@genieincodebottle

Instagram | GitHub | Medium | YouTube

GenAI in Production - Enterprise Ready Checklist

Product & UX

1. **Latency vs Accuracy** - Users won’t wait long for responses, balance speed & quality using prompt optimization, smaller models & caching while still meeting task requirements (SLA/SLO).

2. **Context Management** - Effective RAG depends on chunking, embedding quality, metadata filters, re-ranking & hybrid (sparse + dense) retrieval to deliver relevant context.

3. **Prompt Drift** - Prompts that perform well in dev/test may fail in production, continuously track prompts & responses to catch degradation early.

4. **User Behaviour** - Real world queries are longer, noisier & less structured than test data, build adaptive systems that can handle messy input.

5. F**allback Strategy** - Design graceful degradation with cached results, smaller/faster models or deterministic rule based responses when the primary LLM or retrieval fails.

6. **Conversation Memory** - Maintain dialogue continuity with summarisation or compression to control token costs & safeguard against memory poisoning from malicious or irrelevant content.

7. **Accessibility & Streaming U**X - Improve usability with streaming outputs, partial answers, retries, user interrupt options & screen reader friendly interfaces.

8. **Human-in-the-Loop Escalation** - In high-stakes domains (e.g. finance, healthcare), allow handoff to humans with full context, citations & audit trails for accountability.

Technical & Engineering

9. **Embedding Drift** - Updating embedding models changes vector space, always version embeddings & store both raw text + vectors for re-encoding if needed.

10. **Chunk Size Sensitivit**y - Optimal chunk size varies by domain, too small loses context, too large inflates cost & retrieval noise. Needs real world tuning.

11. **Vector DB Pitfalls** - ANN search libraries (FAISS, Milvus, Pinecone, etc.) have recall/latency trade-offs, misconfiguration leads to missed or irrelevant results.

12. **Hybrid Search + Rerankers** - Combining dense embeddings with sparse (keyword/BM25) & re-ranking (cross-encoder, FlashRank) improves factuality & precision.

13. **Tokenization Surprises** - Emojis, punctuation, or whitespace can inflate tokens unexpectedly, breaking context windows & increasing cost.

14. **Async Orchestration** - When chaining RAG, tools & APIs, use retries, timeouts, correlation IDs & idempotency to handle race conditions.

15. **Agent Loop Guardrails** - Agents can get stuck in infinite loops, enforce max steps, iteration limits & cost ceilings.

16. **Cold Starts** - Serverless platforms introduce latency, mitigate with provisioned concurrency, warm pools or containerisation.

17. **Structured Output** - Enforce JSON schema or constrained decoding for predictable responses, use auto-repair & validators to handle malformed output.

18. **Model Versioning & Rollback** - Model updates change outputs, deploy with canary/shadow testing & use feature flags for instant rollback.

19. **Multi-Vendor Abstraction** - Build a model/router layer to switch providers seamlessly, avoid vendor lock-in & enable failover

20. **Batching & Multiplexing** - Batch embedding generations & multiplex queries to reduce cost & improve throughput without hurting tail latency.

Security, Privacy & Compliance

21. **Prompt Injection Defense** - Scrub & sanitize both user inputs & retrieved documents to block jailbreaks or malicious instructions.

22. **PII & Compliance** - Run DLP scans, redact/mask sensitive info & enforce GDPR/HIPAA/SOC2 standards from day one.

23. **Auditability & Provenance** - Log prompts, retrieved docs, model configs & responses for traceability in audits.

24. **Data Residency & Isolation** - Enforce tenant-level isolation with RLS, geo-fencing & data sovereignty controls.

25. **Continuous Red Teaming** - Run regular adversarial tests, security isn't a one-off pen test.

26. **Source Trust Scores** - Maintain whitelists/blacklists & assign trust levels to data sources to avoid ingestion of malicious content.

27. **Cache Poisoning Protection** - Continuously monitor caches/embeddings, invalidate stale or manipulated entries quickly.

Evaluation & Observability

28. **Evaluation Beyond Metrics** - Numbers (BLEU, ROUGE etc) aren't enough, combine domain-specific eval sets & human-in-the-loop scoring.

29. **Eval-as-Code in CI/CD** - Automate eval pipelines with golden sets, regression checks & drift alerts in PR builds.

30. **Bias/Fairness Checks** - Audit outputs across languages, demographics & modalities to avoid hidden discrimination.

31. **Monitoring SLIs/SLOs** - Track latency (p50, p95), retrieval hit rate, grounding % & tool call success in prod.

32. **Logging Strategy** - Capture selective logs, redact PII & protect sensitive metadata while ensuring enough detail for debugging.

33. **Reproducibility** - Control seeds, temperature, decoding parameters etc to replicate results across environments.

34. **Business Goal Alignment** - Continuously refresh eval criteria as business objectives evolve (conversion, trust, engagement).

Scalability & Ops

35. **A/B Testing** – Test prompts, retrieval configs or model choices side-by-side in prod before global rollout.

36. **Cost Guardrails** - Monitor token usage, set quotas, anomaly alerts & prevent runaway costs.

37. **Rate-Limiting Layers** - Apply limits across LLM APIs, RAG retrievers, DB calls, Agents call & external integrations.

38. **Infra Throttling & Autoscaling** - Use circuit breakers, scale horizontally & fail gracefully during traffic spikes.

39. **Data Freshness** - Re-ingest, re-embed & re-index regularly, stale knowledge is worse than no knowledge.

40. **Operational Playbooks** - Maintain on-call runbooks, dashboards & kill-switches for emergency rollback.