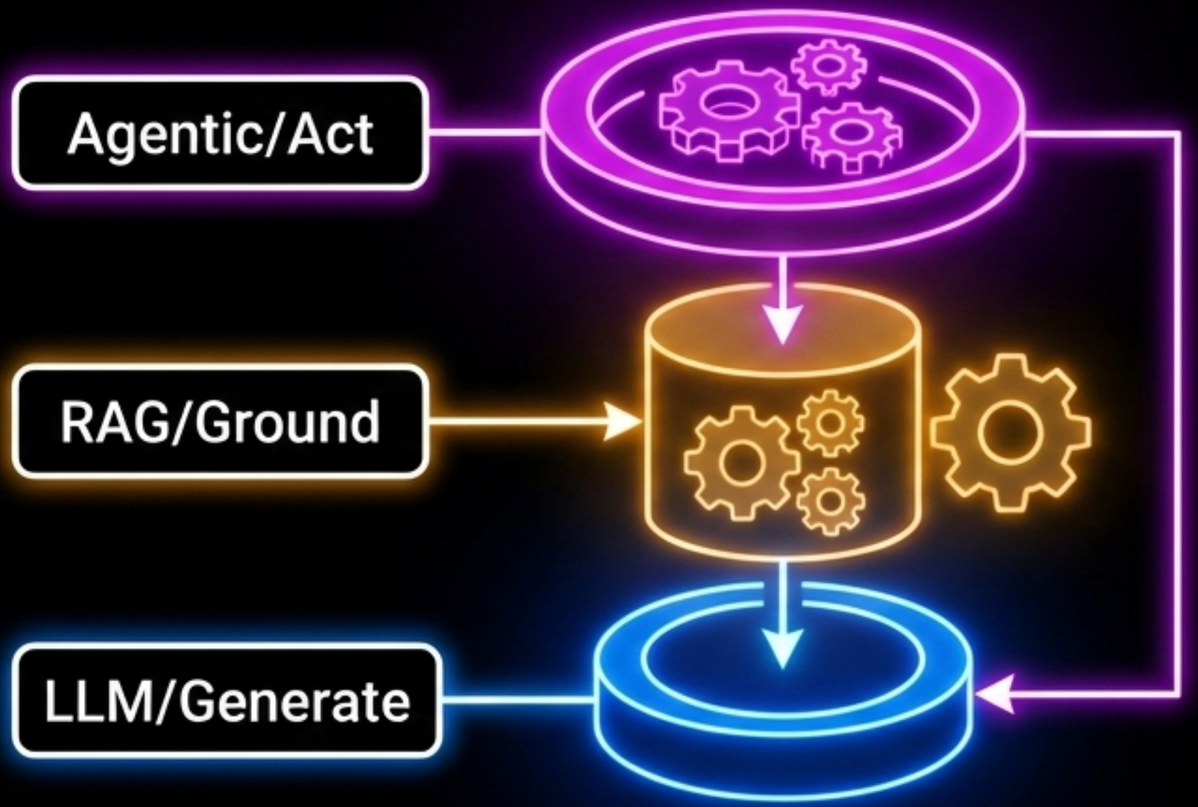
The Enterprise AI Blueprint

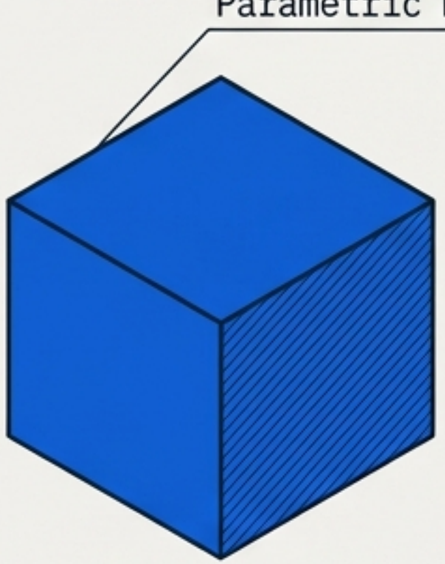
Diagnosing Workflows to Build the Smallest **Sufficient AI Pattern**



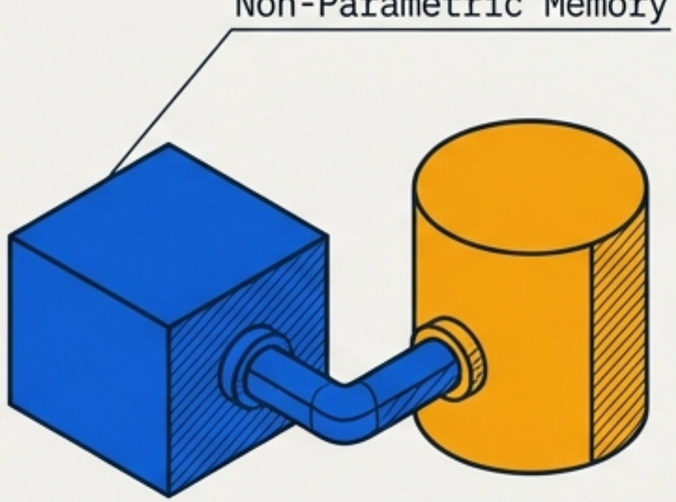
A Practitioner's Guide to LLMs, RAG, and Agentic Orchestration

The enterprise AI challenge is an architectural choice.

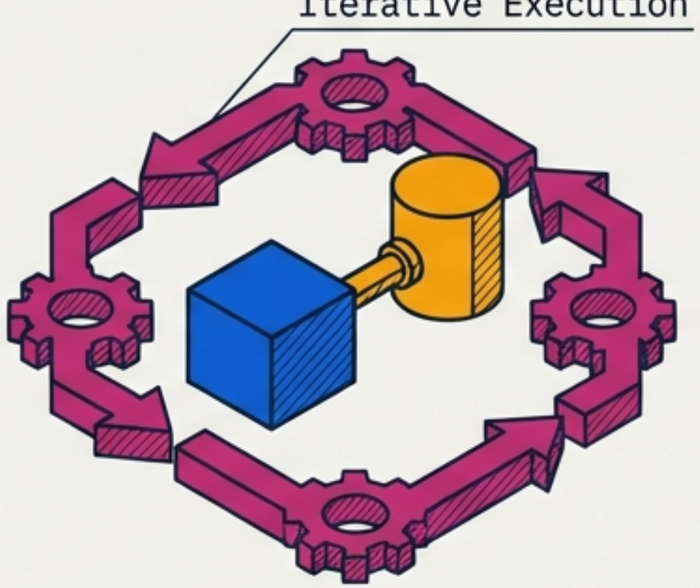
Enterprise generative AI adoption can be reduced to a single strategic question: Does the workflow need to Generate, Ground, or Act? The highest-leverage approach is to select the smallest sufficient pattern to minimize complexity, token cost, and risk.



LLM (Generate):
Linguistic transformation.
Lowest complexity.

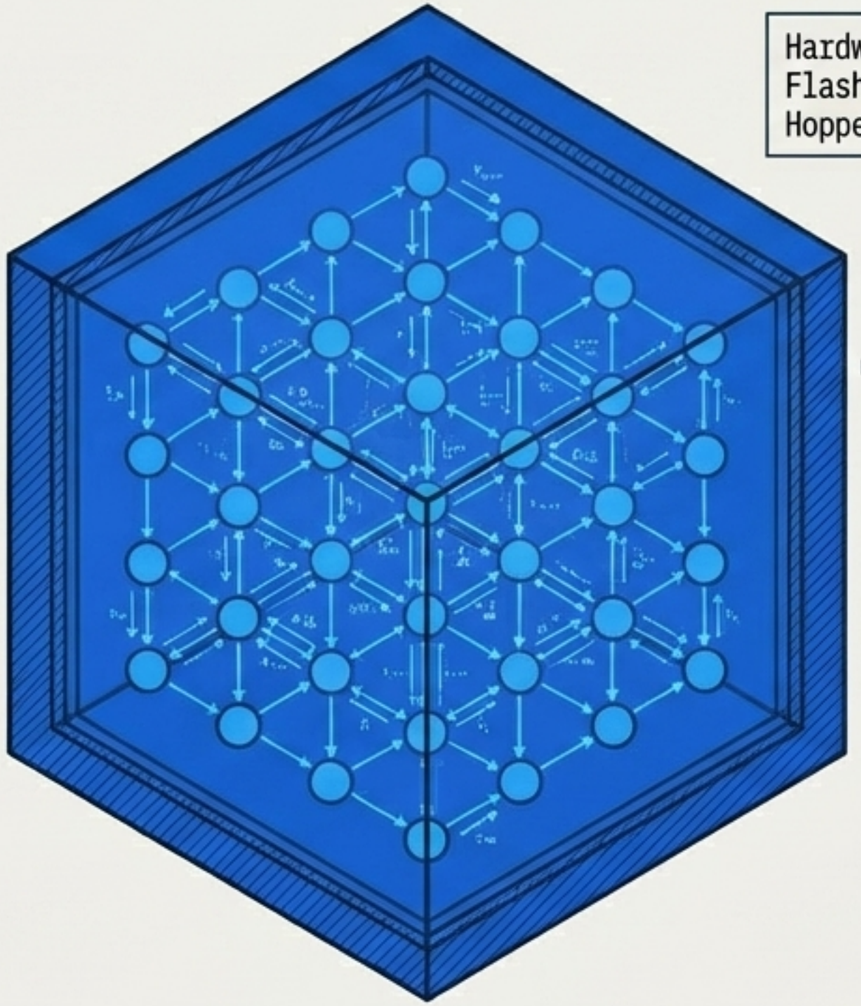


RAG (Ground):
Factuality and provenance.
Medium complexity.



Agentic AI (Act):
Multi-step task execution.
Highest complexity.

The Foundation: Language Transformation via Parametric Memory



Hardware/Software Co-design: FlashAttention-3 on Hopper GPUs

Core Mechanism

LLMs utilize Transformer architecture (attention-based sequence modeling) to predict tokens. Knowledge is "frozen" in the model's weights (parametric memory).

Primary Value

Few-shot performance for drafting, summarizing, and reformatting unstructured text into schemas like JSON.

The Hardware Reality

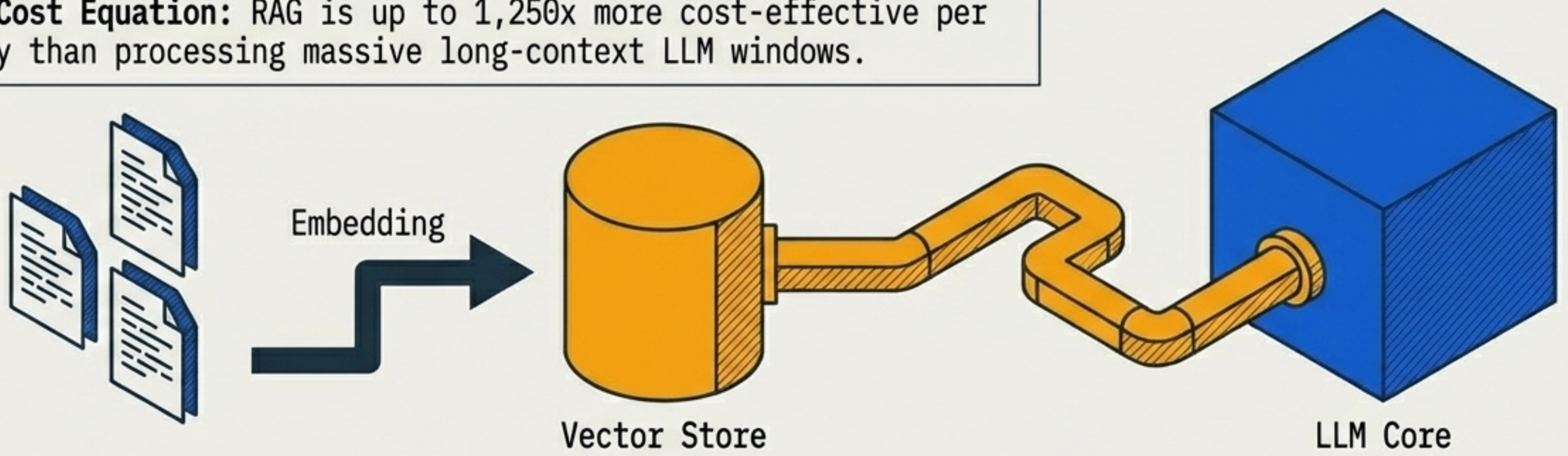
Innovations like GQA and FlashAttention-3 expand context windows up to 1,000,000 tokens by reducing high-bandwidth memory bottlenecks by 4x.

Key Failure Mode & Mitigation

Confident hallucination and prompt injection. **BA Control Rule:** Use only when the user can manually review and verify the generated draft.

Grounding: Injecting Enterprise Truth at Runtime

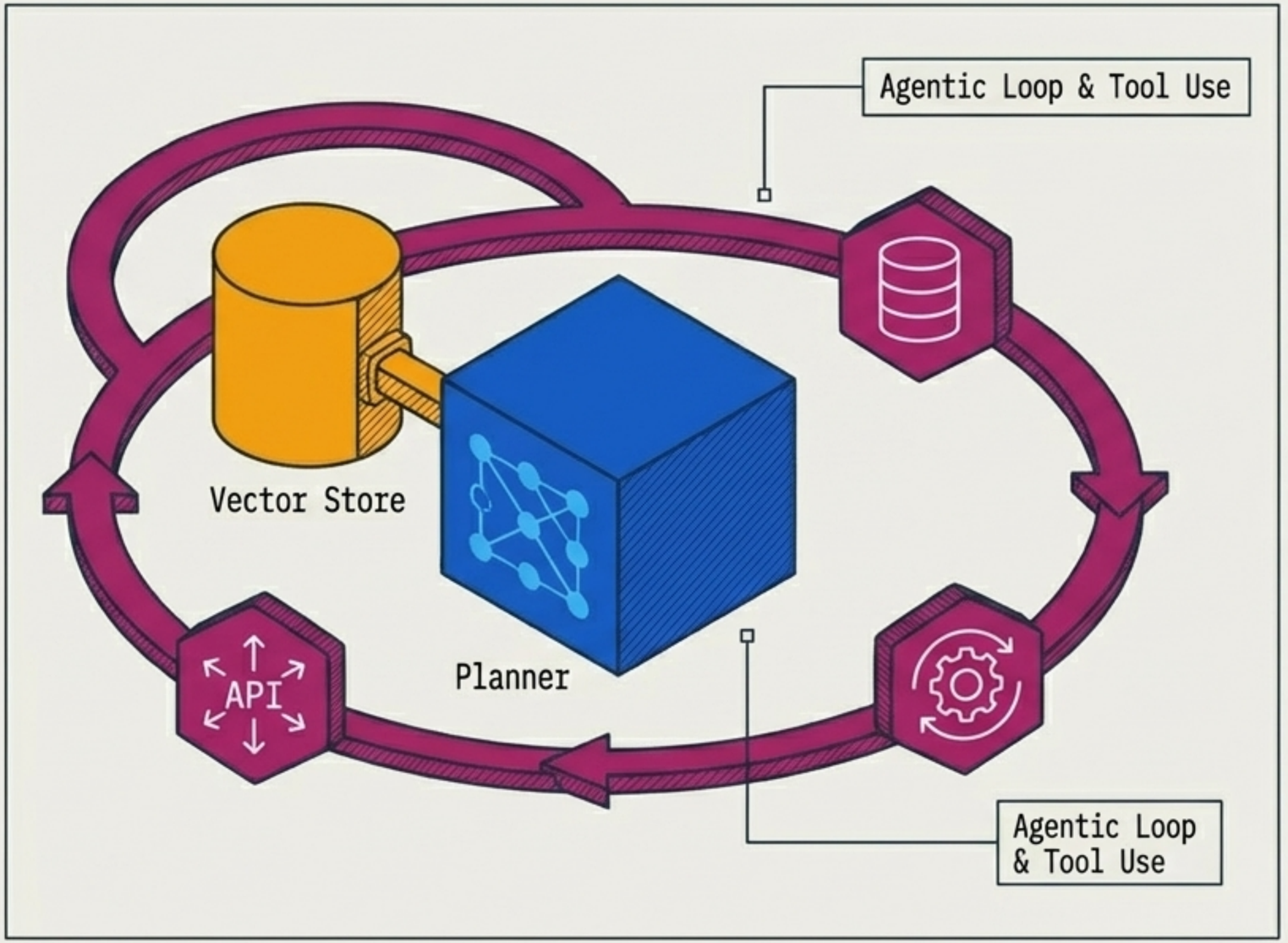
The Cost Equation: RAG is up to 1,250x more cost-effective per query than processing massive long-context LLM windows.



Simple	Advanced	Complex
Standard RAG (FAQ automation)	HyDE (guess-based), Corrective RAG (source double-checking)	Graph RAG (knowledge graphs for entity linking)

BA Control Rule: Mandatory citation enforcement. If the model cannot cite the retrieved chunk, it must respond: “Unknown based on approved sources.”

Autonomous Action: Multi-Step Execution and Tool Use



Core Mechanism

Agentic systems use the LLM to reason, formulate a plan, and call external tools in iterative loops until a stop condition is reached (ReAct/Toolformer patterns).

Multi-Agent Orchestration (MAS)

Specialized digital symphonies orchestrating parallel tasks. At real-world scale, MAS processes 1.4 trillion lines of journal entry data across 160,000 engagements.

Key Failure Mode

Tool failure cascades and infinite reasoning loops.

BA Control Rule

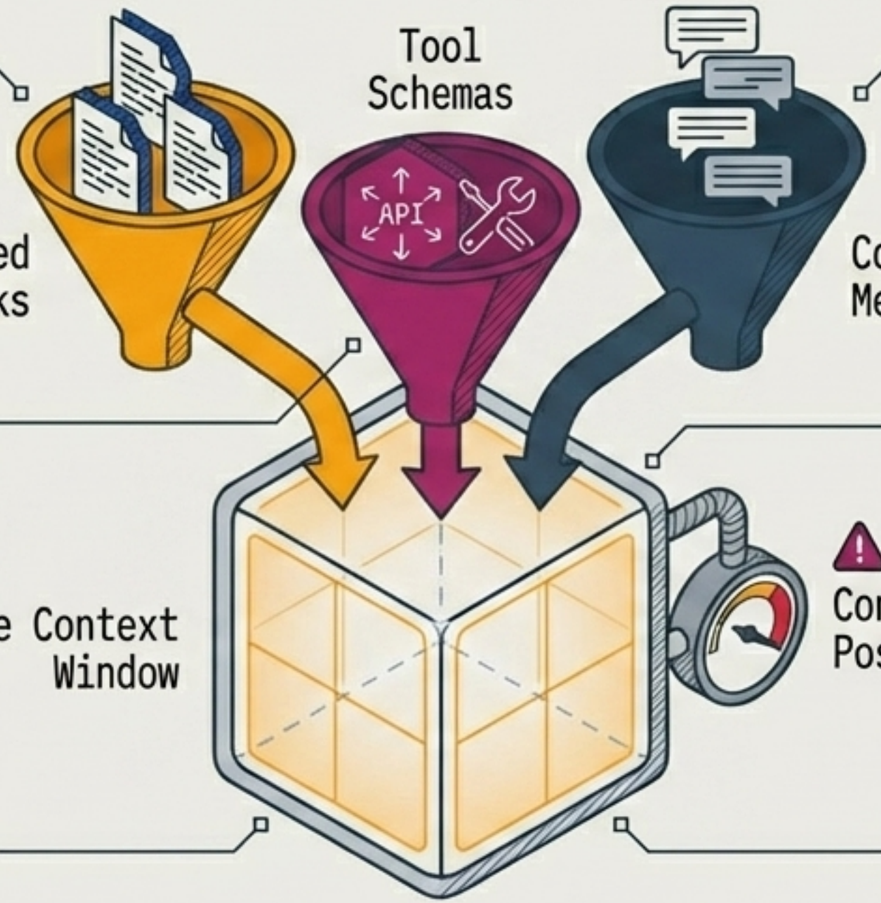
Default to 'Dry Run' mode. Agents propose multi-step actions; human reviewers approve execution.

The Universal Bottleneck: Context Engineering

All AI architectures fiercely compete for the exact same finite resource: the model's attention budget.

1. **Retrieval:** Supplying high-signal external chunks.

2. **Memory:** Preserving short/long-term session history.



6. **Tools:** Injecting API descriptions so the model knows what it can do.

3. **Augmentation:** Expanding queries via HyDE.

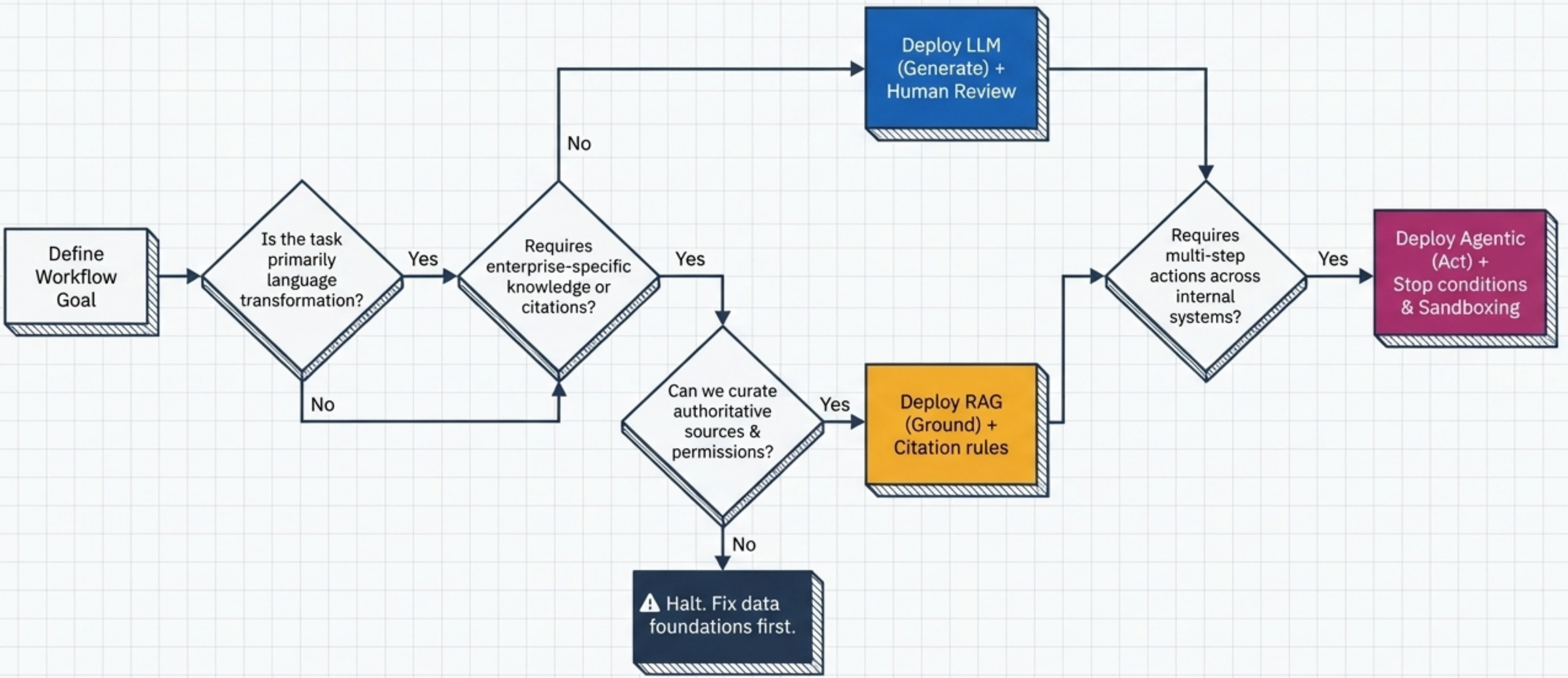
5. **Compression:** Summarizing past exchanges to save token space.

4. **Prompting:** Setting rules, guardrails, and boundaries.

The Diagnostic Matrix: Evaluating Enterprise Trade-offs

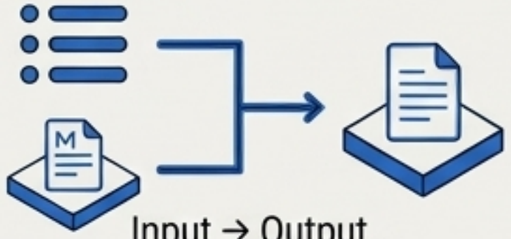
Attribute	Generate (LLM)	Ground (RAG)	Act (Agentic)
Primary Value	Linguistic Transformation	Factuality & Provenance	End-to-End Task Execution
Latency Drivers	Model latency	Model + Search/Rerank	Multi-step tool loop compounding
Cost Drivers	Token usage	Tokens + Index Build	Tokens + Orchestration Runtime
Auditability	Low (Logs only) 	High (Direct Citations) 	Moderate (Step-by-step logs) 
When It Fails	Domain knowledge required	Poor chunking / Messy corpus	Unreliable tools / Weak governance

The Smallest Sufficient Pattern: BA Decision Tree



Applying the Patterns in Regulated Environments

Case A: Generate (LLM)

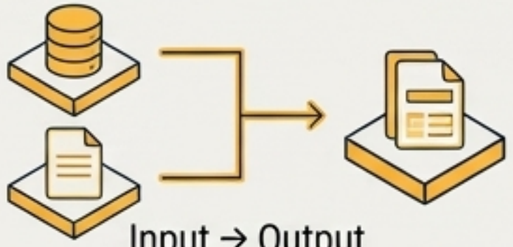


The Audit Memo

Workflow: Creates a first-draft summary from bulleted inputs and meeting transcripts.

BA Control
Mandatory firm templates and automated tone checks.

Case B: Ground (RAG)

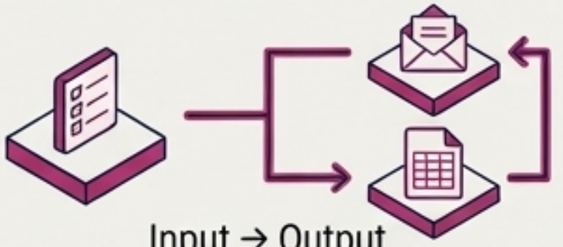


Medical Prior Authorization

Workflow: Retrieves patient EHR data and aligns it with insurance policy guidelines. Achieves 89.4% reasoning accuracy in clinical admin tasks.

BA Control
Strict RBAC permissions and mandatory source citations.

Case C: Act (Agent)

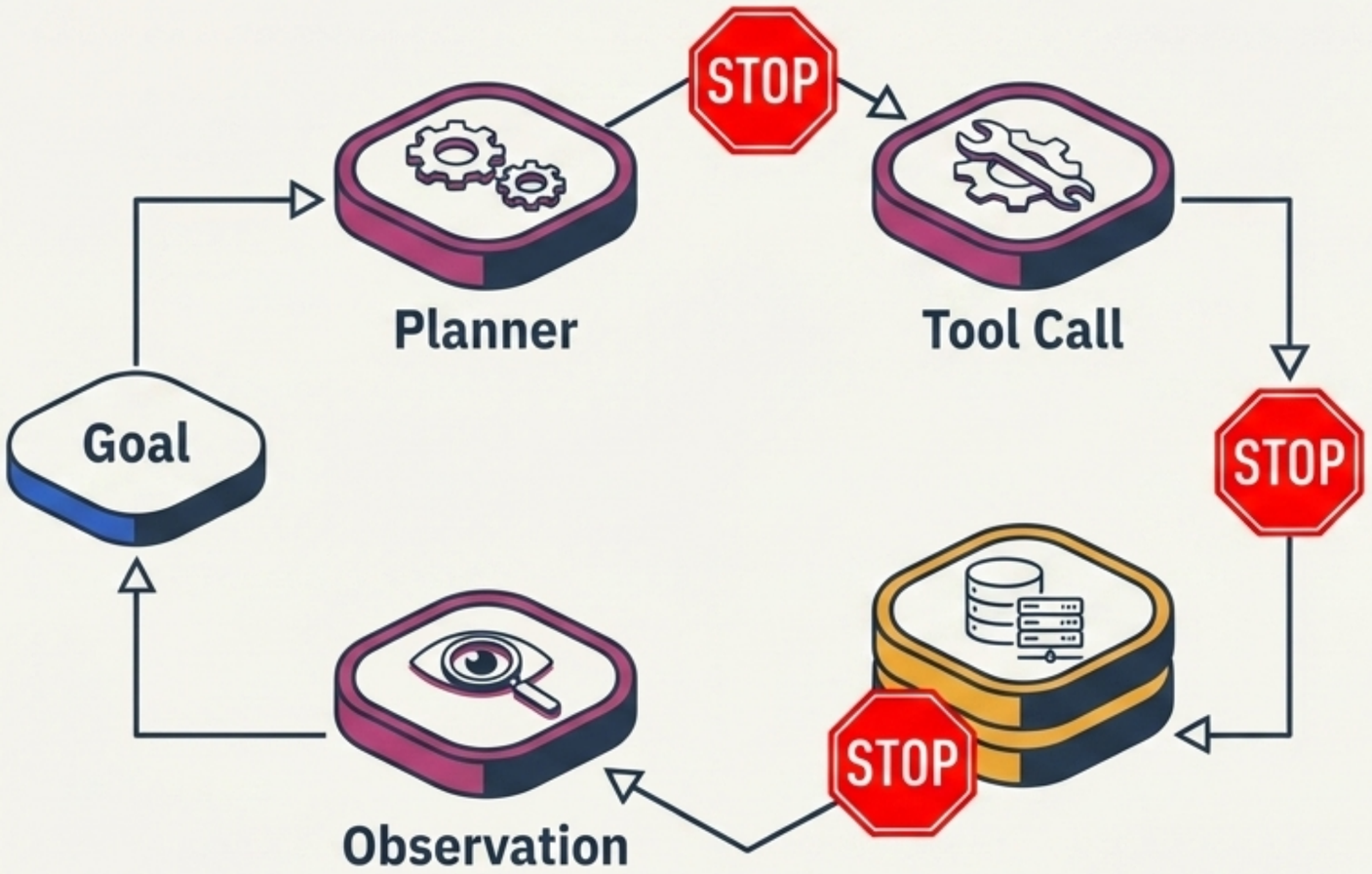


PBC Request Tracking

Workflow: Reads audit status, drafts follow-up emails to clients, and autonomously updates tracker spreadsheets.

BA Control
Executes only after human reviewer approves the batch (Dry-Run mode).

The Anatomy of an Agent (and Where to Install the Brakes)

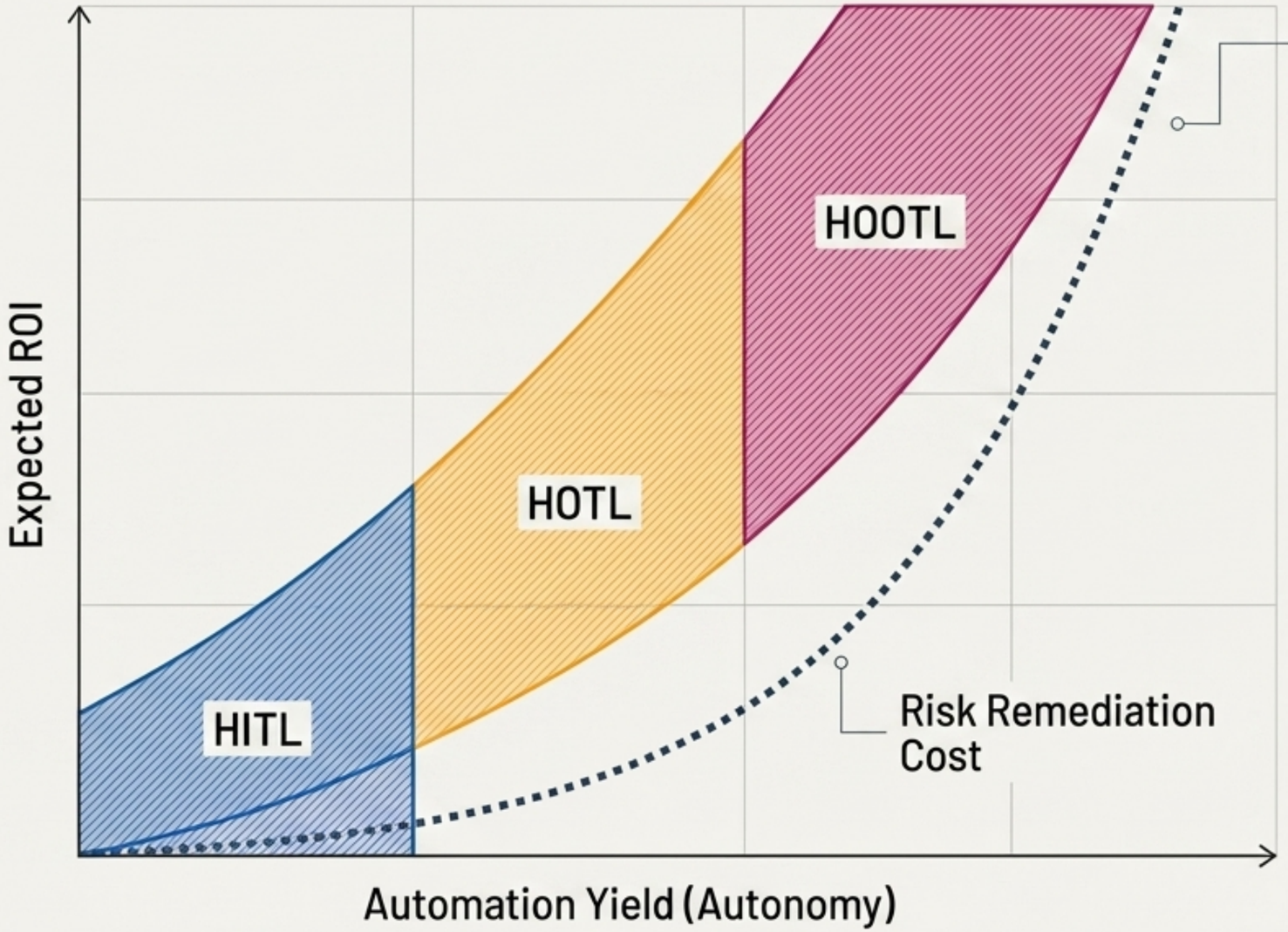


The Brakes (Governing the Loop)

The model interleaves "reasoning traces" with "actions". It thinks, acts, observes the result, and loops. Brakes are mandatory.

- **Iteration Limits:** Hard cap at 5 loop steps to prevent infinite reasoning cycles.
- **Budget Caps:** Hard token spend limits per workflow run.
- **Least Privilege:** Scoped service accounts for tools; strict API allowlists.
- **Immutable Logging:** Capturing prompt, tool call, output, and decision for auditability.

The Economics of Autonomy: Scaling ROI Safely



Baseline: $110M Call Center Model

- **Human-In-The-Loop (HITL):**
  Human instigates, AI assists.
  25% efficiency boost = **$25M** ROI.
  Zero systemic risk.
- **Human-On-The-Loop (HOTL):**
  AI acts, human reviews.
  40% efficiency boost = **$40M ROI** + moderate risk management costs.
- **Human-Out-Of-The-Loop (HOOTL):**
  Fully autonomous batch processing.
  90% boost = **$90M ROI**, offset by massive remediation costs.

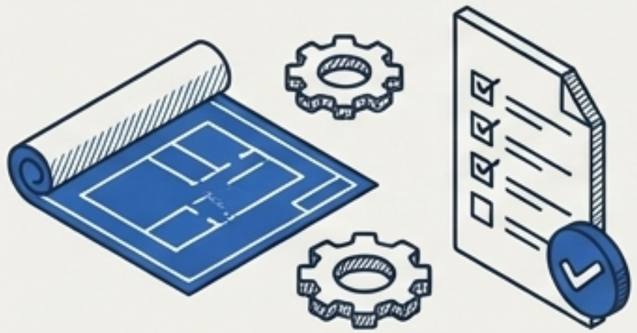
The CFO Rule: High-performing enterprises hit 6–12 month payback by using RAG with HITL controls.

The BA Validation Dashboard

NIST AI RMF Alignment: Measure & Manage

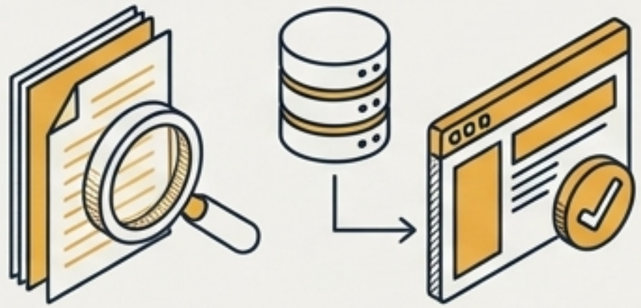
LLM-Only Pipeline

- **Test Target:** Prompt adherence and format compliance.
- **Metric:** Rubric-based quality scores.
- **UAT Strategy:** "Golden set" regression testing; measure rework rate.



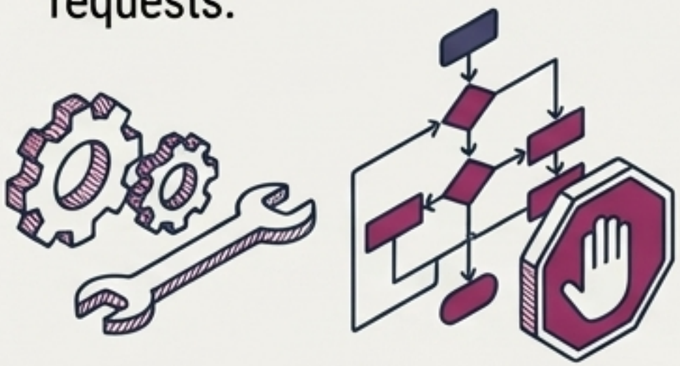
RAG Pipeline

- **Test Target:** Context precision and faithfulness.
- **Metric:** Top-K hit rate (did it find evidence?) and hallucination rate.
- **UAT Strategy:** "Trust tests"—do users accept citations without secondary search?



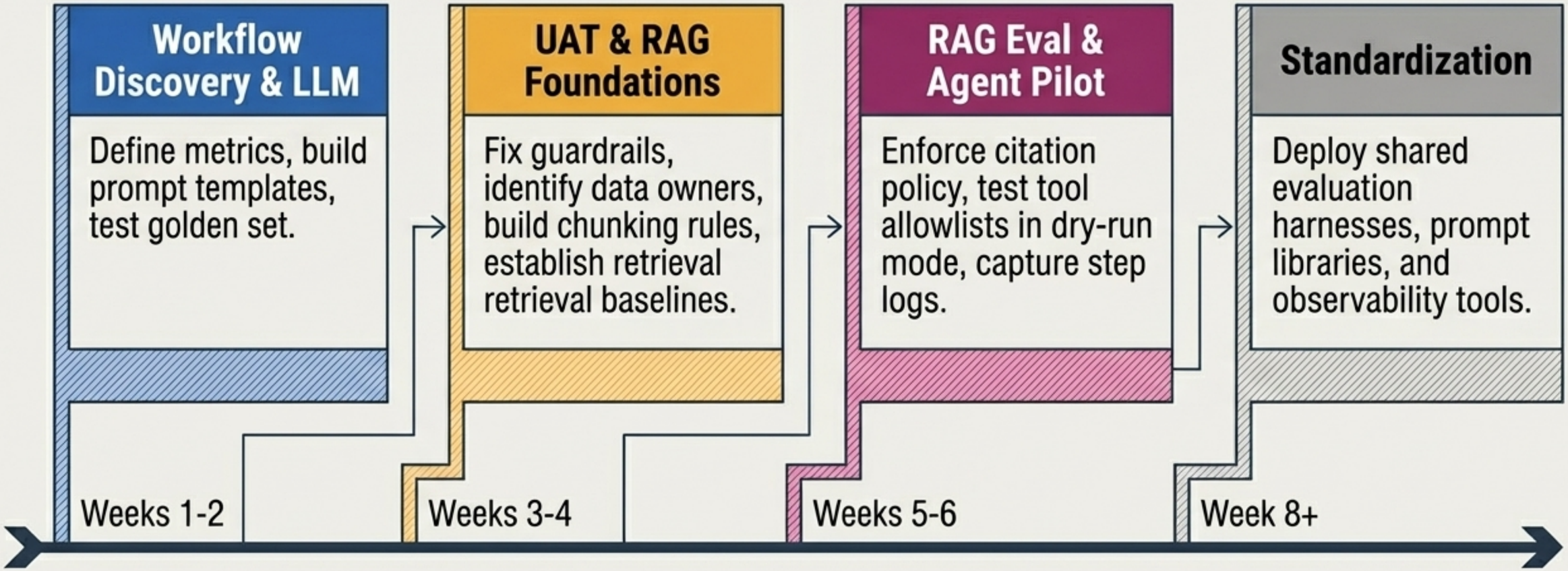
Agentic Systems

- **Test Target:** Tool call accuracy and loop safety.
- **Metric:** Task completion rate and human override rate.
- **UAT Strategy:** Scenario testing of tool failures and ambiguous requests.

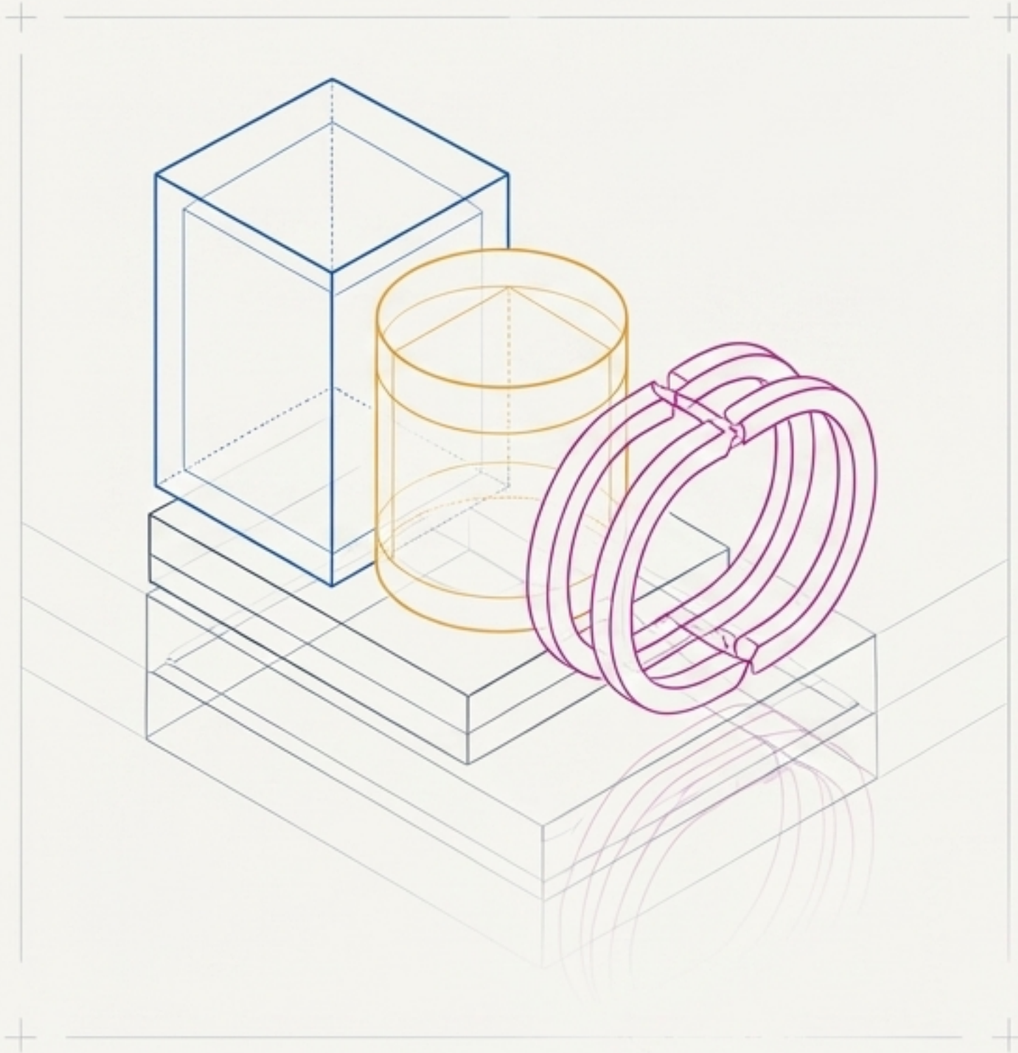


The Execution Path: MVP to Platform Scale

The MVP Scope Rule: Start with one persona, one narrow workflow lane, clear input constraints, and 2-3 measurable KPIs.



The Architecture Defines the Outcome



1. **Complexity is a Cost:** Resist the urge to over-engineer. Default to **LLM generation** (in Precision Cobalt Blue), elevate to **RAG** (Structured Technical Amber) only for enterprise truth, and permit **Agentic** (Strategic Deep Magenta) action only for multi-step execution.
2. **Manage the Attention Budget:** Context engineering is the ultimate constraint. Manage retrieval, memory, and tools with extreme prejudice.
3. **Scale Controls with Autonomy:** As systems move from generation to action, governance must evolve from prompt guardrails to immutable logs and explicit stop conditions.

The most successful enterprises don't deploy the most complex AI—they deploy the most precisely governed AI.