Microsoft

AI-Pex Project

Enterprise Partner Solutions - Asia

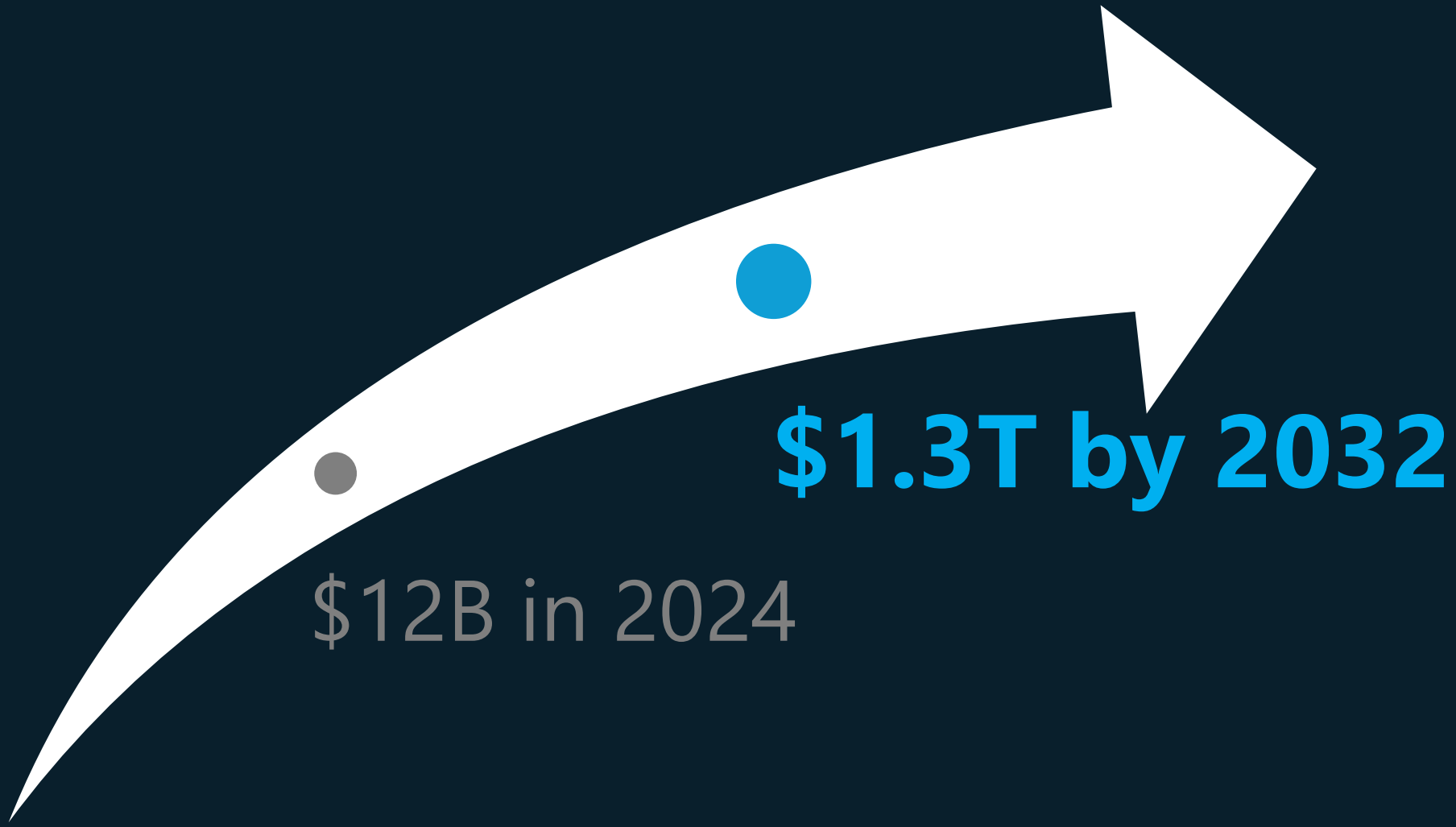
Partner Solution Architect Team

Agenda

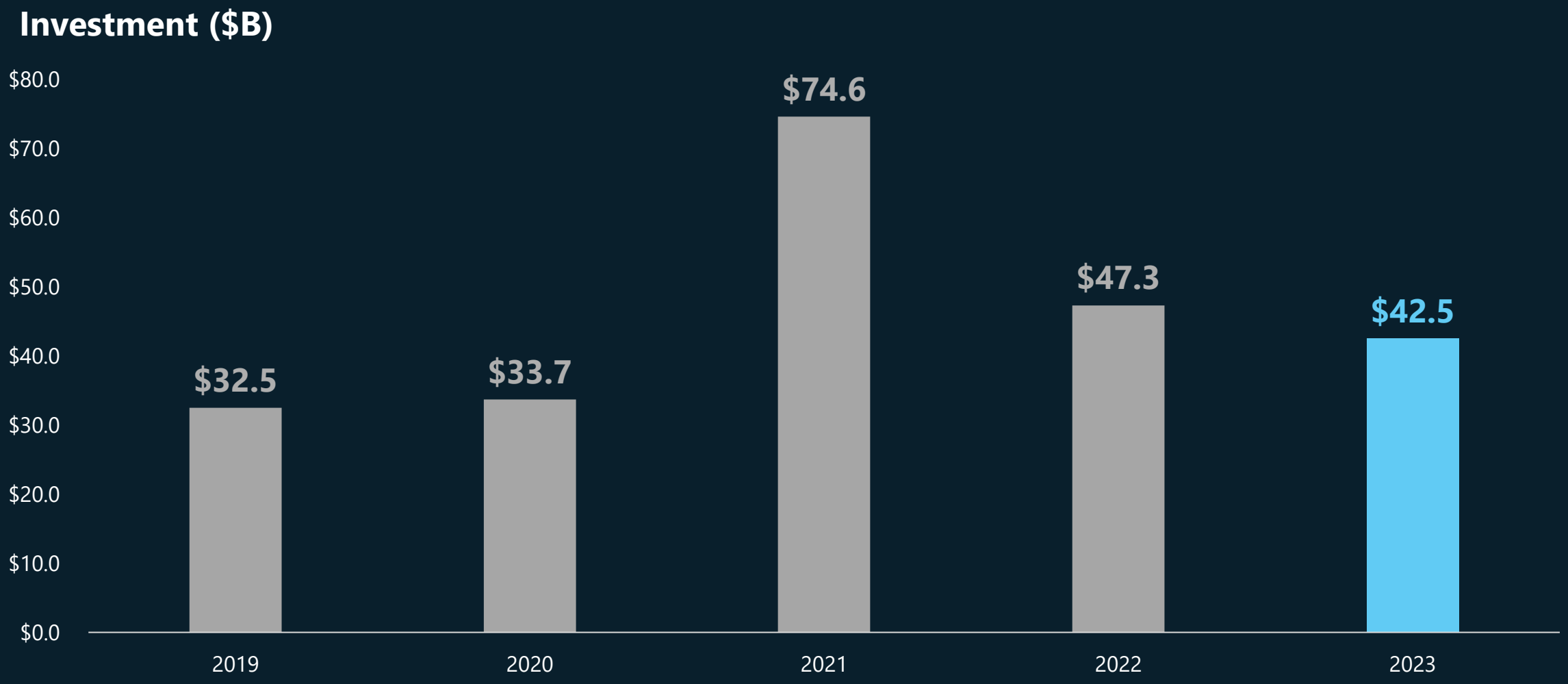
1. How impactful is AI SaaS for Business Growth?
2. Why AI-Pex is Important for ISV partners?
3. What does Microsoft Offer for AI SaaS?

How impactful is AI SaaS for Business Growth?

Generative AI Market in 2024



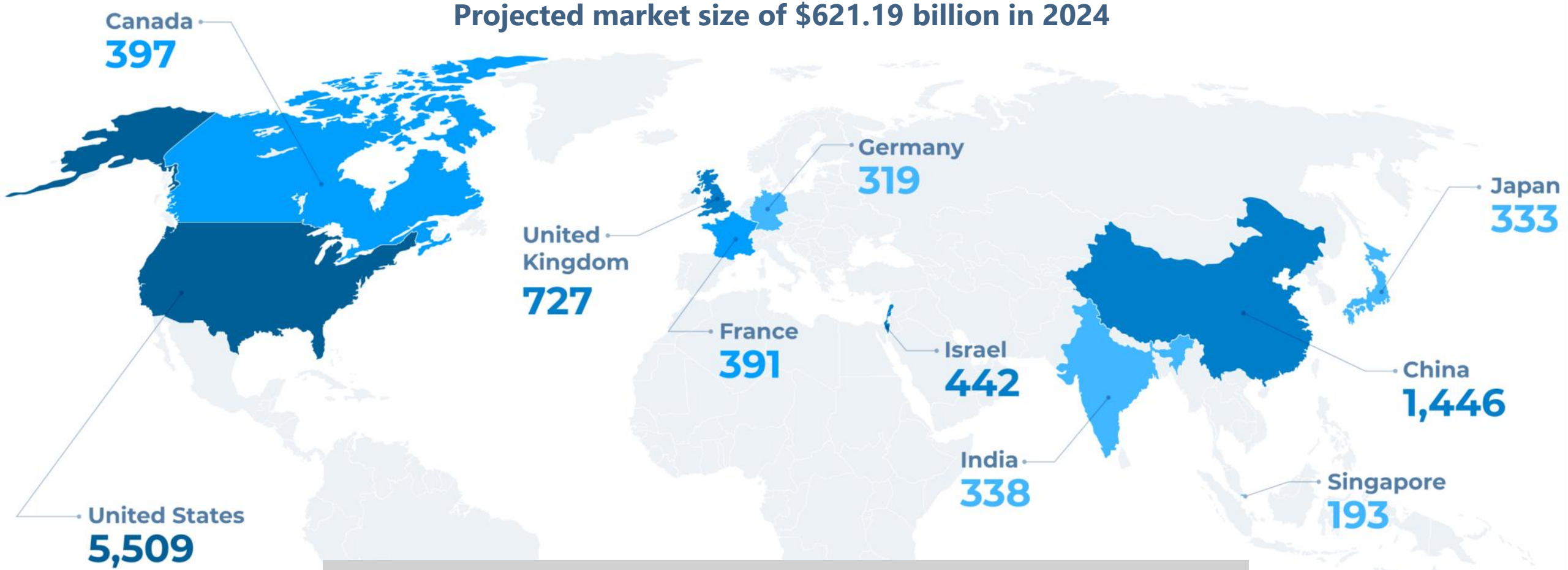
$42.5 billion invested



Source: State of AI 2023 Report - CB Insights Research

AI Startups by Country

Projected market size of $621.19 billion in 2024



+70,000 AI companies in the world

Source: Quid (2023)

*Newly funded AI startups that secured over $1.5 million in private investment between 2013 and 2023.

Ascendix

Driving the Generative AI Market

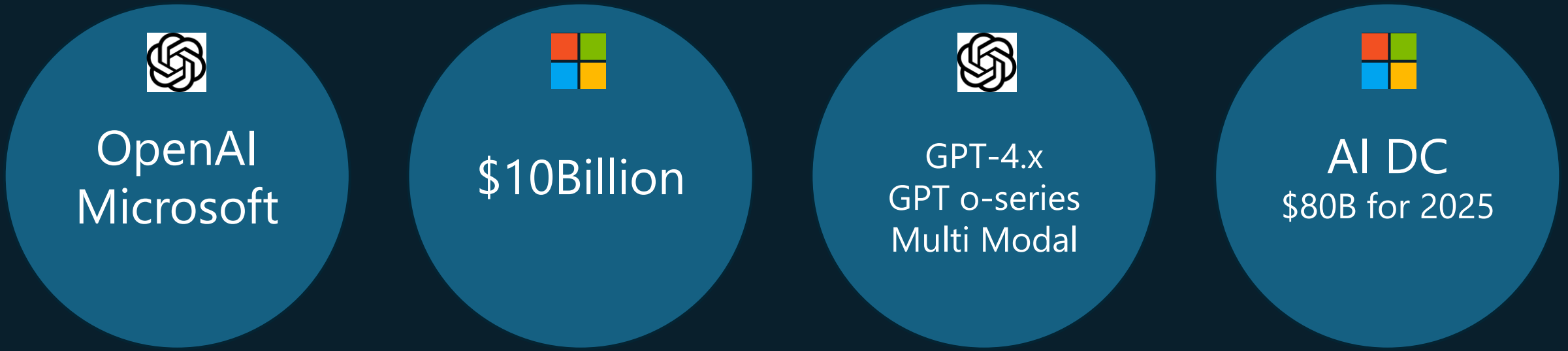
OpenAI & Microsoft are driving the growth of Gen AI market.

Ensure that Artificial General Intelligence (AGI) benefits all of humanity.

+

empower every person and every organization on the planet to achieve more

Key Factors



- OpenAI continues to drive the growth of the generative AI market.
- Microsoft has made a significant investment in OpenAI, amounting to $10 billion.
- GPT models are tailored variants of GPT-4.1 and GPT-40, optimized for diverse tasks from lightweight, low-latency operations to advanced reasoning and code generation.
- Microsoft's $80 Billion Investment in AI Data Centers for 2025.

Why AI-Pex is Important for ISV partners?

AI-Pex Overview

- **Business Growth:**

  AI app development is essential for ISV partners' business growth, and Microsoft offers AI-related workloads on Azure.

- **Talent Development:**

  Success in AI app development requires cultivating human resources as an 'AI App Center of Excellence', with Microsoft providing content and training resources.

- **Goal:**

  The 'AI-Pex' project aims to accelerate AI application development and training for ISV partners, covering Azure AI, Azure App, and Azure Data workloads, including some security content.

Goal of AI-Pex

- **Deliver AI contents**

  Accelerate AI application development and talent training with AI Intelligence Application content.

- **AI SaaS Development with Microsoft**

  Develop AI SaaS that can win customer deals in 5 weeks with Microsoft's phased content delivery and support from sales and architects.

How can we accelerate your idea to reality motion faster?

AI SaaS Development Approach

- We provide free support from the Envision phase to the phase where PoC and MVP can be built. Specifically, we start with mutual understanding between the partner and Microsoft, and provide technical support until the MVP is built.

- We offer this continuous support for 5 weeks (which can be adjusted based on the ISV Partner's situation), assisting in deploying the AI SaaS Application to production.
- Our approach is flexible, allowing adjustments around ISV Partners' sprints and maturity. Microsoft aims to help key individuals from ISV partner companies become an **AI App Development Centre of Excellence**.

Week 1 : Setup & Background

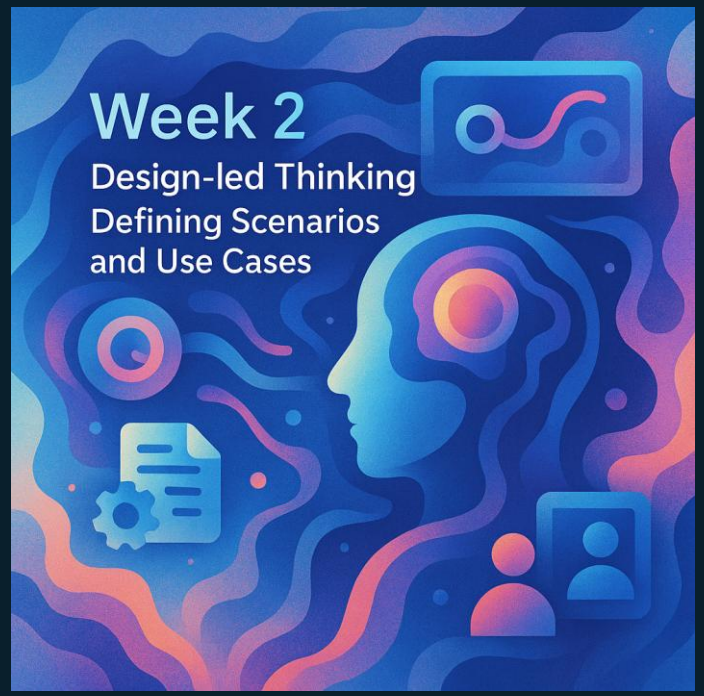
Key Points of Week 1

- Mutual Understanding of Goals
- Define Success criteria & Measurement
- Introduction to Azure Workloads

Before diving into that, we engage in conversations to understand the ISV partner's business, what their customers need, and what the ISV partner defines as success.

Week 2

Design-led Thinking
Defining Scenarios and Use Cases



- Focus on ISV Partner and Customer Needs
- Facilitated by Microsoft Team
- Prepare for MVP Development in Week 3

Week 3

Building MVP
Hands-On Support and Workshops

- Develop MVP with Microsoft Azure
- Tailored Scenarios for ISV Partner Needs
- Rapid Implementation with Common Architectures

Week 4

Scaling the MVP

Aligning with Customer Needs

- Discuss Architecture and Scaling Methods
- Evolve MVP to Production-Ready Solution

Week 5

Preparing for Production Deployment

Ensuring Quality and Security

- Implement Monitoring Methods
- Discuss Security Measures

Let’s get start: Week 1

Topics for us at first

1. **Understanding Your Business Goal**
   - We want to know more detail of ISV partner’s business goal/opportunity to develop AI SaaS application.
   - In this week 1, we can talk about the current situation and challenges of ISV partner.
2. **Define Success criteria & Measurement**
   - Decide what is success for ISV partner
   - Talk about how we do measure to achieve the success

Blank sheet; please use Topics 1 & 2

1. **Understanding Your Business Goal**

2. **Success criteria & Measurement**

Week 1 : Microsoft AI

Generative AI trends

93%
organizations are experimenting with multiple models[1]

61%
people are wary about trusting AI systems[3]

50%
generative AI will launch agentic AI pilots or POC by 2027[2]

30%
or fewer generative AI experiments moved to production by most respondents[4]

1. 16 Changes to the Way Enterprises Are Building and Buying Generative AI | Andreessen Horowitz
2. Autonomous generative AI agents | Deloitte Insights
3. Trust in artificial intelligence – 2023 Global study on the shifting public perceptions of AI, KPMG
4. GenAI and the future enterprise | Deloitte Insights

Organizations face significant challenges with complex, manual, and time-consuming workflows and minimal automation

- Inefficiency and High Operational Costs
- Lack of Visibility & Control
- Human Error
- Scalability Issues

The opportunity for autonomous AI is growing...

By 2028, at least **15% of day-to-day work decisions** will be made autonomously through agentic AI, up from 0% in 2024.*

* Gartner® 2025 Top Strategic Technology Trends E-Book, Gene Alvarez and Tom Coshow, 2024, emt.gartnerweb.com/ngw/globalassets/en/information-technology/documents/trends/2025-top-tech-trends-ebook.pdf. GARTNER is a registered trademark and service mark of Gartner, 2nc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

Strengths of Generative AI (LLM)

Ability to communicate in a variety of languages

- Support for more than 50 natural languages such as Japanese and English
- Computer languages can also be interpreted and generated
- Databases, APIs, etc.
  Interact with digital tools in natural language

Instant and extensive reading comprehension

- Instantly comprehends texts exceeding 100,000 characters
- Can convey complex content such as documents, web pages, and work instructions

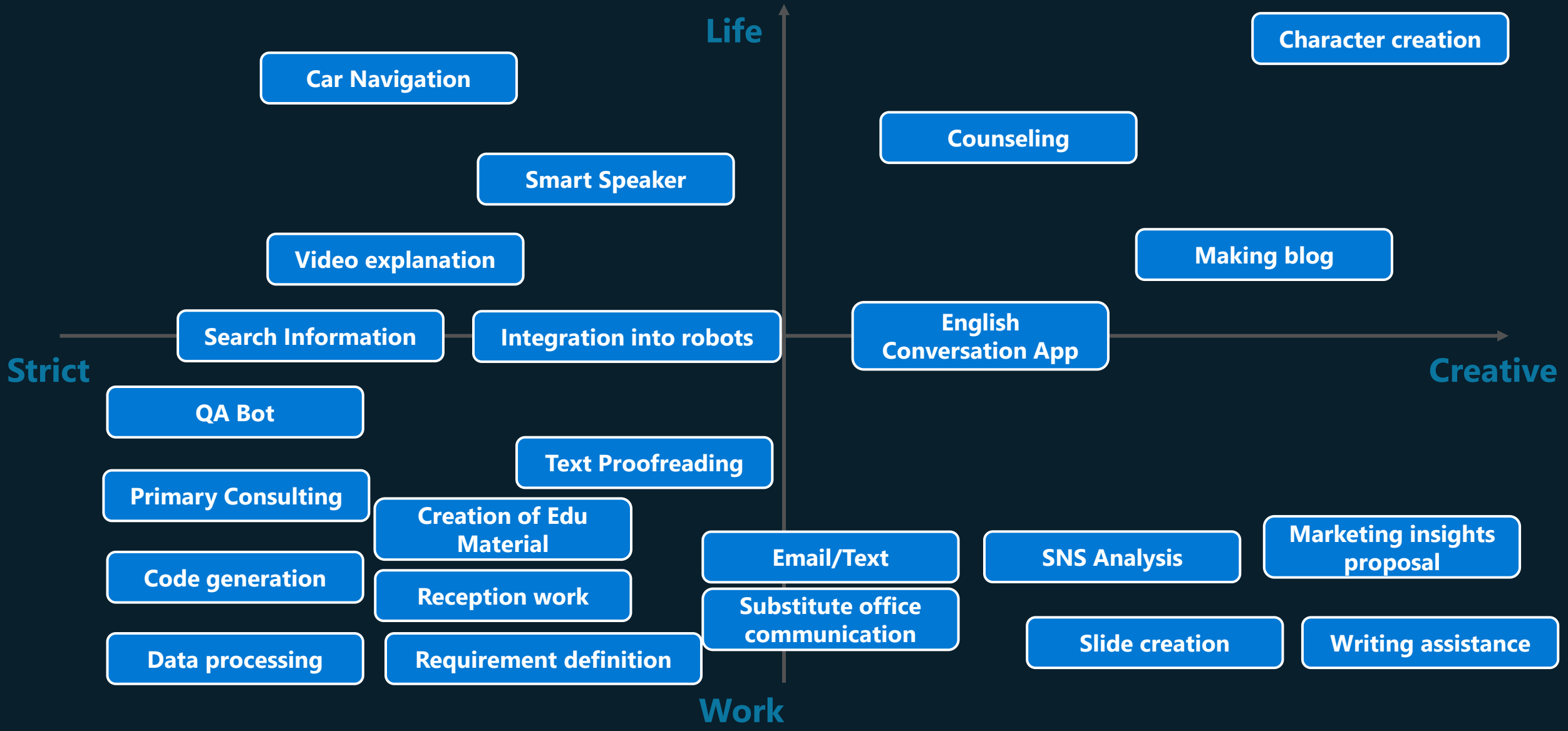
Scalable and knowledgeable

- Diverse information on the world's web in internal storage. It is being learned.
- In combination with external memory
  It is also possible to acquire new knowledge.
- With an open model, it is possible to update internal knowledge.

Flexible role and task changes

- Can change roles confidently according to user's instructions
- Can understand and respond to not only behavior but also detailed task instructions on the spot"

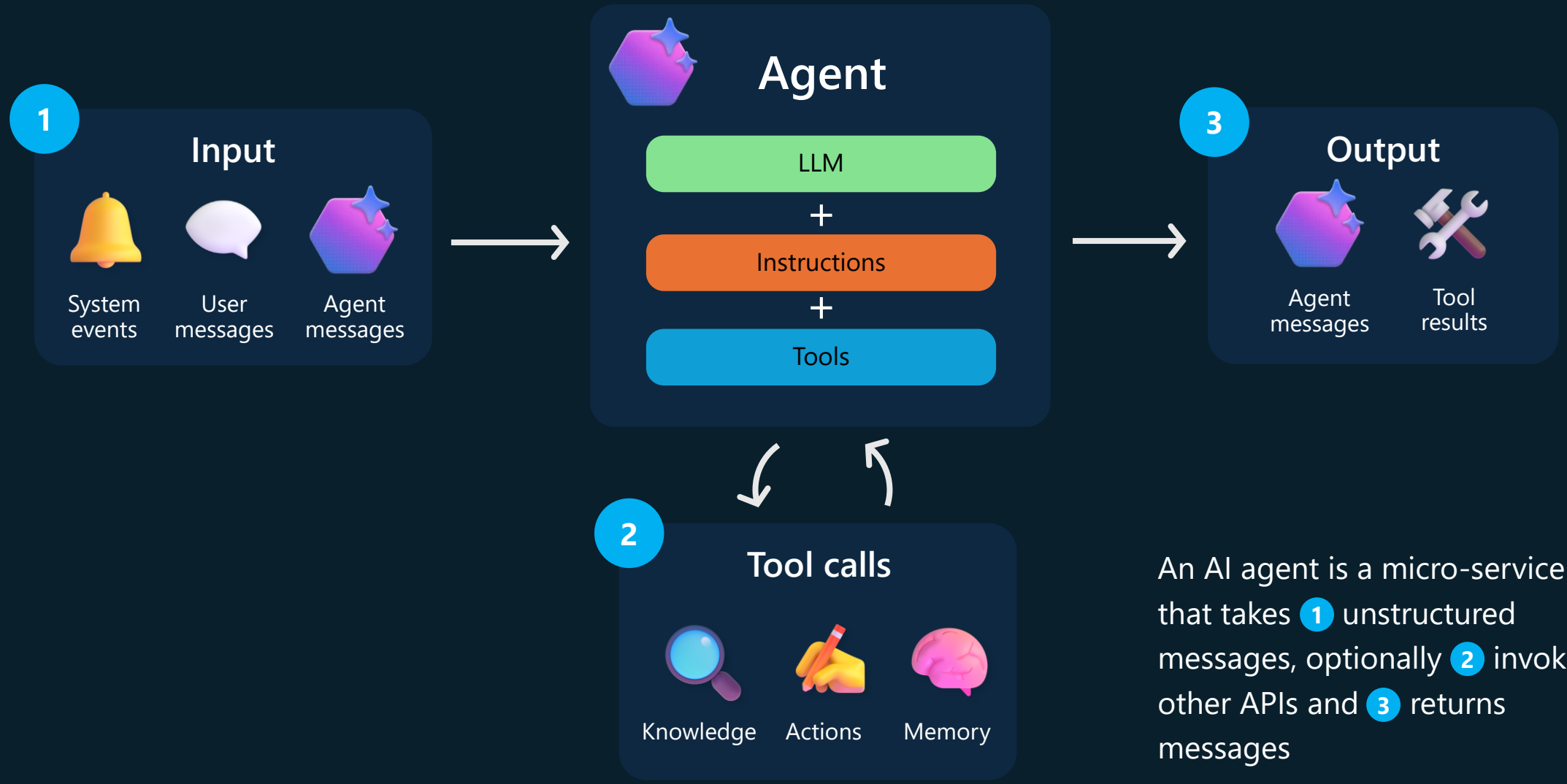
Mapping Use Cases & Scenarios



A new frontier:

Agentic AI

What is an AI agent?



An AI agent is a micro-service that takes 1 unstructured messages, optionally 2 invokes other APIs and 3 returns messages

What are agents?

AI designed to perform a task

Tasks can vary in level of complexity and capabilities depending on your need

Simple → Advanced

Generation

Generate summaries, images, audio, and more with an AI model and inputs.

Generally available

Retrieval

Retrieve information from grounding data, reason, summarize, and answer user questions

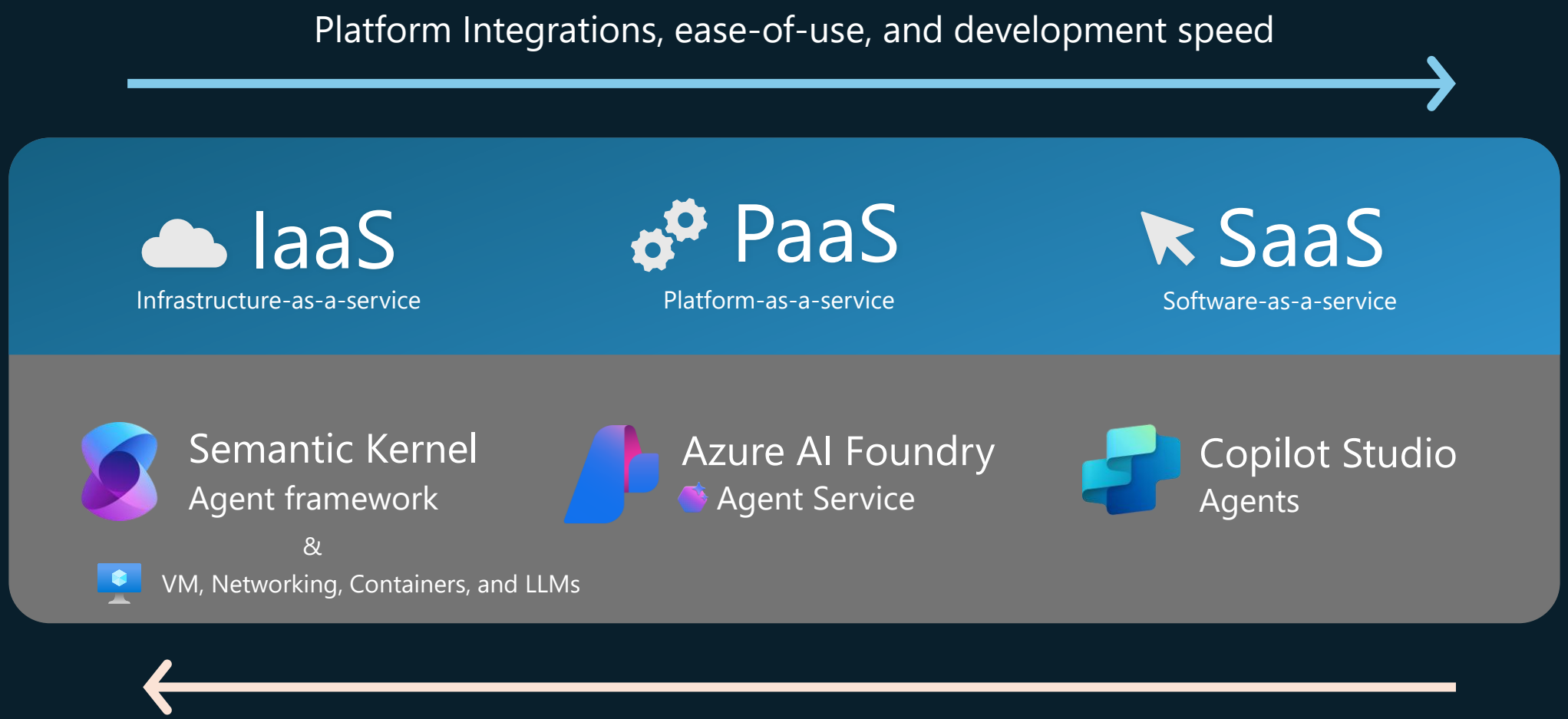
Generally available

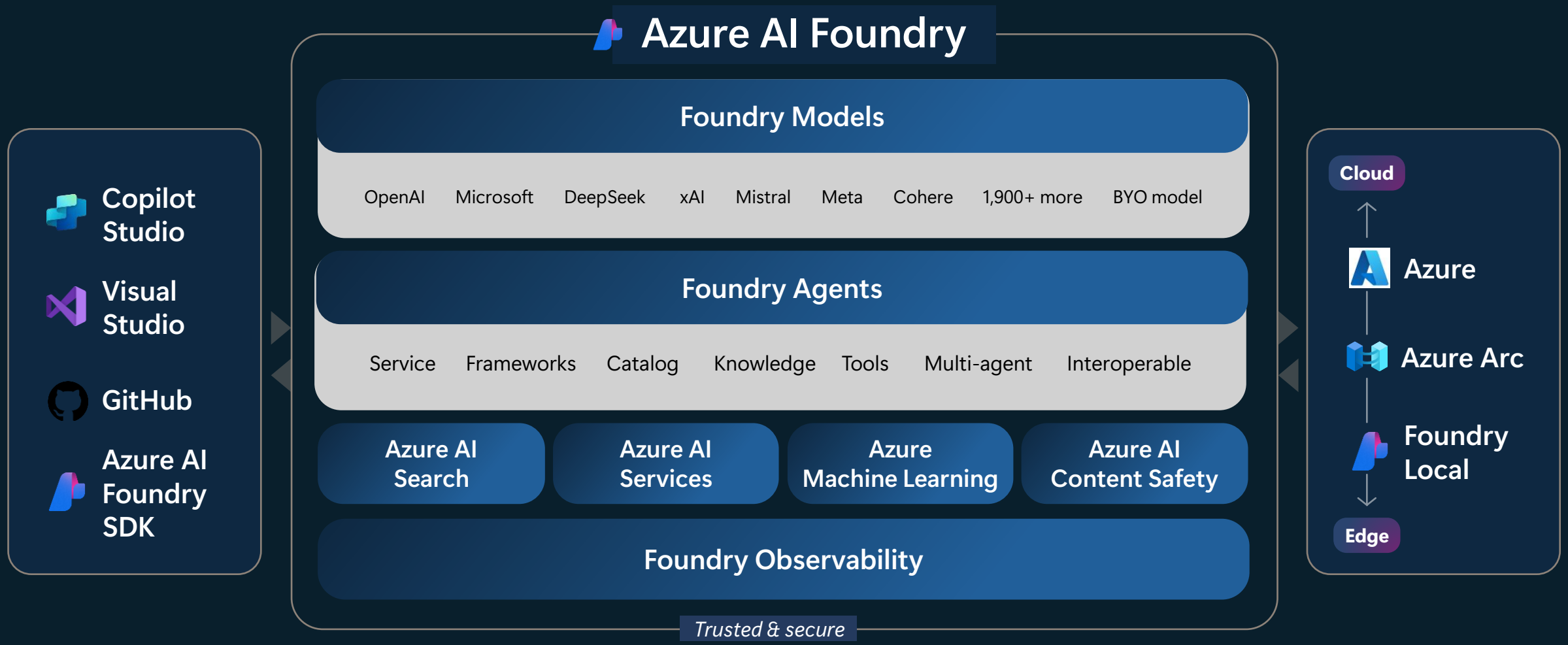
Action

Take actions to automate workflows, and replace repetitive tasks for users

Generally available

Microsoft has different ways of building agents




Azure AI Foundry
Foundry Models
OpenAI
Microsoft
DeepSeek
xAI
Mistral
Meta
Cohere
1,900+ more
BYO model
Foundry Agents
Service
Frameworks
Catalog
Knowledge
Tools
Multi-agent
Interoperable
Azure AI Search
Azure AI Services
Azure Machine Learning
Azure AI Content Safety
Foundry Observability
Trusted & secure
Copilot Studio
Visual Studio
GitHub
Azure AI Foundry SDK
Cloud
Azure
Azure Arc
Foundry Local
Edge

Azure AI Foundry

AI Agent Ecosystem

Authoring Tools

- Visual Studio
- GitHub

Knowledge

- Microsoft Fabric
- Azure Cosmos DB
- SharePoint
- Microsoft Graph
- Azure AI Search
- Microsoft Bing

Enterprise Trust

- VNet deployments
- OBO Auth
- BYO-resources
- Managed Identities

Agent catalog

Ready-made agents to kickstart your agent workforce

Agent frameworks

Client SDKs for enterprise and production agentic systems

- Agent tools
- Orchestration
- Channels

Foundry Agent Service

Deploy and manage agents with fully-managed runtime

- Multi-agent workflows
- Built-in threads
- Long-term memory
- Evaluation
- Tracing & Monitoring
- Governance + Safety

Foundry Models

Azure Direct Models

OpenAI, Llama, Grok, Flux, Mistral

Azure Ecosystem Models

Phi, Cohere, Paige, Stability, Hugging Face, Industry Models

Foundry Labs

- Muse
- Magentic-One
- Aurora
- OmniParser

Open Ecosystem

Protocols

A2A, MCP, OpenAPI

Connectors

Google VertexAI, Amazon Bedrock, CrewAI, LangChain, ElasticSearh, Pinecone, Logic Apps, Azure Functions

No Gating for gpt4.1 models!

New Release

Generally Available

GPT-4.1, 4.1-mini, and 4.1-nano

Sign up to access in Azure AI Foundry or talk to your representative

https://aka.ms/new -models

Generally Available

GPT-4.1

- Enhanced coding
  - Consistently produces outputs that compile and run successfully
- Longer context input
  - 1m tokens
- Improved instruction following
  - Especially those containing multiple requests

Generally Available

GPT-4.1-mini

- Reduced latency
  - Nearly half the latency of GPT-4o
- Cost efficiency
  - Reduces computational cost by 83%
- Performance:
  - Matches or exceeds GPT-4o intelligence

Generally Available

GPT-4.1-nano

- Ultra-low latency
  - Fastest responses in the 4.1 series
- High efficiency
  - Lowest computational cost
- Performance
  - Scores 80.1% on MMLU, 50.3% on GPQA, and 9.8% on Aider polyglot coding.

1m token context Length*

Improved coding for agentic workflows

Improved instruction following

June 2024 knowledge cutoff

GPT Models

Recent Releases as of Apr 2025

	Use Cases	Key Features	Advantages
GPT-4.1	• Customer Service Chatbots • Data Analysis Tools • Machine Learning Applications	• Enhanced Coding • Longer context input • Improved instruction following	• Consistently produces outputs that compile and run successfully • Generates cleaner, simpler front-end code • Understands extensive context in a single interaction
GPT-4.1-mini	• Customer Service Chatbots • Data Analysis Tools • Machine Learning Applications	• Reduced Latency • Cost Efficiency • Performance	• Faster response times • Lower computational costs • Balanced performance
GPT-4.1-nano	• Instantaneous Data Classification • Autocompletion Tools • Basic Content Routing	• Ultra-Low Latency • High Efficiency • Performance	• Immediate feedback • Highest efficiency of the 4.1 series • Quick data processing
GPT-4o-mini	• Customer support chatbots • Applications that chain or parallelize multiple model calls • Handling large volumes of context	• Multimodal inputs • Context window of 128K tokens Strong textual intelligence and multimodal reasoning	• Outperforms GPT-3.5 Turbo and other small models on academic benchmarks • Strong in reasoning tasks, math, and coding Enhanced long-context performance
GPT-4o	• Customer support automation • Content creation • Translation • Code generation • Programming assistance	• Enhanced natural language understanding • Improved context-awareness and coherence • Expansion of fine-tuning capabilities	• Increased accuracy and fluency • Greater versatility across multiple applications • Improved performance in low-resource languages

Azure OpenAI Service models

Models	Description
GPT-4.1 series	Latest model release from Azure OpenAI
computer-use-preview	An experimental model trained for use with the Responses API computer use tool.
GPT-4.5 Preview	The latest GPT model that excels at diverse text and image tasks.
o-series models	Reasoning models with advanced problem-solving and increased focus and capability.
GPT-4o & GPT-4o mini & GPT-4 Turbo	The latest most capable Azure OpenAI models with multimodal versions, which can accept both text and images as input.
GPT-4	A set of models that improve on GPT-3.5 and can understand and generate natural language and code.
GPT-3.5	A set of models that improve on GPT-3 and can understand and generate natural language and code.
Embeddings	A set of models that can convert text into numerical vector form to facilitate text similarity.
Image generation	A series of models that can generate original images from natural language.
Audio	A series of models for speech to text, translation, and text to speech. GPT-4o audio models support either low-latency, "speech in, speech out" conversational interactions or audio generation.

Note: As of May 2025

Key Differences : OpenAI and Azure

Azure OpenAI Services is similar to OpenAI's API, with added features for easier enterprise use in Microsoft Azure.

Features in Microsoft Azure	Generation and fine-tuning are possible simply by making requests to the endpoint. **The API specifications and libraries are essentially standardized with OpenAI's API.**
SLA & Support	Establish an SLA guaranteeing over 99.9% uptime, with Azure support services available. Licensing Documents (microsoft.com)
Microsoft Entra ID	Authentication using Microsoft Entra ID (formerly Azure AD) is available in addition to key-based authentication. Authentication in Azure AI services - Azure AI services | Microsoft Learn
Contents Filtering	Detection of harmful expressions, LLM hijacking, and existing code or text. Use content filters (preview) with Azure AI Foundry - Azure OpenAI | Microsoft Learn
Integration of Private Network	High-security request configurations within a closed virtual network are possible. Configure Virtual Networks for Azure AI services - Azure AI services | Microsoft Learn
Multi Region	Available in multiple regions, ensuring availability through decentralization and ample rate limits. Azure OpenAI Service quotas and limits - Azure AI services | Microsoft Learn
Monitoring	Equipped with a logging mechanism for monitoring requests. Monitor Azure OpenAI Service - Azure AI services | Microsoft Learn
Purchasing Through Put	Stable throughput ensured by pre-purchasing PTUs. Azure OpenAI Service Provisioned Throughput Units (PTU) onboarding - Azure AI services | Microsoft Learn
Deployment RAG App with Low-Code	Rapid development of RAG mechanisms and deployment of chat UIs in combination with Azure AI Search. Using your data with Azure OpenAI Service - Azure OpenAI | Microsoft Learn
GPT-4V support Japanese	Advanced OCR and video analysis capabilities through extensions combined with Azure AI Service. How to use vision-enabled chat models - Azure OpenAI Service | Microsoft Learn
Customer Copyright Commitment	Protection against claims related to specific third-party intellectual property rights associated with output content, provided certain usage conditions are met. Customer Copyright Commitment Required Mitigations | Microsoft Learn

Wide range Managed AI Services on Azure

Azure AI Foundry

Integrated platform on Azure for deploying, managing, and developing AI models, supporting development from low-code to SDK-based approaches.

Azure ML Studio

Cloud-based development environment platform where data scientists can create, train, evaluate, and deploy machine learning models.

Azure AI Search

Powerful search service that quickly retrieves necessary information from large datasets, enhancing search accuracy with AI.

Document Intelligence

AI service that automatically extracts text and tabular data from documents to streamline business processes.

Azure AI Translator

Real-time translation across multiple languages, easily translating text and speech.

Azure AI Speech

Speech recognition and synthesis service that converts speech to text and text to speech.

Azure AI Language

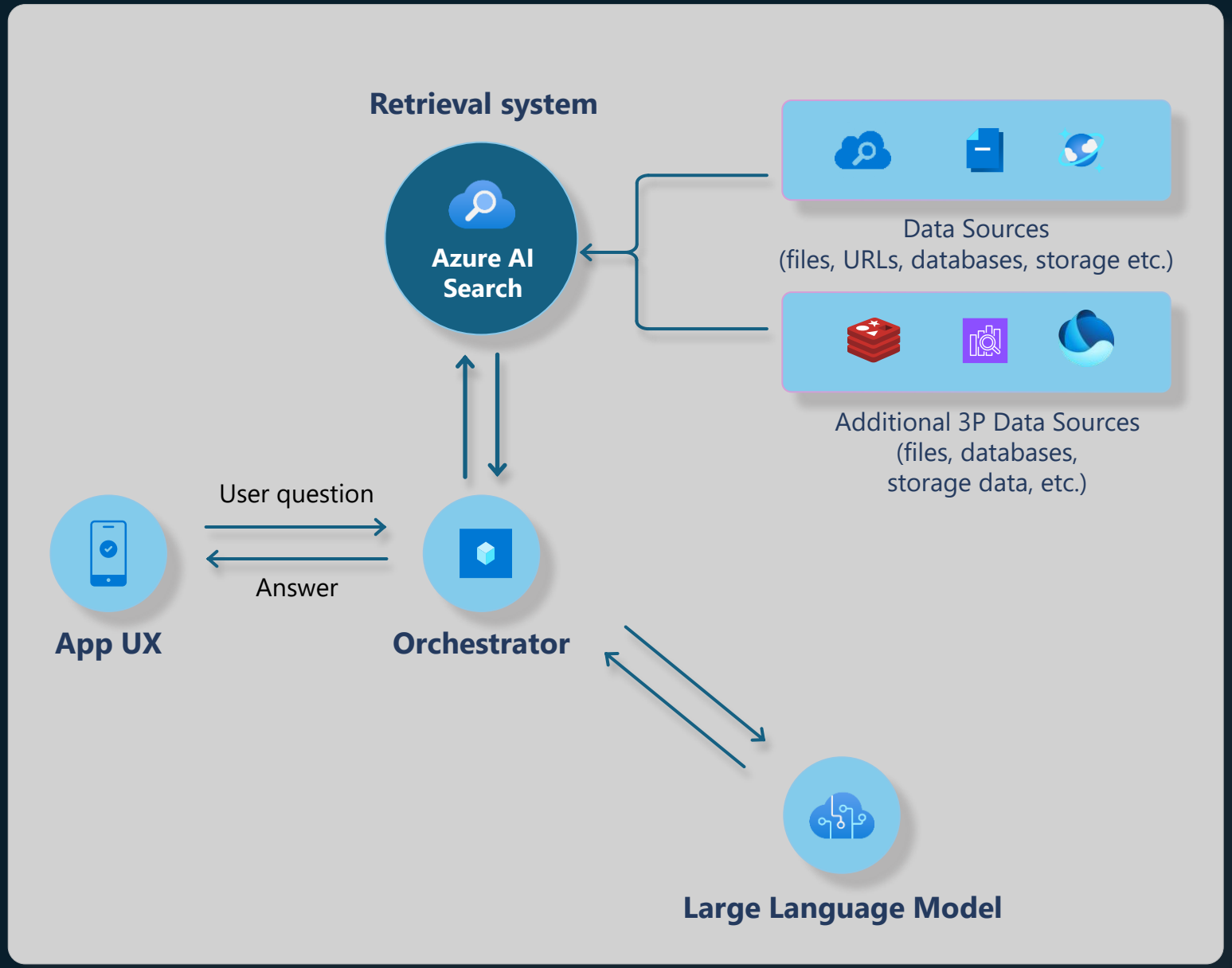
Performs sentiment analysis, topic classification, and summarization of text through natural language processing.

Computer Vision

Recognizes and analyzes objects and text in images and videos, aiding in the understanding and processing of visual data.

Retrieval-Augmented Generation

RAG Architecture for LLM to provide more accurate answers to prompts by referencing external information.
Azure AI Search and other **Azure AI Services** can help to build RAG application in AI SaaS Application.



Azure AI Search

Feature-rich vector database

Optimized vector storage

Seamless data & platform integrations

State-of-the-art search technology

Enterprise-ready foundation

Expanded storage and vector index size

Azure AI Search

Revolutionary retrieval

Quality

Build better apps with better search

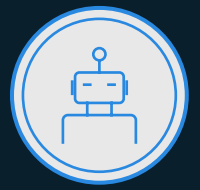
Agility

Flexibility with integrated platform and ecosystem

Scale

Perform RAG at scale with an enterprise-ready platform

Top Use Case in Azure AI Search



Conversational search & insights
Better knowledge mining

AI assistant
Better analytics and service

Automation
Data and document processing

Content generation
Personalized recommendations

Example of Data Source for RAG

Although the combination with search engines tends to attract attention, anything that augments prompts with information from any data source can be considered RAG.

Azure AI Search
etc...

Documents Knowledge

Extract text information from internal PDFs, PowerPoint presentations, Excel files, etc., and search in response to GPT requests. Return information that is close to the question content as a response. Full-text search engines or vector stores are often used.

bing API
etc...

Web Search

Execute web search APIs like Bing or Google, extract information from each hit web page, and use it to respond. Microsoft Copilot uses this mechanism. The information from web pages needs to be retrieved and formatted as HTML each time.

SQL DB
etc..

Database

Query databases such as RDB and NoSQL DB to reference user information and conversation history data. There are methods to generate SQL itself, as well as to retrieve information from fixed SQL queries.

Other

Additionally, there are methods to obtain related information and facts from Knowledge Graphs, as well as methods to combine information retrieval with recommendation engines.

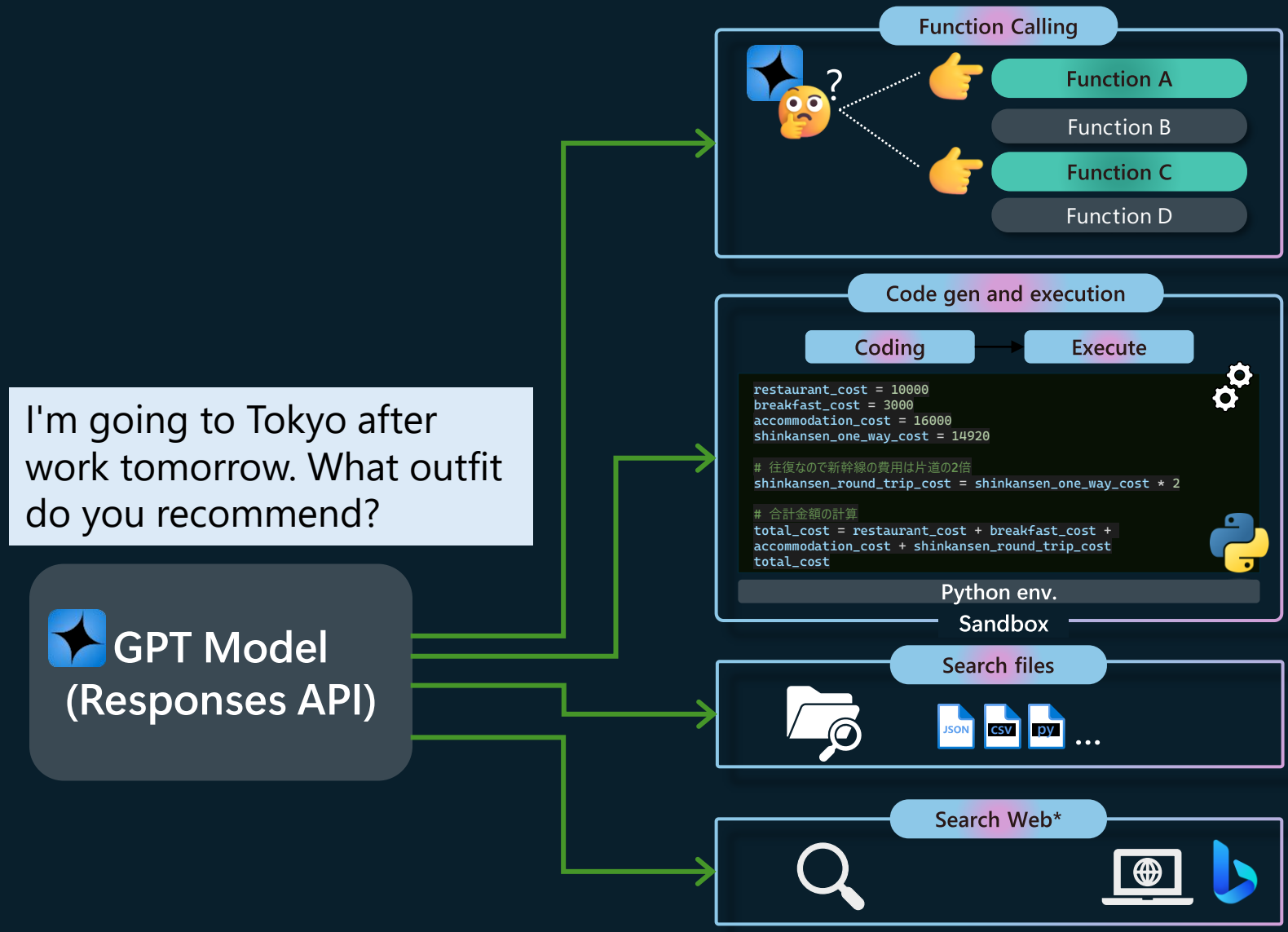
AI Agent

The definition is ambiguous, but the following features are often referred to as AI agents.

Feat. 1	Autonomy	Unlike regular AI chat services that operate based on user instructions, **AI agents act independently based on given goals, minimizing user intervention.**
Feat. 2	Goal-oriented	Unlike regular AI chat services that focus on answering user questions, **AI agents emphasize planning and acting towards achieving specific goals or tasks.**
Feat. 3	Advanced reasoning	AI chat services are generally limited to simple one-question-one-answer interactions. **AI agents, on the other hand, can handle tasks within complex and continuous dialogues. In some cases, multiple agents can be coordinated to solve problems as needed.**
Feat. 4	External integration	AI chat services have minimal integration with external tools and systems. **However, AI agents actively engage in external integrations because they need to autonomously execute tasks.**

[2408.02479] From LLMs to LLM-based Agents for Software Engineering: A Survey of Current, Challenges and Future (arxiv.org)

From RAG to AI Agent system



Function A; To get tomorrow's weather

Function C; To check tomorrow's schedule

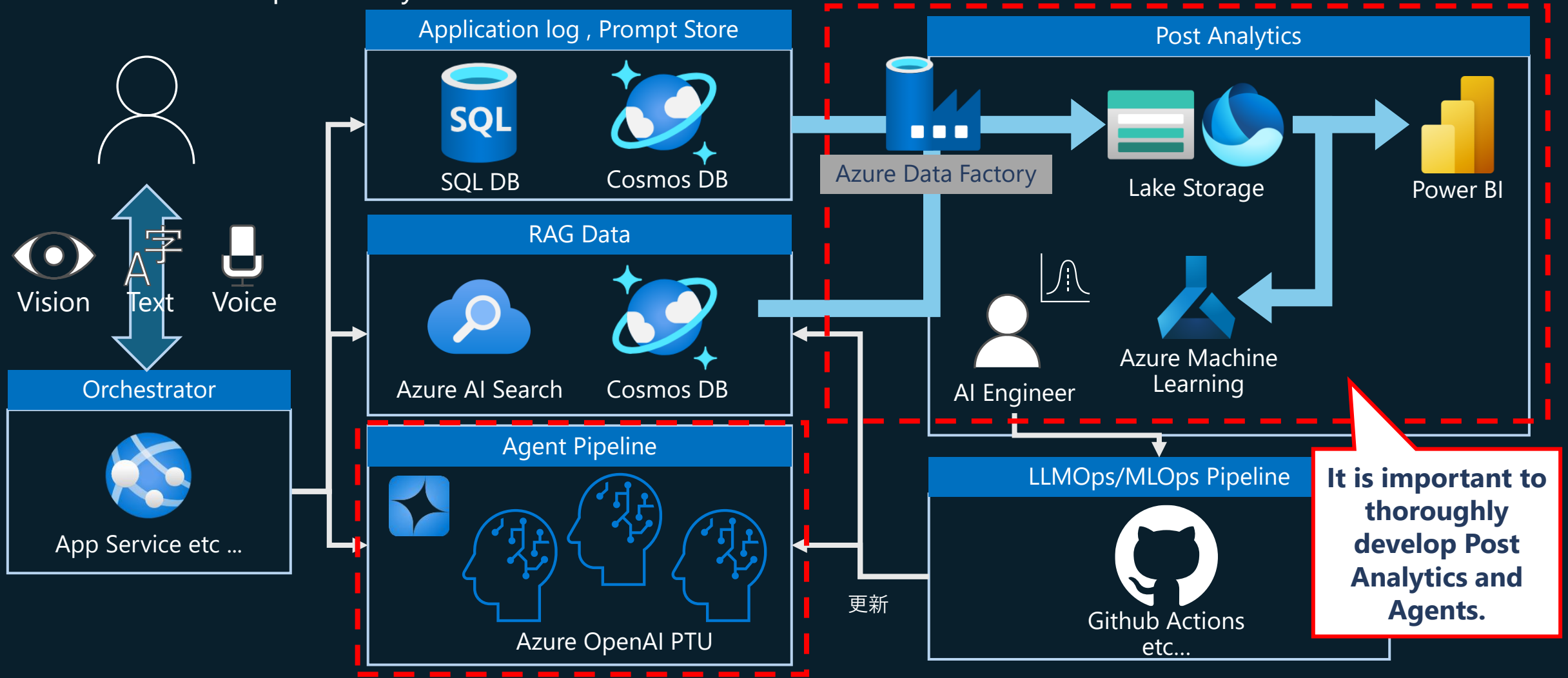
Calculate the total price of clothing

Materials related to temperature and clothing

Search web for recommended clothes

AI Agent Architecture

The architecture is similar to general LLM systems, but managing multiple agents requires careful consideration of post-analytics.



Resource

- GitHub Repo - microsoft/Solution-Accelerators
  - This repository provides solution accelerators and tools to enhance business efficiency using AI, including content processing, code modernization, conversational search, and custom AI assistant creation.
- https://ai.azure.com/code
  - **Creating and deploying AI applications using Azure AI Foundry**: Learn how to build AI agents and process multimodal content efficiently.

Closing

Key Takeaway

- **Generative AI Market Growth**
  The generative AI market is rapidly expanding, with significant investments and a projected market size of $1.3 trillion by 2032.

- **Microsoft's Role and Support**
  - Microsoft, in collaboration with OpenAI, is driving the growth of the generative AI market through substantial investments and advanced AI models like GPT-4o.
  - Additionally, Microsoft's AI SaaS development approach provides continuous support from the envision phase to production deployment, ensuring that ISV partners can efficiently build and scale their AI solutions over a 5-week period.

- **Responsible AI Practices**
  Microsoft emphasizes responsible AI practices, ensuring fairness, reliability, safety, privacy, security, accountability, transparency, and inclusiveness in all AI services.

Thank you.

Microsoft