

Deployment and Monitoring Checklist

Release readiness

- ☐ The approved model, prompt, data and code versions are recorded.
- ☐ Automated tests and evaluation thresholds pass.
- ☐ The deployment environment matches the tested environment.
- ☐ Secrets, authentication and permissions are protected.
- ☐ A rollback or safe shutdown path is tested.

Serving and performance

- ☐ Input validation and output schema checks exist.
- ☐ Latency, throughput, availability and cost targets are defined.
- ☐ Timeouts, retries and error handling are tested.
- ☐ Logs avoid unnecessary personal or confidential data.

Model and data monitoring

- ☐ Input quality and Data Drift signals are monitored.
- ☐ Prediction distribution and failure rates are monitored.
- ☐ Ground truth or human feedback is collected where possible.
- ☐ Performance is reviewed across important groups and conditions.

Generative Ai controls

- ☐ Prompt versions and model routes are recorded.
- ☐ Retrieval quality and groundedness are evaluated.
- ☐ Hallucination and unsafe-output signals are monitored.
- ☐ Tool actions use restricted permissions and approval gates.

Ownership and response

- ☐ Alert thresholds and responsible people are named.
- ☐ Incident, escalation and user-communication procedures exist.
- ☐ Retraining or update triggers are evidence-based.
- ☐ Scheduled review date is recorded.