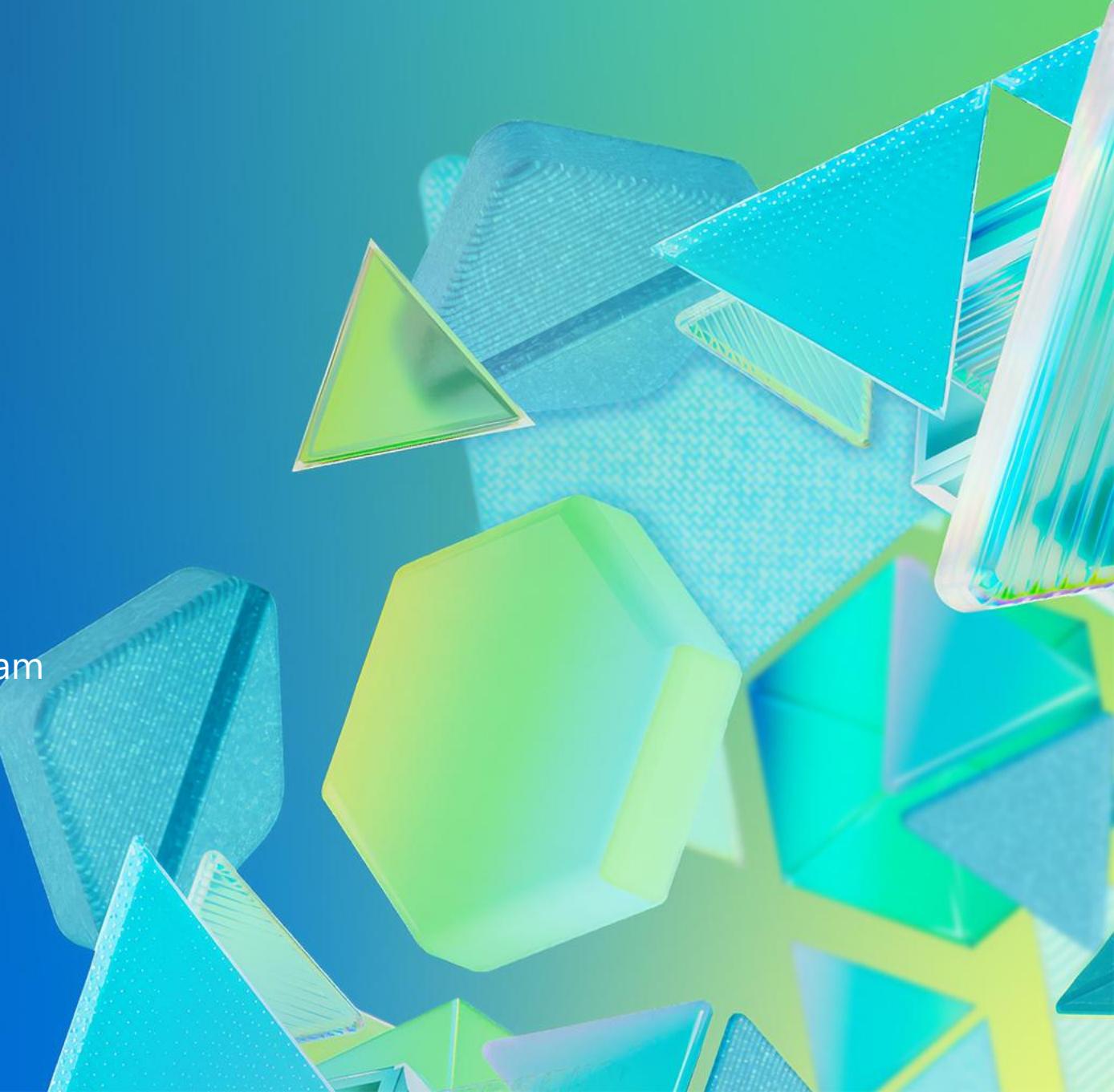




AI-Pex Project

Global Partner Solutions Asia
ISV Recruit Partner - Partner Solution Architect Team



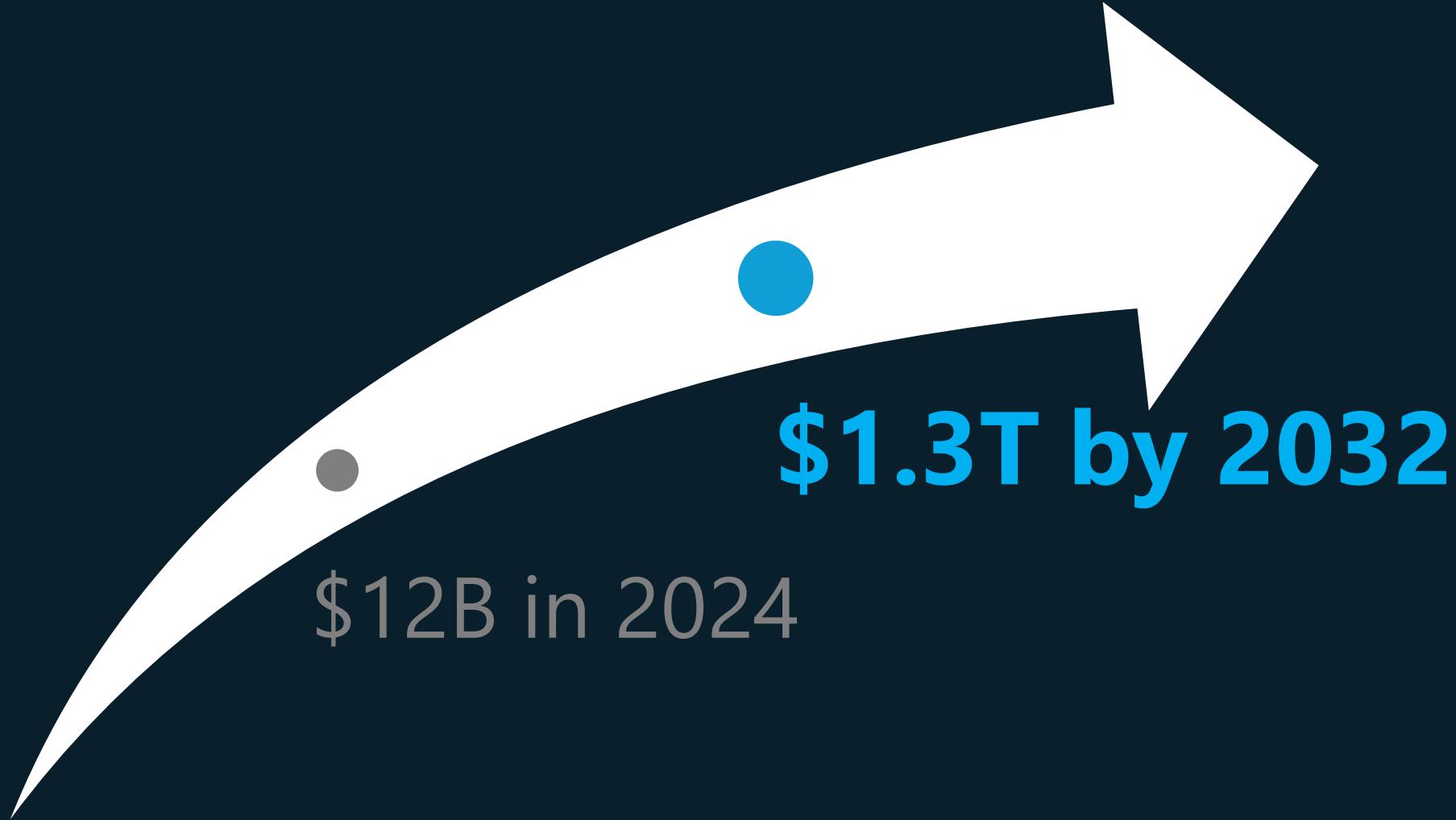
Agenda

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?

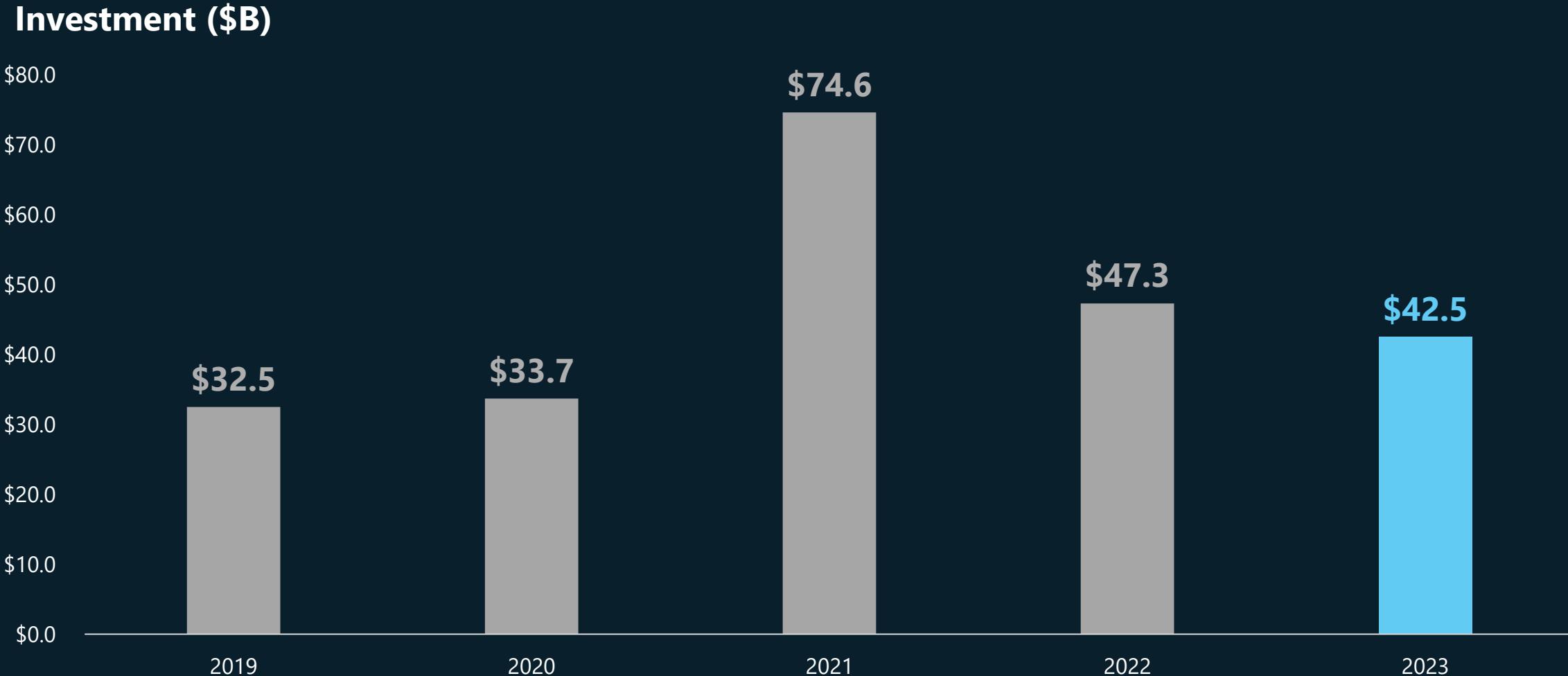


How impactful is AI SaaS for Business Growth?

Generative AI Market in 2024



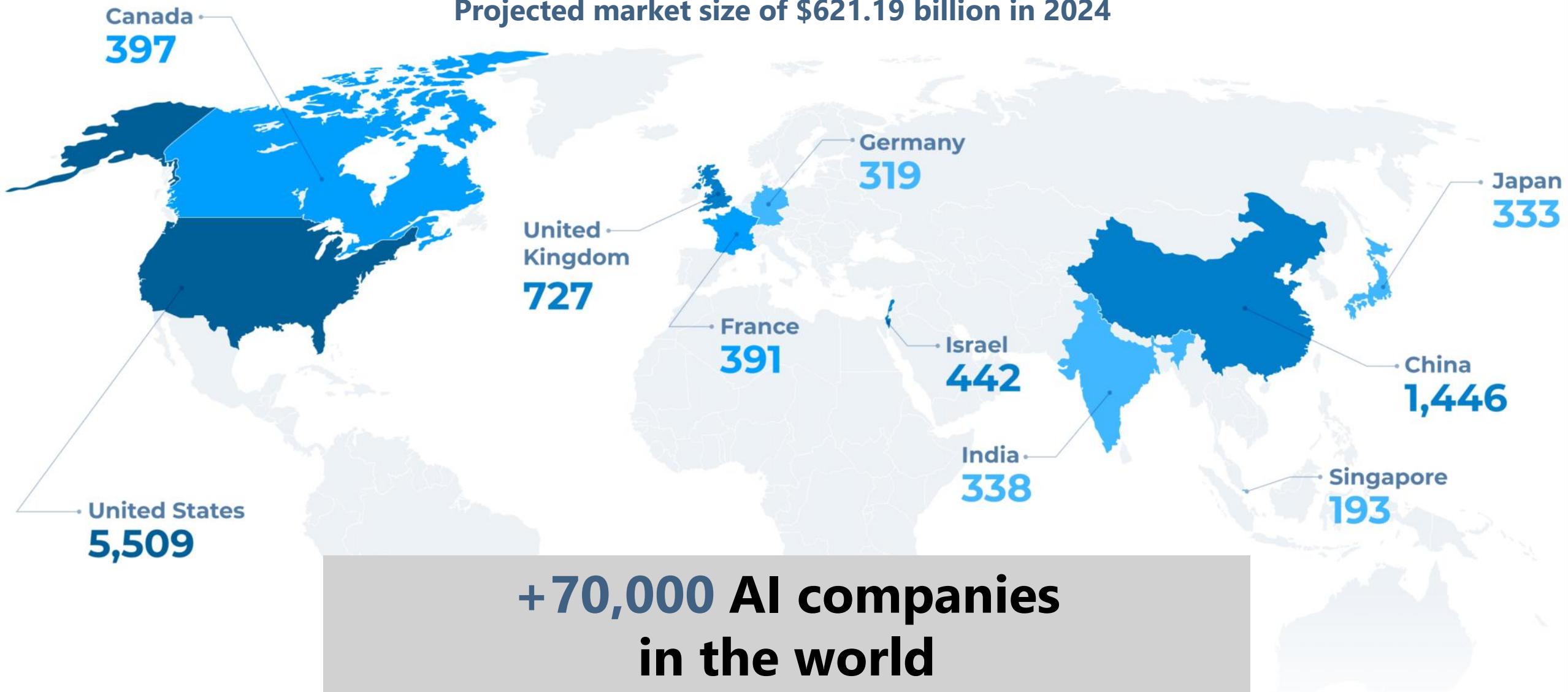
\$42.5 billion invested



Source: State of AI 2023 Report - CB Insights Research

AI Startups by Country

Projected market size of \$621.19 billion in 2024



Source: Quid (2023)

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

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
Microsoft



\$10Billion



GPT-4o
Multi Modal



AI DC
\$80B for 2025

- OpenAI continues to drive the growth of the generative AI market.
- Microsoft has made a significant investment in OpenAI, amounting to \$10 billion.
- GPT-4o model can accept text, audio, images, and video as inputs, recently GPT-o1/GPT-o3 announced.
- Microsoft's \$80 Billion Investment in AI Data Centers for 2025.

Generally available

GPT-4o

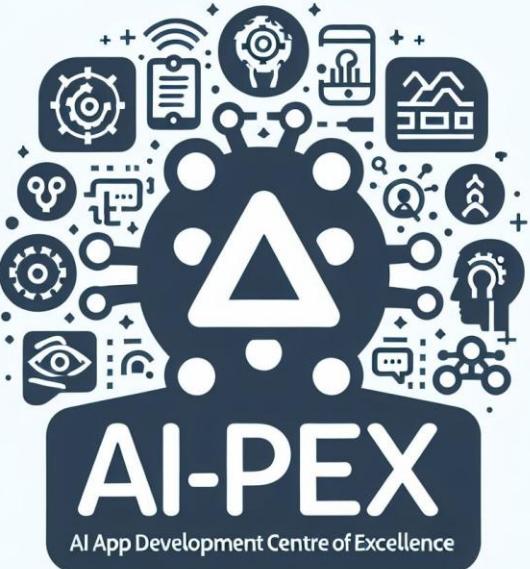
Multimodal
input and output

Real-time
response

Safe
by design

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

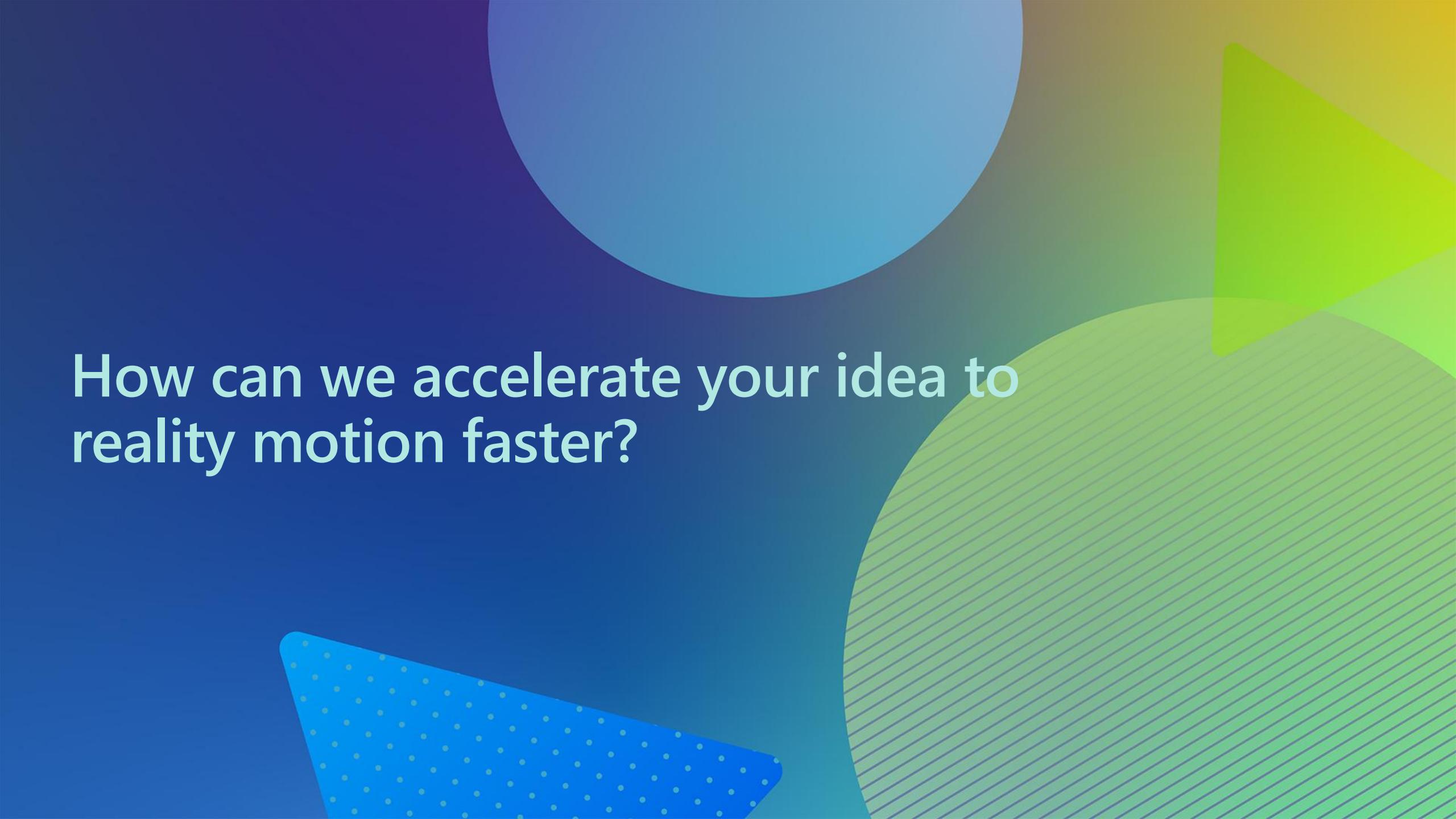
- **Deliver AI contents**

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

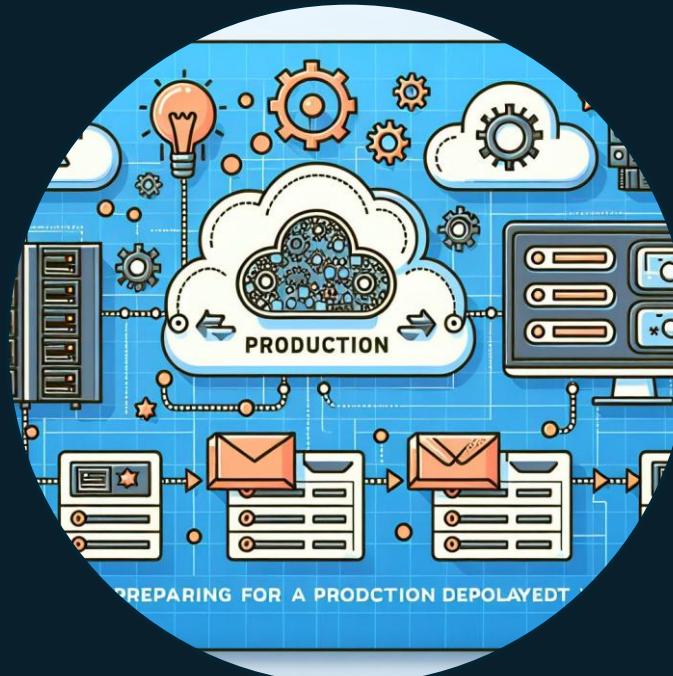
Scaling the MVP Aligning with Customer Needs



- 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's AI Initiatives and Priorities

Our Priority



Copilot for every role, on every device



AI Design Win for every partner & customer



Security foundation for every partner & customer



Deployment of M365 Core Services

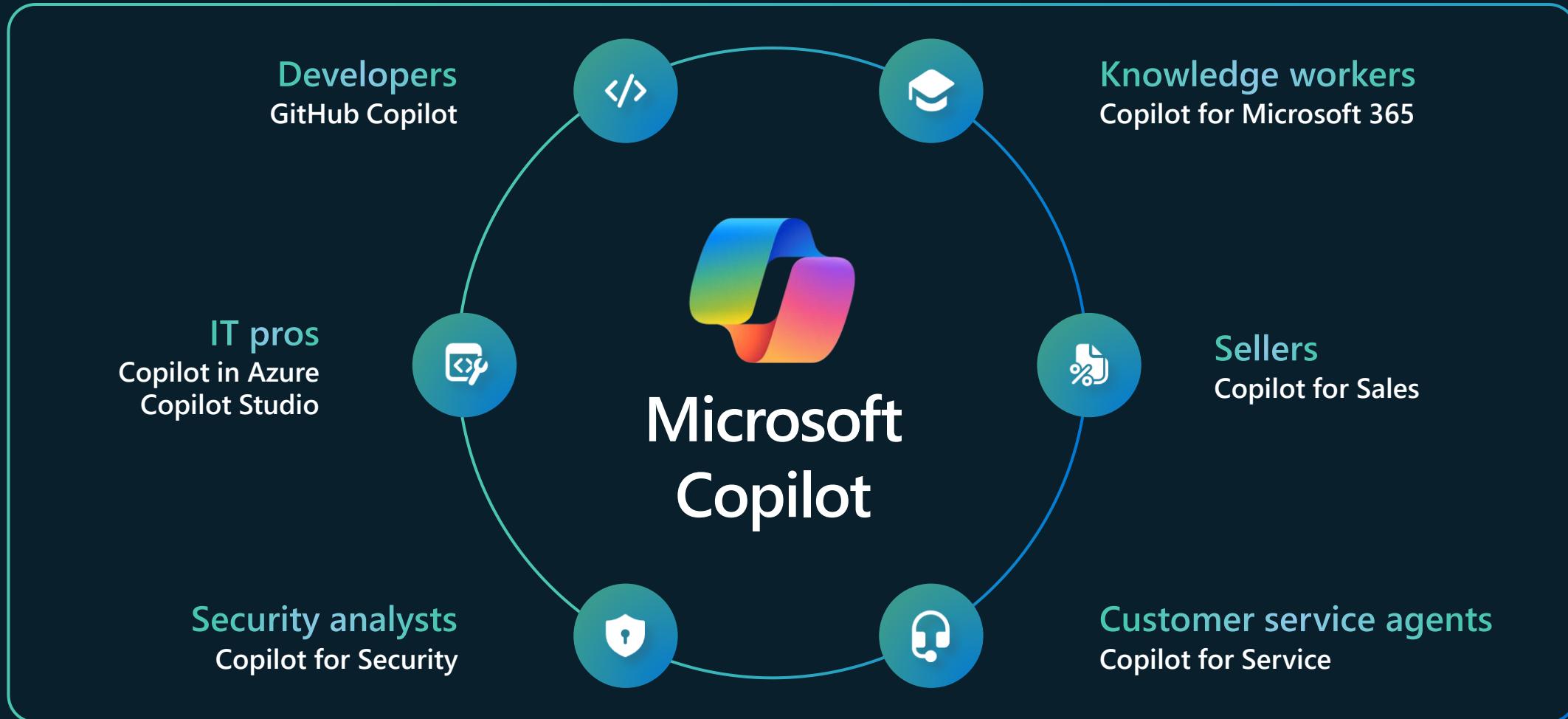


Migration, Migration, Migration



Copilot

Transform work and enhance productivity with AI





AI Design Win

New concepts for expanding AI business

Developer Productivity



Visual Studio

GitHub Copilot



Copilot Studio



Azure AI Foundry

AI design win solutions

AI Platform



Azure OpenAI



Model as a Service (MaaS)



Azure Machine Learning



Azure AI Search



Azure AI Services



Azure AI Foundry

App Platform



Azure Kubernetes Service



Azure Container Apps



Azure App Service



Azure API Management



Azure Functions

Data Platform



Azure SQL Database



Azure Cosmos DB



OSS Databases



Microsoft Fabric



Databricks



Security, Governance, AI Safety



Security Foundation

Provision of End-to-End Security Foundation

Security is 1st Priority



Secure by design

Protect tenants and isolate
Production systems



Secure by default

Protect identities
and secrets



Secure operations

Protect
networks

Protect engineering
systems

Monitor and
detect threats

Accelerate response
And remediation

Responsible AI

Microsoft reviews and updates its AI services based on the following aspects. All AI service documentation includes information on Responsible AI.

Fairness

AI Systems should treat all people fairly.

Reliability and safety

AI systems should perform reliability and safely.

Privacy and security

AI systems should be secure and respect privacy.

Inclusiveness

AI systems should empower everyone and engage people.

Transparency

AI system should be understandable.

Accountability

People should be accountable for AI systems.

Reliability of AI

Microsoft Cloud

You own your data

**Prompts are NOT used to train,
retrain, or improve Azure OpenAI
Service foundation models.**

**Data is protected by advanced
enterprise compliance and
security controls.**

Data is encrypted with customer-managed keys, VNET, RBAC, SOC2, ISO, HIPAA, and CSA STAR compliance.

Week 1 :

Microsoft AI

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

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.



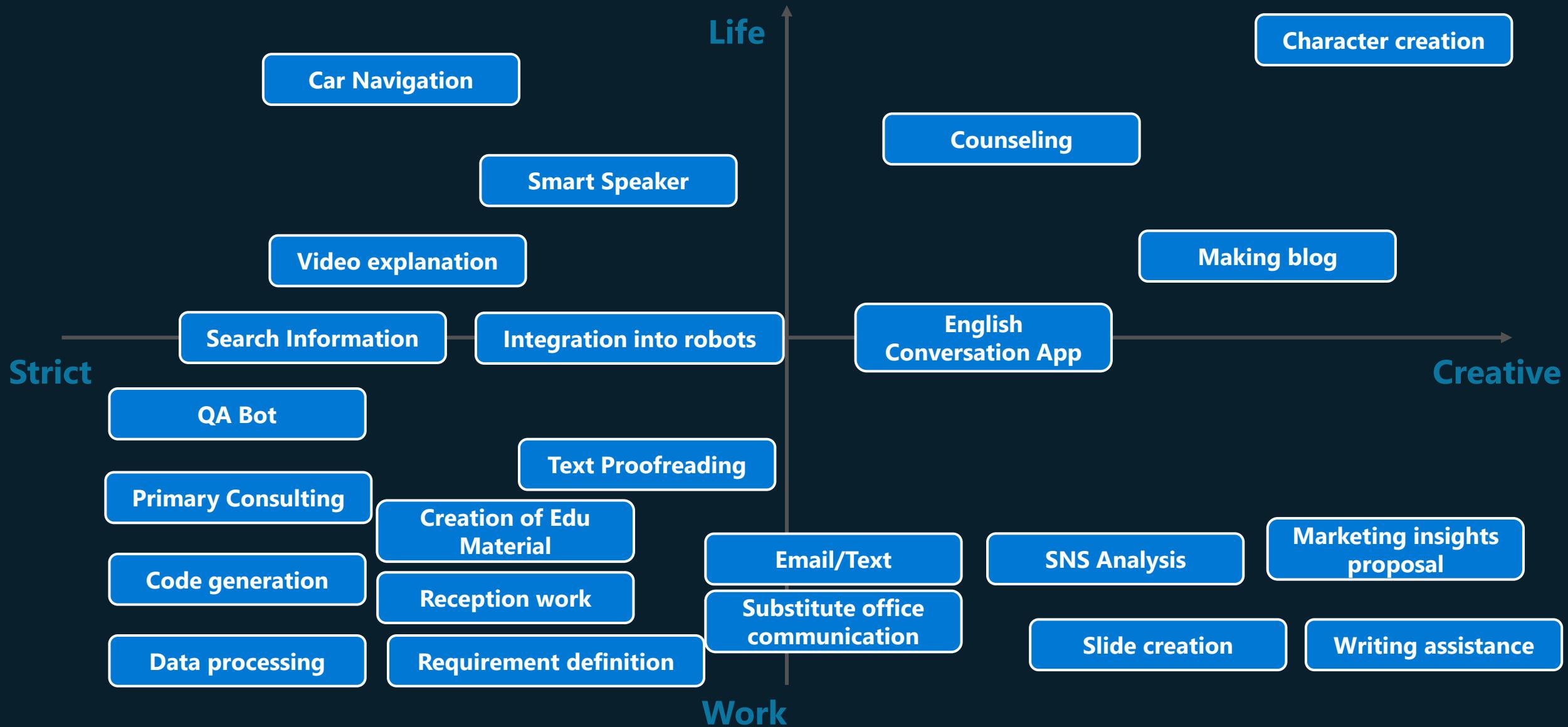
Instant and extensive reading comprehension

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

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



Trend of Generative AI

Before 2024

After 2024

Transformation

Utilization of large language models centered on text input and output



Utilizing **Multimodal** models that support audio, images, and videos

Model size

Large language models with a vast number of parameters (LLM: GPT-4, etc.)



Small Language Models that can operate with fewer parameters and lower computational resources (**SLM**: Phi-3, Phi-Silica)

Trend of Generative AI

After 2024

What is required for AI Engineer?



Utilizing Multimodal models
that support audio, images,
and videos



New application UI to
leverage multimodal
models

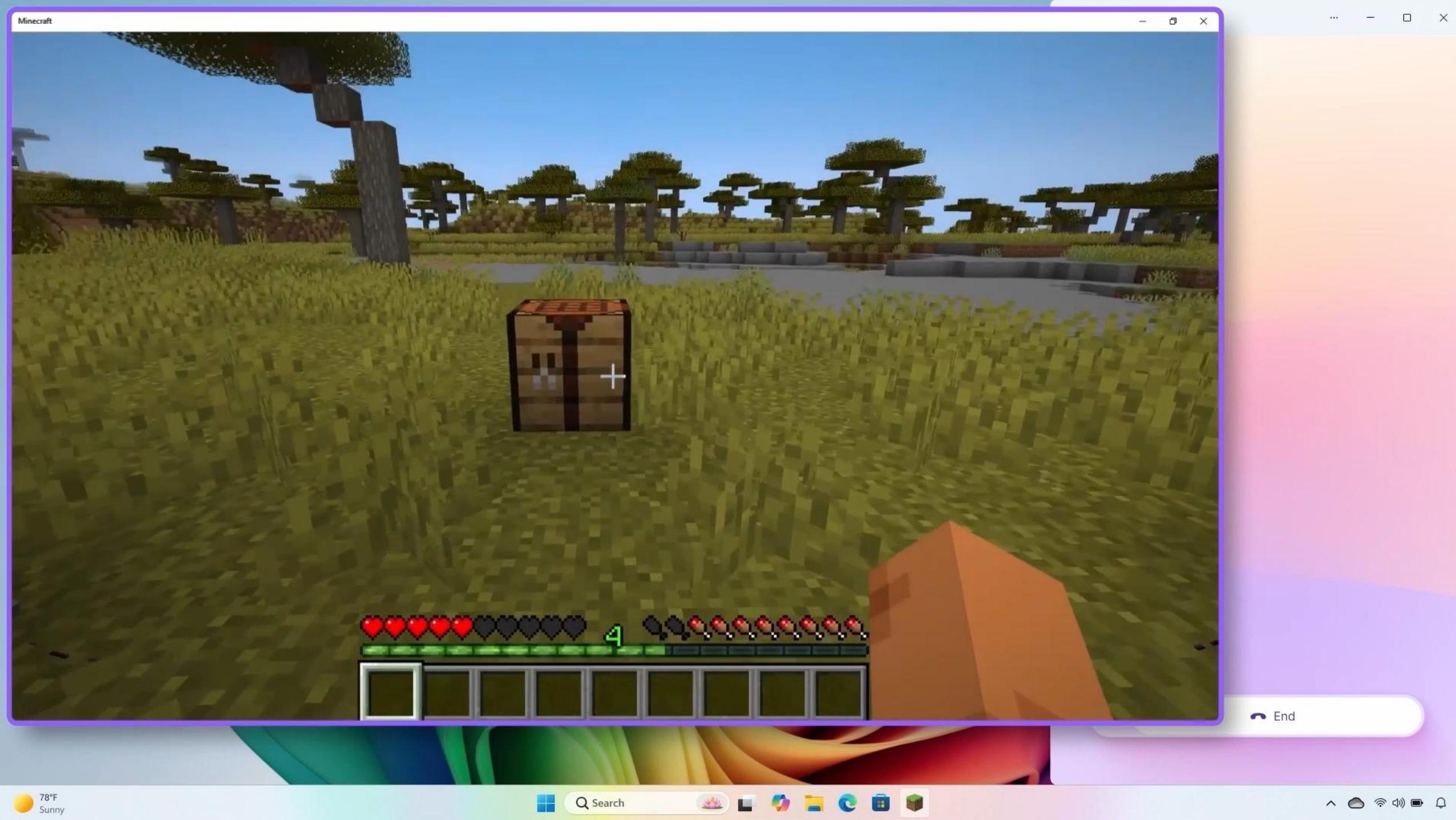


Small Language Models
that can operate with fewer
parameters and lower
computational resources
(SLM: Phi-3, Phi-Silica)



Appropriate use of
models and utilization
of specialized services

Demo Future Vision: Utilizing Multi-modal and Agent



End



Microsoft
Azure



[Get started with GPT-4o with audio in Azure OpenAI Service](#)

Partnership with OpenAI and Microsoft



OpenAI

Ensure that artificial general intelligence (AGI) benefits humanity



Microsoft

Empower every person and organization on the planet to achieve more

Azure OpenAI Service

LLM

Multi-Modal

Fine-tuning

Image

Transcription/Translation

GPT-4,
4-Turbo
3.5-Turbo

GPT-4o
GPT-4-Turbo
with Vision

GPT-4
GPT-3.5-Turbo

DALL·E 3

Whisper

On Your Data GA

Azure AI Studio GA

Assistants API Preview

Provisioned Throughput

Azure OpenAI Service models

Model	Description
<u>o1 & o1-mini</u>	Limited access models, specifically designed to tackle reasoning and problem-solving tasks with increased focus and capability.
<u>GPT-4o & GPT-4o mini & GPT-4 Turbo</u>	The latest most capable Azure OpenAI models with multimodal versions, which can accept both text and images as input.
<u>GPT-4o-Realtime-Preview</u>	A GPT-4o model that supports low-latency, "speech in, speech out" conversational interactions.
<u>GPT-4</u>	A set of models that improve on GPT-3.5 and can understand and generate natural language and code.
<u>GPT-3.5</u>	A set of models that improve on GPT-3 and can understand and generate natural language and code.
<u>Embeddings</u>	A set of models that can convert text into numerical vector form to facilitate text similarity.
<u>DALL-E</u>	A series of models that can generate original images from natural language.
<u>Whisper</u>	A series of models in preview that can transcribe and translate speech to text.
<u>Text to speech (Preview)</u>	A series of models in preview that can synthesize text to speech.

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

Model Catalog: +1,700 models are available



Azure OpenAI Service



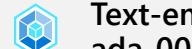
GPT-4



GPT-4-32k



GPT-4V



Text-embedding-
ada-002



GPT-3.5-Turbo



- Llama-2-70b / 70b-chat
- Llama-2-13b / 13b-chat
- Llama-2-7b / 7b-chat
- CodeLlama 7b/13b/34b/
Python/Instruct



Hugging Face



Falcon/TII



Stable Diffusion/
Stability AI



Dolly/Databricks



CLIP/OpenAI



- Nemotron-3-8B-4k
- Nemotron-3-8B-Chat-
SFT/ RLHF/ SteerLM
- Nemotron-3-8B-QA



Microsoft Research



microsoft-phi-1-5, 2



microsoft-Orca-2-7b, 13b



microsoft-phi-3-mini

NEW!!



Deci AI

- DeciLM 7B, 7B instruct
- DeciCoder-1b
- Decidiffusion-v1-0



Mistral AI

- Mixtral-8x7B, Instruct
- Mistral-7B, Instruct
- Mistral-large, small



- Cohere-command-r, r-plus
- Cohere-embed-v3-english
- Cohere-embed-v3-multilingual

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.



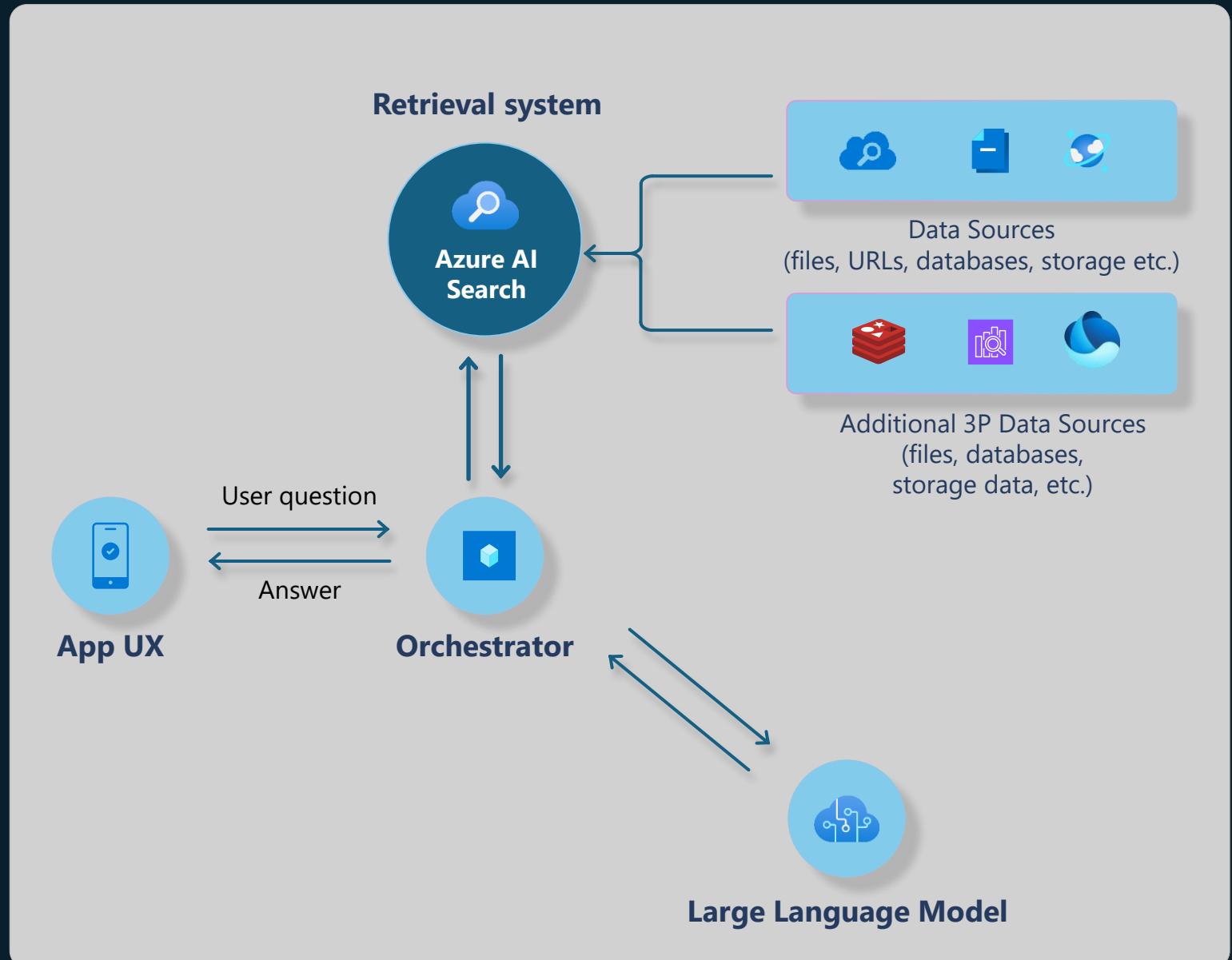
Computer Vision

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

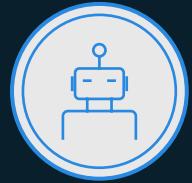
Flexibility with integrated platform and ecosystem



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...	<h2>Documents Knowledge</h2>	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...	<h2>Web Search</h2>	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..	<h2>Database</h2>	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.
	<h2>Other</h2>	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

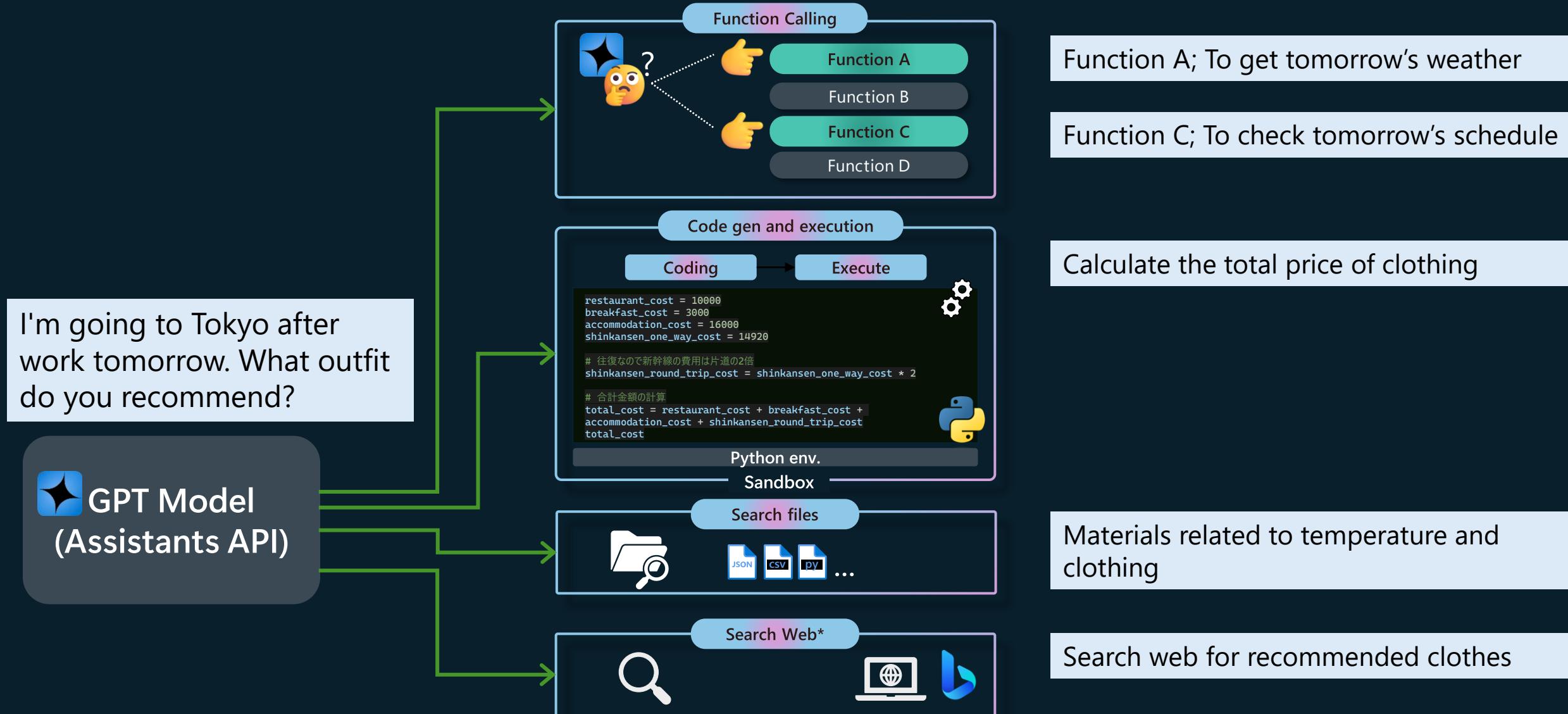
AI chat services are generally limited to simple one-question-one-answer interactions. **AI agents, on the other hand, can handle tasks within complex and continuous dialogues. In some cases, multiple agents can be coordinated to solve problems as needed.**

Feat.
4

External
integration

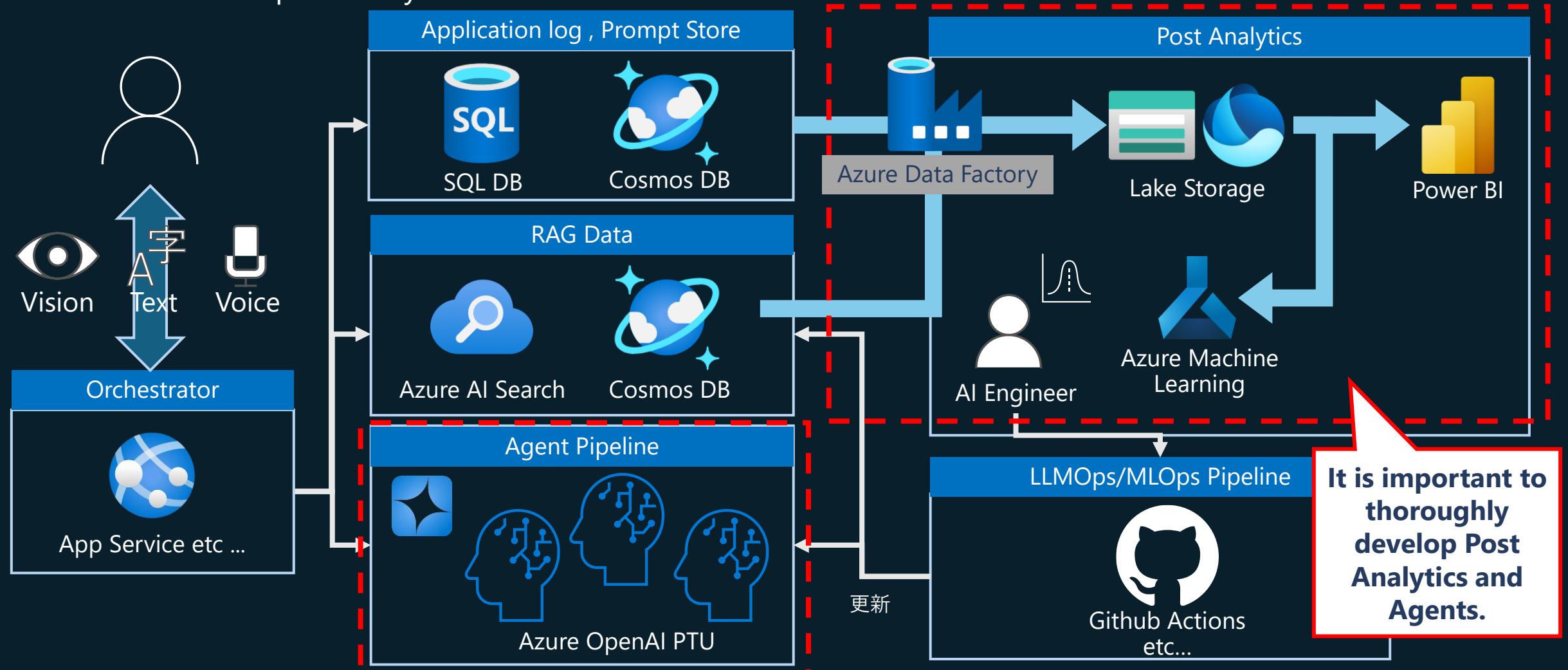
AI chat services have minimal integration with external tools and systems. **However, AI agents actively engage in external integrations because they need to autonomously execute tasks.**

From RAG to AI Agent system



AI Agent Architecture

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



Appendix : Success Stories from ISV Partners

Enhancing customer experience and streamlining support operations

The AI assistant improves customer experience and support efficiency by providing natural language access to extensive knowledge repositories in fashion supply chain.



Coats Digital launches AI assistant in just six weeks

Use Case Example

AI Assistant for Enhanced Customer Support

- Scenario: A new customer, Jane, has recently started using Coats Digital's production planning tools. She needs help understanding how to optimize the tool for her specific needs.
- Challenge: Jane finds it difficult to navigate through the extensive documentation and training materials available, leading to delays in her workflow.
- Solution: Jane uses the AI assistant integrated into Coats Digital's knowledge repository. She types her query in natural language: "How can I optimize production planning for seasonal demand?"
- Outcome: The AI assistant quickly understands Jane's query and provides her with the most relevant training materials and documentation. Jane receives step-by-step guidance tailored to her needs, significantly reducing the time she spends searching for information. This enhances her overall experience and allows her to utilize the tool more effectively.

Benefits

Improved Customer Experience

Enhanced Support Efficiency

Scalability and Security

Official information is here ;
[Coats Digital launches AI assistant in just six weeks using Azure OpenAI Service | Microsoft Customer Stories](#)

Assisting in Job Posting Creation with AI

Providing AI-Assisted Job Posting Creation Functionality on Job Medley, Japan's Largest Medical and Nursing Job Site.

MEDLEY, INC.

Use Case Example

- Medley Inc. has started offering a job posting creation assistance feature using the Azure OpenAI Service's GPT model on "Job Medley," Japan's largest medical and nursing job site. With this feature, businesses can generate attractive job postings by simply answering a few questions, significantly reducing the effort required to create job postings. Even businesses without dedicated recruitment staff can efficiently carry out recruitment activities, enabling quick personnel acquisition.
- Job Medley is utilized by approximately 340,000 businesses nationwide, and with the help of generative AI, it has become easier to create job postings that comply with laws and guidelines. By using the Azure OpenAI Service, enterprise-level security, availability, and stability are ensured, allowing users to use it with confidence.
- Additionally, Medley offers features utilizing generative AI in their online video training service for nursing and disability welfare businesses, "Job Medley Academy." By simply entering the training name and goal, recommended lecture videos are suggested. Medley aims to continue promoting the active use of AI technology to contribute to solving healthcare challenges.

ジョブメドレー

求人の作成を、
AIで補助。

医療介護求人サイト「ジョブメドレー」



Benefits

Cost
Reduction

Operational
Efficiency

New Value
Creation

Official information is here ;

<https://prtimes.jp/main/html/rd/p/000000099.000013108.html>

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 CONFIDENTIAL

This document is intended for informational purposes only, and the information contained herein reflects Microsoft's views as of the date of this document. The content is subject to change due to various circumstances. The terms and conditions, including prices, mentioned in this document will only be finalized through a valid contract with your company. Until then, nothing is confirmed or applicable. Additionally, Microsoft makes no express, implied, or statutory warranties regarding the information in this document. © 2024 Microsoft Corporation. All rights reserved.