Prompt LOGIN

■ 2025 LLM Integration Checklist

Premium Edition – Brought to you by PromptLogin

Foundation Setup

- ■ Select your LLM provider (ChatGPT, Claude, Gemini, DeepSeek, or open-source).
- ■ Configure API authentication securely.
- ■ Standardize prompt formats for consistency.
- ■ Add basic error-handling (timeouts, retries).
- ■ Test sample workflows with real queries.

RAG (Retrieval-Augmented Generation)

- ■ Choose and deploy a vector database (Pinecone, Weaviate, Qdrant, pgvector).
- ■ Create embedding pipeline (document → chunk → embed → store).
- ■ Validate retrieval accuracy with test queries.
- ■ Connect retriever → LLM for grounded responses.
- ■ Monitor for hallucinations and tune retrieval size.

Performance & Reliability

- ■ Implement rate-limit handling (exponential backoff).
- ■ Enable token streaming for faster response times.
- ■ Add circuit breakers and multi-vendor fallback.
- ■ Cache frequent queries and embeddings.
- ■ Track latency and uptime in dashboards.

Compliance & Security (2025 Focus)

- ■ Apply PII redaction and data minimization.
- ■ Maintain audit logs of all interactions.
- ■ Align with EU AI Act (Aug 2, 2025 obligations).
- ■ Review GDPR/HIPAA/SOC-2 requirements.
- ■ Add human-in-the-loop for sensitive decisions.

Cost & ROI Management

- ■ Monitor token usage per feature.
- ■ Forecast monthly costs (API + vector DB + infra).
- ■ Use smaller models for lightweight tasks.
- ■ Negotiate enterprise pricing or volume discounts.
- ■ Track ROI (CSAT, resolution time, revenue impact).

Implementation Roadmap (180 Days)

- ■ Phase 1 (0–2 months): Basic API + prototype
- ■ Phase 2 (3–4 months): Add RAG + monitoring
- ■ Phase 3 (5–6 months): Real-time + agentic workflows

■ How to Use This Checklist

1. Print or save as PDF — tick items as you progress.
2. Prioritize compliance — EU AI Act obligations start Aug 2025.
3. Iterate fast — launch pilots, measure results, then scale.

© 2025 PromptLogin – All Rights Reserved