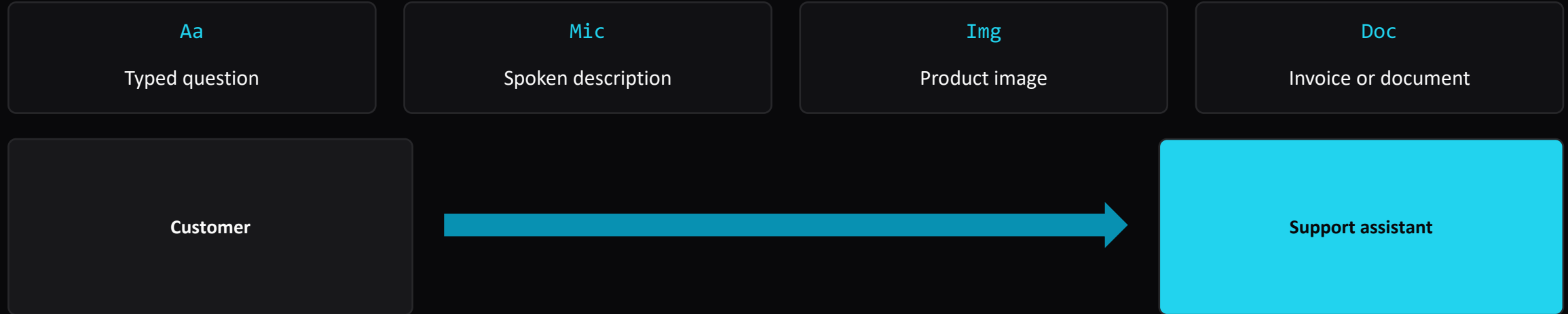


The Contoso customer-support assistant

A customer submits:

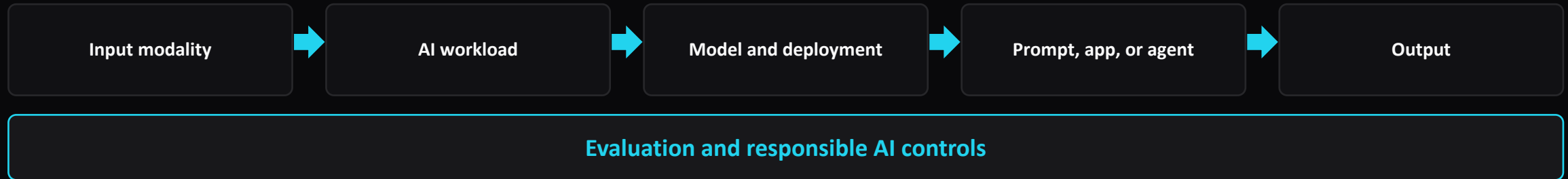


Example request

“Why was I charged twice, and can you correct it?”

The customer attaches an invoice and a screenshot.

Seven parts of an AI solution



- | | | | |
|----------|--|----------|--|
| 1 | Input modality
What the user provides | 5 | Prompt or application
Instructions and orchestration logic |
| 2 | AI workload
What the system must accomplish | 6 | Output
What the solution returns or does |
| 3 | Model
The intelligence selected for the task | 7 | Evaluation and controls
How the team measures quality and manages risk |
| 4 | Deployment
An accessible instance of the model | | |

The input does not determine the workload by itself

Customer input	Required workload
Typed complaint	Text analysis or generative AI
Spoken question	Speech recognition, followed by language processing
Product photograph	Computer vision or multimodal generation
Invoice document	Information extraction
Request for account action	Agentic AI

Key distinction

Modality describes the form of the data. Workload describes what the solution must do with it.

A model becomes usable through a deployment

Model

- Defines the available capabilities
- May support text, vision, audio, or multiple modalities
- Varies in quality, speed, context capacity, and cost

Deployment

- Makes a selected model available to an application
- Provides an endpoint or named target
- Applies configuration and access controls
- Can use parameters that affect model behavior

Simple analogy

The model is the design. The deployment is the running instance the application can call.

Model catalog



Deployed endpoint



Application

The application turns model capability into a solution

System prompt

Defines the assistant's role, boundaries, and expected behavior.

"You are a customer-support assistant. Use verified account information and explain any uncertainty."

User prompt

- Contains the customer's current request and relevant context.

Application logic

- Collects and prepares the inputs
- Calls models or Foundry Tools
- Supplies retrieved or extracted information
- Invokes an agent when the task requires action
- Presents the result to the customer

The agent manages a multi-step support request

Customer goal

“Explain the duplicate charge and correct it if appropriate.”

Possible agent behavior

- Read the customer’s request.
- Use extracted invoice data.
- Retrieve the relevant account transactions.
- Compare the invoice with the account record.
- Request an approved refund action when warranted.
- Explain the result to the customer.

Agent components

- Instructions
- Model
- Knowledge or context
- Tools
- Conversation state

Agent

Invoice data

Account records

Refund tool

A successful response must be useful and controlled

Output

“The invoice contains one order, but the account shows two completed charges. I prepared a refund request for the duplicate charge. Please confirm before submission.”

Evaluate the solution

- Did it identify the duplicate correctly?
- Did it use the supplied evidence?
- Did it choose an authorized action?
- Did it explain the result clearly?
- How often does it produce unsupported claims?

Responsible AI controls

Concern	Example control
Fairness	Test performance across supported users and languages
Reliability and safety	Validate extracted values before taking action
Privacy and security	Limit access to account and document data
Transparency	Tell the customer when AI generated the response
Accountability	Log decisions and require approval for refunds
Inclusiveness	Support accessible text and speech interactions

Every AI-901 scenario can be unpacked by asking seven questions: What entered the system, what workload was required, which model handled it, where was it deployed, what directed it, what came out, and how was the result controlled?