AI Agents vs RAG: Which Architecture Wins for Enterprise in 2026?

AI Pinnacle — aipinnacle.pk

The "agents vs RAG" debate is mostly a false binary. In 2026 the winning enterprise architecture is hybrid — RAG for retrieval, agents for orchestration.

When Pure RAG Wins

- Single-document Q&A; (legal, policy, knowledge base)
- Sub-second latency requirements
- Deterministic citations are non-negotiable (regulated industries)

When Agents Win

- Multi-step workflows that cross 3+ systems
- Tasks where the plan is dynamic
- Long-horizon work

Our Reference Architecture

1. LangGraph or OpenAI Agents SDK as orchestrator
2. pgvector RAG layer
3. Strict tool budget per task
4. Langfuse for trace observability
5. Human-in-the-loop checkpoint for irreversible actions

Agents cost 4–8x more per task than RAG. Only deploy them where the workflow value justifies the spend.