



AI-901: Responsible AI and Generative Models

AI-901 domain 1, Identify AI concepts and capabilities (40-45 %), skills measured as of April 15, 2026 | Six responsible AI principles, how generative models work, choosing a model, deployment options and configuration parameters

AI-901 Domain 1 Responsible AI LLM Foundry Models

✓ The six principles of responsible AI

Fairness

treat all people equitably; no group gets worse outcomes
e.g. a loan model approves fewer women with equal incomes → rebalance data, measure per group

Reliability and safety

perform as intended, also on unusual input; fail safely
e.g. test edge cases, content filters, human review for medical advice

Privacy and security

protect personal data and resist attacks
e.g. no PII in prompts or logs, Entra ID instead of keys, private endpoints

Inclusiveness

empower everyone, including people with disabilities
e.g. speech input, alt-text, several languages and accents

Transparency

people understand that AI is used, how and its limits
e.g. label AI-generated content, publish limitations, show citations

Accountability

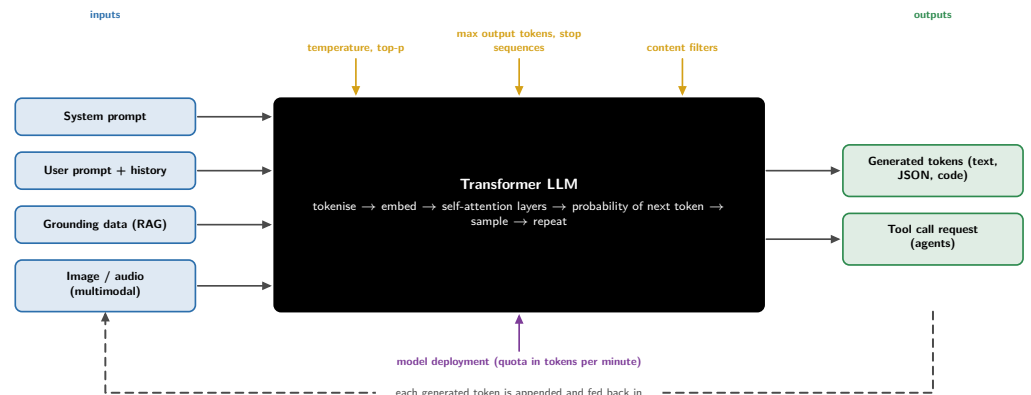
people stay answerable for AI systems
e.g. named owner, governance board, audit logs, human approval for actions

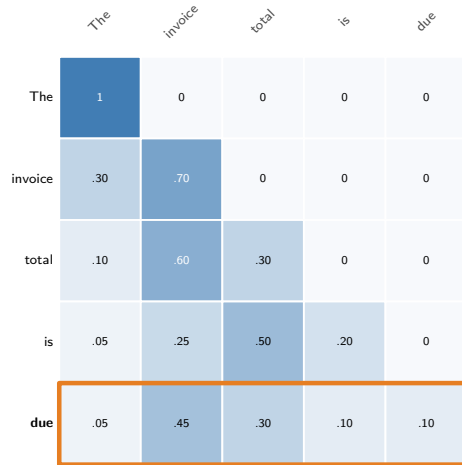
Scenario wording	Principle	Why
model error rate higher for one accent	Fairness (also inclusive-ness)	unequal quality of service across groups
app must work for blind users	Inclusiveness	design for people with disabilities
users must know they talk to a bot	Transparency	disclose AI use
explain why an application was rejected	Transparency	understandable decisions
chat history contains customer addresses	Privacy and security	personal data protection
self-driving shuttle meets snow for the first time	Reliability and safety	behaviour under unexpected conditions
who answers when the agent refunds the wrong client	Accountability	humans own outcomes
agent may email customers only after approval	Accountability (+ safety)	human oversight of actions

📄 EXAM TIP

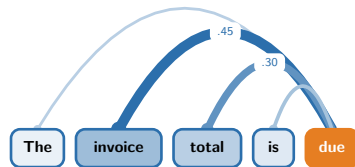
Two pairs confuse candidates: **fairness** (equal outcomes between groups) vs **inclusiveness** (no-body excluded from using it); **transparency** (people understand the system) vs **accountability** (people are responsible for it). In Foundry, the principles become concrete tools: content filters and guardrails (reliability and safety), evaluations (fairness, groundedness), Entra ID and private networking (privacy and security), transparency notes and citations (transparency).

🔧 How generative models work





rows: query token · columns: key token it attends to



attention from "due" (self .10 · arc width ∝ weight)

Self-attention with causal mask: 'due' attends mostly to 'invoice' (what is due); illustrative weights

- **Training** pre-training predicts the next token on huge text corpora; instruction tuning and RLHF (reinforcement learning from human feedback) make it follow instructions and refuse harm
- **Inference** the model only predicts likely tokens: it does not look facts up unless you give it grounding data or a search tool
- **Tokens** approx. 4 English characters or 0.75 word per token; input and output tokens are billed separately, output usually costs more
- **Embeddings** separate embedding models (text-embedding-3-small / -large) turn text into vectors for semantic search
- **Knowledge cutoff** a model knows nothing after its training date: use RAG or web grounding for fresh facts

© Choose a model by capability

Need	Model family (Foundry catalogue, as of 2026)	Why
general chat, summaries, cheap volume	GPT-4.1-mini / GPT-4o-mini, GPT-5-mini, Phi (SLM)	fast, low price per token
hard multi-step reasoning, code	o-series (o3, o4-mini), GPT-5	reasoning tokens before answering
read images, charts, screen-shots	GPT-4o, GPT-4.1 (multimodal)	image input in the prompt
spoken conversation, audio in/out	GPT-4o realtime / audio models; Azure Speech	low-latency speech
create images	gpt-image-1, FLUX (Black Forest Labs)	text → image, edits
create short videos	Sora	text → video
semantic search, RAG	text-embedding-3-small / -large	vectors, not text
open weights, other vendors	Llama, Mistral, DeepSeek, Grok, Cohere	choice, price, licence

Option	Image input	Cost per token	Latency	Reasoning depth	Weighted total (/5)
	38 %	25 %	25 %	12 %	
GPT-4.1-mini multimodal, small	5	5	5	2	4.62
GPT-4.1 multimodal, large	5	3	3	4	3.88
o4-mini reasoning	4	4	2	5	3.62
Phi-4-mini SLM, text only	1	5	5	2	3.12

GPT-4.1-mini leads GPT-4.1 by 0.75 (15 % of the scale)

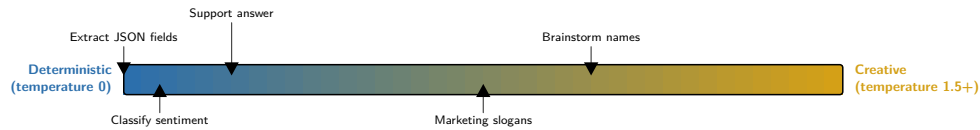
Choosing a model for a receipt-photo chat assistant: image input dominates (illustrative scores 1-5)

💡 Compare before you pick

Foundry's model catalogue shows model cards, the **leaderboards** (quality, cost, throughput benchmarks) and a side-by-side compare. The **model router** deployment picks a model per prompt automatically to trade cost against quality.

⚙️ Deployment options and configuration parameters

Deployment type	Billing	Where data is processed	Choose when
Global Standard	per token	any Azure region	default; highest quota; no residency need
Data Zone Standard	per token	inside EU or US data zone	EU data processing requirement
Standard (regional)	per token	the resource's region	strict regional residency
Provisioned (global / data zone / regional)	per PTU per hour (reservations possible)	as named	steady high volume, predictable latency
Global / Data Zone Batch	per token, about 50 % less	as named	offline jobs, results within 24 h
Serverless API / managed compute	per token / per VM hour	region of deployment	partner and open-weight models



Set temperature by task; change temperature or top-p, not both

Parameter (playground / API)	Effect	Typical value
temperature	randomness of token choice	0-0.3 facts, 0.7-1 creative
top_p	sample from the top cumulative probability p	1 (leave) or 0.9
max output tokens (max_tokens, max_output_tokens)	caps answer length and cost; answer cut when reached	300-1,000
stop sequences	stop generating at a string	"###"
frequency / presence penalty	discourage repeated tokens / topics	0-1
past messages included	how much chat history is resent	10 messages
reasoning effort (reasoning models)	low / medium / high thinking budget; temperature not supported	medium

PROBLEM A chat app gets $R = 20,000$ requests per month, each with $T_i = 1,500$ input and $T_o = 250$ output tokens. Prices (illustrative, not a quote): $P_i = 0.40$ USD and $P_o = 1.60$ USD per million tokens. Monthly cost?

- GIVEN**
- $R = 20,000$ requests/month
 - $T_i = 1,500$, $T_o = 250$ tokens
 - $P_i = 0.40$, $P_o = 1.60$ USD per 1M tokens
 - $C = R \times (T_i \times P_i + T_o \times P_o) / 10^6$

1 Input cost

$20,000 \times 1,500 \times 0.40 / 10^6 = 12.00$ USD

Why: 30 M input tokens

2 Output cost

$20,000 \times 250 \times 1.60 / 10^6 = 8.00$ USD

Why: 5 M output tokens at 4x the price

3 Total

$C = 12 + 8 = 20$ USD per month

ANSWER About 20 USD per month; 60 % comes from input tokens (system prompt + history)

CHECK ☒ Cutting history from 1,500 to 750 input tokens saves 6 USD (30 %)

⚠ Quota

Each deployment gets a **tokens-per-minute (TPM)** quota (and derived requests per minute) per subscription, region and model. Exceeding it returns **HTTP 429 Too Many Requests**: retry with backoff, raise the deployment's TPM, or use Global Standard for more capacity.

🚩 Self-test and recap

Statement	True	False
1. An LLM looks facts up in a database when answering → It predicts tokens; facts come from training or from grounding you supply	☐	☒
2. Temperature 0 makes outputs more repeatable → the most likely token is chosen almost always	☒	☐
3. Data Zone Standard keeps processing in the EU or US zone → Global Standard may process anywhere	☒	☐
4. Making an app usable by deaf users is mainly fairness → it is inclusiveness	☐	☒
5. Provisioned deployments cost nothing when idle → PTUs are billed per hour	☐	☒
6. Reasoning models accept a reasoning effort setting → low / medium / high	☒	☐

Q1 Name the six Microsoft responsible AI principles.

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

Q2 Disclosing that a reply was written by AI serves which principle?

A Transparency.

Q3 What does self-attention let a transformer do?

A Weigh every earlier token when predicting the next one, so meaning depends on context.

Q4 Which parameter caps answer length and cost?

A Max output tokens (max_tokens / max_output_tokens).

Q5 Which deployment type for an offline job of 1 million prompts at lowest price?

A Global (or Data Zone) Batch: about 50 % cheaper, results within 24 hours.

Q6 Model type for an app that reads photos of receipts?

A A multimodal model such as GPT-4o / GPT-4.1 (image input).

Q7 What is an SLM good for?

A Cheap, fast narrow tasks or edge devices (for example Phi models).

Q8 HTTP 429 from a deployment means?

A Rate limit or TPM quota exceeded: back off and retry or raise quota.

Q9 What reduces hallucinations most at design time?

A Grounding the model in your data (RAG) plus low temperature and evaluation.

TOPIC

Domain 1, part 1

HEADLINE 1

Six principles, two confusable pairs

Why: fairness vs inclusiveness, transparency vs accountability

HEADLINE 2

LLMs predict tokens

Why: ground them for facts

HEADLINE 3

Pick the model by capability

Why: multimodal, reasoning, SLM, image, embedding

HEADLINE 4

Deployment type = billing + data location

Why: Global, Data Zone, regional, provisioned, batch

HEADLINE 5

Temperature low for facts

Why: max tokens caps cost