MockCert

Study guide | Microsoft | AI-103

Microsoft Certified: Azure AI Apps and Agents Developer Associate

Notes for every domain of the Microsoft Certified: Azure AI Apps and Agents Developer Associate exam, organized by the official exam guide's domains and weights.

By MockCert about 65 min read Updated November 23, 2026

CONTENTS

DOMAIN 1 · 25–30%
Plan and manage an Azure AI solution 2

DOMAIN 2 · 30–35%
Implement generative AI and agentic solutions 14

DOMAIN 3 · 10–15%
Implement computer vision solutions 24

DOMAIN 4 · 10–15%
Implement text analysis solutions 32

DOMAIN 5 · 10–15%
Implement information extraction solutions 38

MockCert is an independent certification preparation platform. It is not affiliated with, endorsed by or operated by Microsoft. Exam domains and weights follow the official exam guide. These notes were last updated on November 23, 2026; the notes at mockcert.com are kept up to date as the exam changes.

Domain 1 | 25–30% of the exam | 19 min read

Plan and manage an Azure AI solution

Overview

Every point below comes from Microsoft's official documentation; each section lists its sources.

Choose the appropriate Foundry services for generative AI and agents

Start here

THINK FIRST

How do you decide whether an agent's knowledge belongs in memory, a knowledge base, file search or a live tool?

Why it matters: Questions describe an agent need and ask which Foundry service, retrieval method or tool integration fits it.

Each Foundry service grounds an agent on a different kind of knowledge. Memory keeps what the agent learns about a user, a knowledge base holds curated content, and file search covers user documents. Pick by where the knowledge comes from and how long it must last.

Key terms

Term	Meaning
Vector store	The store of chunked files that the file search tool searches.
Knowledge base	A Foundry IQ object that grounds an agent on curated organizational content.
Memory scope	The partition of a memory store that holds one user's memories, often set to `{{$userId}}`.
Toolbox	An MCP-compatible endpoint that holds tools such as Web Search or File Search; agents use its consumer endpoint.

Memory, knowledge base or file search

Match the grounding service to where the knowledge comes from. Ask two questions: who supplies the knowledge, and how long must the agent keep it? The answers point to one of three services.

Three ways to ground an agent

Service	Use it for	Who supplies it
Agent memory	User-specific context that persists over time	The agent, which learns it
Foundry IQ knowledge base	Grounding on curated content	Your organization
File search tool	Searching during one interaction	The user, who provides the documents

How file search works

File search rewrites the query, runs hybrid searches and reranks the results. Complex queries split into parallel searches. Files are cut into **800**-token chunks with **400** tokens of overlap. They are embedded with text-embedding-3-large at **256** dimensions, and at most **20** chunks reach the model context.

EXAM TRAP

Basic setup stores files in storage and search resources Microsoft manages. Only standard setup keeps your files in your own Blob Storage and Foundry IQ resources.

File search limits

An agent and a conversation can each have only one vector store attached. Each store holds up to **10,000** files, each up to **512** MB and **5,000,000** tokens. Stores made with conversation helpers expire by default **seven** days after last use. Responses wait up to **60** seconds for a conversation store still processing files, but not for the agent's store.

Foundry IQ tool and knowledge bases

One Foundry IQ tool can target only one index. It defaults to `top_k` of **5** and `query_type` of `vector_semantic_hybrid`. Over a private network, authenticate with project managed identity, because keys aren't supported there. Low or medium planning and answer synthesis need an LLM. A knowledge base over an existing search index or file source can use minimal effort with `extractiveData` and no model.

EXAM TRAP

If Microsoft Retriever 1 is a knowledge base's only model, it can't synthesize answers. It supports minimal reasoning and extractive output.

Agent memory

A memory store holds up to 100 scopes and 10,000 memories per scope. Writes are debounced until `update_delay` of inactivity passes, **300** seconds by default. With scope set to `{{$userId}}`, the user comes from `x-memory-user-id` or the caller's Entra token. Default options such as TTL, the time to live after which memories expire, can only be set when the store is created.

EXAM TRAP

A remember command doesn't override TTL. A memory added that way can still expire.

Web grounding choices

Start web grounding with Web Search; it needs no extra Azure resources. Its general search runs on Web IQ, which uses Bing. Domain-restricted search runs on Grounding with Bing Custom Search. Grounding with Bing Search adds parameters and supports non-OpenAI models on Azure. The Data Protection Addendum doesn't cover data sent to either Bing grounding service.

EXAM TRAP

Web grounding tools act as public endpoints. They don't respect VPN or private endpoints, even in a network-secured project.

SharePoint as a knowledge tool

The SharePoint tool supports only user identity authentication. It returns text extracts, not images or charts. Each call returns at most **25** results, and each user gets **200** requests per hour. Queries can be up to **1,500** characters.

EXAM TRAP

The SharePoint tool stops working when the agent is published to Microsoft Teams.

Language agents and tools

Two Azure Language agents cover routing and FAQ answers. Both need a Custom keys connection: the key name is `Ocp-Apim-Subscription-Key` and the value is your resource key. For other language tasks, the Azure Language MCP server exposes features such as entity recognition, sentiment and PII redaction.

The two Azure Language agents compared

Feature	Intent routing agent	Exact question answering agent
Best for	Multi-intent conversations with deterministic routing and fallback	FAQ scenarios that need consistent, verbatim answers
Built on	Conversational Language Understanding (CLU) and Custom Question Answering (CQA)	Custom Question Answering (CQA)
Setup effort	Medium: deploy CLU and CQA projects	Low: deploy a CQA project, no code

Key takeaways

- Use memory for learned user context, a knowledge base for curated content, and file search for user documents.
- One Foundry IQ tool targets one index.
- Connect agents to the toolbox consumer endpoint, which always serves the default version.
- A toolbox endpoint is MCP-compatible, so MCP-capable runtimes can use it too.

Sources: What is Memory? - Microsoft Foundry | Microsoft Learn (opens in a new tab) · File search tool for

Microsoft Foundry agents - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Overview of web grounding capabilities in Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Vector stores for file search in Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Create a Knowledge Base - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Create and use memory in Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Connect agents to MCP server endpoints - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Create and manage a toolbox in Microsoft Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Set up AI solutions in Foundry

Start here

THINK FIRST

How do you pick the right deployment type, agent setup and network design before an AI app ever reaches production?

Why it matters: Questions describe a cost, compliance, network or release need and ask which deployment option, agent setup or pipeline step fits it.

Requests reach a model through its deployment name, and the deployment type shapes price, latency variance and where data is processed. The agent setup decides where agent state lives. Some network and resource settings are fixed at creation, so plan them first.

Key terms

Term	Meaning
Deployment name	The name a request uses to reach a model; one model can have several.
Managed compute	A managed GPU service that runs open-source or custom-weight models on GPUs reserved for you.
Standard agent setup	An agent setup that keeps agent state in customer-managed, single-tenant Azure resources.
Gated release	A release where production stays pinned until the tested candidate version is approved.

Choosing a deployment type

Start most workloads on Global Standard. When latency variance must stay low, choose a Provisioned type instead. Instant access (preview) skips deployment entirely: you call supported models by name.

Common deployment types and when to pick each

Type	When to choose it	Key fact
Global Standard	Most workloads	Broadest regions; new models launch here first

Type	When to choose it	Key fact
Data Zone	Data must stay inside one zone	Processes data only in the US, EU or Asia Pacific zone
Global Batch	Asynchronous groups of requests	**24**-hour target, **50%** less than Global Standard
Developer	Evaluating fine-tuned models	No SLA; deleted after a fixed **24**-hour lifetime

Serverless API or managed compute

Serverless API is the preferred option and supports the most capabilities. Managed compute runs open-source or custom-weight models on dedicated GPU capacity, with no VMs or clusters to own. Its quota is per accelerator SKU per region, separate from VM quota. Partner and community models also need Azure Marketplace access.

EXAM TRAP

Managed compute deployments have no content filtering. Serverless API deployments come with built-in filtering you can customize.

Configuring deployments

Requests reach a model through its deployment name. One model can be deployed several times under different names, for example to test content filter settings. A capacity of **1** equals **1,000** tokens per minute. Creating deployments needs Cognitive Services Contributor or equivalent. On the `/openai/v1/` route, put the deployment name in `model` and omit `api-version`.

Basic or standard agent setup

Standard setup keeps agent state in Azure resources you control. There, Cosmos DB holds messages, conversation history and agent metadata, and Foundry IQ holds vector stores. In either setup, data is isolated between projects.

Basic and standard agent setup compared

Feature	Basic setup	Standard setup
Where agent state lives	The platform's built-in storage	Customer-managed, single-tenant Azure resources
Notable point	Compatible with OpenAI Assistants	Use it for data sovereignty, enterprise compliance or project isolation

Networking for agents

Set the virtual network when you create the Foundry account. A BYO network needs a subnet delegated to `Microsoft.App/environments`, at least `/27` and used by one Foundry resource only. A managed virtual network gives full isolation without planning IP ranges. Prompt agents use a static pool of about **10** IP

addresses per project.

EXAM TRAP

With public egress, a private endpoint protects only the inbound path. Agent egress still isn't isolated.

Settings fixed at creation

Some resource and project settings can't change after creation. Create the resource with `--assign-identity`, or project creation fails. `--allow-project-management` can't be changed later. A project inherits capability settings it doesn't set, but later account changes don't reach it. To change a project's resources, delete and recreate it.

EXAM TRAP

A project can have its own access controls, but it shares network security, deployments and tool integration with its parent resource.

Infrastructure as code

`azd ai agent init` **writes no infrastructure files unless you ask.** Add `--infra` or `--infra=bicep` for Bicep, or `--infra=terraform` for Terraform. Terraform output has no remote-state backend, so set one up in Azure Storage before CI/CD. The templates don't create Application Insights, Log Analytics, Foundry IQ, a storage account or Bing grounding.

CI/CD for hosted agents

`azd pipeline config` **detects the Git provider and creates a service principal.** It also configures repository secrets and variables. The pipeline deploys to existing infrastructure, so run `azd provision` and `azd deploy` first. The workflow signs in with GitHub OIDC, deploys, shows status and sends a smoke test. Code deployment needs Foundry User and Contributor on the project.

EXAM TRAP

An endpoint follows the latest version by default. For a gated release, pin production first, test the new version, and promote it only after approval.

Key takeaways

- Start on Global Standard; choose Data Zone when data must stay in one zone.
- Managed compute has no content filtering; Serverless API has it built in.
- Plan the virtual network up front: it is set when you create the Foundry account.
- Pin production before deploying, so tests run before traffic moves to a new version.

Sources: Deploy models using Azure CLI and Bicep - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Understanding deployment types in Microsoft Foundry Models - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Set up standard agent resources for Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Set up your environment for Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Networking options for Foundry Agent Service - Microsoft Foundry |

Microsoft Learn (opens in a new tab) · Create a project - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Deployment overview for Microsoft Foundry Models - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Set up CI/CD for hosted agents with the Azure Developer CLI - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Manage, monitor, and secure AI systems

Start here

THINK FIRST

How do you keep AI workloads within quota, watch their health, and control who and what can reach them?

Why it matters: Questions describe a capacity, monitoring or security problem and ask which quota action, status signal or access setting resolves it.

Quota tiers and rate limits decide how much traffic a workload can send. Indexer status and query metrics show whether search stays healthy. Keys, roles, managed identities and IP rules decide who can connect.

Key terms

Term	Meaning
Quota tier	A level from Free Tier to Tier **6** that raises quota automatically as usage grows.
PTU	Provisioned Throughput Units: dedicated resources with guaranteed capacity and predictable latency.
Throttled query	A search query that is dropped instead of processed.
Admin key	A Foundry IQ key sent only in the request header; it can be disabled but not deleted.
Indexer	A Foundry IQ object whose runs index documents, for example from blob storage; a run can log errors per document.

Quota tiers and increases

Foundry Models has seven quota tiers: Free Tier and Tiers 1 to 6. Quota grows automatically as usage grows, and Tier **6** gives the most. Free tier subscriptions can't request increases. Requests are handled in order, with priority for customers who actively use their quota. An approved request adds quota but keeps the current tier.

EXAM TRAP

Quota isn't the recommended way to manage costs, even though some customers use it to cap spending.

Quota pools and PTU

Global Standard deployments of one model version share a single quota pool. That pool spans every region in the subscription. Usage tiers cover Standard, Data Zone Standard and Global Standard, but not global batch or provisioned throughput. For latency-critical or high-volume work, PTU offers dedicated resources and predictable latency.

EXAM TRAP

PTU quota is a limit, not a reservation. Approved quota doesn't guarantee that capacity exists in the region you asked for.

Rate limits and service limits

A 429 means a tokens- or requests-per-minute limit was exceeded. Retry with exponential backoff and wait for the `Retry-After` value. Agent file, message and tool limits are fixed and can't be raised.

Foundry limits that shape a design

Limit	Value	What to know
Foundry resources per region per subscription	**100**	Default limit
Projects per resource	**250**	Default limit
Model deployments per resource	**32**	Default limit
Connections per project	**120**	Maximum per project
Versions per agent	**1,000**	Creating another version then fails

EXAM TRAP

The error `regional_session_quota_exceeded` calls for a backoff retry or a different region. Idle or stopped sessions don't count toward that quota.

Reading indexer status

An indexer status of running means it is ready, not that a run is active. A run can report Success with failed documents, as long as errors stay below Max failed items. Execution history keeps up to the **50** most recent runs. S3 HD and Serverless services give all their indexers a shared **48** runtime hours in each **24**-hour UTC window.

Query health and throttling

Throttled queries are dropped, and some throttling is normal. If throttling climbs during indexing, postpone indexing until query traffic eases. The portal keeps query metric history for **30** days. Diagnostic logging gives longer retention and reports on query strings. Logged query requests land in the `AzureDiagnostics` table.

Keys and role-based access

A Foundry IQ service uses key-based authentication by default. Until you choose Both or Role-based access control, every role-based request is denied. Choose Both only while you move clients off keys. A request with a key and role credentials is authenticated by the key. Roles that stand in for an admin key are Search Service Contributor and Search Index Data Contributor.

EXAM TRAP

Regenerating both admin keys at the same time makes every client request that uses them fail with **403** Forbidden.

Managed identities

Each Foundry IQ service has one system-assigned managed identity for life. For blob indexing, give it Storage Blob Data Reader on the storage account. For a direct Azure OpenAI connection, give it Cognitive Services OpenAI User on the hosting resource. Connection details and permissions are checked at run time, while the indexer executes.

Private networking

IP firewall rules control inbound requests only. Callers outside the allowed list get **403** Forbidden. For private endpoints, indexers must run in the service's own private execution environment. A shared private link keeps the Foundry IQ connection to a Foundry billing resource private. It needs a managed identity in the skillset. With private networking, deploy the Foundry resource in its virtual network's region.

EXAM TRAP

Shared private links work only between Azure services. Connections to OpenAI or another external model provider must use the public internet.

Key takeaways

- Handle a **429** with exponential backoff and the `Retry-After` value.
- PTU quota is a ceiling, not a reservation of capacity.
- An indexer run can report Success while some documents fail.
- Foundry IQ starts with keys; switch to role-based access or role requests are denied.

Sources: Microsoft Foundry Models quotas and limits - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Monitor Indexer Status and Results - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Monitor Queries - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Connect Using Azure Roles - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Connect Using API Keys - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Configure a Managed Identity - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Configure Network Access - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Azure OpenAI in Microsoft Foundry Models Quotas and Limits - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Implement responsible AI across generative AI and agentic systems

Start here

THINK FIRST

How do you stop harmful content and risky agent actions without blocking the work your users actually need?

Why it matters: Questions describe a safety, audit or oversight need and ask which guardrail setting, detector, evaluator or control meets it.

Guardrails scan inputs and outputs for risks and decide how to respond. Safety evaluators score how severe harmful content is, and domain experts review the results before deployment. Signed provenance metadata signals where media came from, and task adherence signals let you block or escalate risky tool calls.

Key terms

Term	Meaning
Guardrail control	One rule in a guardrail: the risk to detect, where to scan and how to respond.
Blocking level	The setting that decides which severities a core-harm filter blocks.
Prompt Shields	A detector that scans text for attacks on a large language model.
Spotlighting	Encoding document content so the model trusts it less than direct prompts.
Provenance	Signals about where media came from, such as signed C2PA metadata in the file.

How guardrails work

A guardrail is a set of controls, each with a risk, a scan point and a response. Default safety policies cover all Foundry Models except transcription models. The response depends on where a control acts. When input is blocked, the API sends back an error. When output is blocked, you may get a filtered completion instead.

EXAM TRAP

Not every detection starts on. Document attack detection stays disabled until you enable it, unlike core-harm filters and user prompt attack detection.

Blocking levels

Blocking levels set how much of each core-harm category is blocked. Content at the Safe level is always annotated and never blocked, at any level. Guardrails in Foundry moderate the first **1000** text characters for free.

Core-harm blocking levels compared

Level	What it blocks	What to know
Low	High-severity content only	The least restrictive setting
Medium	Medium and high severity	Default for the four core-harm categories
Off	Nothing in that category	On output, needs approval for modified guardrails

Prompt attacks and spotlighting

Prompt Shields look for user input attacks and document attacks. For agents, they can also scan tool responses from supported tools (preview). The standalone API takes one `userPrompt` and up to **five** documents, and leaves blocking to your app. To cut false positives, switch a control from block to annotate. Spotlighting base-64 encodes documents, but only for models used through Chat Completions.

EXAM TRAP

Prompt Shields only detect manipulation attempts. They don't judge whether content is harmful, accurate, grounded, compliant or on task.

Provenance metadata

C2PA defines signed metadata that can be embedded in media files. Content Provenance Detection accepts media under **100** MB: JPEG, PNG, GIF and WebP images, MP3, WAV and MP4-audio files, and MP4 video. Treat what it finds as a signal of origin. It never proves that content is true, accurate or safe.

EXAM TRAP

A result with no provenance information doesn't prove that the content wasn't AI-generated.

Groundedness detection

Groundedness detection checks text-based summaries and QnA answers. Its domains are MEDICAL and GENERIC. With correction turned on, the response adds a corrected Text field. The feature retires on **April 30, 2027**.

The two groundedness detection modes

Feature	Non-Reasoning mode	Reasoning mode
What it returns	A binary grounded or ungrounded result	An explanation of why content is flagged
When to use it	Real-time apps where latency matters	Development and debugging

Safety evaluators

Content safety evaluators score severity from 0 to 7. A result passes at or below the threshold, which defaults to **3**. These evaluators run on the hosted Foundry Evaluation service, so they need no `deployment_name`. The AI Red Teaming Agent uses them in automated scans. The defect rate is the share of instances above a severity threshold.

EXAM TRAP

The prohibited actions and sensitive data leakage evaluators (preview) work only on agent targets. They aren't available for dataset or model evaluations.

Keeping people in the loop

Automated safety checks support human judgment; they don't replace it. Before release, have domain experts review what automated evaluations covered. Safety evaluations also don't replace manual red-teaming aimed at your domain and users. Use Content Safety to help people decide quickly and accurately, rather than to automate moderation completely.

Governing agent tools

Task adherence flags tool use that is misaligned, unintended or premature. Its signal can block a tool call or escalate to human-in-the-loop review. Routing MCP tools through a gateway is your job, because Foundry doesn't register the server or create the route. A correlation ID policy stamps each tool call with a request ID for later tracing.

Key takeaways

- Default guardrails filter core harms and user prompt attacks; enable document attack detection yourself.
- Medium is the default blocking level and blocks medium and high severity.
- By default, a content safety evaluator passes at a severity of **3** or lower on the **0** to **7** scale.
- Task adherence signals when to block a tool call or escalate it to a person.

Sources: What is Azure AI Content Safety? - Azure AI services | Microsoft Learn (opens in a new tab) · Default safety policies in Microsoft Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Risk and Safety Evaluators for Generative AI - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Microsoft Foundry risk and safety evaluations Transparency Note - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Provenance detection overview - Azure AI services | Microsoft Learn (opens in a new tab) · Task Adherence in Azure AI Content Safety - Foundry Tools | Microsoft Learn (opens in a new tab) · Harm categories in Azure AI Content Safety - Azure AI services | Microsoft Learn (opens in a new tab) · Prompt Shields in Azure AI Content Safety - Azure AI services | Microsoft Learn (opens in a new tab)

Domain 2 | 30–35% of the exam | 15 min read

Implement generative AI and agentic solutions

Overview

Every point below comes from Microsoft's official documentation; each section lists its sources.

Build generative applications by using Foundry

Start here

THINK FIRST

How would you deploy, call, ground and evaluate a generative app that runs on a Foundry project?

Why it matters: Questions describe an app's needs and ask which deployment, endpoint, retrieval approach, integration or evaluator fits them.

An app reaches a model through a named deployment and grounds answers with retrieved content. Evaluators then score the quality of its answers. Most questions turn on picking the right deployment, retrieval style, endpoint and evaluator.

Key terms

Term	Meaning
Deployment name	The name an app passes in the `model` parameter to reach one specific deployment.
Grounding data	Retrieved content the app adds to the user's question so the model answers from it.
Agentic retrieval	A model splits a complex input into focused subqueries, runs them in parallel and returns structured results.
Project endpoint	The `/api/projects/{projectname}/` address, used for stateful APIs such as the agent service.
Groundedness	A quality evaluator that checks whether a response stays within the context it was given.

Deployment paths and types

Serverless API is the preferred, most capable deployment path. Global Standard gives the highest default quota, so you skip load balancing across resources. Batch suits work that can wait: it has no real-time SLA and its own enqueued token quota.

Serverless API and managed compute compared

Feature	Serverless API	Managed compute
Typical models	Foundry Models sold by Azure and select partner and community models	Open-source and custom models
Deployment types	Standard, provisioned throughput, batch and developer	Not used

EXAM TRAP

Batch does not guarantee results within **24** hours. It only aims for that window and might take longer.

Resources, roles and quota

Deploying needs the right role, enough quota and the right resource. Creating deployments takes Cognitive Services Contributor, or equal rights, on the Foundry resource. Deployments and inference draw on tokens-per-minute (TPM) quota, set per region and per model. Partner and community models need an Azure Marketplace subscription first; models sold by Azure don't. An Azure OpenAI resource lists only Azure OpenAI models, so upgrade to a Foundry resource for the full catalog.

EXAM TRAP

You can't deploy through a Foundry or Azure OpenAI resource that sits in another subscription. Cross-subscription connections aren't supported for deployments.

Calling a deployment

Pass the deployment name in the `model` parameter. One model can be deployed several times under different names, for example to compare content filter settings. With Entra ID through a project, point the client at the project endpoint and append `/openai/v1` to the base URL.

The two Foundry endpoints compared

Feature	Foundry endpoint	Project endpoint
Path	`/api/` on the resource	`/api/projects/{projectname}/` on the resource
Serves	Stateless APIs, such as model inference	Stateful APIs, such as the agent service

EXAM TRAP

For a project call with Entra ID, the resource endpoint is the wrong base. The client must use the project endpoint plus `/openai/v1`.

RAG or fine-tuning

RAG retrieves content, adds it to the prompt, then generates. Choose RAG when answers must reflect private or fast-changing data. Choose fine-tuning to change behavior, style or task performance. For new Foundry IQ (Azure AI Search) builds, start with agentic retrieval.

Classic RAG and agentic retrieval compared

Feature	Classic RAG	Agentic retrieval
Querying	The app sends one query and hands the results to the LLM itself	A model splits the input into focused subqueries run in parallel
Semantic ranking	Optional	Built in
Strength	Fewer parts, and faster because no LLM plans the query	Structured grounding data that suits chat completion models

EXAM TRAP

Fine-tuning is not the way to add fresh knowledge. When answers depend on private or changing data, use RAG instead.

Building the retrieval layer

Agents get retrieval as a tool; full RAG apps use the Foundry SDK. An agent can use a tool such as file search. Agentic retrieval adds **3** objects. Knowledge sources each point at searchable content, such as an index or SharePoint site. A knowledge base unifies them, and app code calls the retrieve action.

Grounded and secure answers

Enforce access control when content is retrieved. With Foundry IQ, document-level security filters limit what each user can retrieve. To curb hallucination, turn on citations and tell the model to stay with retrieved content. In production, prefer Entra ID; API keys are a development convenience.

EXAM TRAP

Grounding alone doesn't stop hallucination. Citations plus clear system messages that keep the model on retrieved content are the usual fix.

SDKs, connectors and gateways

Pick the integration that matches where the app runs. Power Apps, Power Automate and Logic Apps use the Azure OpenAI connector. Azure API Management can act as an AI gateway, one entry point across model hosts. In `langchain-azure-ai`, a project endpoint needs Entra ID and the Foundry User role.

EXAM TRAP

`init_embeddings`, unlike `init_chat_model`, has no automatic fallback to `DefaultAzureCredential`. Pass it a credential explicitly.

Evaluating quality

Groundedness checks whether answers stay within the given context. AI-assisted quality evaluators require a deployed GPT model, reached through an Azure OpenAI connection. Agent evaluators run only at the Individual turns scope. A Dataset target scores existing outputs without re-running anything. Run comparisons use t-testing for more reliable results.

EXAM TRAP

A surprisingly low Groundedness score often points to incomplete retrieval or context. Check how the context is built and how long retrieval takes.

Key takeaways

- Start new Foundry IQ RAG builds with agentic retrieval; classic RAG is simpler and faster but sends a single query.
- Adding a project connection takes Foundry User, Foundry Owner or Azure Contributor, or a higher role.
- A Foundry resource holds only **1** Azure Key Vault connection at a time.
- Prompt reasoning models with simple instructions, skip chain-of-thought, and set `max_completion_tokens` high enough for reasoning and answer.
- When content filtering blocks a call, catch HTTP **400** errors and the `content_filter` finish reason; never resend it unchanged.

Sources: Deploy Microsoft Foundry Models in the Foundry portal - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Retrieval augmented generation (RAG) and indexes in Microsoft Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Run evaluations from the Microsoft Foundry portal - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Integrate Microsoft Foundry with your applications - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Deployment overview for Microsoft Foundry Models - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Understanding deployment types in Microsoft Foundry Models - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Add a new connection to your project - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Deploy models using Azure CLI and Bicep - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Build agents by using Foundry

Start here

THINK FIRST

How do you give an agent the right tools, knowledge and memory while keeping a person in control of risky actions?

Why it matters: Questions describe an agent requirement and ask which tool, setting, approval control or orchestration pattern meets it.

An agent can call tools such as function tools, OpenAPI tools, web grounding and MCP servers. Memory carries context across sessions. The `require_approval` setting keeps MCP calls under a person's control, and Workflows orchestrate several agents.

Key terms

Term	Meaning
Function tool	A tool you define by schema; your code runs it and returns the output to the agent.
Vector store	A store of files, up to 10,000 of them, that you attach to an agent or to a conversation.
Knowledge base	A Foundry IQ object that defines which knowledge sources to query and how retrieval behaves.
Memory scope	A division of a memory store, such as one end user's memories, set with the tool's `scope`.
Managed MCP server	The MCP server Foundry creates and runs for you when you add a catalog connector.

Function tools

Set function definitions with the SDK or REST API, not the portal. The portal can run agents with function tools but can't change their definitions. To make the schema enforce required fields, add a `required` array. Runs expire **10** minutes after creation, so submit tool outputs before then.

EXAM TRAP

A function call with no final answer means the output never reached the model. Run the function, then call `responses.create` with its output and the conversation ID.

API and custom tools

Each OpenAPI operation needs an `operationId` of letters, `-` and `_` only. OpenAPI tools don't support OAuth identity passthrough. Pick anonymous access, a key or the project's managed identity instead. A custom tool hosted on Azure Functions lets agents call it without direct access to the databases, APIs or networks behind it.

EXAM TRAP

The Azure Functions tool doesn't work with the basic agent setup. It needs the queue resources and managed network integration that the standard setup provides.

File search and vector stores

Attach at most one vector store to an agent and one to a conversation. Deleting a file object removes it from every vector store across the organization. Response generation waits up to **60** seconds for a conversation's store that is still processing files.

File search defaults and limits

Setting	Default or limit	What to know
Chunking	**800**-token chunks, **400**-token overlap	Default chunking settings
Embedding model	text-embedding-3-large, **256** dimensions	Up to **20** chunks go into context
Files per vector store	**10,000**	At most one store per agent and one per conversation
File size	**512** MB and up to **5,000,000** tokens	Tokens are counted when you attach the file
Conversation store expiry	**Seven** days after last activity	Runs on that conversation then fail

Knowledge with Foundry IQ

A knowledge base decides which knowledge sources an agent queries. The Foundry IQ tool defaults to `top_k` of **5** and `query_type` of `vector_semantic_hybrid`. With a private virtual network, authenticate with project managed identity, because keys aren't supported there. Low or medium query planning and answer synthesis need an LLM.

EXAM TRAP

A **401** or **403** with keyless auth means roles are missing. Give the project's managed identity Search Index Data Contributor and Search Service Contributor.

Web grounding

Web grounding gives agents real-time public web data beyond the model's cutoff. Start with Web Search, which needs no extra Azure resources. Grounding with Bing Search gives you more parameters. It also works with non-OpenAI models deployed directly on Azure. If the agent never uses the tool, set `tool_choice="required"` to force it.

EXAM TRAP

Web grounding tools behave as public endpoints. They ignore VPN and private endpoints, even in a network-secured Foundry project.

Conversation memory

Short-term memory covers one session; Foundry memory is long-term. Each memory store pairs a chat model with an embedding model to extract and index user-profile memories. It holds up to **100** scopes and **10,000** memories per scope. A `default_ttl_seconds` of **0** means memories never expire. For per-user

memory, set the scope to `{{$userId}}` and send `x-memory-user-id`.

> **EXAM TRAP**
> Memory writes failing with **401** and an empty store point to a missing Cognitive Services OpenAI User role.

MCP tools and approvals

`require_approval` **defaults to** `always`**, so every call needs approval.** Set it to `never` only when no approval is wanted. Leaving `allowed_tools` out, or empty, exposes every tool on the MCP server. A catalog connector becomes a managed MCP server scoped to its project, with actions only. Foundry doesn't create gateway routes for MCP servers for you.

> **EXAM TRAP**
> Don't let the model decide permissions for A2A agents. Enforce permissions and approvals independently, and treat provider descriptions as untrusted input.

Multi-agent orchestration

The classic Connected Agents tool is gone from the new Agent Service. Use Workflows to orchestrate agents declaratively: in sequence, as a group chat, or with a human in the loop. Or use the A2A tool to reach any A2A-compatible endpoint. The `a2a` type uses protocol **1.0** and is generally available; `a2a_preview` uses **0.3**. Tracing confirms whether a tool call really happened.

Key takeaways

- Edit function definitions in code; the portal can only run agents that use them.
- `require_approval` stays `always` unless you set it to `never`.
- Use Workflows or the A2A tool where classic Connected Agents used to fit.
- Web grounding tools act as public endpoints, even in a network-secured project.

Sources: Use function calling with Microsoft Foundry agents - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Connect OpenAPI tools to Microsoft Foundry agents - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Overview of web grounding capabilities in Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Connect agents to MCP server endpoints - Microsoft Foundry | Microsoft Learn (opens in a new tab) · What is Memory? - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Connect to an A2A agent endpoint from Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in a new tab) · File search tool for Microsoft Foundry agents - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Vector stores for file search in Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Optimize and operationalize generative AI systems

Start here

THINK FIRST

How do you tune a model's output, improve its instructions and see what it does once it runs in production?

Why it matters: Questions describe a quality, cost or latency problem and ask which parameter, prompt technique or monitoring signal addresses it.

You set how random output is with temperature, then let Prompt Optimizer improve your instructions. Connecting Application Insights turns on server-side tracing. One endpoint lets an app switch between models without code changes.

Key terms

Term	Meaning
Temperature	A parameter from **0** to **2** that sets how random or focused the output is.
Chain-of-thought prompting	Asking a model to work step by step and show every step in its response.
Server-side tracing	Traces Foundry records once Application Insights is connected to the project.
Deployment	An alias that gives a model a name and configuration; requests use that name.
Defect rate	The share of test instances whose severity passes a threshold in a safety evaluation.

Temperature and top_p

Temperature accepts values from 0 to 2. A higher value such as **0.7** makes responses more random and divergent. A lower value such as **0.2** keeps them focused and concrete. Top_p also controls randomness, in a different way. Adjust one of the two at a time, not both.

Prompt structure

The end of a prompt can sway the output more than the start. That is recency bias, so test repeating your instructions at the end. A system message steers the model but doesn't guarantee compliance, so layer it with filtering and evaluation. Citations placed inline curb false content better than citations at the end.

EXAM TRAP

Classic prompt engineering techniques aren't recommended for reasoning models, such as gpt-5 and the o-series models.

Chain of thought

Chain-of-thought prompting asks the model to show every step. Seeing the steps lowers the chance of an inaccurate outcome. It also makes the response easier to assess. The technique applies only to non-reasoning models. For a reasoning model, the reasoning summary parameter is the supported way to see its reasoning.

EXAM TRAP

Pulling out a reasoning model's reasoning by any other method isn't supported. It may break the Acceptable Use Policy and lead to throttling or suspension.

Prompt Optimizer

Prompt Optimizer rewrites an agent's system instructions for you. It applies prompt engineering best practices and explains each change paragraph by paragraph. You need a prompt or workflow agent and a model deployment in a supported region. Results aren't kept, so select Use prompt before closing the dialog. Each reoptimization starts from the latest optimized text.

EXAM TRAP

Prompt Optimizer works only on text instructions. It doesn't optimize tool definitions or knowledge sources.

Optimization runs

The agent optimizer compares candidates with your current agent. Candidates can improve instructions, function-calling tool descriptions or model selection. Turn on Compare across models to test several deployments. If every candidate scores below the baseline, keep the current agent and revise the dataset, evaluators or settings. Treat the cost estimate as a modeled range: it is neither a cap nor the final charge.

EXAM TRAP

Optimization runs exercise the agent's tools. When those tools can modify production systems, a run can cause unintended or irreversible changes.

Tracing in Foundry

Server-side tracing starts once you connect Application Insights. It needs no code changes, and traces cover the past **90** days. Prompt, hosted and external agents have it in general availability; workflow agents have it in preview. Querying traces needs Log Analytics Reader, plus Privileged Monitoring Data Reader for protected tables. Search by Trace ID, Response ID or time range.

EXAM TRAP

If spans appear without input and output attributes, enable content recording with `OTEL_INSTRUMENTATION_GENAI_CAPTURE_MESSAGE_CONTENT=true`.

Signals to monitor

Each signal comes from a different place. Safety evaluations check output for hate, sexual, violent and self-harm content, jailbreaks and protected material. The defect rate is the share of instances above a severity threshold.

Where each observability signal comes from

Signal	Where to find it	What to know
Agent traces	Traces view in the Foundry portal	Past **90** days, out of the box
Resource logs	Azure Monitor logs	Saved only after you add a diagnostic setting
Token counts	`usage` object in each Responses reply	`input_tokens`, `output_tokens`, `total_tokens`
Safety signals	Risk and safety evaluations	Risk level or True/False label, with reasoning
Startup latency	`x-ms-debug-latency-container-readiness-ms`	Send `x-ms-debug-latency-enabled: true` and compare

EXAM TRAP

Once a provisioned-managed deployment's utilization reaches **100%** or more, its calls are throttled and return a **429**.

Orchestrating several models

Foundry Models reaches many providers through one endpoint and credential set. So an app can switch models without code changes. Each request names a deployment. Choose the AI Gateway option for token limits, governance or routing. In code, a connected model looks like `central-gateway/approved-chat`: connection name, then model name.

EXAM TRAP

Connected models don't work everywhere. Foundry IQ, agent memory, guardrails and real-time audio don't support them.

Key takeaways

- Tune temperature or top_p, but change only one of them at a time.
- Use chain-of-thought prompting with non-reasoning models only.
- Connect Application Insights to get server-side traces with no code changes.
- A deployment that rejects the Responses API with **400** Model not supported needs Chat Completions.

Sources: Prompt engineering techniques - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Optimize agent prompts with Prompt Optimizer - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Set Up Tracing for AI Agents in Microsoft Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Endpoints for Microsoft Foundry Models - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Microsoft Foundry risk and safety evaluations Transparency Note - Microsoft Foundry | Microsoft Learn (opens in a new tab) · System message design for Azure OpenAI - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Quickstart: Optimize a prompt agent - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Quickstart: Trace your hosted agent - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Domain 3 | 10–15% of the exam | 13 min read

Implement computer vision solutions

Overview

Every point below comes from Microsoft's official documentation; each section lists its sources.

Design and implement image- and video-generation solutions

Start here

THINK FIRST

When should you generate new images or video, when should you edit what you already have, and which controls shape the result?

Why it matters: Questions describe an image or video task and ask which API, parameter value or editing workflow fits it.

Image models generate from a text prompt, and Sora 2 can also take input images or video. Parameters such as `quality`, `size` and `seconds` shape what comes back. Edits and remixes change an existing image or video instead of starting from scratch.

Key terms

Term	Meaning
Mask	A PNG whose fully transparent areas show the edit API where to change the image.
Remix	A targeted change to a completed Sora 2 video, made from its ID and an updated prompt.
`input_reference`	The Sora 2 parameter that turns an existing image into video at a matching resolution.
Partial images	Interim images sent while an image generation streams, from **0** to **3** per request.

Generating images from text

An image prompt can run up to 32000 characters. Set `n` between **1** and **10** for the number of images. `quality` defaults to `auto`, and GPT image models also accept `high`, `medium` and `low`. Generation time varies with the quality setting and prompt complexity, so `low` is suggested for time-sensitive apps.

EXAM TRAP

`style` and `response_format` are legacy parameters that GPT image models don't support. Describe the style in the prompt; the images always come back base64-encoded.

Output controls and their rules

Each output control has its own range and a condition to meet. Check the condition before you combine controls. GIF output, for example, can't be streamed and can't use a mask.

Image output controls, their values and the condition each needs

Control	Values	Condition
`output_compression`	**0** to **100%**, default **100**	Only with webp or jpeg output
Transparent background	Transparent	Output format must be png or webp
`moderation`	`low` or `auto` (default)	`low` filters less strictly
`partial_images`	**0** to **3**	Streaming only; **0** sends one image in one event
GIF `frame_count`	**2** to **16**, default **9**	No streaming, partial images or masks

Custom image sizes

Newer GPT image models accept custom `WIDTHxHEIGHT` sizes. This applies to gpt-image-2 and the gpt-image-2.5 models. Both dimensions must divide evenly by **16**. The aspect ratio must sit between **1:3** and **3:1**. The largest size is **3840x2160**, and anything above **2560x1440** is experimental.

Editing images with masks

The edit API takes up to 16 png, webp or jpg images, each under 50 MB. A mask marks the area to edit with fully transparent pixels. It must be a PNG under **4** MB with the same dimensions as the image. When you send several images, the mask applies to the first one.

EXAM TRAP

Asking for image variations won't work. The variations endpoint is retired and no longer accepts any requests.

Images inside an agent

The agent tool generates images from text prompts during a conversation. It needs two deployments in the same Foundry project: a model for the agent, such as gpt-5, and an image generation model. For editing, masks or partial-image streaming, call the Azure OpenAI Image API directly instead.

EXAM TRAP

When agent image generation fails, check the `x-ms-oai-image-generation-deployment` header. It must be on the Responses request and match your image deployment's name.

Generating video with Sora 2

Sora 2 creates video from text, input images or an input video. You can give it up to **two** images, and the video interpolates between them. Or give it one video of up to **five** seconds. The table shows each parameter's allowed values and default.

Sora 2 video parameters and their defaults

Parameter	Allowed values	Default
`model`	`sora-2`, `sora-2-pro`	`sora-2`
`seconds`	**4**, **8**, **12**	**4**
`size`	**720x1280**, **1280x720**, **1024x1792**, **1792x1024**	**720x1280**

EXAM TRAP

With `input_reference`, the image and the video must have the same resolution, and only **720x1280** or **1280x720** are supported.

Remixing and editing video

A remix adjusts an existing video instead of regenerating it. Call `POST /openai/v1/videos/{video_id}/remix` with a completed video's ID and an updated prompt. Limit each remix to one clearly stated change for the best fidelity. The edit operation, `POST /openai/v1/videos/edits`, takes a video or a completed video's ID plus an edit prompt.

EXAM TRAP

Sora 2 can't generate real people, including public figures. Input images that show human faces are currently rejected too.

Key takeaways

- GPT image models don't support legacy `style` or `response_format`; put the style in the prompt.
- A mask is a same-size PNG under **4** MB; transparent pixels mark the edit.
- Use the Image API, not the agent tool, for edits, masks and partial-image streaming.
- Remix a completed Sora 2 video with one precise change to keep the most fidelity.

Sources: Create an image | Microsoft Learn (opens in a new tab) · Sora 2 video generation overview (preview) - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Create a video | Microsoft Learn (opens in a new tab) · Create an image edit multipart Create an image edit json | Microsoft Learn (opens in a new tab) · Use the image generation tool (preview) in Foundry Agent Service - Microsoft Foundry | Microsoft Learn (opens in

a new tab) · Create an image variation | Microsoft Learn (opens in a new tab) · Create a video remix | Microsoft Learn (opens in a new tab) · Create a video edit | Microsoft Learn (opens in a new tab)

Design and implement multimodal understanding workflows

Start here

THINK FIRST

How would you design a Foundry workflow that turns images and video into captions, answers, and extracted visual details you can trust?

Why it matters: Questions describe an image or video task and ask which Content Understanding setting, limit, or behavior fits it.

Content Understanding needs a chat completion model and an embeddings model to analyze visual input. It answers only from the content and context you supply. Your main design levers are the fields you define, the frames you send, and when output goes to human review.

Key terms

Term	Meaning
Classify field	A field kind that picks a value from the input, such as whether an image shows a defect.
Video segment	A meaningful part of a video, returned with its transcription and a scene description.
Confidence threshold	A cutoff score; output such as a caption that scores below it goes to further review.
Face grouping	Groups of similar faces with no further analysis, offered only for video.
Grounding	Page numbers and bounding boxes that show where an extracted value came from.

Models behind the service

Content Understanding needs a chat completion model and an embeddings model. Have both deployments ready before you analyze images or video. The service stays grounded: it builds output only from the content and context you give it. It is not meant for ungrounded inference.

EXAM TRAP

Content Understanding does not fill gaps with outside knowledge. Its output comes only from the input and context you supply.

Image input and captions

Image Analysis takes JPEG, PNG, GIF, or BMP files under 4 MB. Each image must also be larger than **50 x 50** pixels. For automatic captions, set a minimum confidence threshold. A caption that scores below it goes to further review instead of straight to users.

EXAM TRAP

A tiny thumbnail fails even in a valid format, because images must be larger than **50 x 50** pixels.

Extracting chart details

Content Understanding reads a chart in the context of its page or image. You can pull out the chart's type, a summary, and its overall meaning. To label the chart type, add a field that uses the `classify` method. Give that field an enum of values such as bar, line, and pie.

Video analysis settings

Video analysis splits a video into meaningful segments. Each segment comes with a transcription and a scene description. Field extraction can add scene-level summaries and key topics. Keep frames at HD or smaller, at **30 FPS**, with no more than **10** people in a frame.

Sending frames to a model

Send about 1 frame per second to an AI model. Processing **10** or more frames can delay the result. Camera angle matters too. Video object detectors were trained on frontal views, so they may misclassify objects seen from overhead.

EXAM TRAP

An overhead camera is not a neutral choice. Detectors trained on frontal views may misclassify objects seen from above.

Faces in images and video

Face detection marks where faces are, not who they are. It returns bounding boxes and finds no individually identifying features. Faces are blurred before images or video reach the vision models for processing. So emotion inference doesn't work.

How faces are handled in images and video

Feature	Images	Video
Faces before model processing	Blurred	Blurred
Face grouping	Not supported	Supported, with no further analysis
Emotion inference	Doesn't work	Doesn't work

EXAM TRAP

Do not plan emotion detection on faces. Blurring stops that kind of inference on both images and video.

Grounding and review

Bounding-box grounding of extracted values covers documents only. For now, images and video don't get page numbers or bounding boxes for extracted values. Plan human review of visual output another way, for example with a caption confidence threshold.

EXAM TRAP

Grounding with page numbers and bounding boxes is not yet available for images or video. It currently covers the document modality only.

Key takeaways

- Send about **1** frame per second; **10** or more frames can delay the result.
- Only video supports face grouping, and blurred faces block emotion inference.
- Use a `classify` field with an enum to label chart types such as bar, line, and pie.
- Grounding with bounding boxes currently covers documents, not images or video.

Sources: Model deployment options for Content Understanding analyzers - Foundry Tools | Microsoft Learn (opens in a new tab) · Transparency Note and use cases for Content Understanding - Foundry Tools | Microsoft Learn (opens in a new tab) · Characteristics and limitations for Image Analysis - Foundry Tools | Microsoft Learn (opens in a new tab) · Build a retrieval-augmented generation solution with Azure Content Understanding in Foundry Tools - Foundry Tools | Micr (opens in a new tab)

Implement responsible AI for multimodal content

Start here

THINK FIRST

How do you keep harmful images, hidden instructions and misused brands out of an app that accepts visual content?

Why it matters: Questions describe a visual content risk and ask which detector, setting or policy feature handles it.

Content Safety classifies harmful images, alone or with their text. Prompt Shields detect instructions hidden in documents, and spotlighting makes the model trust document content less. Brand detection names the brand of a detected logo, and the Detect API can give information about the system behind a watermark.

Key terms

Term	Meaning
Severity level	A harm rating of **0**, **2**, **4** or **6**; a higher value means more severe content.
Document attack	Hidden instructions in documents, emails or web pages that try to take over the model session.
Spotlighting	Base-64 encoding of document content so the model trusts it less than direct prompts.
Invisible watermark	A signal an AI generator embeds in content that can't be seen or heard.

Which detector to use

Pick the detector by what you send and what you need back. Content Safety rates harm with severity values of **0**, **2**, **4** or **6**. A configurable match severity level exists only in Content Safety Studio.

Three Content Safety detectors compared

Feature	Image Detection	Multimodal API	Prompt Shields
Input	Images	Image and text content together	A user prompt and up to **five** documents
What it finds	Hate, sexual, violence, self-harm	Harmful user- or AI-generated content	Attempts to manipulate the model

Images that contain text

The Multimodal API analyzes an image and its text together. Set `enableOcr` to `true` and it reads at most **1000** characters from the image, truncating the rest. If you list no categories, all four are checked. Images must be between **50 x 50** and **7,200 x 7,200** pixels, up to **4** MB. The preview retires on **April 30, 2027**.

> **EXAM TRAP**
> An animated image isn't checked frame by frame. The service extracts only the first frame for analysis.

Adult, racy and gory images

Add `Adult` to `visualFeatures` in Analyze Image 3.2. The response then holds three booleans: `isAdultContent`, `isRacyContent` and `isGoryContent`. Each comes with a matching score between zero and one, such as `adultScore`.

Hidden instructions in content

Document attacks hide instructions in documents, emails or web pages. In Foundry, user prompt attack detection starts on, but document attack detection must be enabled. For agents, Prompt Shields can also scan tool responses from supported tools (preview). Each control returns `detected` and `filtered` values.

EXAM TRAP

Prompt Shields only detect manipulation attempts. They don't decide whether content is harmful, accurate, grounded or compliant with policy.

The standalone Prompt Shields API

The standalone API returns `attackDetected` for the prompt and each document. It never blocks or routes content itself; your application decides what to do with the result. If false positives are a problem, switching from block to annotate mode logs detections without filtering.

Spotlighting documents

Spotlighting makes the model trust document content less. It is off by default; turn on its toggle in the document attack controls. It works only for models used through Chat Completions. It has no direct cost, but base-64 adds tokens, raising total cost and possibly pushing large documents past input limits.

EXAM TRAP

A response may say that a document was base-64 encoded even though no prompt asked. That's a known side effect, not a fault.

Brands and logos

Add `Brands` to `visualFeatures` to detect logos. Each detected logo comes back with the brand name, a confidence score and a bounding box. If the brand you need isn't recognized, train your own logo detector with Custom Vision. Content Safety custom categories also support brand protection rules.

Watermarks and provenance

Invisible watermarks can't be seen or heard, but they can be detected. When the Detect API finds a supported watermark, it can give information about the generation system behind it. Content Provenance Detection accepts media under **100** MB, including JPEG, PNG, GIF and WebP images.

Key takeaways

- Turn on document attack detection; only user prompt attack detection is on by default.
- Set `enableOcr` to check up to **1000** characters of image text together with the image.
- Spotlighting is off by default and works only through Chat Completions.
- Use `Brands` for known logos and a Custom Vision detector for brands it misses.

Sources: Transparency Note and use cases for AI Content Safety - Foundry Tools | Microsoft Learn (opens in a new tab) · Quickstart: Analyze multimodal content - Azure AI services | Microsoft Learn (opens in a new tab) · Prompt Shields in Microsoft Foundry - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Brand detection - Azure Vision in Foundry Tools - Foundry Tools | Microsoft Learn (opens in a new tab) · Provenance detection overview - Azure AI services | Microsoft Learn (opens in a new tab) · Prompt Shields in Azure AI Content Safety - Azure AI services | Microsoft Learn (opens in a new tab) · Detect Adult, racy, or gory content - Azure Vision in Foundry Tools - Foundry Tools | Microsoft Learn (opens in a new tab)

Domain 4 | 10–15% of the exam | 9 min read

Implement text analysis solutions

Overview

Every point below comes from Microsoft's official documentation; each section lists its sources.

Apply language model text analysis

Start here

THINK FIRST

When should you shape text output with a prompt and a schema, and when should you call a dedicated language or translation feature?

Why it matters: Questions describe a text task and ask which feature, parameter or translation option fits it.

Structured outputs make a model follow the JSON Schema you supply. Dedicated features handle summaries, sentiment, safety checks and translation, each with its own settings. Your job is to match each text task to the feature that fits it.

Key terms

Term	Meaning
Structured outputs	A mode where the model's JSON must follow the JSON Schema you supply, not just be valid JSON.
Abstractive summary	A summary the model writes in its own words by paraphrasing the key information.
Confidence score	A value from **0** to **1** for how sure the model is about a sentiment label.
AdaptCT	Adaptive custom translation: steers LLM output with **5** to **10,000** example segment pairs, through few-shot retrieval at inference.
NMT	Azure Translator's standard translation model, the option you pick instead of an LLM deployment.

Strict JSON from a model

Structured outputs make the model follow your JSON Schema exactly. Both Chat Completions and the Responses API support it. For Chat Completions, the schema goes in `response_format`. For the Responses API, it goes in `text.format`.

EXAM TRAP

JSON mode is the older option. It guarantees valid JSON, but it can't guarantee that the output sticks to your schema.

Schema rules and limits

Every object in the schema needs `additionalProperties: false`**.** That setting blocks keys the schema doesn't define. A schema can hold up to **100** object properties in total, with up to **five** levels of nesting. Structured outputs don't work with parallel function calls, so set `parallel_tool_calls` to `false`.

Two kinds of summary

Extractive summaries pick sentences; abstractive summaries paraphrase. Document summarization takes at most **25** documents per request. Larger documents make each call slower.

Extractive and abstractive summaries compared

Feature	Extractive	Abstractive
What you get back	Sentences taken from the input	New, succinct sentences that paraphrase it
What to watch	Chronological order unless you set `sortBy`	Some information or accuracy can be lost

EXAM TRAP

Summarization doesn't check facts or verify content. Your application needs its own guard against spreading false information.

Key phrase extraction

Key phrase extraction surfaces the main concepts in a text. It retires from Azure Language on **March 31, 2029**. Send new projects to Microsoft Foundry, and move existing workloads before that date to avoid disruption.

Reading sentiment results

Sentiment returns a label and 0 to 1 scores for each of three labels. You get positive, neutral and negative scores for each document and each sentence. By default, the label is the highest score, but you can set your own thresholds. Optional opinion mining adds aspects, such as product attributes, with their opinion words.

EXAM TRAP

A score near **1** means the model is confident in that label. It says nothing about how intense the sentiment is.

Safety checks on text

Content Safety text detection covers hate, sexual, violence and self-harm. Each category comes back with a risk level on a multi-severity scale. Custom categories add finer control on top of those defaults; they don't replace them. Supported languages are currently Chinese, English, French, German, Italian, Japanese, Portuguese and Spanish.

Translator request options

Azure Translator can run standard NMT or an LLM deployment. LLM-based translation needs a Microsoft Foundry resource and adds formal, informal or neutral tone plus gender controls. Repeat `to` to get several languages from one request. Leave out `from` and the input language is detected automatically. For profanity, `profanityAction` takes `NoAction` (default), `Marked` or `Deleted`.

EXAM TRAP

The dynamic dictionary doesn't work with autodetection. Pass `from` explicitly, and remember that the dictionary is case-sensitive.

Customizing translation

Custom Translator trains a model; AdaptCT steers the LLM at inference. Choose Custom Translator for a dedicated production model built from a large parallel corpus. Choose AdaptCT to influence LLM style and terminology with no large training run or separate deployment.

Two ways to tailor translations to your domain

Feature	Custom Translator	AdaptCT
How it adapts	Full model training and deployment	Few-shot retrieval at inference
Data you supply	A large parallel corpus	**5** to **10,000** segment pairs, each up to **250** characters
Time to update	Retrain and redeploy, up to about **48** hours	Adaptive dataset ready in minutes
How a request uses it	Category ID in `category` (default `general`)	Up to **five** reference translations or a dataset index ID

EXAM TRAP

With `allowFallback=false`, Translator uses only custom systems for the requested category. If none exists, the request fails with a **400** status code.

Key takeaways

- Structured outputs enforce your schema; the older JSON mode only guarantees valid JSON.
- By default, the sentiment label is the highest confidence score, unless you set thresholds.
- LLM translation needs a Microsoft Foundry resource and adds tone and gender controls.

- AdaptCT builds its dataset in minutes; a Custom Translator retrain can take about **48** hours.

Sources: How to use structured outputs with Azure OpenAI in Microsoft Foundry Models - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Characteristics and limitations for summarization - Foundry Tools | Microsoft Learn (opens in a new tab) · Transparency note - Sentiment Analysis feature of Azure Language in Foundry Tools - Foundry Tools | Microsoft Learn (opens in a new tab) · Transparency Note and use cases for AI Content Safety - Foundry Tools | Microsoft Learn (opens in a new tab) · What is Azure text translation in Microsoft Foundry? - Microsoft Foundry | Microsoft Learn (opens in a new tab) · Foundry Tools adaptive custom translation - Foundry Tools | Microsoft Learn (opens in a new tab) · What is key phrase extraction in Azure Language in Foundry Tools? - Foundry Tools | Microsoft Learn (opens in a new tab) · Translator Translate Method - Foundry Tools | Microsoft Learn (opens in a new tab)

Implement speech solutions

Start here

THINK FIRST

How would you give an agent the ability to listen, speak and translate, and when would you need a custom speech model?

Why it matters: Questions describe a voice scenario and ask which speech feature, setting or custom model fits it.

Speech to text turns a user's spoken words into text that a dialog system or analytics can use. The Speech Translation API transcribes audio, then translates it into the user's language, and text to speech can voice the result. You can customize recognition at runtime with a phrase list or ahead of time with custom speech.

Key terms

Term	Meaning
Speech MCP server	The remote MCP server that gives a Foundry agent speech to text and text to speech.
Diarization	Telling speakers apart in a transcript, labeled GUEST1, GUEST2 and so on.
Phrase list	A runtime way to customize speech to text, with no model training ahead of time.
Voice sync for avatar	A custom voice like the avatar talent's, created with a custom avatar and used only with it.
ITN rules	Display text formatting rules for speech to text; Microsoft's default rules run before any custom ITN model.

Speech as an agent tool

Connect Azure Speech to an agent by adding a remote MCP server. The agent's model must be able to call tools, as gpt-4o and later models can. MCP tools need the Agent Service Enterprise tier. In `X-Blob-Container-Url`, give a SAS URL with read and write permissions; audio output files are stored

there.

> **EXAM TRAP**
> The Speech MCP tool doesn't work with network-secured Microsoft Foundry.

Audio in and out

For the best recognition, send WAV audio at 16 kHz and 16-bit depth. Spoken output from the agent's Speech tool is saved as a WAV file in the blob container, and an audio link appears in the chat. Transcripts can feed a dialog system or analytics such as summarization and sentiment.

> **EXAM TRAP**
> Speakers aren't separated unless you ask. Diarization is off by default; turn it on to get a speaker label on each utterance.

Text to speech over REST

`POST /openai/v1/audio/speech` **turns up to 4096 characters into audio.** The required `model` field takes one of the available TTS models, such as `tts-1`, `tts-1-hd` or `gpt-4o-mini-tts`. The `input` field holds the text to speak.

> **EXAM TRAP**
> The `instructions` parameter for shaping the voice doesn't work with `tts-1` or `tts-1-hd`.

Customizing speech to text

Use a phrase list at runtime, or custom speech ahead of time. For custom display formatting, Microsoft's default ITN rules run first. Their output then feeds your custom ITN model, which matches regardless of case. A deployed model bills for endpoint uptime, whether or not anyone calls it.

> **EXAM TRAP**
> Deleting a deployment stops hosting charges but doesn't delete the model. To use it again, create a new deployment.

Translating speech

The Speech Translation API transcribes audio, then translates it. It is a cascade of Speech Services and Text Translator, and the user chooses the target language. The translated result can stay as text, or text to speech can turn it into audible speech.

Custom neural voice

Custom neural voice is a Limited Access feature that needs registration. Only customers managed by Microsoft are eligible. The lite project type (preview) builds a demo model from **20** to **50** samples recorded in Speech Studio. Every deployed custom voice must disclose that it is synthetic and offer users a feedback

channel.

Custom avatars

Custom avatar supports only standard (S0) Foundry or Speech resources. Training usually takes **20** to **40** compute hours. All training data must match your chosen setting, such as **4K** footage for **2160p**. An avatar name allows letters, numbers, hyphens and underscores, and must be unique in the resource.

> **EXAM TRAP**
> A voice sync for avatar voice can't be reused elsewhere; it serves only its own avatar. Your consent statement must cover both.

Storage and encryption

Set `userOwnedStorage` when you create the Foundry resource. Speech outputs then go to your bound storage containers. That covers batch and real-time speech to text, Custom Speech, text to speech and Custom Voice.

> **EXAM TRAP**
> Without `userOwnedStorage`, Speech uses Microsoft-managed storage, which your customer-managed key doesn't cover.

Key takeaways

- Agents reach Speech through a remote MCP server and need a tool-calling model.
- Send **16** kHz, **16**-bit WAV audio for the best recognition.
- The `instructions` voice control doesn't work with `tts-1` or `tts-1-hd`.
- A deployed custom model bills for uptime, even when nobody calls it.

Sources: Use cases for Speech to text - Foundry Tools | Microsoft Learn (opens in a new tab) · Connect Azure Speech in Foundry Tools to an agent - Microsoft Foundry | Microsoft Learn (opens in a new tab) · How to create a custom text to speech avatar - Speech service - Foundry Tools | Microsoft Learn (opens in a new tab) · Display text format training data - Speech service - Foundry Tools | Microsoft Learn (opens in a new tab) · Create a speech | Microsoft Learn (opens in a new tab) · Limited Access - Foundry Tools | Microsoft Learn (opens in a new tab) · Connect your own storage to Speech/Language - Microsoft Foundry | Microsoft Learn (opens in a new tab)

Domain 5 | 10–15% of the exam | 9 min read

Implement information extraction solutions

Overview

Every point below comes from Microsoft's official documentation; each section lists its sources.

Build retrieval and grounding pipelines

Start here

THINK FIRST

How would you turn a pile of PDFs and images into search results an agent can ground its answers on?

Why it matters: Questions describe a grounding need and ask which indexer setting, skill, query type or agent connection fits it.

An indexer cracks each document, runs a skillset to enrich it and writes fields to an index. Queries then search that index by keywords, vectors or both. A knowledge base hands those results to agents.

Key terms

Term	Meaning
Skillset	The enrichment steps an indexer runs on cracked content, such as OCR, Text Merge, Text Split or embedding.
Hybrid query	One request with both `search` and `vectorQueries`, run in parallel and merged by Reciprocal Rank Fusion (RRF).
Semantic ranker	A reranking step after the first ranking that rescores the top results and returns `@search.rerankerScore`.
Knowledge base	The object that groups one or more knowledge sources and connects them to Foundry Agent Service.

Indexer pipeline order

Field mappings run before the skillset; output field mappings run after it. Document cracking puts file text in `/document/content` and images in `/document/normalized_images/*`. Skills read those paths. Anything a skill creates exists only after enrichment, so it reaches the index through an output field mapping.

The two kinds of indexer mapping compared

Feature	fieldMappings	outputFieldMappings
When it runs	Before the skillset	After the skillset
What it maps from	Fields in the data source	Fields made by cracking or enrichment
When you can skip it	Names and types already match the index	Never, if enriched output must reach the index

EXAM TRAP

Mapping an OCR or key phrase result with `fieldMappings` fails, because those fields don't exist yet when field mappings run.

Text from images with OCR

Set `imageAction` to `generateNormalizedImages` and add OCR plus Text Merge. Together they pull text out of images and merge it into the content field. That covers images inside a PDF and standalone image files such as JPGs. When you need embeddings for the images themselves, the recommended choice is the Foundry Tools Image Retrieval Vectorize Image API.

Chunking large content

Run Text Split before skills that take at most 50,000 characters. Key phrase extraction and entity recognition have that input limit. Text Split defaults to `pages` mode, which builds chunks of several sentences; `sentences` mode makes one-sentence chunks. A sound starting point is **512**-token chunks with **25%** overlap, which is **128** tokens. By default, `maximumPagesToTake` is **0**, meaning every chunk is kept.

EXAM TRAP

`pageOverlapLength` can't be half the maximum page length or more. It must stay below half.

Integrated vectorization

Built-in chunking and embedding needs an indexer, a data source and a skillset. You supply endpoints and connection details, for example for Azure OpenAI, and the service makes the calls. Built-in enrichment includes **20** free transactions per indexer. Beyond that, the skillset needs an attached Microsoft Foundry resource, billed at Standard pricing.

Keyword, vector or hybrid

A hybrid query runs full-text and vector search in parallel, then merges them. Keyword search suits product codes, jargon, dates and people's names, because it finds exact matches. To filter a vector query, add a filterable text or numeric field. The filter can run before or after the vector query.

The two halves of a hybrid query

Feature	Full-text query	Vector query
Ranking function	BM25	HNSW or exhaustive KNN (eKNN)

Feature	Full-text query	Vector query
Filterable field type	Text or numeric fields can be filterable	A vector field isn't filterable
Setting that widens or weights it	`maxTextRecallSize`, default **1,000**	Vector weight, default **1.0** (no weighting)

How RRF merges results

RRF scores each result as `1/(rank + k)`. Here rank is the result's position in each list and k is a constant. The algorithm performs best with a small k, such as **60**. With the default vector weight of **1.0**, the vector score goes into RRF as it is.

EXAM TRAP

The RRF constant k is not the vector query's `k`. When semantic ranker is used, set the vector query's `k` to **50** to give it the most input.

Semantic ranker on top

Semantic ranking runs after RRF and rescores only the top 50 results. Its `@search.rerankerScore` runs from **4** down to **0**, and higher means more relevant. That score is always reported separately. An empty search (`search=*`) isn't charged, even with `queryType` set to semantic.

EXAM TRAP

Semantic captions and answers are not written by a generative model. They are always verbatim text taken from your index.

Grounding agents

Agents reach an index through a knowledge source inside a knowledge base. Create an indexed knowledge source over a vector index, parent it to a knowledge base, then connect that to Foundry Agent Service. Every agentic retrieval scenario requires one knowledge base plus one or more knowledge sources. Knowledge sources over Blob Storage, ADLS Gen2 or OneLake can create the indexer, skillset and index for you.

EXAM TRAP

Agentic subqueries don't run one after another. They run at the same time, as keyword, vector or hybrid searches, and each one is semantically reranked.

Key takeaways

- Use output field mappings, not field mappings, for any field a skill creates.
- Text inside images needs `generateNormalizedImages` plus the OCR and Text Merge skills.
- Hybrid search merges BM25 and vector rankings with RRF; semantic ranker then rescores the top **50**.
- An agent gets search results through a knowledge base that holds at least one knowledge source.

Sources: Tutorial: Skillsets - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Hybrid Search Overview - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Vector Search Overview - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Hybrid Search Scoring (RRF) - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Semantic Ranking Overview - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Agentic Retrieval Overview - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab) · Chunk Documents - Foundry IQ (Azure AI Search) | Microsoft Learn (opens in a new tab)

Extract content from documents

Start here

THINK FIRST

How do you turn messy documents into clean Markdown or structured fields that an agent or RAG pipeline can trust?

Why it matters: Questions describe a document workload and ask which analyzer, extraction setting or field method gives the needed output.

Content Understanding reads a document with OCR and layout recognition, then returns it as Markdown or as JSON fields. You run it by naming an analyzer ID and a file. Opt-in confidence scores from **0** to **1** estimate how reliable each field value is.

Key terms

Term	Meaning
Analyzer	What you name by its ID, with the file, to run content extraction; a custom one sets a base in `baseAnalyzerId`.
Field schema	An optional list of fields to extract, each using the extract, classify or generate method.
Confidence score	An opt-in reliability estimate from 0 to 1 for each extracted field value.
Selection mark	A checkbox-style mark on a form, shown as ? when selected and - when cleared.
Span offset	The character position of extracted content, returned with a length in Unicode code points by default.

What content extraction does

Content extraction turns unstructured files into structured text and metadata. It runs OCR, finds selection marks and barcodes, detects formulas, and recognizes paragraphs, sections and tables. You don't need a field schema for this: the analyzer ID and the file are enough.

EXAM TRAP

Each Excel sheet and each PowerPoint slide maps to a page. A Word or HTML file, by contrast, is treated as one single page.

Choosing a prebuilt analyzer

Match the prebuilt analyzer to the file and the job. A scanned or image-based file calls for prebuilt-read or prebuilt-layout. In production, work from a copy of a prebuilt analyzer, because prebuilt definitions can change between API versions.

Three prebuilt document analyzers compared

Analyzer	What it does	Use it when
prebuilt-read	Basic OCR with no layout analysis	Plain text is enough; no language or embedding model needed
prebuilt-digitalParse	Reads the file's internal structure and encoding	The file is machine-readable, not a scan
prebuilt-documentSearch	Chunked output ready for embedding, plus figure descriptions	You ingest documents for RAG

Extraction mode for RAG

`extractionMode` **chooses layout-aware or faster text-only extraction.** Pick layout-aware extraction when tables, sections and reading order matter. Pick text-only extraction for clean Markdown in RAG ingestion. RAG analyzers return content with layout as Markdown and add semantic analysis to improve retrieval.

Reading the Markdown output

Output is GitHub Flavored Markdown that keeps content and layout. Tables use HTML table markup, so merged cells can span rows and columns. Page numbers, headers, footers and breaks become HTML comments. Selected marks appear as ? and cleared ones as -. `config.chartFormat` turns charts into Chart.js code blocks or Markdown tables.

Custom analyzers

A custom analyzer derives from one of four base analyzers. Name it in `baseAnalyzerId`: prebuilt-audio, prebuilt-document, prebuilt-image or prebuilt-video. Turn on segmentation into logical sections with `enableSegment`. For figures, set `enableFigureAnalysis` and `enableFigureDescription` to `true`.

Field extraction methods

Each field uses one method: extract, classify or generate. Define repeated items, such as invoice line items, as arrays of objects. Transcripts, OCR text and video key frames are already in the output, so add fields for them only when they need extra processing.

The three field extraction methods

Method	Use it for	Example
Extract	Values exactly as they appear; documents only	Dates from receipts
Classify	Choosing from predefined options	Document type or category
Generate	Values that need inference or summarization	Risk level or summary

EXAM TRAP

Dates and numbers come back normalized to a canonical format. That normalization can't be configured or switched off.

Key takeaways

- Use Markdown output for search and retrieval, and schema-based JSON for automation.
- Scanned files need prebuilt-read or prebuilt-layout, not prebuilt-digitalParse.
- Copy a prebuilt analyzer for production, since its definition can change across API versions.
- Model repeated items as arrays of objects, not as strings that ask for JSON.

Sources: What is Azure Content Understanding in Foundry Tools? - Foundry Tools | Microsoft Learn (opens in a new tab) · Build a retrieval-augmented generation solution with Azure Content Understanding in Foundry Tools - Foundry Tools | Micr (opens in a new tab) · Azure Content Understanding in Foundry Tools Document Overview - Foundry Tools | Microsoft Learn (opens in a new tab) · Azure Content Understanding in Foundry Tools prebuilt analyzers - Foundry Tools | Microsoft Learn (opens in a new tab) · Best practices for using Content Understanding - Foundry Tools | Microsoft Learn (opens in a new tab) · Document analysis: Markdown representation - Foundry Tools | Microsoft Learn (opens in a new tab) · Document Analysis: Extract Structured Content with Azure Content Understanding in Foundry Tools - Foundry Tools | Micros (opens in a new tab)

MockCert

Take a Microsoft Certified: Azure AI Apps and Agents Developer Associate mock exam the way the real one feels

5 timed mock exams · 60 questions in 120 minutes each · one is free

- Free to start: study notes, explained quizzes, flashcards and the first mock exam
- Timed mock exams built to the official exam guide
- A verifiable Certificate of Assessment when you pass a premium mock

Explore Microsoft Certified: Azure AI Apps and Agents Developer…



Open it on your phone

mockcert.com/microsoft-ai-103/

Scan the code with your phone's camera.