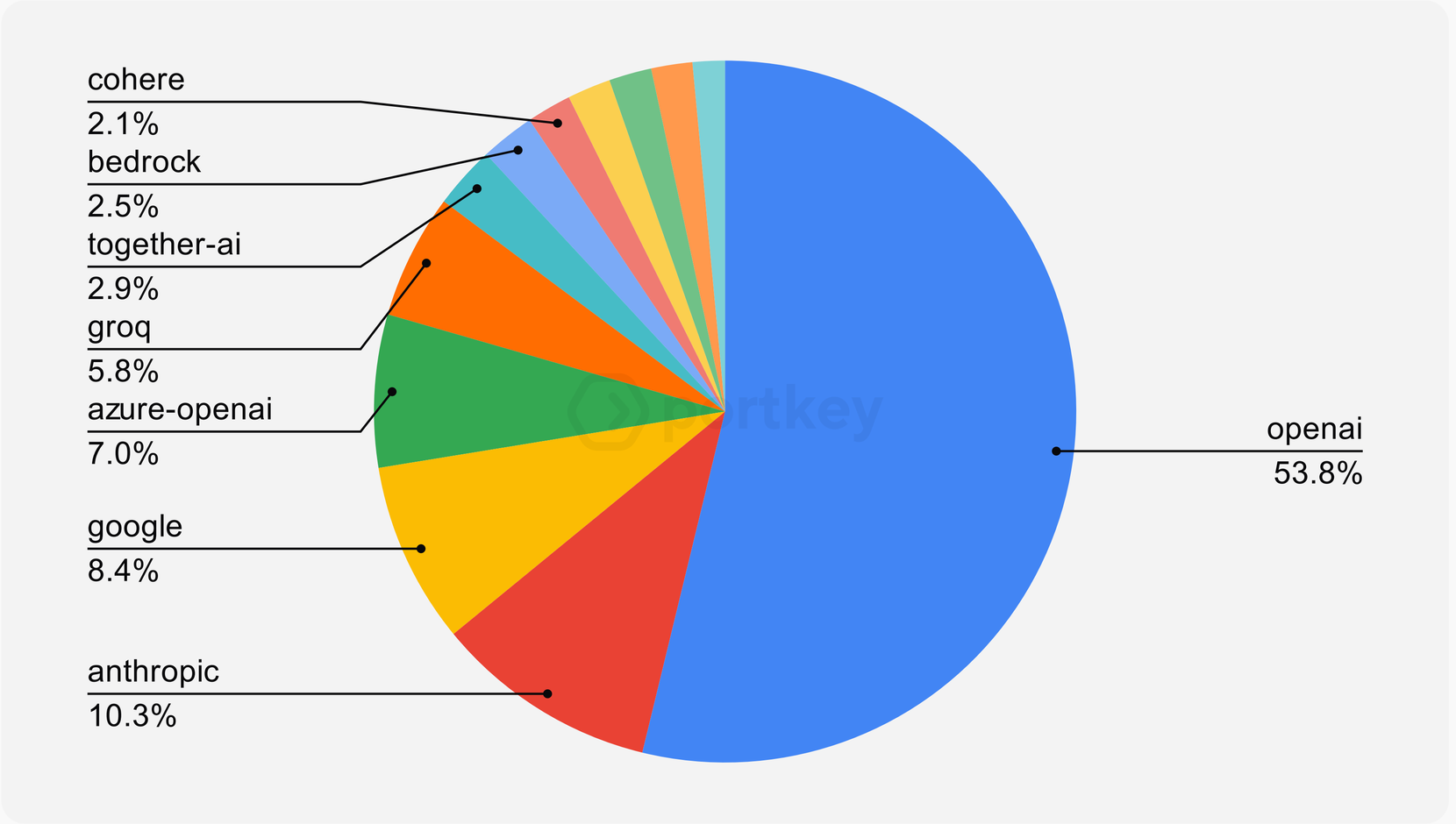
This analysis shows what's working in enterprise AI - from which providers companies trust to how teams are creating more advanced AI applications

1. **LLM Provider Market Share**

Analysis of 2 trillion tokens shows rapid fragmentation of enterprise LLM usage. New providers capture 46.2% of traffic as market matures beyond initial OpenAI dominance.

- OpenAI: **53.8%**
- Anthropic: **10.3%**
- Google: **8.4%**

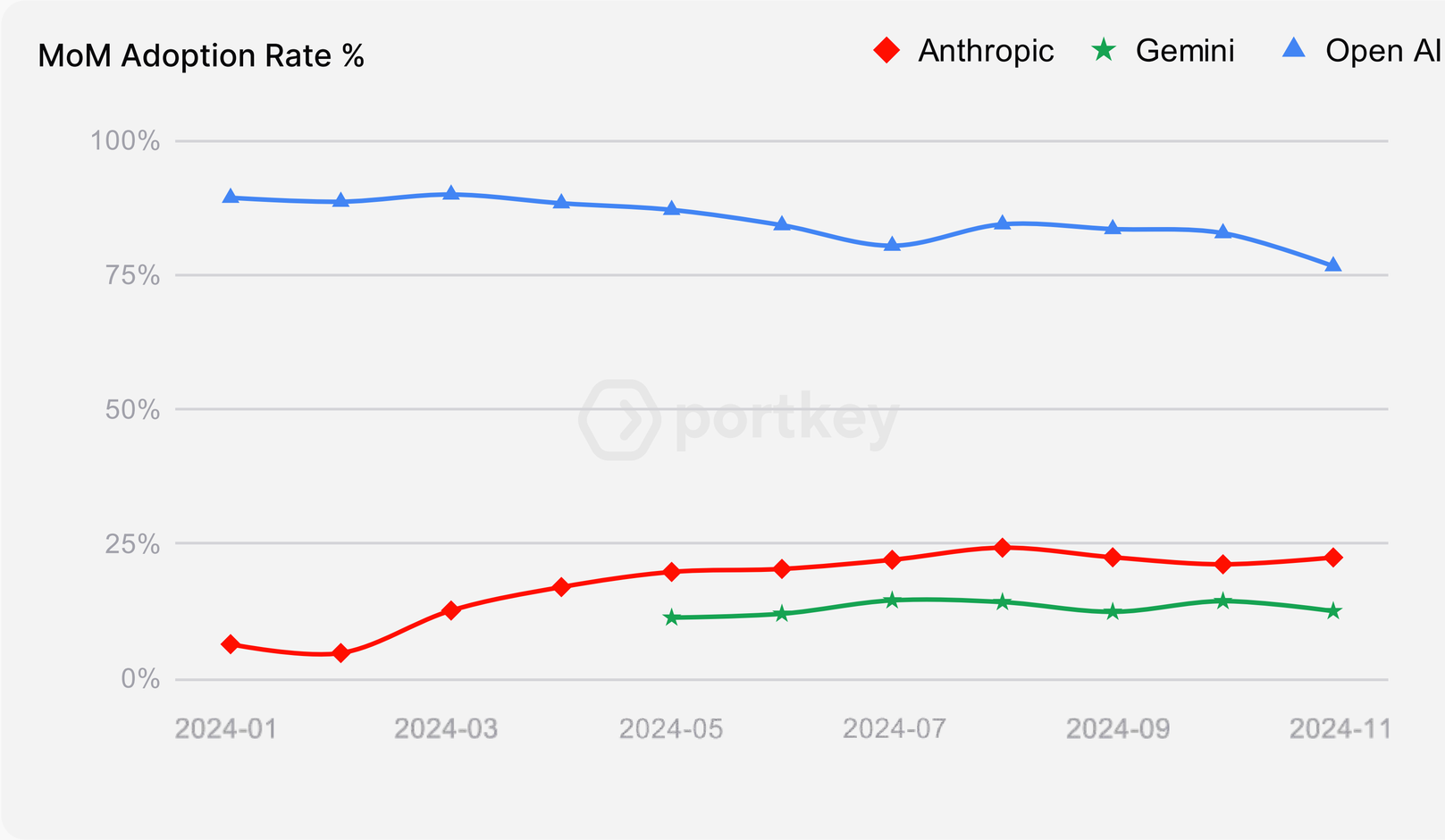
New providers: **27.5%**



2. **LLM Adoption Rates**

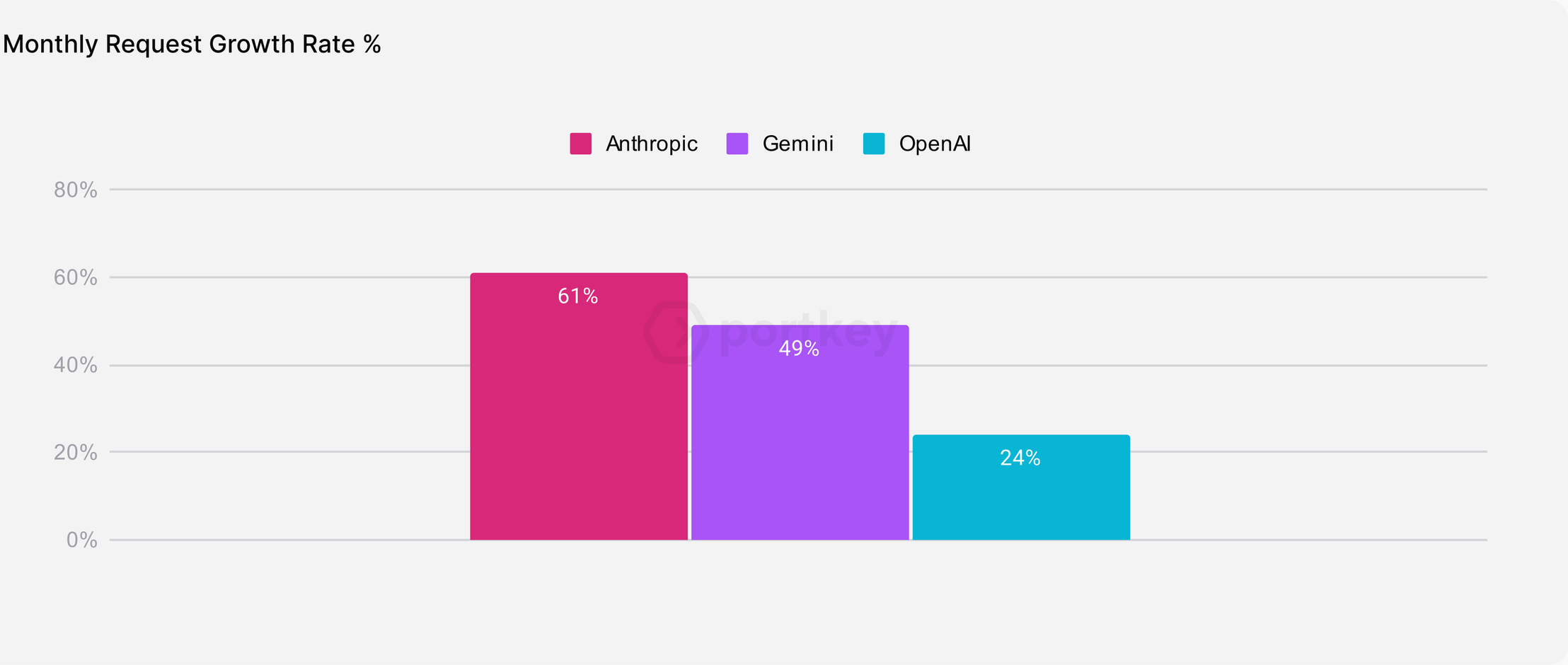
OpenAI's enterprise adoption drops from 89% to 76% as Anthropic and Google show remarkable growth. Market competition intensifies with Anthropic's 61% monthly request growth challenging OpenAI's leadership

- **OpenAI**: **76**% adoption rate, **24%** monthly request growth
- **Anthropic**: **61**% monthly request growth, **23**% market penetration
- **Gemini**: **49**% monthly request growth



3. **LLM Request Growth Rates**

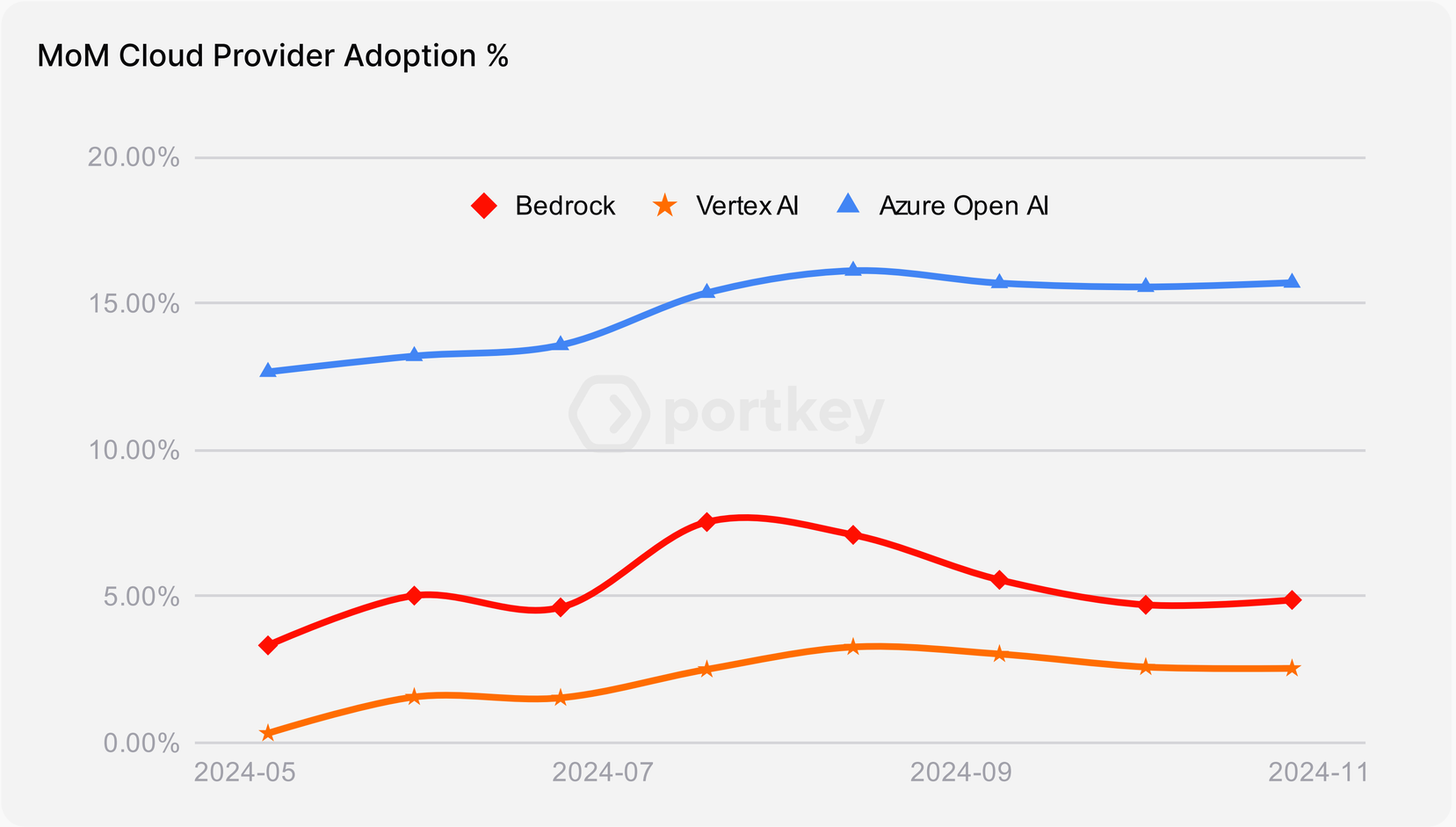
Anthropic and Google show accelerated growth with 61% and 49% monthly request increases respectively, while market leader OpenAI maintains steady 24% growth rate



4. Cloud LLM Adoption Rates

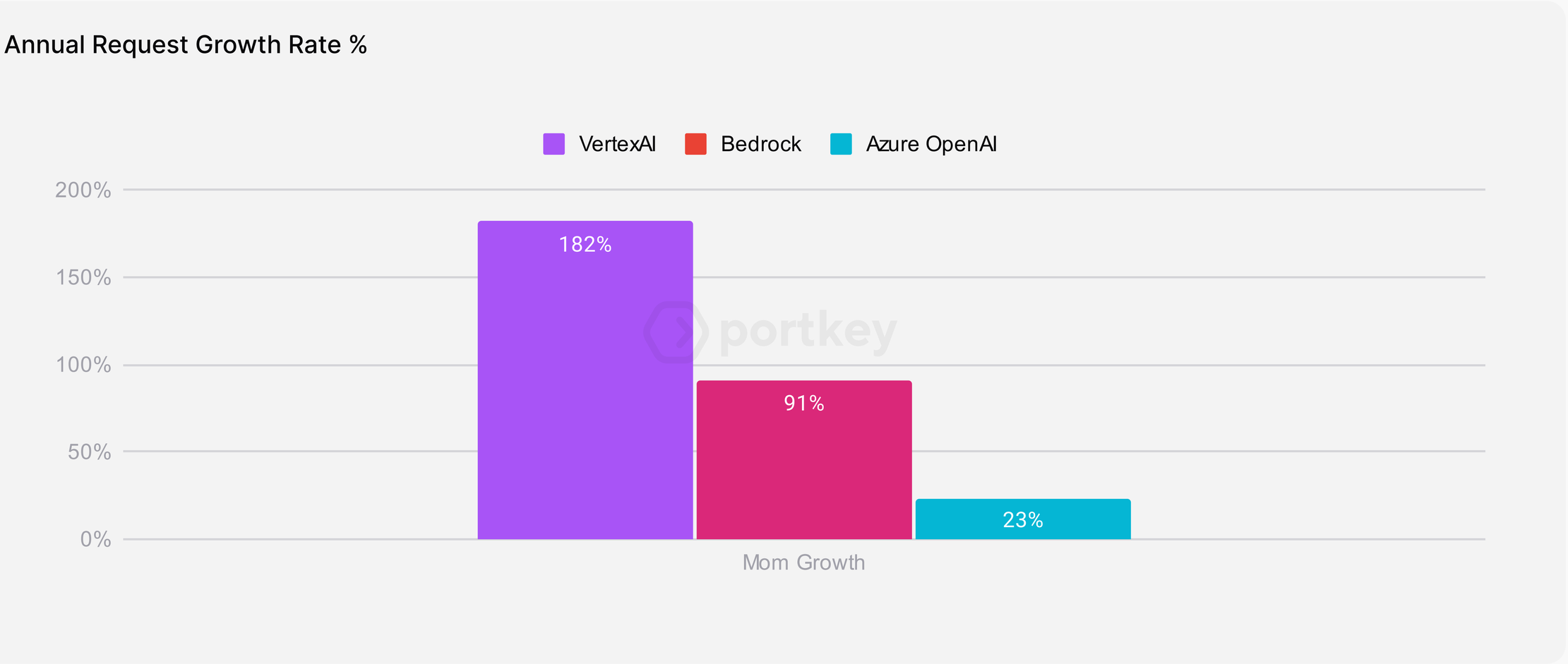
Microsoft's early mover advantage translates to 3x more enterprise adoption than AWS Bedrock. Azure's partnership with OpenAI fuels its dominance in cloud-native LLM offerings

- Azure OpenAI: Leading enterprise adoption
- AWS Bedrock: Strong depth metrics
- Google Vertex AI: **43%** monthly growth



5. **Cloud LLM Request Growth Rates**

Azure's OpenAI integration drove early success, but other cloud providers are gaining momentum with their own LLM offerings



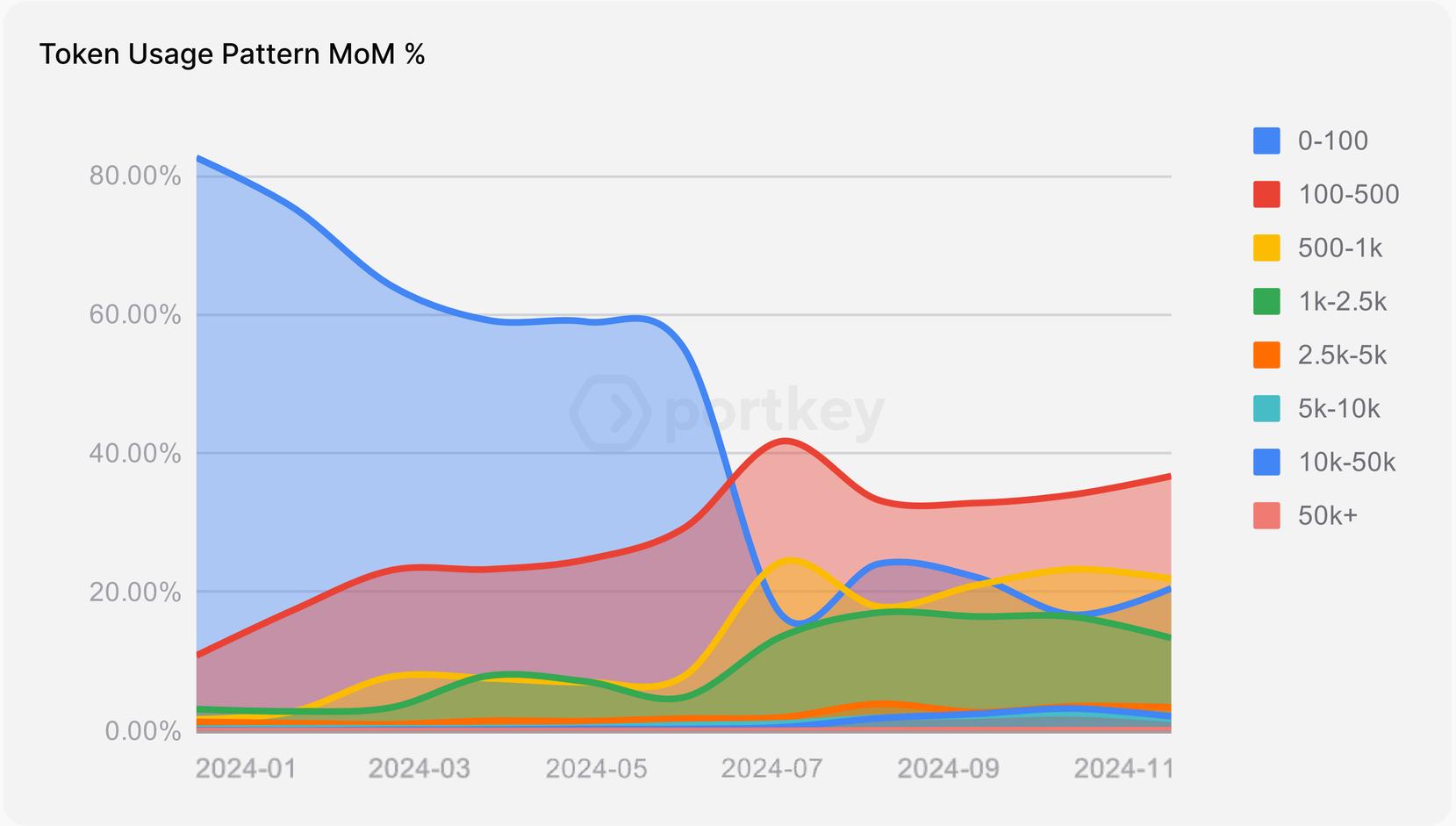
6. LLM Token Usage Patterns

Companies are moving from simple prompts to more complex workflows, prioritizing multi-step chains and advanced orchestration

Simple queries: Dropped from **80% to 20%**

Medium complexity: Now **35%** of total usage

Complex workflows **(500+ tokens)**: Consistent growth trend

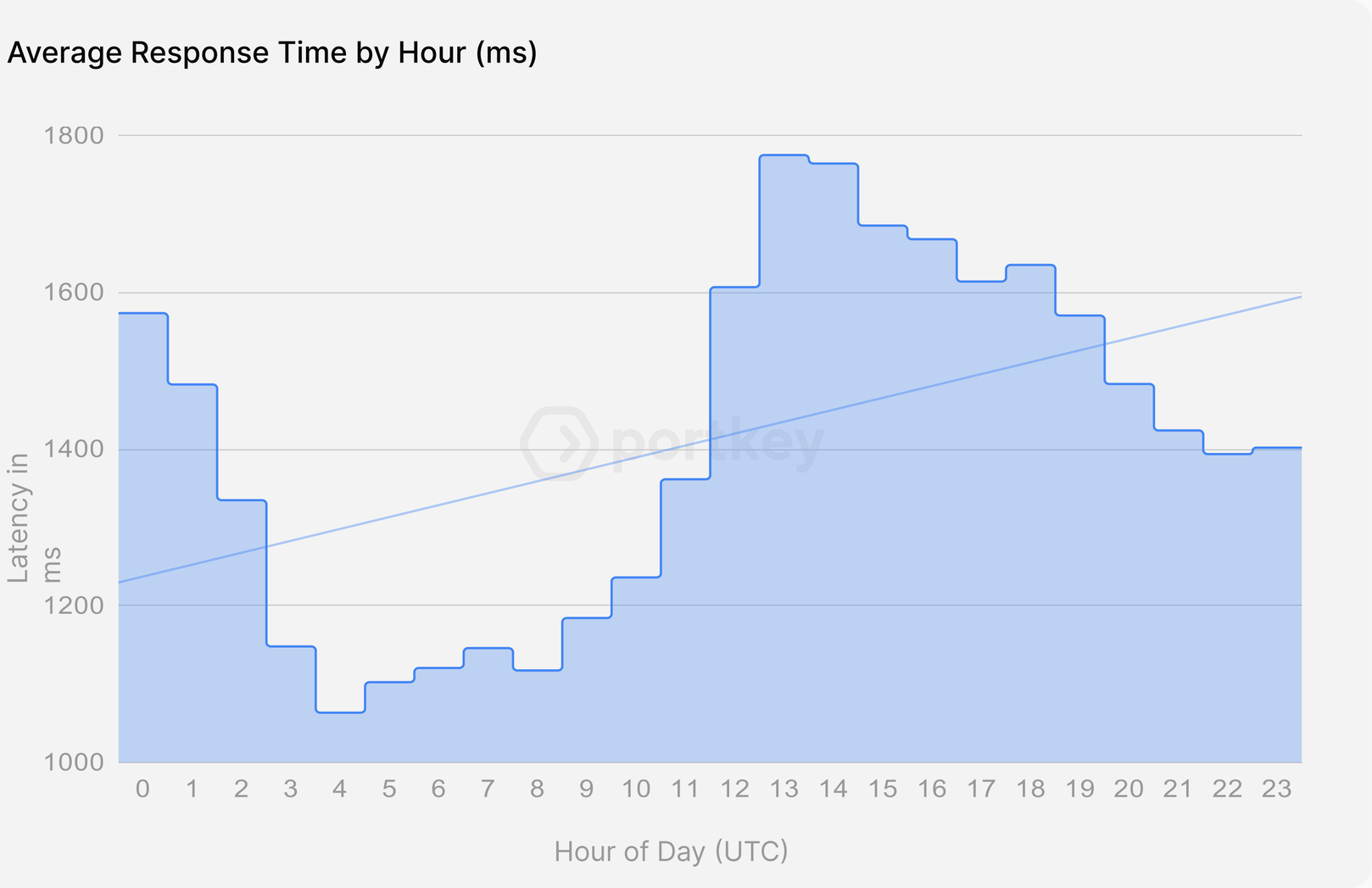


7. **Peak Usage Hours**

Primary Peak: **13:00-15:00** UTC (US/EU overlap)

Secondary peak: **22:00** UTC (APAC operations)

Minimum activity: **03:00-05:00** UTC



Performance: The **Reality of LLMs** in Production

C. **Performance: The Production Reality**

Performance in production isn't just about speed—it's about reliability, consistency, and handling failure gracefully. Our analysis of over 2 trillion tokens processed through Portkey's AI Gateway reveals critical insights about how different providers and models perform in real-world conditions.

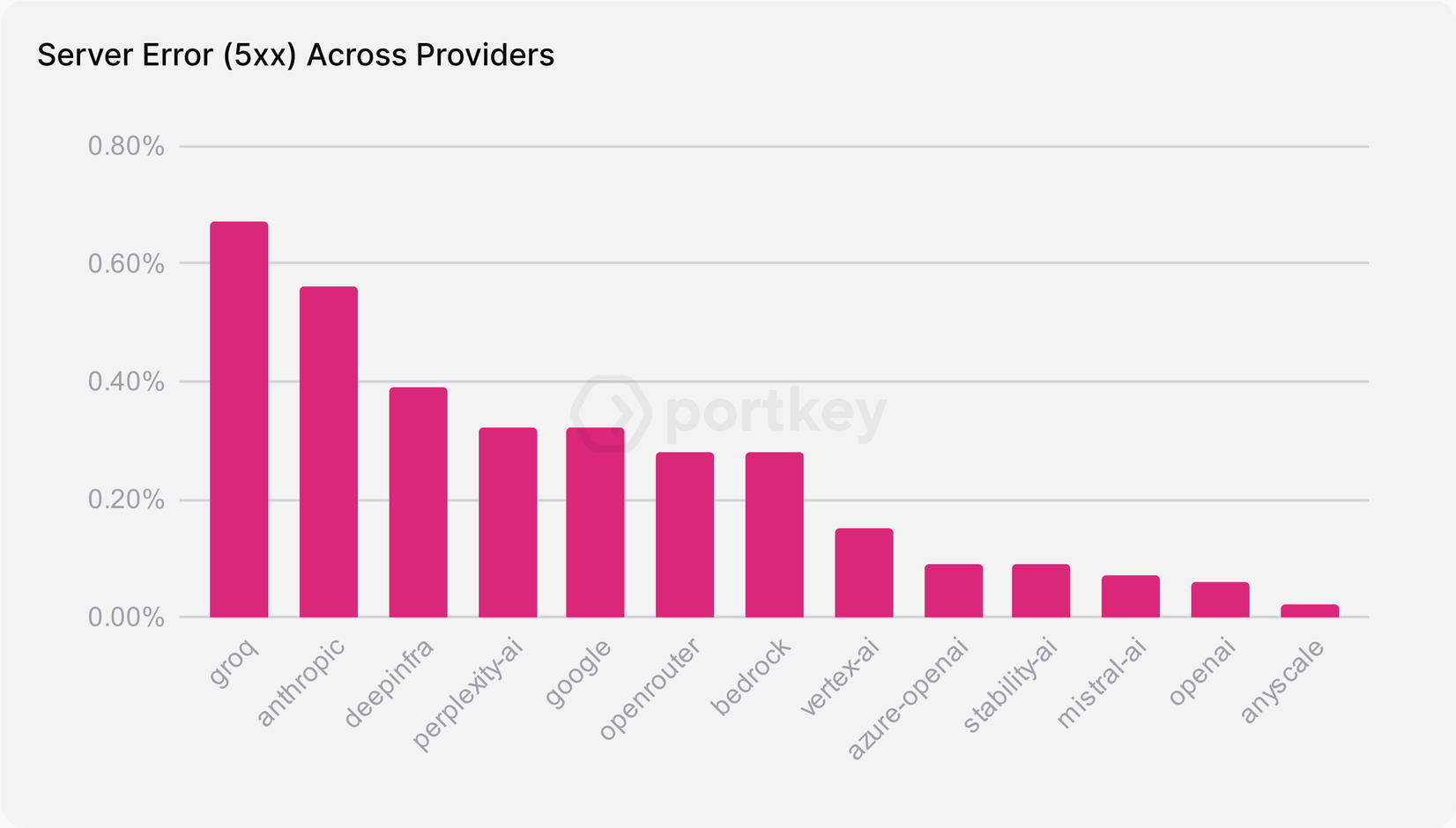
This section examines two key aspects of production performance: error rates and rate limits.

1. **Server Error Rate Across Providers**

Server errors remain a persistent challenge even for enterprise-grade providers, with new entrants showing higher instability. Real-world data shows surprising gaps in perceived vs actual reliability

Provider Error Rates:

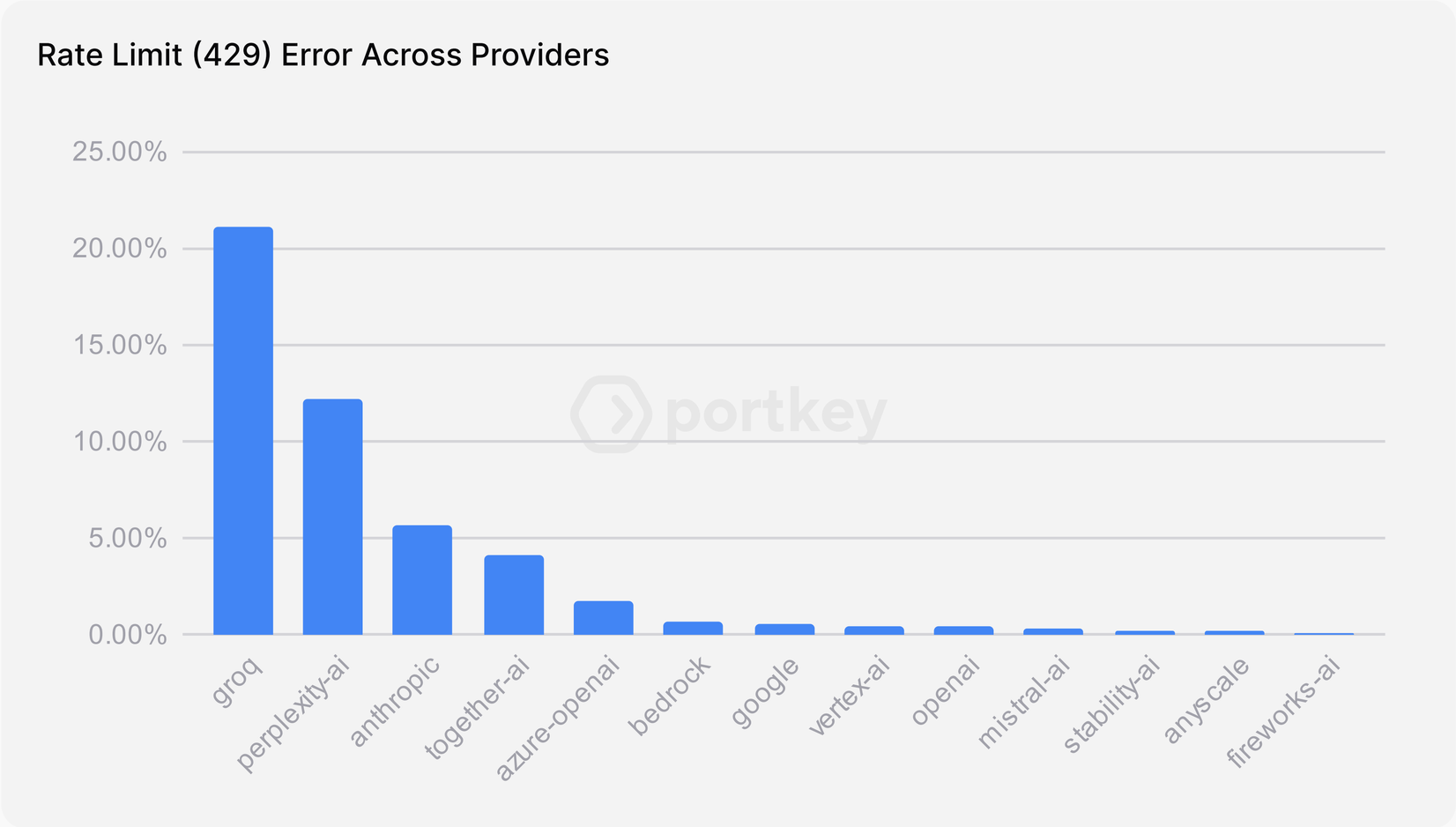
- Groq: **0.67%**
- Anthropic: **0.56%**
- Perplexity: **0.39%**
- Google: **0.32%**
- Bedrock: **0.28%**



2. **Rate Limit Errors Across Providers**

Rate limit errors can silently degrade application performance, with some providers showing alarming spikes of over 20% failed requests. Multi-provider strategy becomes critical for production reliability

- Groq shows highest rate limiting at **20%**
- Large variance: **0.1% to >20%** across providers
- Top 5 providers average **8.6%** rate limit errors

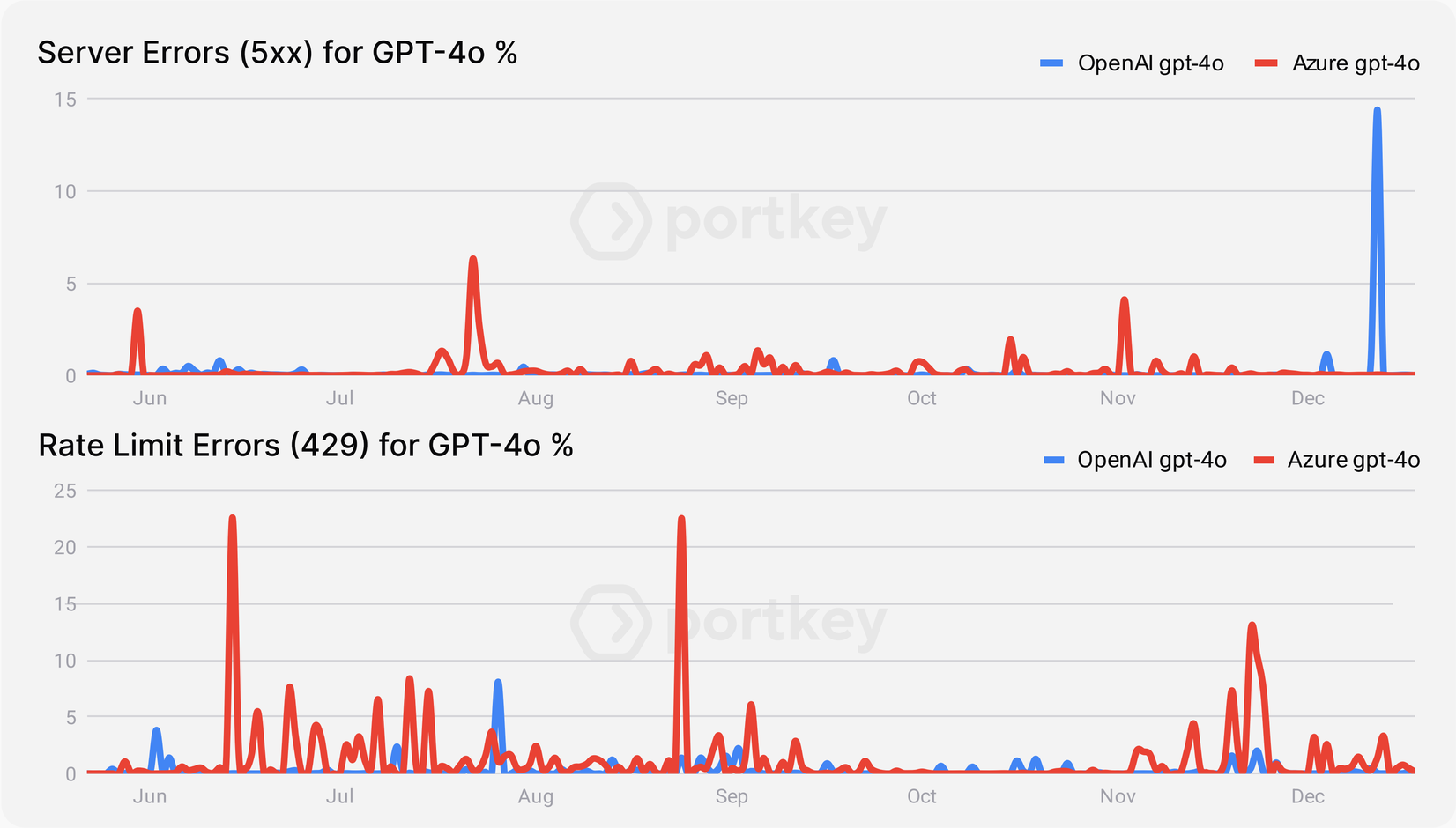


3. OpenAI vs Azure Error Rates

Despite enterprise familiarity with Azure, direct OpenAI performance is more reliable. Comparing GPT-4o models reveals crucial performance differences. showing lower error rates across OpenAI directly

Provider Error Rates:

- OpenAI: Consistent performance patterns
- Azure: Higher but improving stability

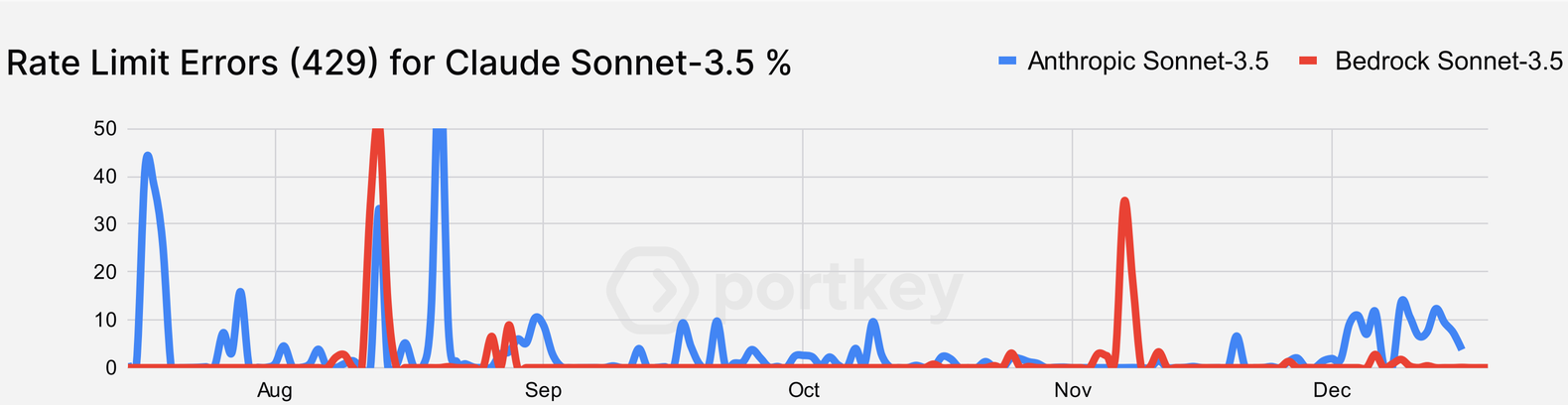


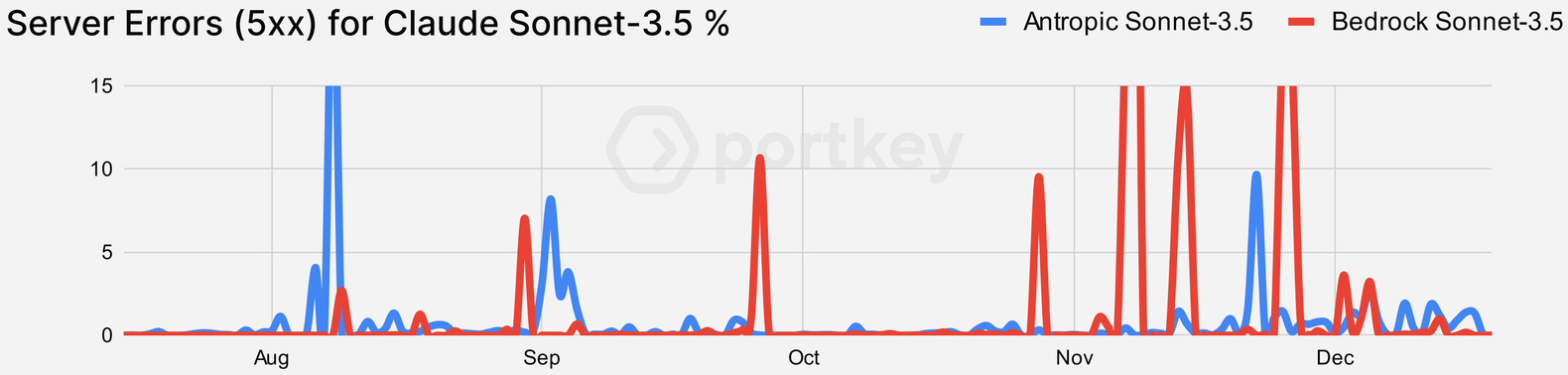
4. Error Rates for Claude Sonnet-3.5

When the same model Claude 3.5 Sonnet runs on different infrastructures, reliability metrics tell a compelling story about AWS platform impact on production performance.

Key Metrics:

- Bedrock (429): **6.2** errors/100
- Anthropic (429): **8.4** errors/100
- AWS advantage: **25%** better reliability





5. **Multi-Provider Adoption**

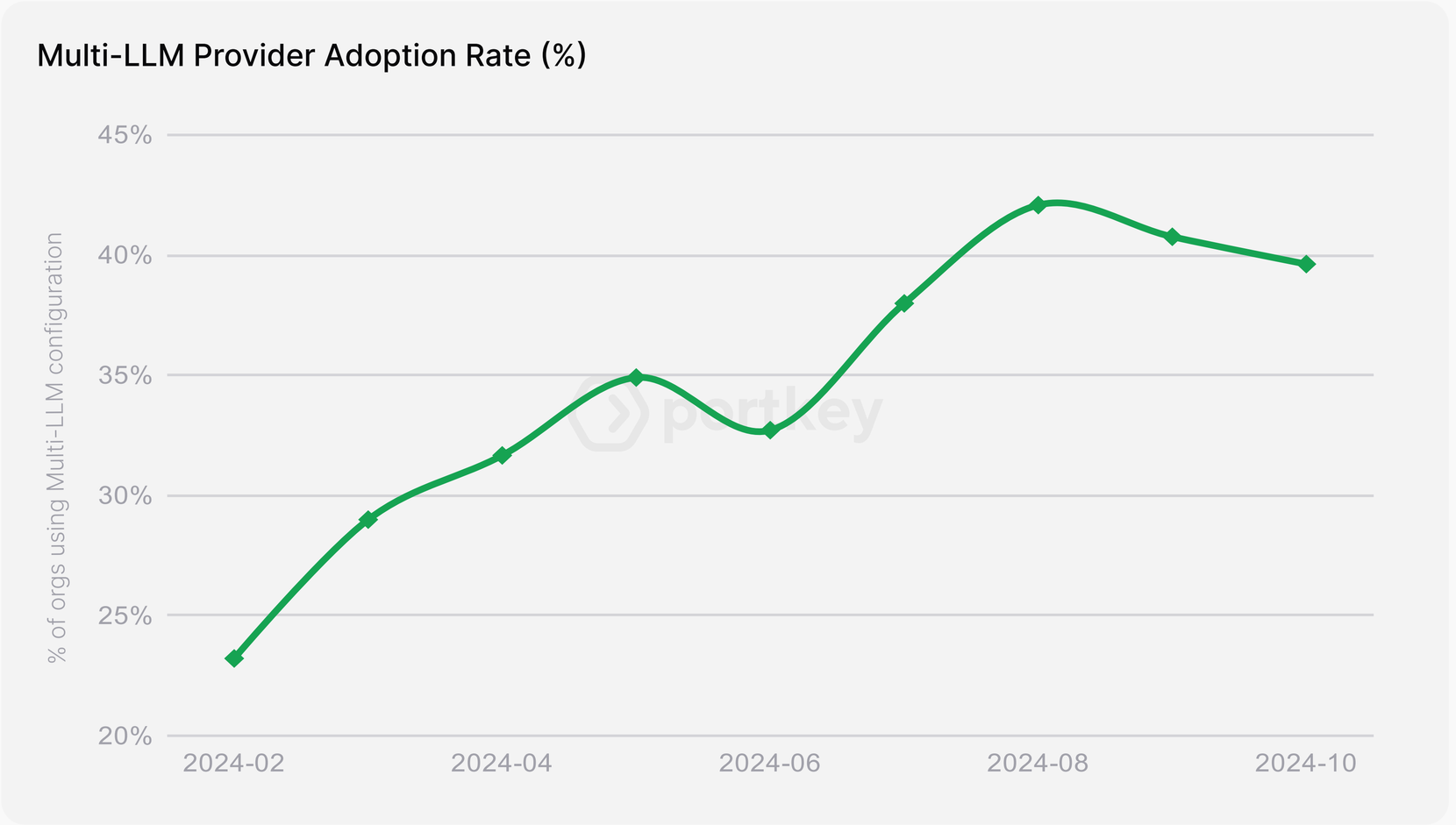
Organizations using multiple providers see concrete reliability improvements, with data showing significant latency and uptime benefits

Multi-provider adoption jumped from **23% to 40%** in 10 months

Key Advantages:

Organizations see **30%** lower P95 latencies

Zero complete outages reported with proper fallbacks



Latency Patterns in Production

D. Latency Patterns in Production

Latency impacts every part of an AI application - from user experience to running costs. This section analyzes latency data across different providers and deployment scenarios to help teams understand:

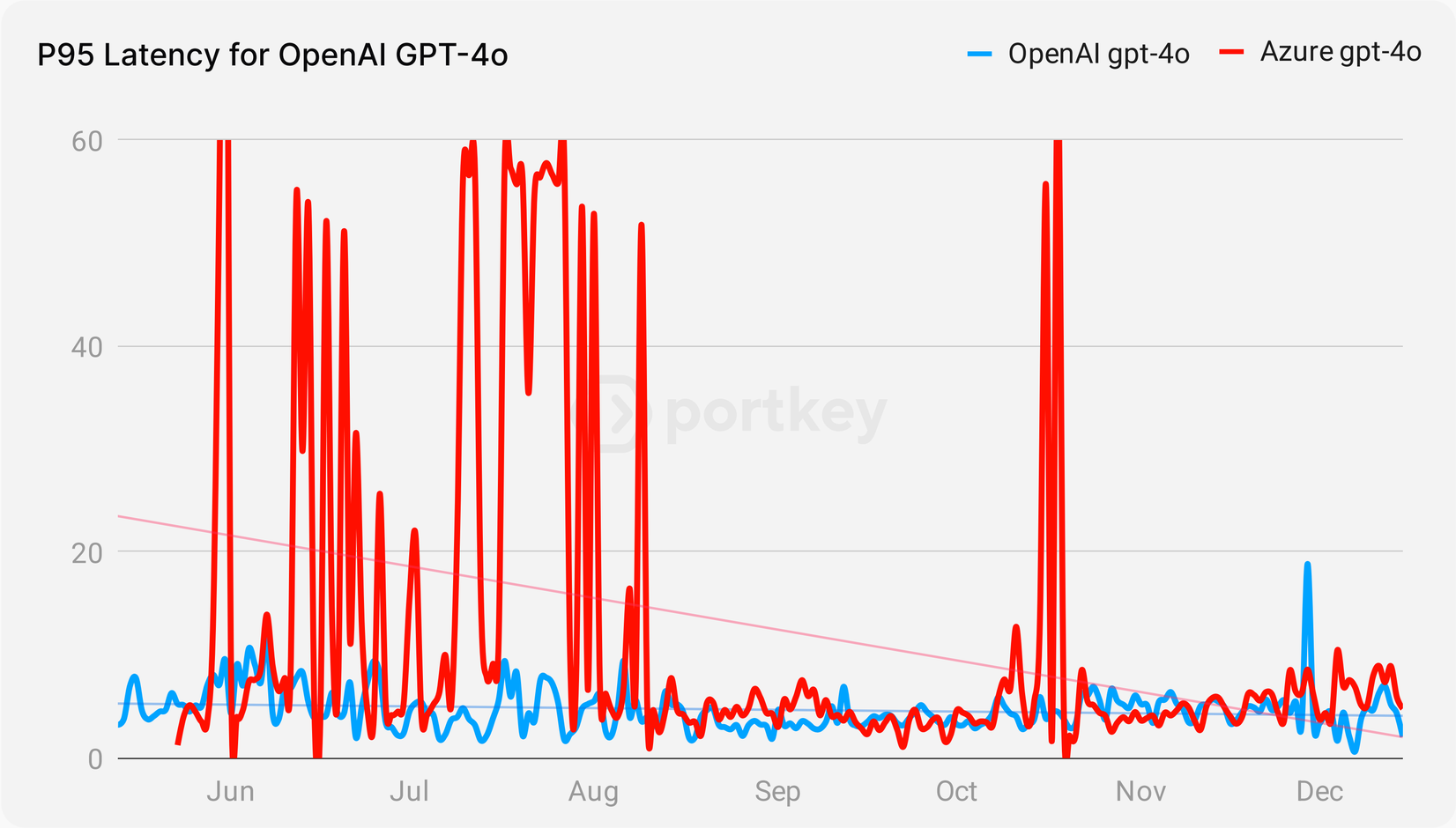
- How provider choices affect response times
- Which setups deliver consistent performance
- How to reduce costs at scale

1. **OpenAI vs Azure Latency Patterns**

OpenAI delivers consistent low-latency performance compared to Azure OpenAI, making it a preferred choice for time-critical applications.

Provider Error Rates:

- OpenAI: **~3s P95** latency, offering faster response times.
- Azure: **~5s** P95 latency, with a stability trade-off.

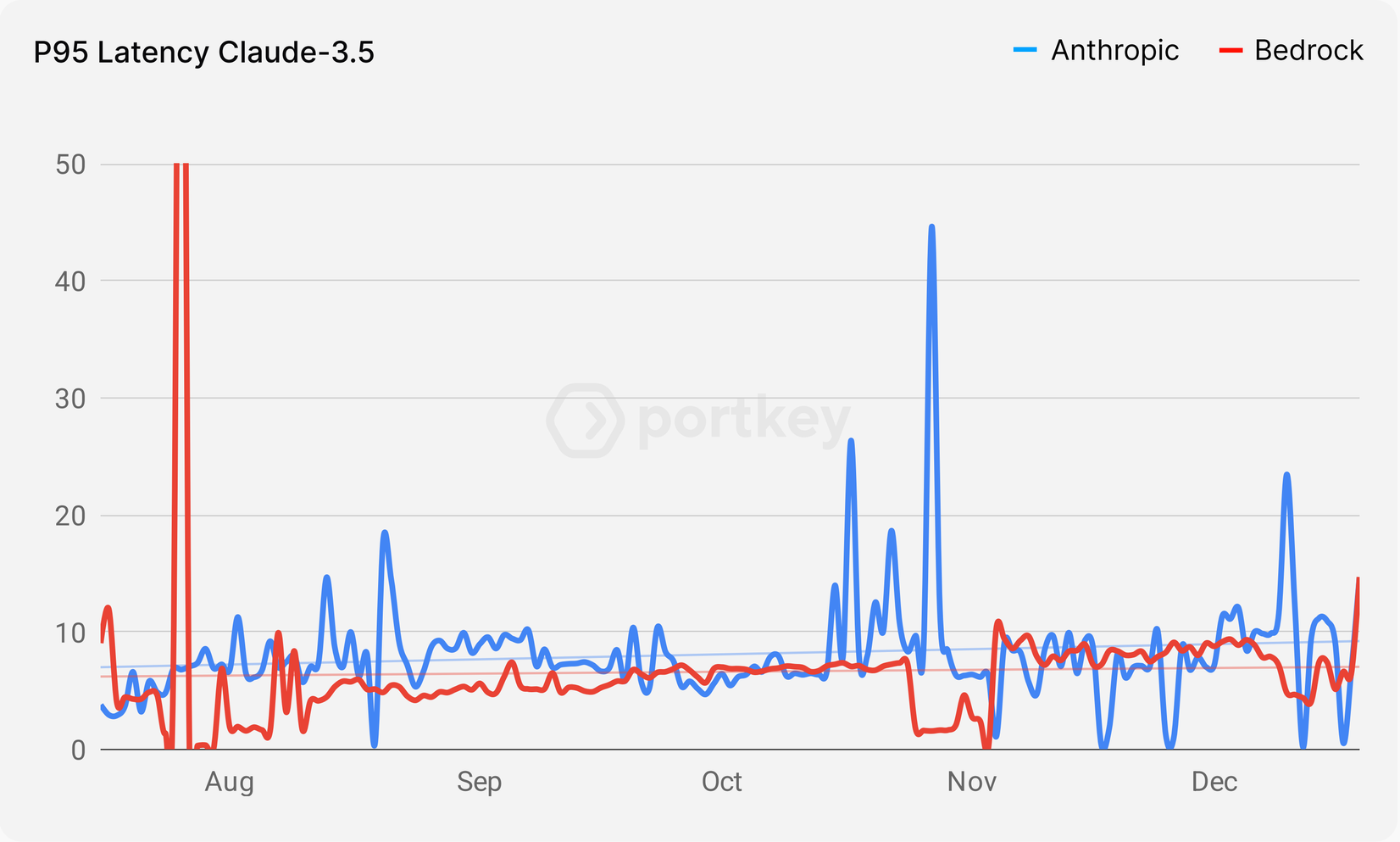


2. **Claude 3.5 Sonnet Latency Analysis**

Claude 3.5 Sonnet P95 latency analysis across Antrhopic and AWS Bedrock

Provider Error Rates:

- Anthropic: **~7.3s** P95 latency
- Bedrock: **~6.5s** P95 latency



3. **Caching**

Caching remains the most effective cost optimization for LLM calls, offering up to 38% in savings

- 36% hit rate on average (highest in Q&A scenarios)
- 30x faster response delivery
- 38% proven cost reduction

Portkey's caching system delivers these benefits through two powerful approaches:

1. Semantic cache
2. Simple cache.

For teams scaling their LLM applications, Portkey's cache serves responses up to 20x faster while eliminating API costs for cached requests. In production environments where every millisecond and API call matters, effective caching isn't just an optimization—it's essential infrastructure

Conclusion

E. **Conclusion**

As we move into 2025, our analysis of 2 trillion tokens reveals a fundamental shift in how organizations are implementing AI at scale. The data highlights critical trends that will shape enterprise AI infrastructure:

Key Findings for 2025:

- Multi-provider adoption surged from **23% to 40%,** becoming the new standard for production reliability
- Complex AI workflows **(over 500 tokens)** now dominate **(80% of traffic**), requiring more sophisticated orchestration
- Caching strategies deliver **38% cost savings and 30x** faster response times
- Organizations using multiple providers see **30% lower P95** latencies and zero complete outages
- Rate limiting remains a critical challenge, with some providers showing >20% failed requests

For technical leaders planning their 2025 AI infrastructure, the focus should shift from basic implementation to building resilient, cost-effective systems that can handle complex workflows while maintaining enterprise-grade reliability.

Have feedback or insights to share? We'd love to hear your thoughts on these findings and discuss how they align with your organization's AI journey.

→ Share Feedback

Production-Grade LLM Apps with Portkey

Turn reliability challenges into advantages with intelligent request routing and implement failover configuration across providers using Portkey's AI Gateway

Instant Multi-Provider Setup:

Single unified API connects to **250+ models**

Smart Reliability Layer:

Fallback, retries, load balancing, and conditional routing to prevent failures.

Performance Optimization:

Semantic & Simple Caching Solution to improve Performance

Production Ready:

Enterprise-grade Observability with full request tracing

portkey

Ready to Build Production-Ready AI Apps?

Observe, govern, and optimize your AI apps across your entire org while mitigating critical errors and scaling LLMs.

2 Trillion+ Tokens	*90+ Regions*	*650+ Customers*

→ Talk to us

→ Share the report

Join 2k+ AI Engineers Community

Follow @PortkeyAI