//ADASTRA | LLM Gateway

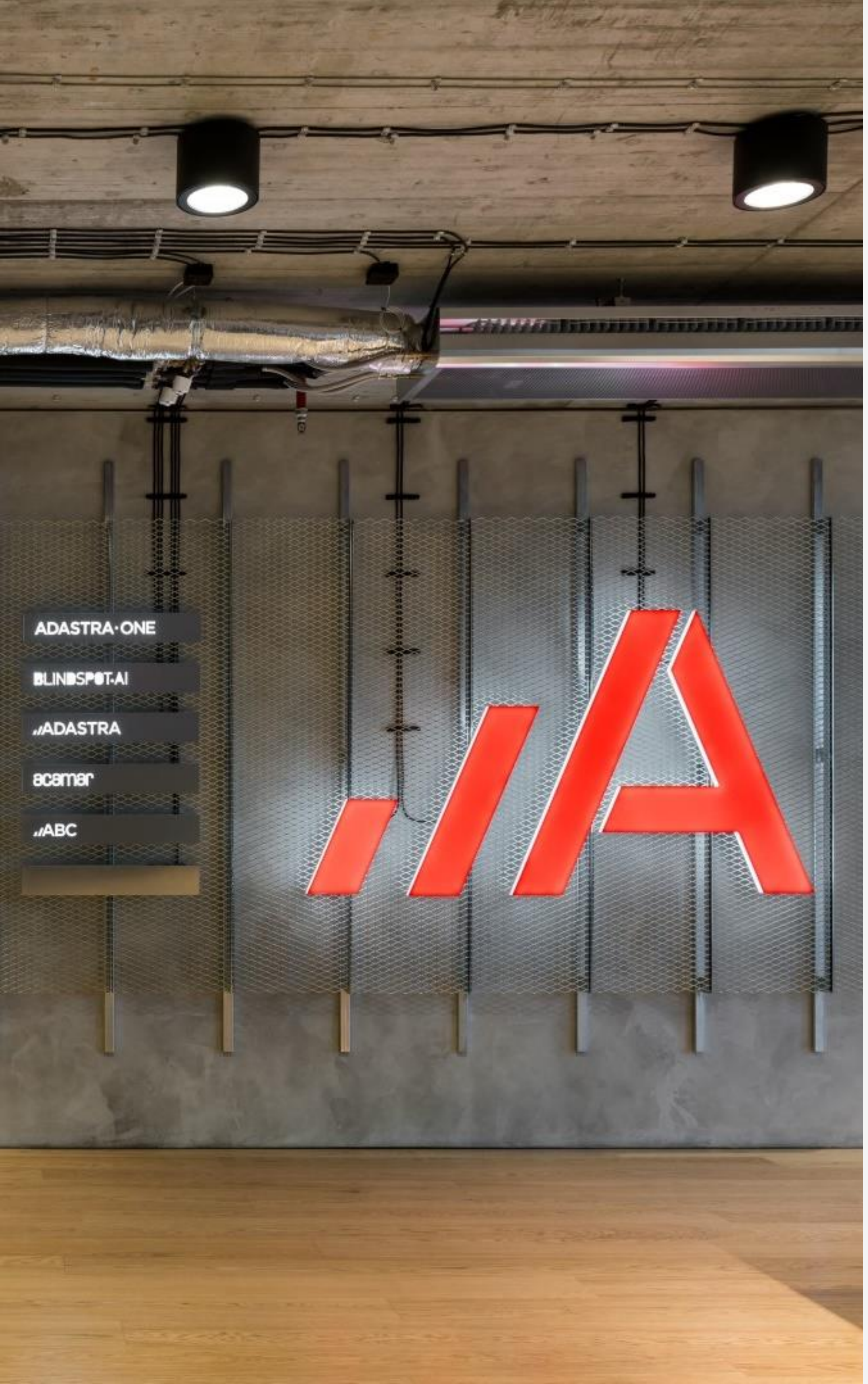
Adastra LLM Gateway:

Gain control over your growing GenAI infrastructure.

Scale GenAI securely and reliably. Eliminate downtimes and control costs.

Adastra AI

Public



Adastra AI

Global AI Practice

20+ years

Delivering AI Solutions

450+ projects

Executed by top AI experts

40+ countries

Where we delivered projects

We help companies harness the power of AI, by using smart products, services, and tailored solutions.

Our end-to-end AI services include:

- AI consulting
- Tailored AI solution development
- Ready-made AI product deployment
- Augmented AI teams

aws PARTNER Advanced Tier Services

Gold Microsoft Partner Microsoft

Google Cloud Partner

Microsoft Awards 2021

Deloitte.

50 Technology Fast 50 2020 CENTRAL EUROPE

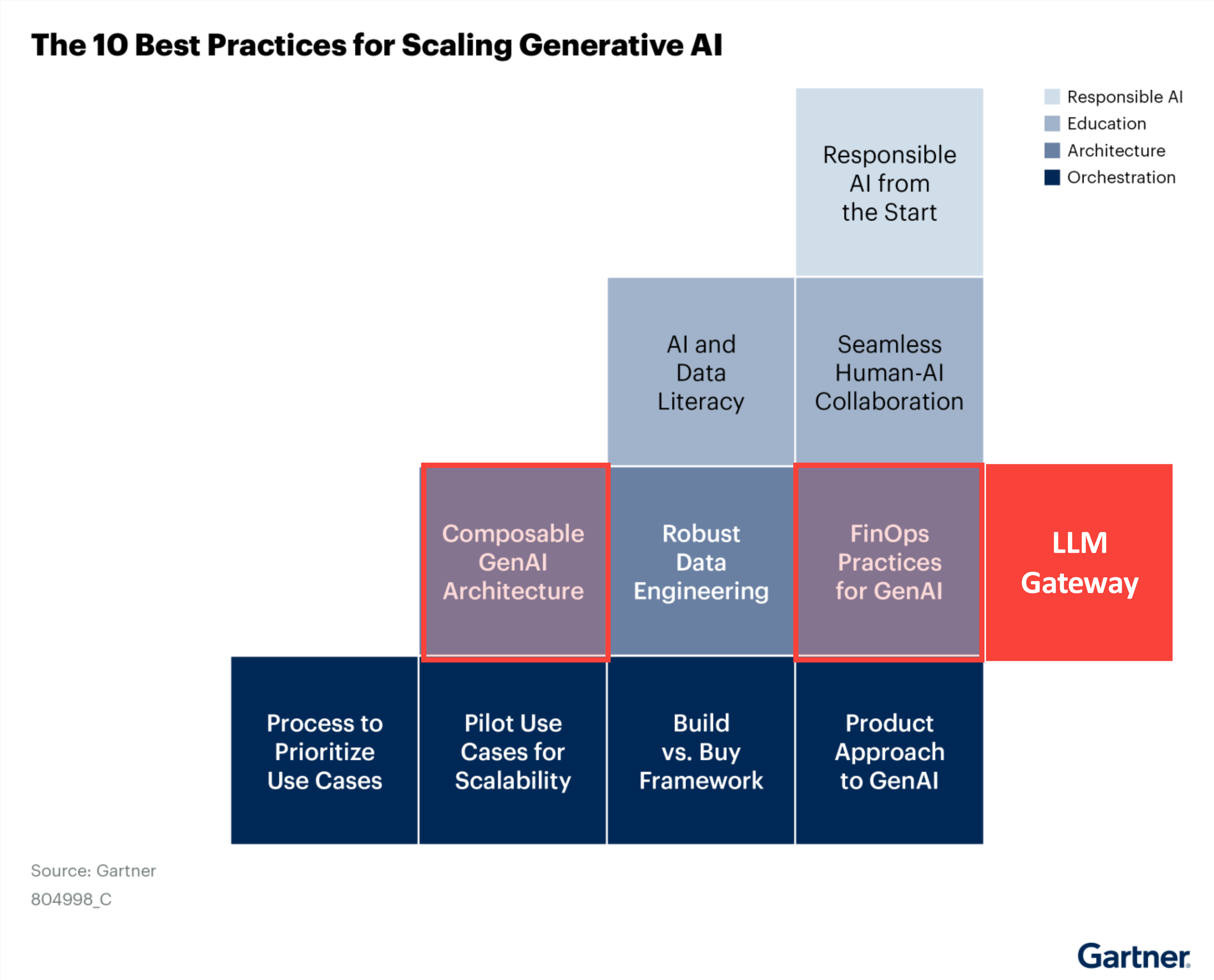
50 Technology Fast 50 2019 CENTRAL EUROPE

//AI

If 2023 was the year the world discovered Gen AI, 2024 is the year organizations truly began using—and deriving business value from—this new technology.

— McKinsey & Company, May 2024

Promoting best practices for scaling GenAI



The 10 Best Practices for Scaling Generative AI

Responsible AI from the Start

AI and Data Literacy

Seamless Human-AI Collaboration

Composable GenAI Architecture

Robust Data Engineering

FinOps Practices for GenAI

LLM Gateway

Process to Prioritize Use Cases

Pilot Use Cases for Scalability

Build vs. Buy Framework

Product Approach to GenAI

- ▫ Responsible AI
- ▫ Education
- ▪ Architecture
- ■ Orchestration

Source: Gartner

804998_C

Gartner.

LLM Gateway addresses key architectural best practices for scaling generative AI across the enterprise.

Scaling GenAI brings major challenges

Performance and Usage Limits

- Varying **response times** across different models and datacenters.
- Strict quotas on model usage

No Monitoring and Auditability

- No out-of-the-box monitoring of usage and other statistics
- No auditability

Cost Management

- No cost management
- Limited cost tracking and optimization

Operational Inefficiencies

- Managing **multiple AI providers** and models
- Managing and standardizing multiple GenAI projects across the company

LLM Gateway features addressing these challenges

- **Load balancing**
- **Smart Fallbacks**
- **Automatic Retries**

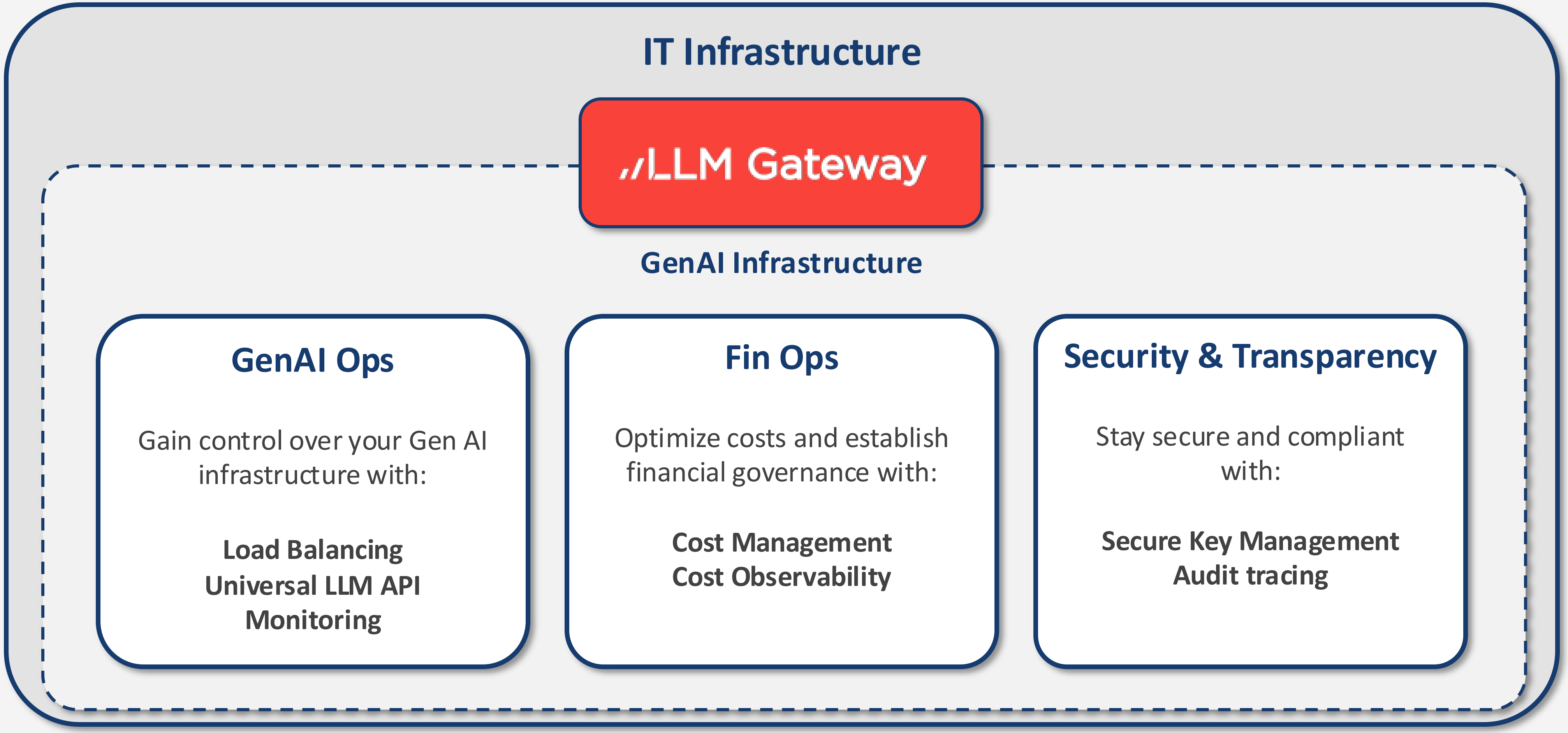
- **Detailed usage, performance and cost monitoring**
- **Audit Trace**

- **PTU allocation and control features**
- **Cost Observability**
- **Fine-grained cost management**

- **Universal API**
- **Unified logging, cost management and monitoring**

What is LLM Gateway?

Core component of your growing GenAI infrastructure helping you to gain control over all key aspects necessary for **scaling GenAI solutions.**



IT Infrastructure

//LLM Gateway

GenAI Infrastructure

GenAI Ops

Gain control over your Gen AI infrastructure with:

Load Balancing
Universal LLM API
Monitoring

Fin Ops

Optimize costs and establish financial governance with:

Cost Management
Cost Observability

Security & Transparency

Stay secure and compliant with:

Secure Key Management
Audit tracing

LLM Gateway helps the client control:

- GenAI Ops
- Fin Ops
- Security & Transparency

By using LLM Gateway they will gain

No unexpected costs

Optimize and limit LLM costs with complete visibility of all LLM usage and caching features.

90% Reduced Downtime

Intelligent fallbacks and automatic retries ensure uninterrupted service and quick recovery from LLM provider failures.

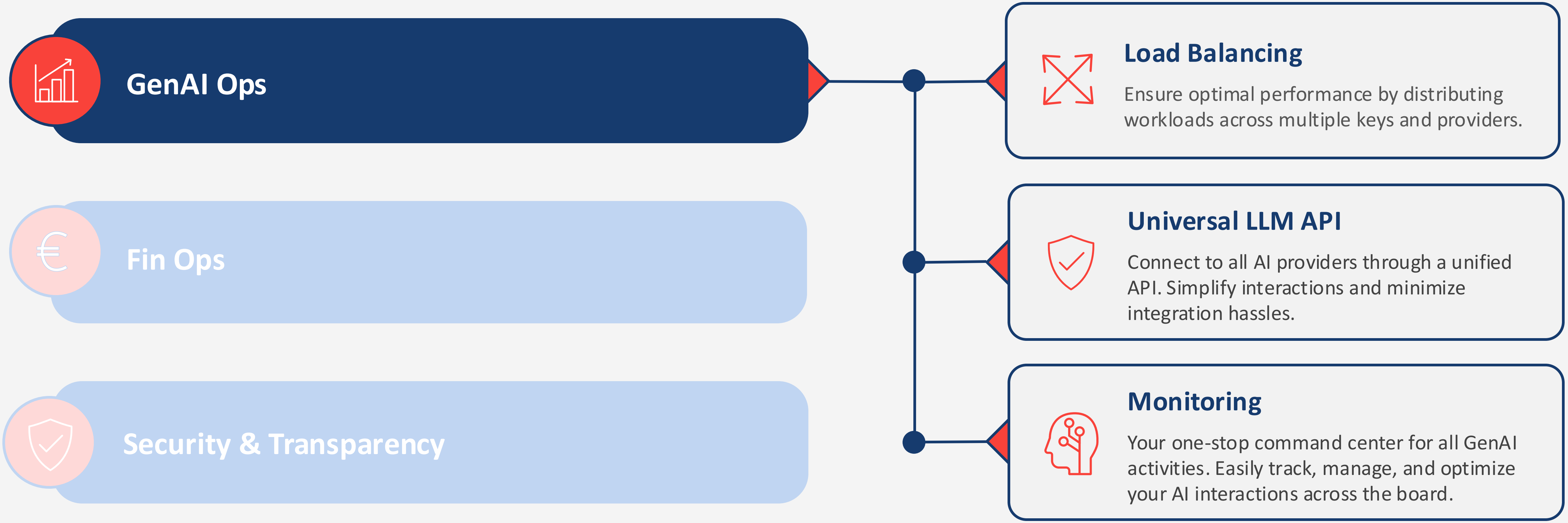
Better response time

Ensure better response time and reduce quota impact with robust load balancing and automatic retries.

Compatibility with GenAI libraries

Allow developers to use the tools best fitting the job by ensuring compatibility with all open source and closed source GenAI libraries and tools.

LLM Gateway helps you control:



GenAI Ops

Fin Ops

Security & Transparency

Load Balancing

Ensure optimal performance by distributing workloads across multiple keys and providers.

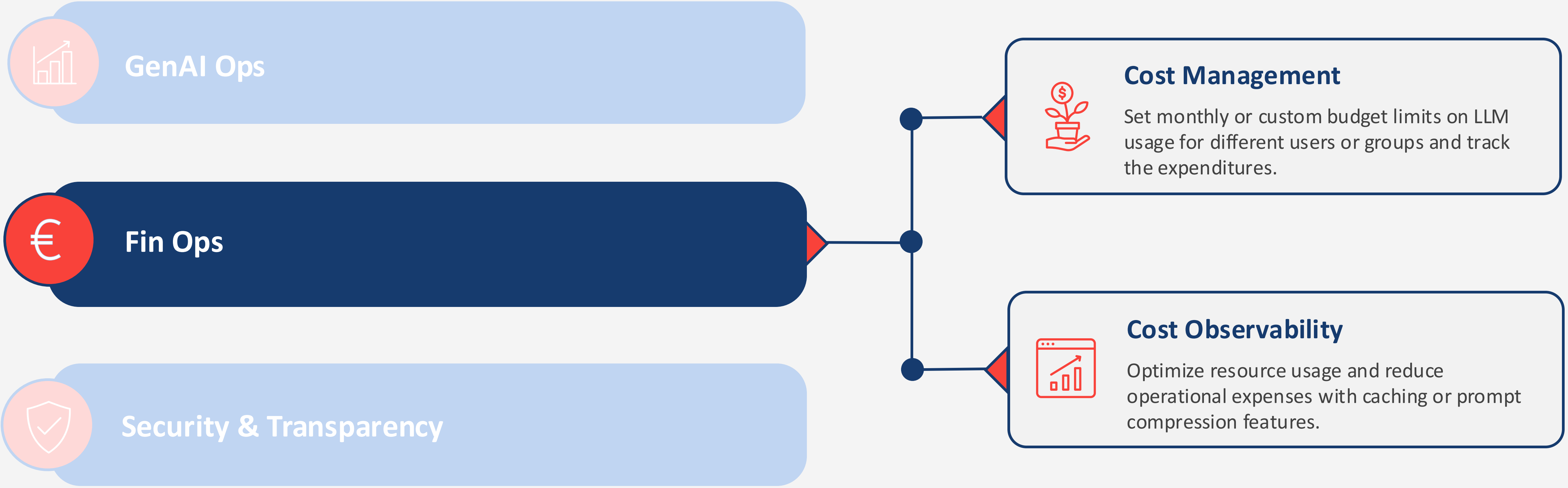
Universal LLM API

Connect to all AI providers through a unified API. Simplify interactions and minimize integration hassles.

Monitoring

Your one-stop command center for all GenAI activities. Easily track, manage, and optimize your AI interactions across the board.

LLM Gateway helps you control:



GenAI Ops

Fin Ops

Security & Transparency

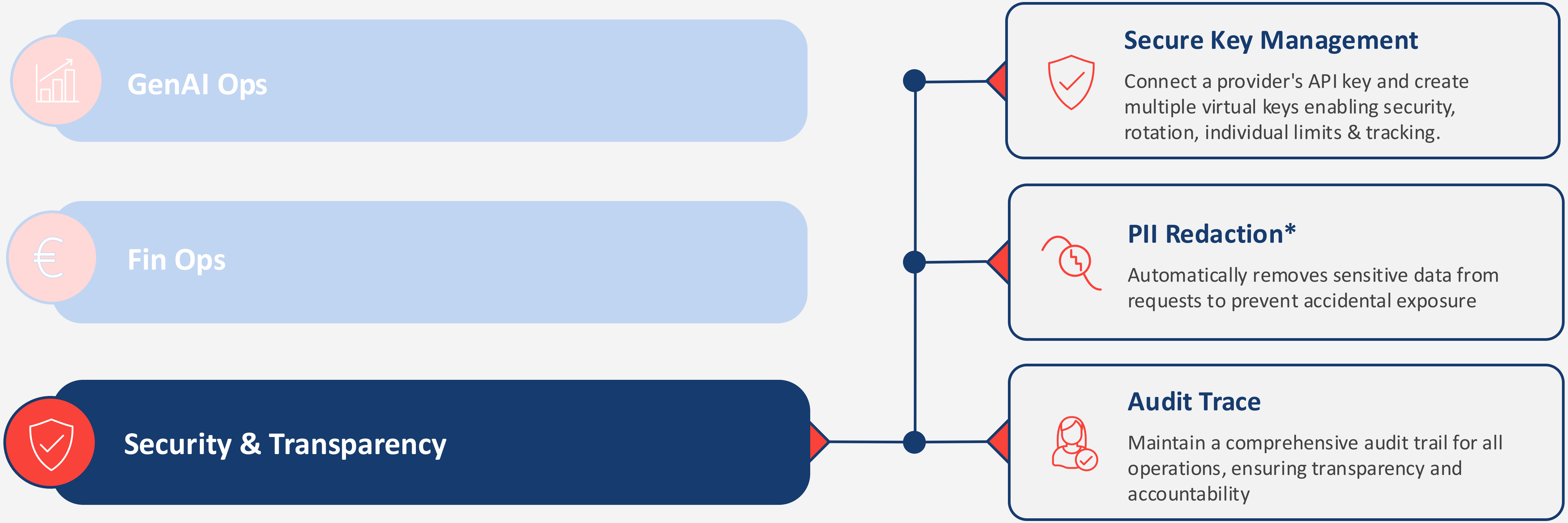
Cost Management

Set monthly or custom budget limits on LLM usage for different users or groups and track the expenditures.

Cost Observability

Optimize resource usage and reduce operational expenses with caching or prompt compression features.

LLM Gateway helps you control:



GenAI Ops

Fin Ops

Security & Transparency

Secure Key Management

Connect a provider's API key and create multiple virtual keys enabling security, rotation, individual limits & tracking.

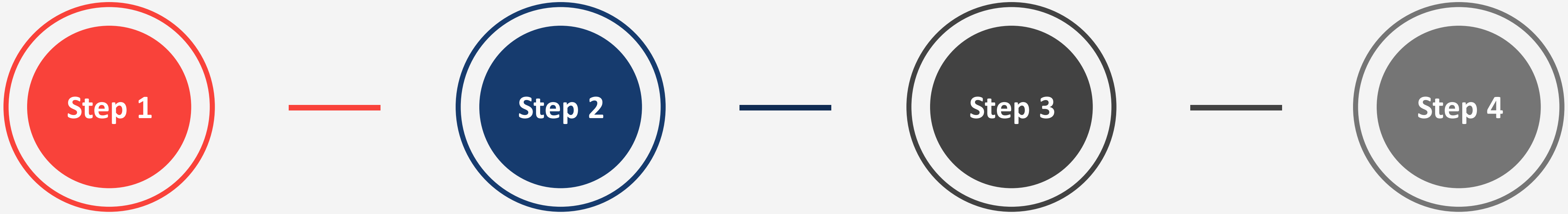
PII Redaction*

Automatically removes sensitive data from requests to prevent accidental exposure

Audit Trace

Maintain a comprehensive audit trail for all operations, ensuring transparency and accountability

Steps of LLM Gateway Deployment



Step 1

Infrastructure Analysis

Assess the client's existing infrastructure to understand compatibility and integration needs.

Step 2

Use Case Identification

Understand LLM Gateway functionalities and how they can benefit your business. Choose a tier that best suits your company‘s needs.

Step 3

Deployment

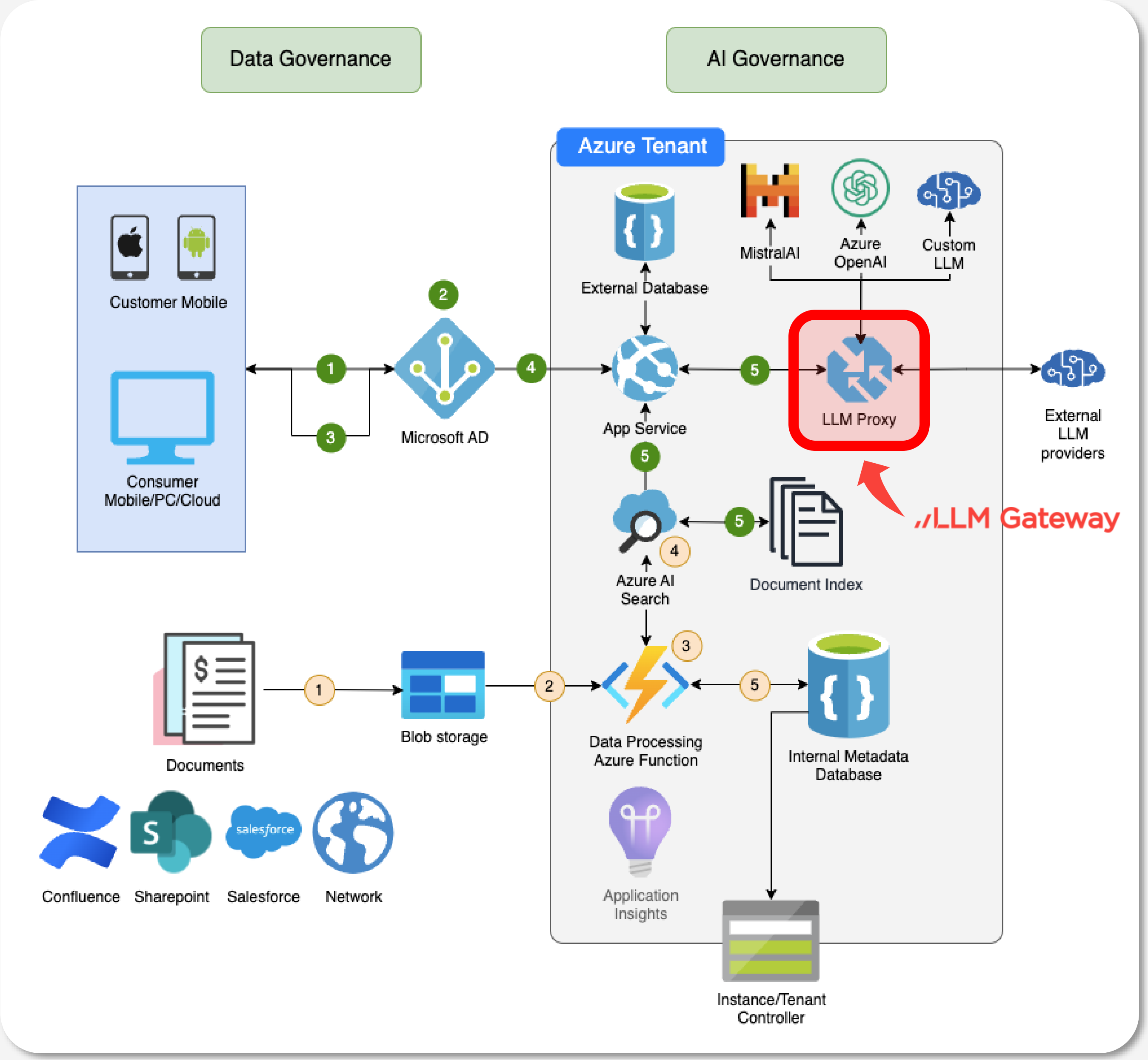
Deployment of LLM Gateway to the selected cloud with the setup of according services.

Step 4

Continuous support

Ongoing support and continuous updates with the newest LLMs.

Powering Your Next-Gen AI Infrastructure on MS Azure



LLM Gateway serves as the backbone of your architecture and ensures there is enough flexibility to effortlessly scale your GenAI infrastructure.

Out-of-the-box

cohere

together.ai

Vertex AI

Custom

Our team can work with you **to create custom integrations** that align with your specific use cases and workflows.

//AI

Comparison with Key Competitors

	//A | LLM Gateway	Microsoft	amazon	
	LLM Gateway	**API Management**	**API Gateway**	**Open-Source Tools**
Target APIs	Managed LLM APIs	Standard Backend APIs		Various, often backend APIs
Configuration for LLM scenarios	Easy	Complex for LLM scenarios		Customizable, often complex
LLM Tooling	Comprehensive	Limited		Variable, often lacking
Maintenance	Easy	Difficult		Variable
Support service	Included in license	Not provided		Not provided
Trustworthiness	High	High		Variable (community / startups)

Adopt the LLM Gateway now and enjoy the Early Bird benefits

Lower price

Propose & Prioritize features

Customize the platform

Terms and Conditions

- ✓ The Client provides a public reference (logo at minimum)
- ✓ The Client engages with the Supplier in a frequent feedback loop
- ✓ No business-critical use cases
- ✓ 8/5 SLA with a deployment cycle of a week

* EA price is only valid for the initial license period (up to 1 year maximum) and for a limited number of customers.

Early Bird Pricing

Tier 1: The Fundamentals

~~3 490 EUR/month~~

1 690 EUR/month

Fundamental Features:

- ✓ Universal LLM API
- ✓ Smart fallbacks
- ✓ Automatic retries
- ✓ Monitoring

Tier 2: Scale Up

~~5 990 EUR/month~~

3 000 EUR/month

All from Fundamentals, plus:

- ✓ Multimodal support
- ✓ PTU support
- ✓ Throttling
- ✓ Automatic load balancing
- ✓ Role-based cost management
- ✓ Use case-based cost management

Tier 3: Enterprise

~~11 990 EUR/month~~

5 500 EUR/month

All from Scale Up, plus:

- ✓ PII filtering*
- ✓ Semantic cache*
- ✓ Prompt compression*
- ✓ GDPR compatibility
- ✓ SOC 2 / NIS 2 / DORA ready
- ✓ Private Cloud Deployment

**Requires custom development*

+ **Infrastructure costs** associated with the clients' individual consumption

//AI

Thank You.

Jiří Čermák

Generative AI Lead

jiri.cermak@adastragrp.com

in

Jan Kalašnikov

Software Product Manager

jan.kalasnikov@adastragrp.com

in

//AI