AI Agents & LLM Selection Guide (Improved Version)

This improved edition enhances clarity, structure, and flow while adding a complete section on Open■Source LLMs and Agent Frameworks.

1. Introduction

AI agents combine models, tools, orchestration, and deployment to autonomously complete tasks using the Think■Act■Observe loop.

2. Agent Levels (0–4)

Level 0: Pure reasoning

Level 1: Tool■connected reasoning

Level 2: Multi■step strategic planning

Level 3: Multi■agent collaboration

Level 4: Self■evolving agents

3. LLM Models Guide (Expanded with Open■Source)

Proprietary Models:

• Gemini 2.5 Pro — heavy reasoning

• Gemini 2.5 Flash — fast + cost■efficient

• Gemini Live — multimodal

Open■Source Models (New):

• Llama 3 / Llama 3.1 — strong reasoning, fully local deployability

• Mistral & Mixtral — high efficiency, excellent throughput

• Qwen 2 — multilingual strength, strong structured reasoning

• Phi■3 — lightweight, great for edge devices

Open■Source Advantages:

• Full data privacy

• Local deployment

• Lower cost at scale

• Customization (fine■tuning, RAG, prompt■routing control)

4. Model Routing

Use fast models for simple tasks, strong models for complex planning.

Hybrid routing example:

• Flash → intent detection

• Llama 3 → SQL generation

• Pro model → strategic reasoning

5. Tools (RAG, SQL, APIs, Code Exec)

RAG: vector DBs (FAISS, Chroma, Weaviate)

NL2SQL: text■to■SQL tools

APIs: external system interaction

Code Execution: secure sandbox for calculations

6. Orchestration Layer (Improved Section)

Use patterns like:

• ReAct

• Plan■Execute

• Coordinator■Worker agents

7. Open■Source Agent Frameworks (New Section)

• LangChain

• LlamaIndex

• CrewAI

• AutoGen

• Haystack

Benefits:

Flexibility, full transparency, customizable routing, OSS ecosystem.

8. Deployment Strategies

Proprietary: Vertex AI Agent Engine

Open■Source Deployment (New):

• Ollama

• vLLM

• LM Studio

• Docker self■hosting

• Kubernetes (GKE/EKS)

9. Best Practices

• Use long■term memory only when needed

• Add human■in■the■loop for high■risk tasks

• Evaluate using LM■as■Judge + test suites

• Use observability: logs, traces, error reports

10. Key Takeaways

• Start simple (Level 1) → evolve

• Use OSS models for cost/privacy

• Use routing for efficiency

• Build strong evaluation + monitoring early