

Responsible AI changes the application architecture

Reference application: a customer-support assistant that accepts text, speech, images, and documents.

Controls surround the complete application

- Inputs and customer data
- Models, prompts, and retrieved information
- Generated responses and agent actions
- Monitoring, escalation, and human review

Key message

The six principles become engineering and operational requirements.

Protect customer data

Privacy and security controls limit what the assistant collects, stores, and reveals.

Example · “A customer uploads an invoice containing an account number and payment details. Those details must not appear in another user’s session or in unnecessary diagnostic logs.”

Privacy and security controls

- Collect only the data required for the support task
- Authenticate users before retrieving account information
- Restrict data access by user and application role
- Remove sensitive information from logs
- Protect credentials used by models and connected tools
- Define retention and deletion policies

Remember

Authentication confirms identity. Authorization determines which information that identity may access.

Test performance across user groups

Fairness evaluation measures more than average accuracy.

Example · “Speech recognition works well overall but frequently misinterprets one regional accent.”

Test consistently across

- Languages and dialects
- Accents and speech patterns
- Assistive technology users
- Different document and image formats
- Customer groups represented in the service population

Measure more than average accuracy

- Task completion rate
- Error and refusal rate
- Escalation rate
- Response quality by cohort

Remember

A strong overall result can conceal serious failures for a smaller group.

Disclose AI involvement and limitations

Transparency controls help users judge when to rely on the assistant.

Example disclosure

“This AI assistant can answer support questions and summarize your account information. Verify important details before acting, or request a support representative.”

Tell users

- They are interacting with an AI system
- Which tasks the assistant can perform
- When answers may be incomplete or incorrect
- When the assistant uses customer data or external sources
- How to reach a human representative

Remember

Avoid presenting generated answers as verified human decisions.

Provide accessible interaction modes

User need	Interaction option
Cannot type easily	Speech input
Cannot hear audio	Text transcript and captions
Cannot see an image	Text description
Uses assistive technology	Keyboard and screen-reader support
Needs language assistance	Supported translation or human escalation

Exam cue

Alternative access methods most directly support inclusiveness.

Design predictable service degradation

Failure condition

Safe application behavior

Model endpoint unavailable

Display a clear status and offer human support

Retrieved evidence is missing

Avoid unsupported claims

Confidence is too low

Ask for clarification or escalate

Tool action fails

Report the failure without claiming success

Unsafe request detected

Refuse safely and explain available options

Key term

Graceful degradation means preserving safe, limited service when full service becomes unavailable.

Monitor failures after deployment

Operational monitoring must lead to defined action, not only accumulated logs.

Response loop · “Detect failure → investigate cause → correct the system → validate the correction”

Track

- Incorrect, harmful, or unsupported responses
- Failed retrievals and tool calls
- Refusals and human escalations
- Latency and availability
- Performance differences between user groups
- Repeated user corrections and complaints

Remember

Predeployment testing cannot represent every real-world interaction.

Establish human ownership and escalation

Responsibility	Example owner
Model and prompt behavior	AI application team
Customer-data controls	Security and privacy teams
Support policies	Customer-support leadership
Production incidents	Operations team
High-impact customer decisions	Authorized human reviewer

Escalate when

1. The customer requests a person
2. The assistant lacks reliable evidence
3. A request involves sensitive or exceptional circumstances
4. The system detects potential harm or policy violations
5. An action exceeds the assistant's authority

Accountability

Assign named owners before deployment. Reviewers need authority, information, and time to intervene.

Responsible AI control check

The support assistant	Primary principle
Prevents customer documents from entering diagnostic logs.	Privacy and security
Compares speech-recognition success across accents.	Fairness
Identifies itself as an AI assistant.	Transparency
Provides captions and keyboard navigation.	Inclusiveness
Escalates when retrieved evidence is unavailable.	Reliability and safety
Records failures and alerts the operations team.	Reliability and safety; accountability
Assigns refund approval to an authorized employee.	Accountability

Exam strategy

One control can support multiple principles. Choose the principle most directly addressed.