



AI-102 Domain 1: Plan and Manage Azure AI

AI-102 skills measured, domain 1 (20-25 %), as of 2026 | Choosing services and resources, Foundry projects, authentication, networking, quotas, monitoring, cost, containers, Content Safety and responsible AI

AI-102 Domain 1 Azure AI Foundry Content Safety Entra ID

☁ Choose the service and the resource

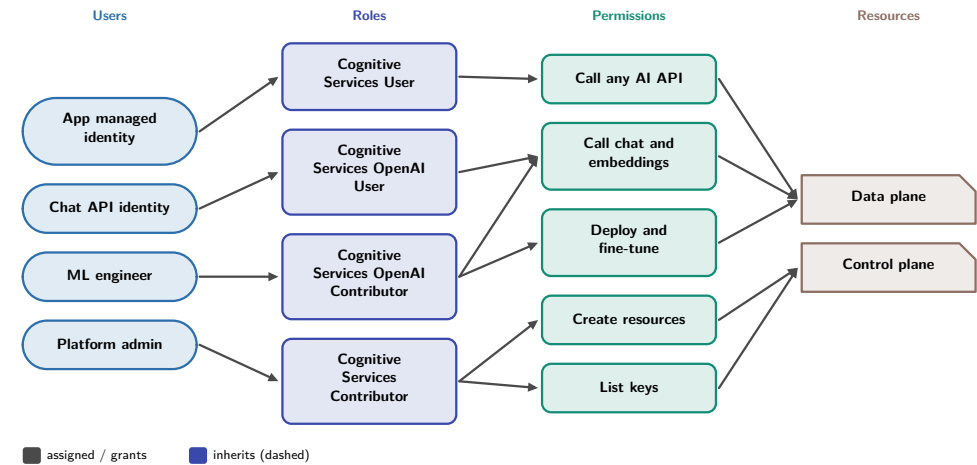
Need	Service	Resource kind / API
chat, reasoning, embeddings, images	Azure OpenAI or Foundry Models	AI Services (or OpenAI) deployment
agents with tools and memory	Foundry Agent Service	Foundry project on AI Services
caption, tag, OCR an image	Azure AI Vision	ComputerVision / AI Services
detect and verify faces	Azure AI Face (Limited Access)	Face
sentiment, PII, NER, CLU, Q&A	Azure AI Language	TextAnalytics / AI Services
translate text or documents	Azure AI Translator	TextTranslation / AI Services
speech-to-text, text-to-speech	Azure AI Speech	SpeechServices / AI Services
extract fields from forms	Document Intelligence	FormRecognizer / AI Services
multimodal extraction (docs, audio, video)	Azure AI Content Understanding	AI Services
full-text, vector, hybrid search	Azure AI Search	Microsoft.Search/searchServices
moderate text and images	Azure AI Content Safety	ContentSafety / AI Services

Multi-service or single-service resource?

Conditions	R1	R2	R3	R4
Several AI services in one app	Y	N	N	--
Separate billing or keys per service	N	Y	--	--
Need the free F0 tier	N	--	Y	--
Need agents or Foundry projects	--	--	--	Y
Actions				
AI Services (Foundry) resource	X			X
Single-service resource		X	X	
Notes	one endpoint, one key pair	cost split per resource	F0 exists only on single-service kinds	projects live under the Foundry resource

Foundry project (current default)	Hub-based project (older)
child of an AIServices resource (accounts/projects)	child of an Azure Machine Learning hub (workspace kind Hub)
models, agents, evaluations, files, connections	hub brings Storage, Key Vault, optional ACR and AI Search connections
RBAC: Azure AI User (renamed Foundry User in 2026) to build, Azure AI Project Manager (renamed Foundry Project Manager in 2026) to manage	needed for prompt flow and managed-compute open models (as of 2026, check)
SDK: azure-ai-projects AIProjectClient(endpoint, credential)	connection string instead of a project endpoint in older SDKs

🔑 Authenticate and authorise



Data-plane roles call APIs without keys; Contributor can list keys, so it is effectively a key holder

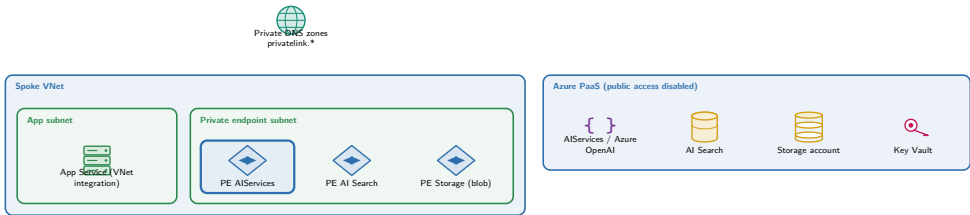
```
Keyless call with DefaultAzureCredential (Python)
from azure.identity import DefaultAzureCredential
from azure.ai.textanalytics import TextAnalyticsClient
cred = DefaultAzureCredential() # managed identity in Azure
client = TextAnalyticsClient(
    endpoint="https://ai102-foundry.cognitiveservices.azure.com/",
    credential=cred)
r = client.analyze_sentiment(["The demo went great"])[0]
print(r.sentiment, r.confidence_scores)
# Azure OpenAI: token scope
# https://cognitiveservices.azure.com/.default
```

Method	Header	Use it when
Resource key (AI services)	Ocp-Apim-Subscription-Key: <key>	quick test; store in Key Vault, rotate key1/key2
Region header (Translator, global endpoint)	Ocp-Apim-Subscription-Region: westeurope	multi-service or regional key on api.cognitive.microsofttranslator.com
Azure OpenAI key	api-key: <key>	legacy apps; disable with disableLocalAuth
Entra ID token	Authorization: Bearer <token>	production: managed identity, no secret to leak
Keys disabled	disableLocalAuth: true	policy requires identity-only access

 **EXAM TIP**

Rotate keys without downtime: switch the app to **key2**, regenerate **key1**, switch back, regenerate key2 (az cognitiveservices account keys regenerate --key-name key1). Better answer when offered: managed identity + Cognitive Services User role and disableLocalAuth.

Network isolation and deployment



Private endpoints + private DNS zones (cognitiveservices, openai, services.ai, search.windows.net, blob)

```
main.bicep: locked-down Foundry resource and a GPT deployment
resource ai 'Microsoft.CognitiveServices/accounts@2025-06-01' = {
  name: 'ai102-foundry'
  location: 'swedencentral'
  kind: 'AIService'
  sku: { name: 'S0' }
  identity: { type: 'SystemAssigned' }
  properties: {
    customSubDomainName: 'ai102-foundry'
    publicNetworkAccess: 'Disabled'
    disableLocalAuth: true
  }
}
```

```
allowProjectManagement: true
networkAcls: { defaultAction: 'Deny' }
}
}
resource gpt 'Microsoft.CognitiveServices/accounts/deployments@2025-06-01' = {
  parent: ai
  name: 'chat'
  sku: { name: 'GlobalStandard', capacity: 50 } // 50K TPM
  properties: { model: { format: 'OpenAI', name: 'gpt-4.1-mini',
    version: '2025-04-14' } }
}
```

- **Firewall rules** networkAcls.defaultAction: Deny + ipRules and virtualNetworkRules (service endpoint Microsoft.CognitiveServices); bypass: AzureServices lets trusted services such as AI Search reach it
- **Private endpoint** DNS must resolve <name>.cognitiveservices.azure.com to the private IP: link the privatelink zone to the VNet, or calls fail with 403 from the public path
- **Managed identity outbound** AI Search indexers and Foundry agents reach Storage and AI services with the resource's identity: grant it Storage Blob Data Reader, Cognitive Services User
- **CI/CD** Bicep or Terraform for resources and deployments, GitHub Actions or Azure Pipelines with OIDC federated credentials; evaluation gate before swapping a model version

Quotas, monitoring and cost

Y

A GlobalStandard deployment has 100K TPM. Each request sends about 1,500 prompt tokens with max_tokens: 500. How many requests per minute before HTTP 429?

Given: TPM = 100,000 tokens per minute; RPM = 6 per 1,000 TPM for most chat models; Rate limiter counts prompt tokens + max_tokens per request

Y

Step 1: What is the request cap?

$RPM = 100,000 / 1,000 \times 6 = 600$ requests/min
derived from the TPM you assign

T

Y

Step 2: How many tokens does one request reserve?

$1,500 + 500 = 2,000$ tokens
the estimate uses max_tokens, not the real answer length

T

Y

Step 3: Token-limited rate?

$100,000 / 2,000 = 50$ requests/min
the smaller of the two limits wins

T

Y

Step 4: Lower max_tokens to 250?

$100,000 / 1,750 = 57$ requests/min
a tight max_tokens raises throughput

T

About 50 requests per minute: the TPM limit binds long before the 600 RPM cap

T

Check: On 429 read the retry-after-ms header, retry with exponential backoff, raise quota or add a second region

Signal	Where	Alert on
Total calls, errors, latency	Azure Monitor metrics of the resource	server errors > 0, latency p95
Processed prompt / generated completion tokens	Azure OpenAI metrics	token burn vs budget
Provisioned-managed utilisation	PTU deployments	> 90 % (429s start at 100 %)
Audit, RequestResponse, Trace logs	diagnostic setting → Log Analytics	auth failures, throttling (KQL)
Spend per resource and tag	Cost Management + budgets	80 % and 100 % of budget
Traces of prompts, tools, tokens	Foundry tracing → Application Insights (OpenTelemetry)	failed runs, slow tools

KQL: throttled calls per hour

AzureDiagnostics

| where ResourceProvider == "MICROSOFT.COGNITIVESERVICES"

| where ResultSignature == "429"

| summarize throttled = count() by bin(TimeGenerated, 1h), Resource

Pricing tiers

F0 free (one per service per subscription, low limits), S0 standard pay-per-call; commitment tiers prepay monthly volumes at a discount

Azure OpenAI

Standard / Global / Data Zone pay per token; Provisioned (PTU) per hour with monthly or yearly reservations; Batch about half the Standard price for 24 h jobs

Cost levers

smaller model (mini/nano), shorter prompts, prompt caching, Batch for offline work, budgets with action groups

Containers and edge

Run the Language container (billing still goes to Azure)

```
docker run --rm -it -p 5000:5000 --memory 8g --cpus 1 \
  mcr.microsoft.com/azure-cognitive-services/textanalytics/language:latest \
  Eula=accept \
  Billing=https://ai102-foundry.cognitiveservices.azure.com/ \
  ApiKey=<key>
# then call http://localhost:5000/text/analytics/v3.1/sentiment
```

- **Three required settings** Eula=accept, Billing (endpoint URI), ApiKey; the container refuses to answer without them
- **What leaves the site** only usage metering to the billing endpoint; text, images and audio stay local
- **Available containers** Language (sentiment, key phrases, NER, PII, language detection, CLU), Speech-to-text and text-to-speech, Translator, Vision Read OCR, Document Intelligence (layout, prebuilt)
- **Disconnected** approval form + commitment tier; licence file downloaded once; for air-gapped sites
- **Hosts** Docker, AKS, Azure Container Instances, Azure Container Apps, IoT Edge

Responsible AI and Content Safety

Mnemonic: **FRPITA**

F

Fairness similar people get similar outcomes; test per group

R

Reliability and safety behaves as designed, fails safely

P

Privacy and security data protected, PII handled

I

Inclusiveness usable by people of all abilities

T

Transparency users know it is AI and its limits

A

Accountability people answer for the system

Content Safety feature	Detects	Output / setting
Text and image analysis	hate, sexual, violence, self-harm	severity 0, 2, 4, 6 (8 levels with outputType: EightSeverityLevels for text)
Content filter on a deployment	same four categories on prompt and completion	block threshold low / medium (default) / high
Prompt Shields	user jailbreak attacks and indirect attacks in documents	attackDetected true/false
Groundedness detection	answer claims not supported by the sources	ungrounded percentage, optional correction
Protected material	known song lyrics, articles, public code	detected + citation for code
Blocklists and custom categories	your own terms or trained category	exact-match block, custom severity

⚠ HTTP 401 or 403 calling the AI resource

HIGHCOMMON

LIKELY CAUSES

1. Key-based call while disableLocalAuth is true

2. Identity lacks Cognitive Services User (or OpenAI User)

3. Regional endpoint used instead of custom subdomain for Entra auth

4. Public network access disabled, call not via private endpoint

HOW TO CONFIRM

Show auth settings and role assignments

```
az role assignment list --scope <ai-resource-id> -o table
```

FIX

✓ Use DefaultAzureCredential and assign the data-plane role

✓ Call https://<name>.cognitiveservices.azure.com

✓ Fix private DNS resolution

⚠ HTTP 429 Too Many Requests

MEDIUMCOMMON

LIKELY CAUSES

1. Deployment TPM or RPM exhausted

2. max_tokens set far above real answers

3. PTU utilisation at 100 %

HOW TO CONFIRM

Read retry-after-ms and the token metrics

```
az cognitiveservices account deployment show -g rg -n ai --deployment-name chat
```

FIX

✓ Backoff and retry

✓ Lower max_tokens

✓ Raise quota, spill over to another region or a Global deployment

MYTH

✗ The default content filter can be switched off by anyone

→

FACT

✓ Lowering or disabling filters needs an approved modified-content-filter request

You can make filters stricter freely; looser needs Microsoft approval (as of 2026).

✗ A private endpoint alone blocks internet access

→

✓ Public network access must also be disabled

Otherwise the public endpoint keeps answering keys from anywhere.

✗ Containers run fully offline by default

→

✓ Standard containers need the billing endpoint

Only approved disconnected containers run air-gapped.

✗ Contributor is a safe role for apps

→

✓ Contributor can list keys

Give apps the data-plane User role only.

🚩 Recap and self-test

Key takeaways

01

AI Services resource

one endpoint for OpenAI, Speech, Language, Vision, agents

02

Identity over keys

managed identity, data-plane role, disableLocalAuth

03

Private endpoint + DNS + public access off

the isolation trio

04

TPM binds with max_tokens

429 → backoff, quota, region spill-over

05

Content Safety

four harm categories, Prompt Shields, groundedness

Q1

Which header carries a key for Azure AI Language? And for Azure OpenAI?

A

Ocp-Apim-Subscription-Key for AI services; api-key for Azure OpenAI.

Q2

Least-privilege role for an app that only calls Azure OpenAI chat?

A

Cognitive Services OpenAI User on the resource.

Q3

Which property forbids key authentication?

A

disableLocalAuth: true on the Cognitive Services account.

Q4

Three mandatory container settings?

A

Eula=accept, Billing=<endpoint>, ApiKey=<key>.

Q5

Which private DNS zone serves Azure OpenAI endpoints?

A

privatelink.openai.azure.com (plus privatelink.cognitiveservices.azure.com for the other services).

Q6

Default content-filter threshold on a new deployment?

A

Medium: blocks medium and high severity in all four harm categories.

Q7

Which Content Safety feature finds instructions hidden in a retrieved web page?

A

Prompt Shields, indirect (document) attack detection.

🌐 Tadata | tadata.fr | contact@tadata.codes

Q8 Where do you send the RequestResponse logs to query them with KQL?

A A Log Analytics workspace through a diagnostic setting.

Q9 Name the six responsible AI principles.

A Fairness, reliability and safety, privacy and security, inclusiveness, transparency, accountability.