ARTIFICIAL INTELLIGENCE

ANDRÉ VALA

ANDRÉ VALA

Sr Cloud Solution Architect @ Microsoft

Based in Lisbon, Portugal

Former Microsoft MVP for Office Apps & Services

Co-organizer of SharePoint Saturday Lisbon, CollabDays Lisbon, TUGA IT Conference

Passionate about technology, but focused on Cloud Solution Architecture

/in/andrevala

https://andrevala.com

@atomicvee

andre.vala@microsoft.com
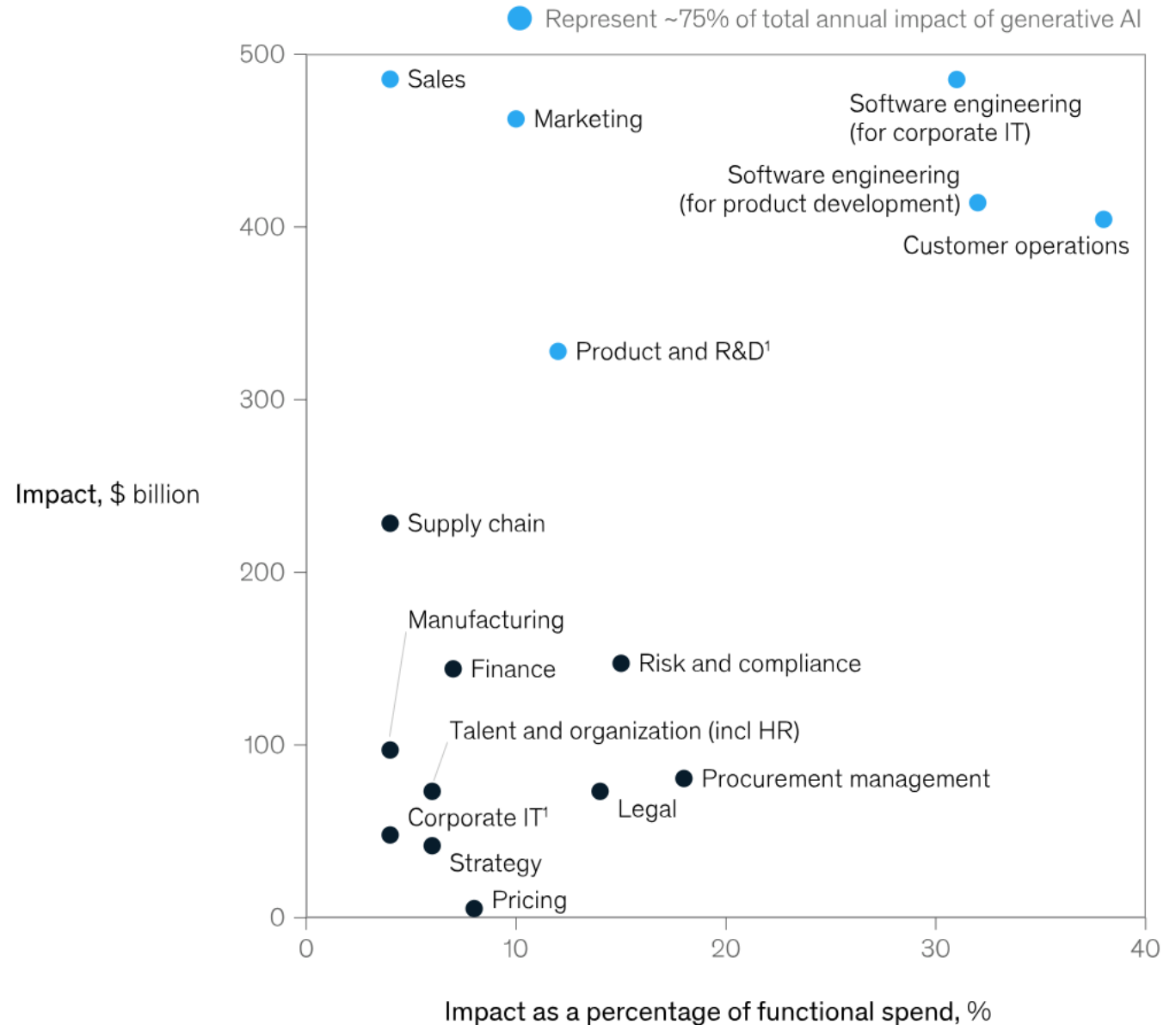
AI IMPACT IN THE WORLD

GENERATIVE AI IMPACT

Generative AI can add

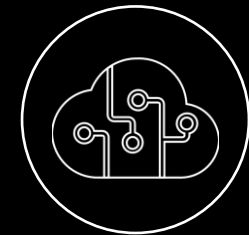## 2.6T to 4.4T USD
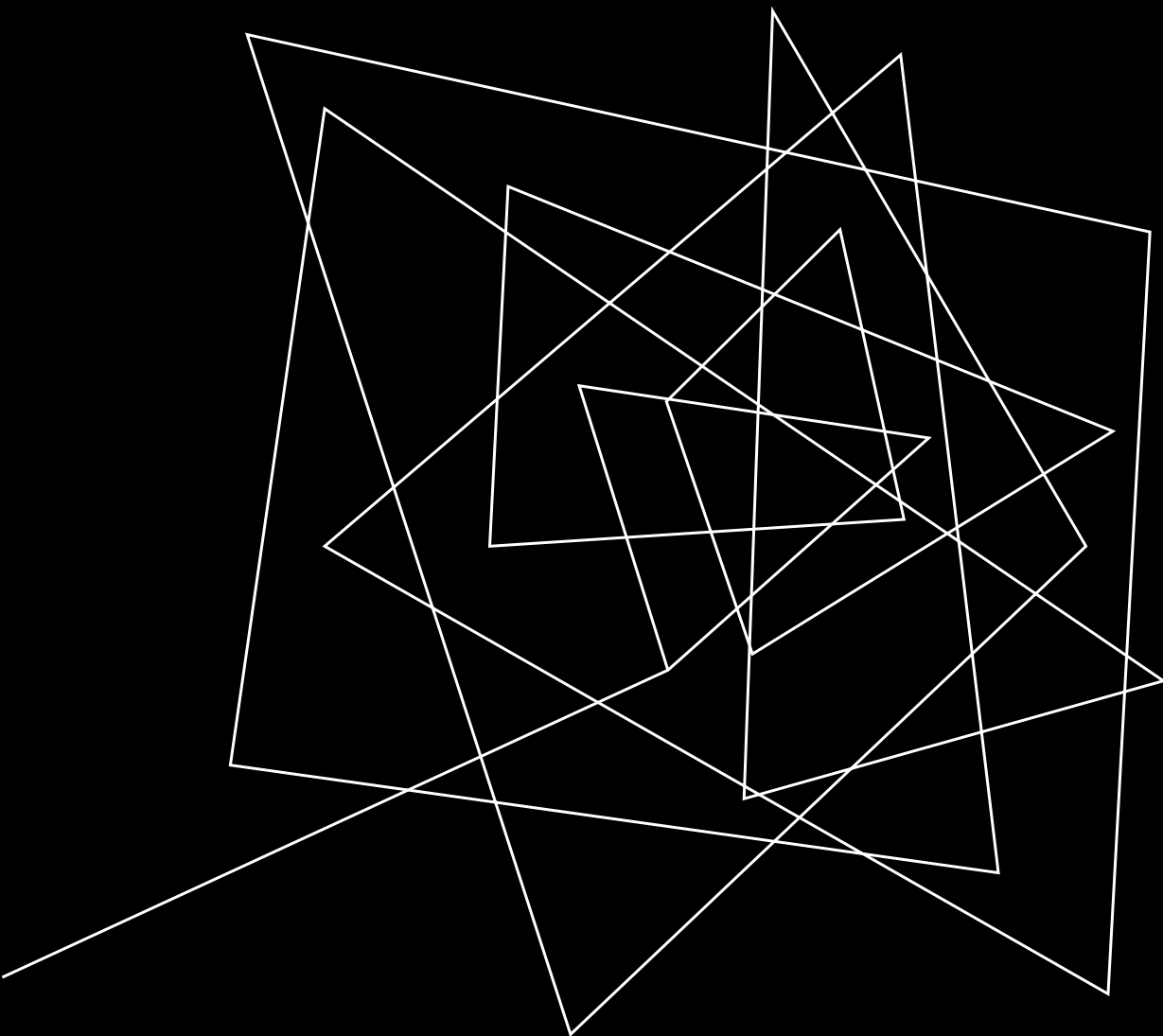
in value to the global economy
across 63 use cases

75%

of the value fall across 4 areas:
customer operations, marketing and
sales, software engineering and R&D

Source: McKinsey Report – June 14, 2023

Represent ~75% of total annual impact of generative AI

Impact, $ billion

- Sales
- Marketing
- Software engineering (for corporate IT)
- Software engineering (for product development)
- Customer operations
- Product and R&D[1]
- Supply chain
- Manufacturing
- Finance
- Risk and compliance
- Talent and organization (incl HR)
- Procurement management
- Corporate IT[1]
- Legal
- Strategy
- Pricing

Impact as a percentage of functional spend, %

500
400
300
200
100
0

0   10   20   30   40

GENERATIVE AI

Artificial Intelligence

Machine Learning

Deep Learning

GENERATIVE AI

1956 Artificial Intelligence
the field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence

1970s Machine Learning
subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions

2010s Deep Learning
a machine learning technique in which layers of neural networks are used to process data and make decisions
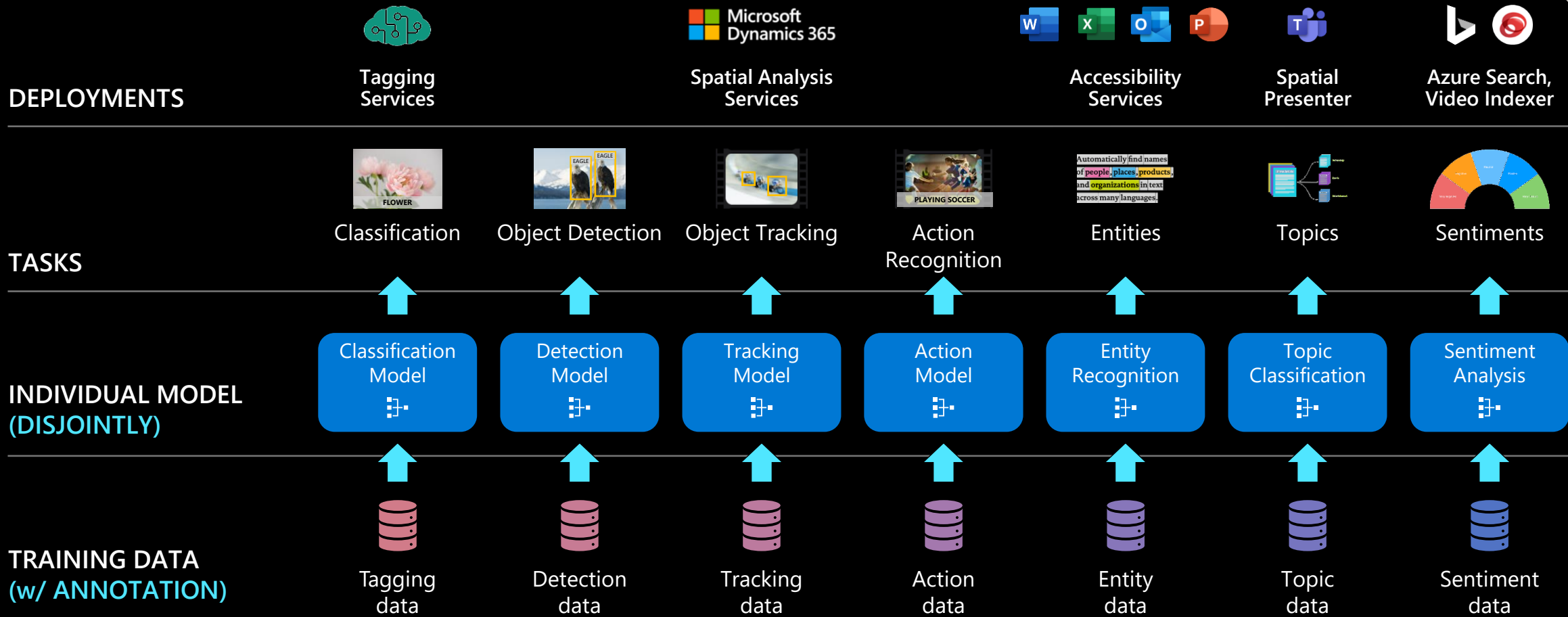
2020s Generative AI
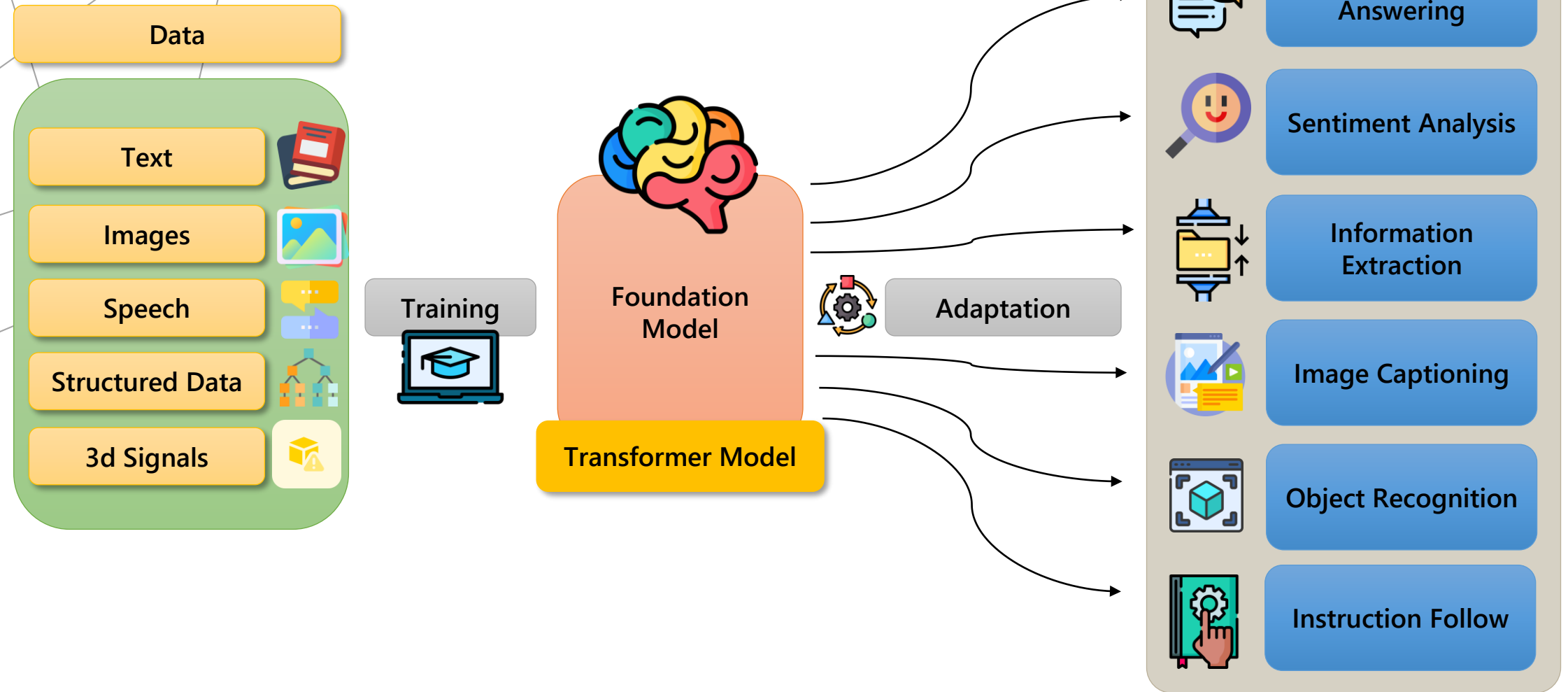Create new written, visual, and auditory content given prompts or existing data.

GENERATIVE AI

TRADITIONAL MODEL DEVELOPMENT

High cost and slow deployment—each service is trained disjointly

DEPLOYMENTS

Tagging Services | Spatial Analysis Services | Accessibility Services | Spatial Presenter | Azure Search, Video Indexer

TASKS

Classification | Object Detection | Object Tracking | Action Recognition | Entities | Topics | Sentiments

INDIVIDUAL MODEL (DISJOINTLY)

Classification Model | Detection Model | Tracking Model | Action Model | Entity Recognition | Topic Classification | Sentiment Analysis

TRAINING DATA (w/ ANNOTATION)

Tagging data | Detection data | Tracking data | Action data | Entity data | Topic data | Sentiment data

OPEN AI / MICROSOFT PARTNERSHIP

Ensure that artificial general intelligence (AGI) benefits humanity

+

Empower every person and organization on the planet to achieve more

GPT-3.5 and GPT-4	ChatGPT	Codex	DALL·E 2
Text	Conversation	Code	Images

AZURE OPEN AI SERVICE

Azure OpenAI Service

GPT-3	GPT-4
DALL·E	ChatGPT

Deployed in your Azure subscription, secured by you, and tied to your datasets and applications
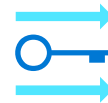
Large, pretrained AI models to unlock new scenarios

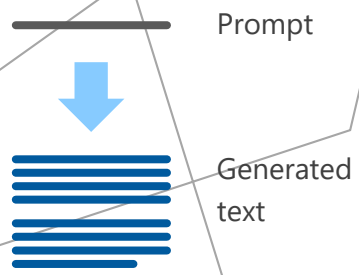AI models, some custom-tunable with your data and hyperparameters

Built-in responsible AI to detect and mitigate harmful use

Enterprise-grade security with role-based access control (RBAC) and private networks

CAPABILITIES

Generation

Prompt

Generated text

Summarization

Large text (prompt)

Summarized text

Rewrite

Source text (prompt)

Rewritten text

Extraction

Source text (prompt)

Extracted text

Search

Query

Embedding

Similarity Search Results

Clustering

Information to be clustered

Clusters

Classification

Samples

Text to be classified
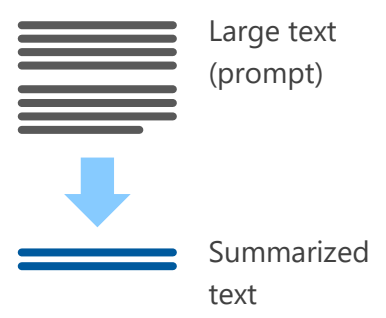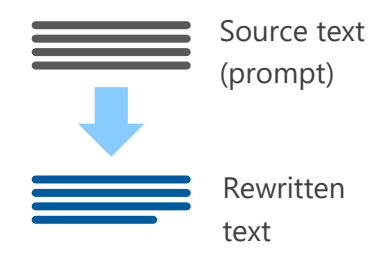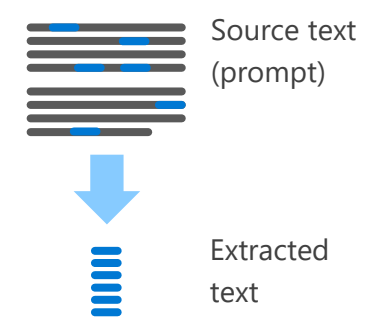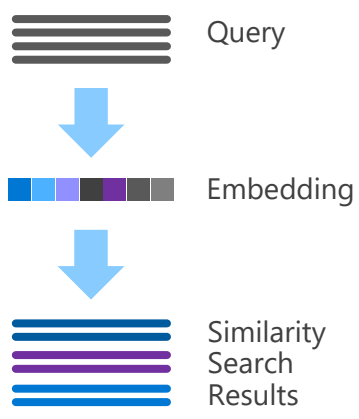
Classification
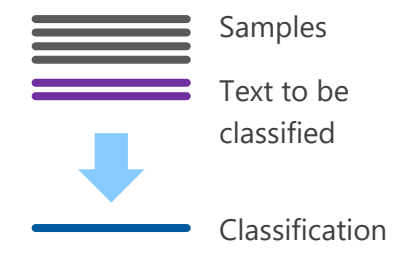
PROMPT ENGINEERING

- Creating the optimal prompt to elicit a particular type of response or steer the model's output toward a desired direction.

- Essential skill when using LLMs. It is how you "program" an LLM.

Sample Prompt

```
The following is a conversation with an AI research
assistant. The assistant tone is technical and
scientific.

Human: Hello, who are you?
AI: Greeting! I am an AI research assistant. How can
I help you today?

Human: Can you tell me about the creation of
blackholes?
AI:
```

---> Instruction or Task

---> Context or Examples

---> Input data
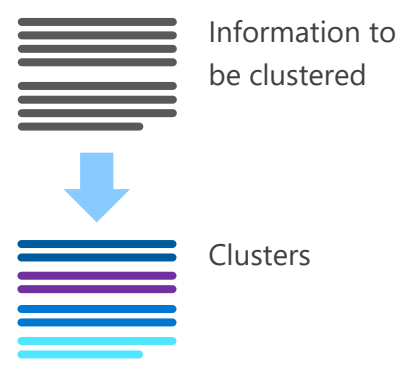
---> Output indicator

Zero-shot learning

```
Translate English to French:
Cheese =>
```

One-shot learning

```
Translate English to French:
Sea otter => loutre de mer
Cheese =>
```

Few-shot learning

```
Translate English to French:
Sea otter => loutre de mer
Peppermint => menthe poivre
Plush giraffe => giraffe peluche
Cheese =>
```

LIMITATIONS
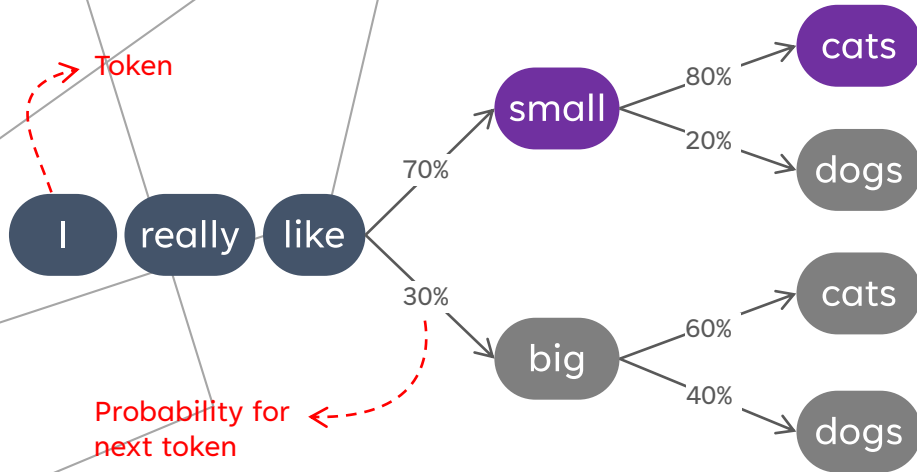
- May generate outputs that are not supported by the input or are factually incorrect. Also known as **hallucination** (or making up information).

- May generate outputs that are inappropriate, **offensive**, **biased** or harmful, and it may be **difficult to monitor** or correct their behavior.

- May become **outdated** or irrelevant over time as new information emerges. They can be retrained, but training requires a lot of computational power and time.

- **Limited token count** means prompts hold limited context which, when in a longer conversation, causes the model to "forget" what was said before.
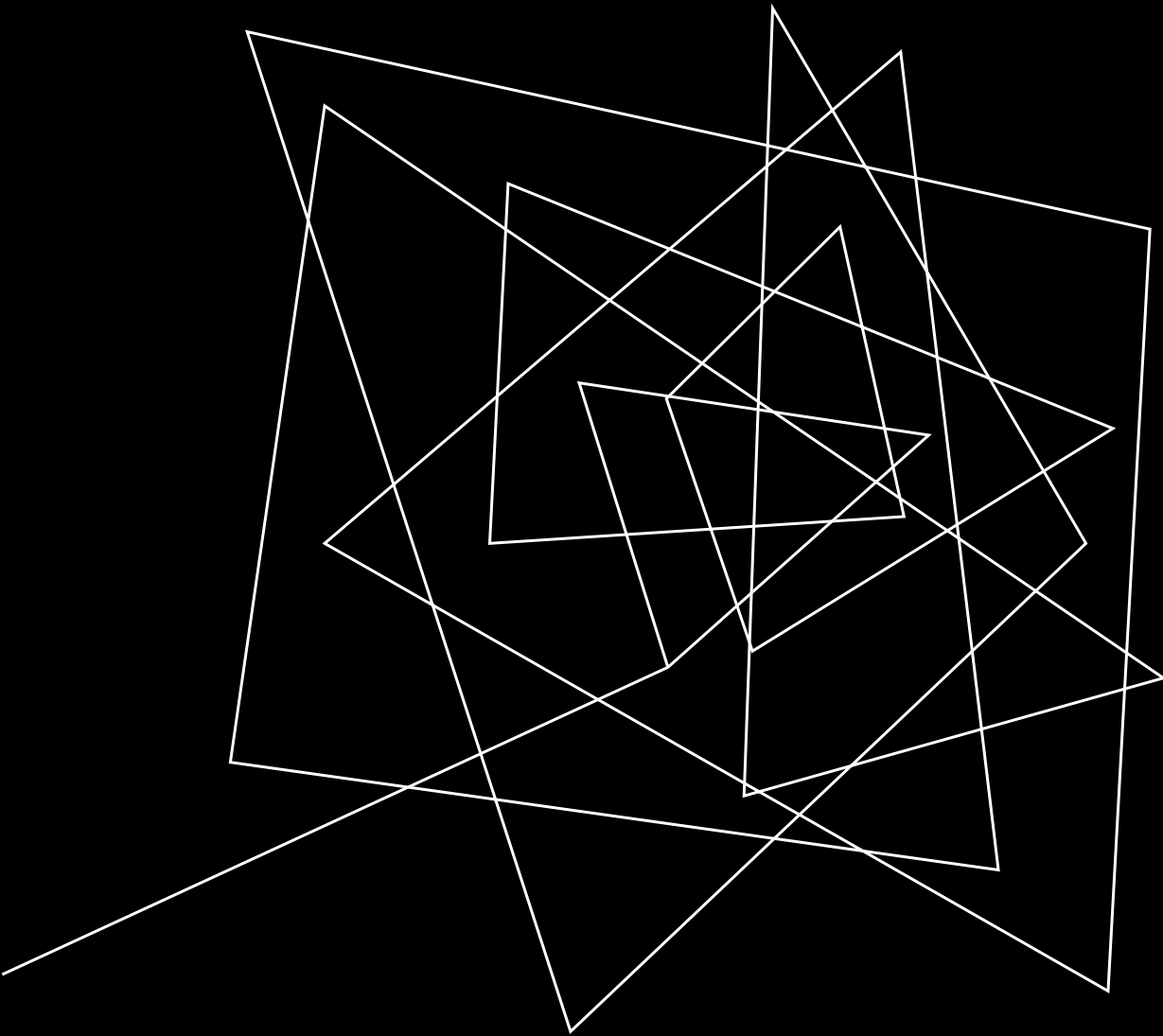
USE CASES

ENTERPRISE KNOWLEDGE Q&A

Natural language interface to surface internal enterprise data

- Leverage internal knowledge from multiple data sources

- Combine results in a human-like natural language answer

- Use citations to explain where information came from and allow user to access source

- Support follow up questions

CONTACT CENTER ANALYTICS

Post-call analysis of transcript

- Summarization (including outcome)

- Sentiment Analysis

- Entity extraction (customer identification, reason for call, campaign/product mentions)

- Automatic call auditing

- Search through call contents
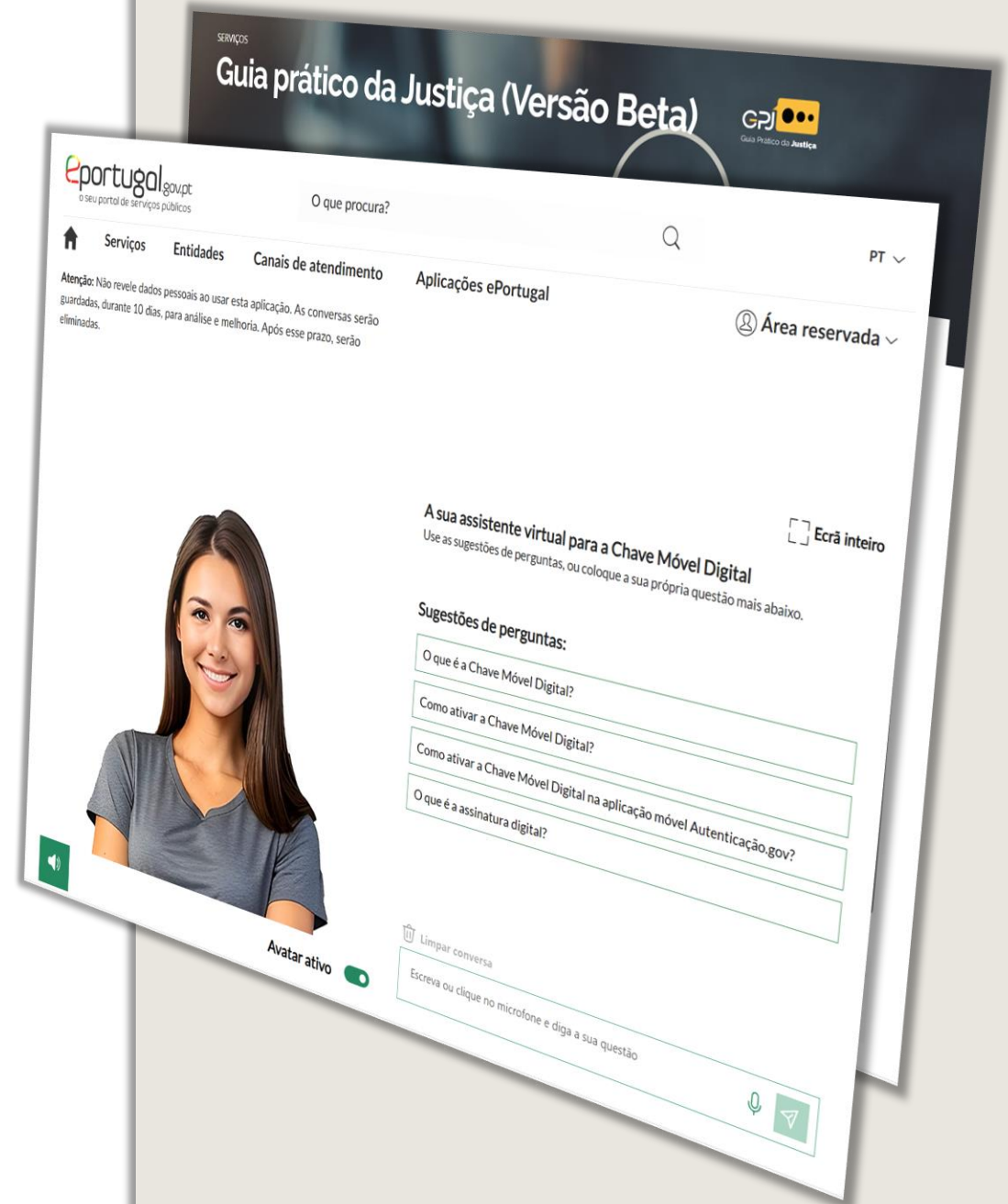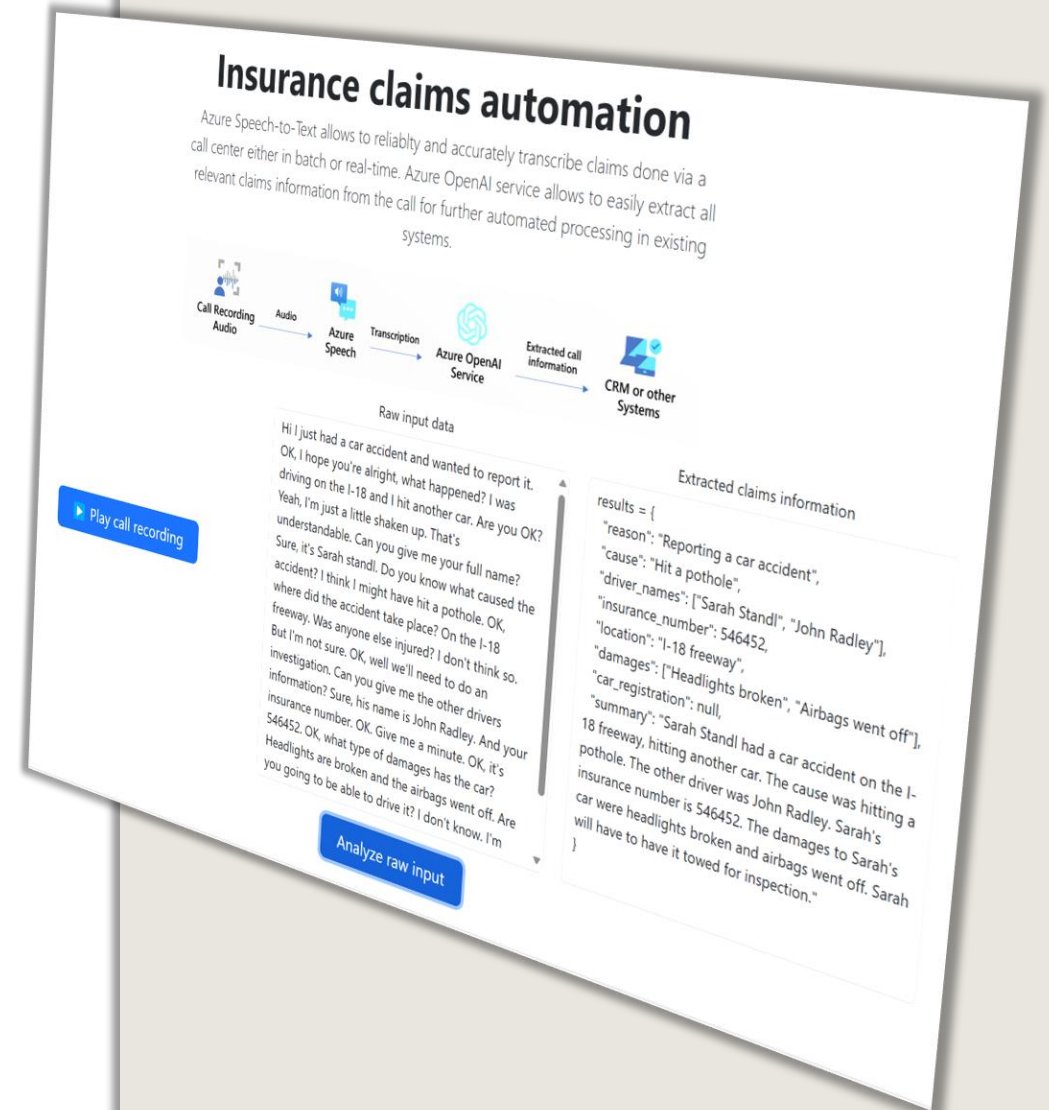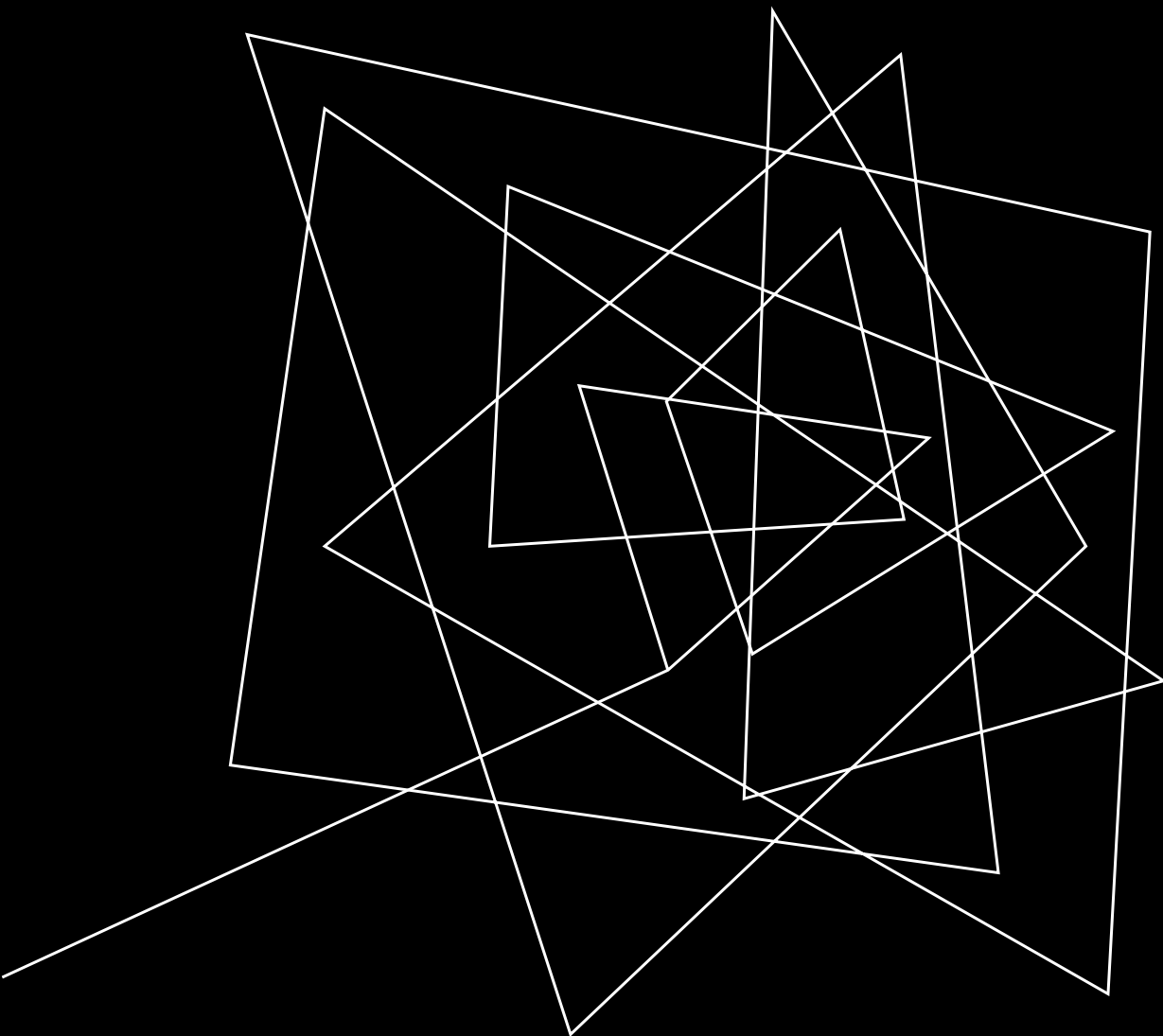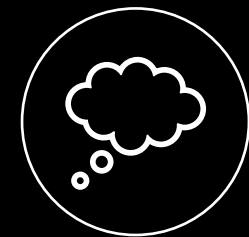
- Reporting/dashboard

Support for individual call analysis or volume call processing.

Can be expanded for real-time in-call analysis scenarios.

Insurance claims automation

Azure Speech-to-Text allows to reliablty and accurately transcribe claims done via a call center either in batch or real-time. Azure OpenAI service allows to easily extract all relevant claims information from the call for further automated processing in existing systems.

Call Recording Audio → Audio → Azure Speech → Transcription → Azure OpenAI Service → Extracted call information → CRM or other Systems

Raw input data

Hi I just had a car accident and wanted to report it. OK, I hope you're alright, what happened? I was driving on the I-18 and I hit another car. Are you OK? Yeah, I'm just a little shaken up. That's understandable. Can you give me your full name? Sure, it's Sarah standl. Do you know what caused the accident? I think I might have hit a pothole. OK, where did the accident take place? On the I-18 freeway. Was anyone else injured? I don't think so. But I'm not sure. OK, well we'll need to do an investigation. Can you give me the other drivers information? Sure, his name is John Radley. And your insurance number. OK. Give me a minute. OK, it's 546452. OK, what type of damages has the car? Headlights are broken and the airbags went off. Are you going to be able to drive it? I don't know. I'm

▶ Play call recording

Analyze raw input

Extracted claims information

```
results = {
    "reason": "Reporting a car accident",
    "cause": "Hit a pothole",
    "driver_names": ["Sarah Standl", "John Radley"],
    "insurance_number": 546452,
    "location": "I-18 freeway",
    "damages": ["Headlights broken", "Airbags went off"],
    "car_registration": null,
    "summary": "Sarah Standl had a car accident on the I-18 freeway, hitting another car. The cause was hitting a pothole. The other driver was John Radley. Sarah's insurance number is 546452. The damages to Sarah's car were headlights broken and airbags went off. Sarah will have to have it towed for inspection."
}
```

CONCLUSIONS

DATA PRIVACY & SECURITY

Your data is **your** data
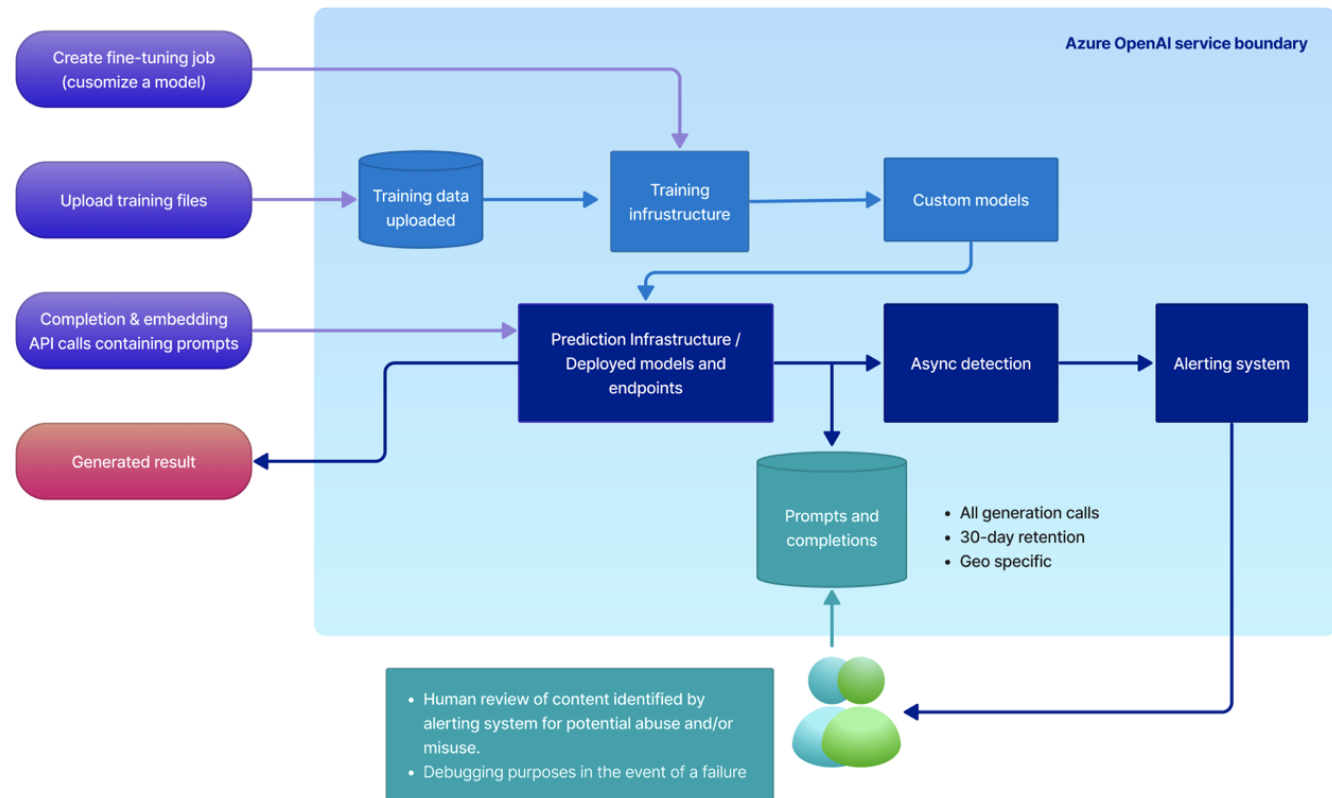
Data is stored encrypted in your Azure subscription

Your data from any fine-tuning is **not used** to train the foundation AI models

- Azure OpenAI Service provisioned in your Azure subscription
- Model fine tuning stays in your Azure subscription and never moves into the foundation AI models

Your data is **protected** by the most comprehensive enterprise compliance and security controls

- Encrypted with Customer Managed Keys
- Private Virtual Networks, Role Based Access Control
- Soc2, ISO, HIPPA, CSA STAR Compliant

Customer action

Azure OpenAI service boundary

Create fine-tuning job (cusomize a model)

Upload training files → Training data uploaded → Training infrastructure → Custom models

Completion & embedding API calls containing prompts → Prediction Infrastructure / Deployed models and endpoints → Async detection → Alerting system

Generated result

Prompts and completions
- All generation calls
- 30-day retention
- Geo specific

- Human review of content identified by alerting system for potential abuse and/or misuse.
- Debugging purposes in the event of a failure

KEY TAKEAWAYS

- Infusing AI into applications has **huge potential**.

- Azure Open AI Service is an **enterprise-ready** offering on top of Open AI models, bringing Azure's security, compliance and Responsible AI principles to solutions.

- Azure Open AI Service is usually just a **piece of the puzzle**. A complete solution will require additional components.

- Large Language Models, such as GPT, can execute **different types of tasks with few or no training**.

- **Prompt engineering** is critical, but it is more of an art than a science. It can yield better results than Fine Tuning at a much lower cost.

THANK YOU

 /in/andrevala

 https://andrevala.com

 @atomicvee

 andre.vala@microsoft.com