

TRADITIONAL RAG VS AGENTIC RAG

Why one search is not always enough

Both ground AI answers in external knowledge—but they retrieve it differently.

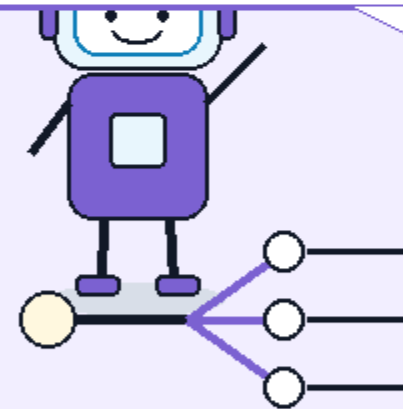
TRADITIONAL RAG



Retrieve once, then answer.

AGENTIC RAG

"Plan, search, inspect, and search again if needed."



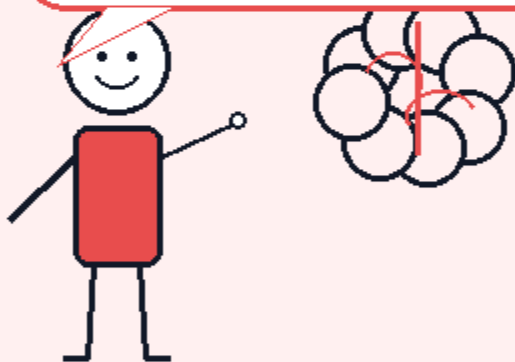
Same goal: ground the answer. Different way of finding the evidence.

RAG = OPEN-BOOK AI

Instead of answering only from training, the model receives relevant external information at answer time.

CLOSED-BOOK ANSWER

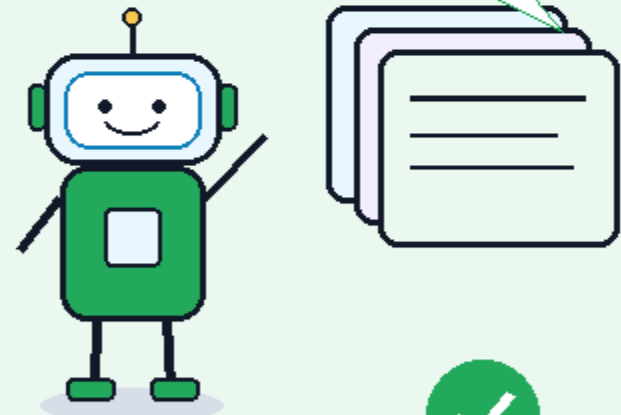
"I'll answer from what I remember."



Knowledge may be outdated or incomplete.

OPEN-BOOK RAG

"Let me retrieve the relevant pages first."

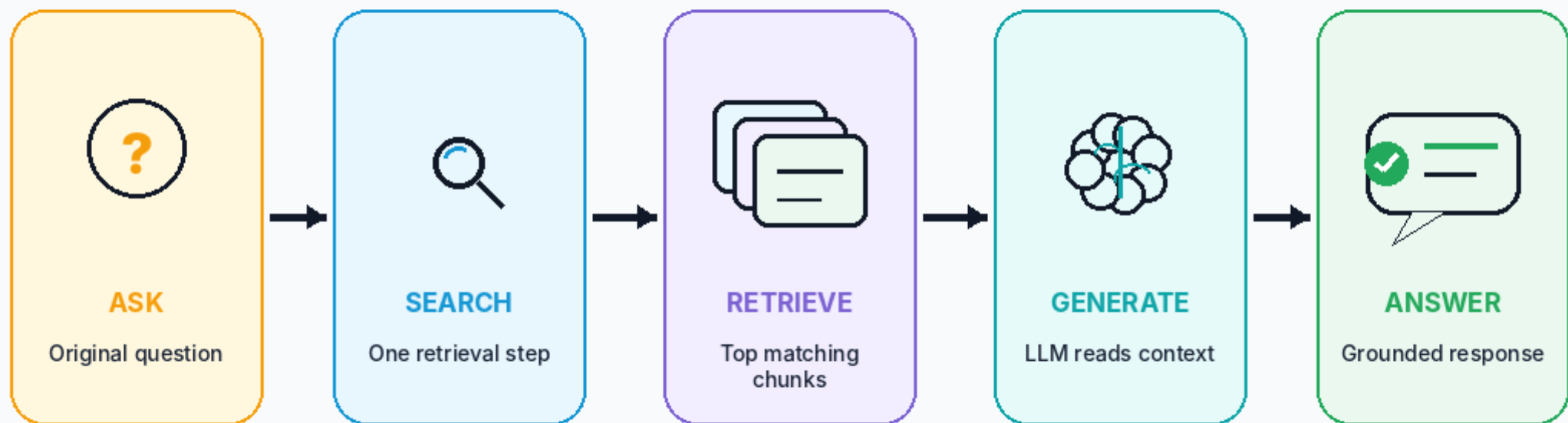


Answer is grounded in retrieved sources.

RAG adds external context; it does not need to retrain the model for every update.

TRADITIONAL RAG: A FIXED PIPELINE

The same retrieval path runs for every question.



- Fast and predictable
- Easy to test and debug
- Usually lower cost and latency

Traditional RAG is excellent when one good search is enough.

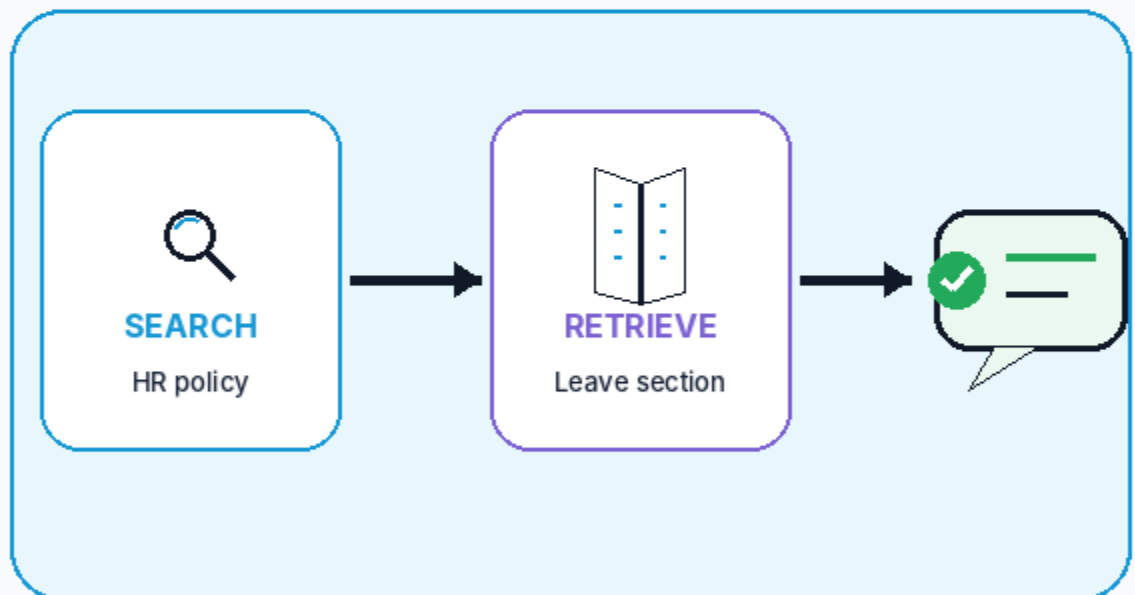
SIMPLE QUESTION?

TRADITIONAL RAG WORKS

Example: an employee asks one clear question about one policy.

GREAT

"How many casual-leave days do employees receive?"



"Employees receive 12 casual-leave days per year." Source: Leave Policy §3.2

Clear question + one authoritative source + one retrieval = a good fit.

BUT SOME QUESTIONS HIDE MANY QUESTIONS

One sentence may contain several separate information needs.

"I'm a contractor in India. Can I carry unused leave into 2027, and does the rule change after I become an employee?"

1 Who is covered by the contractor policy?

2 What is the India carry-forward rule?

3 What changes after employee conversion?

4 Which dates and policy versions apply?



A single search query may retrieve one piece and miss the rest.

WHY ONE-SHOT RETRIEVAL CAN FAIL

The answer can sound confident even when the evidence is incomplete.



AMBIGUOUS WORDS

"Leave" may mean vacation, sick leave, or company exit.



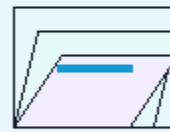
MULTI-HOP FACTS

The answer depends on facts stored in different documents.



WRONG SOURCE

A policy question is sent only to a general vector index.



MISSING CONTEXT

Country, role, date, or product version was not included.



STALE CHUNKS

Older policy text ranks higher than the current version.



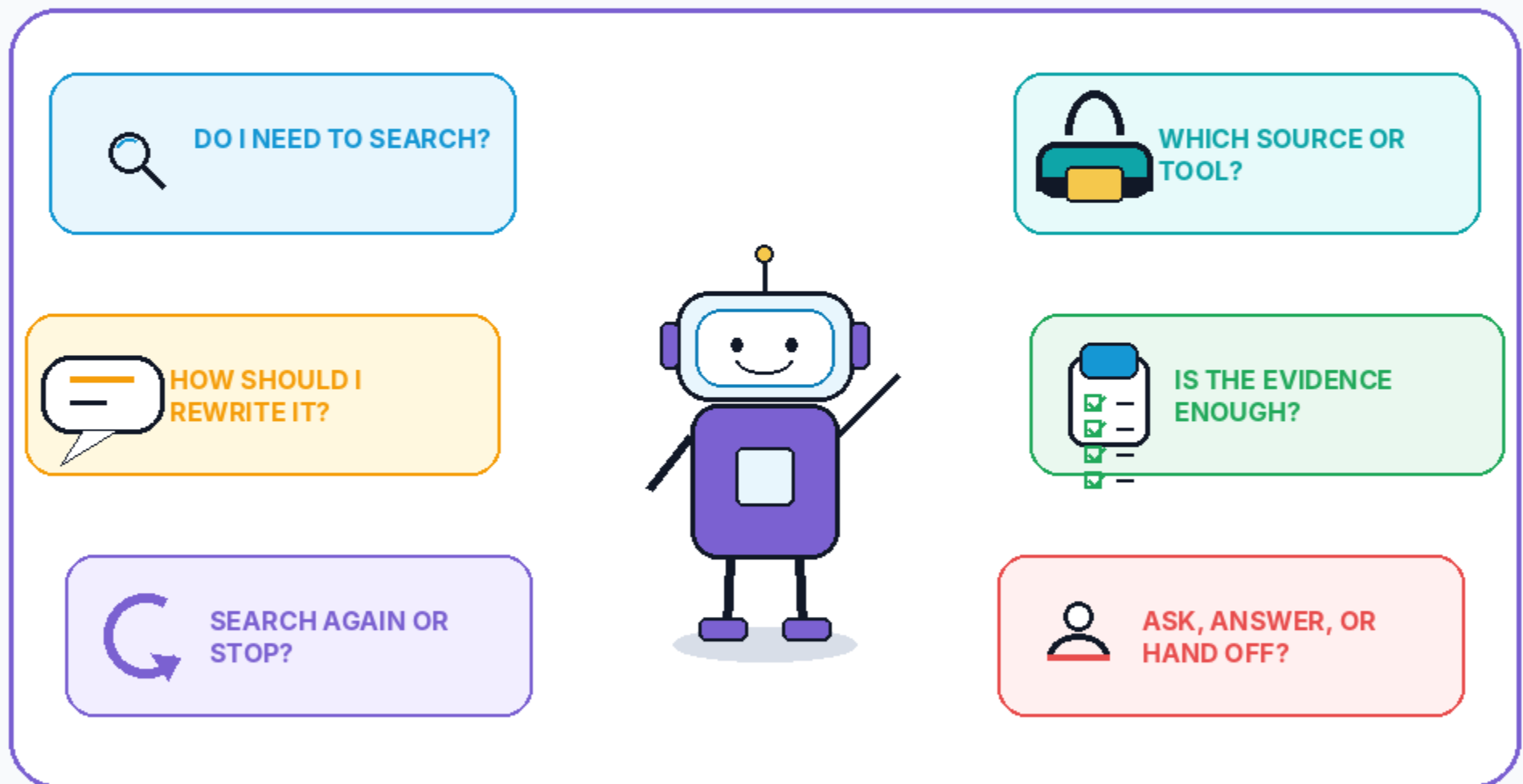
NO QUALITY CHECK

The model answers before verifying coverage or authority.

Retrieval quality sets the ceiling for answer quality.

AGENTIC RAG: RETRIEVAL WITH DECISIONS

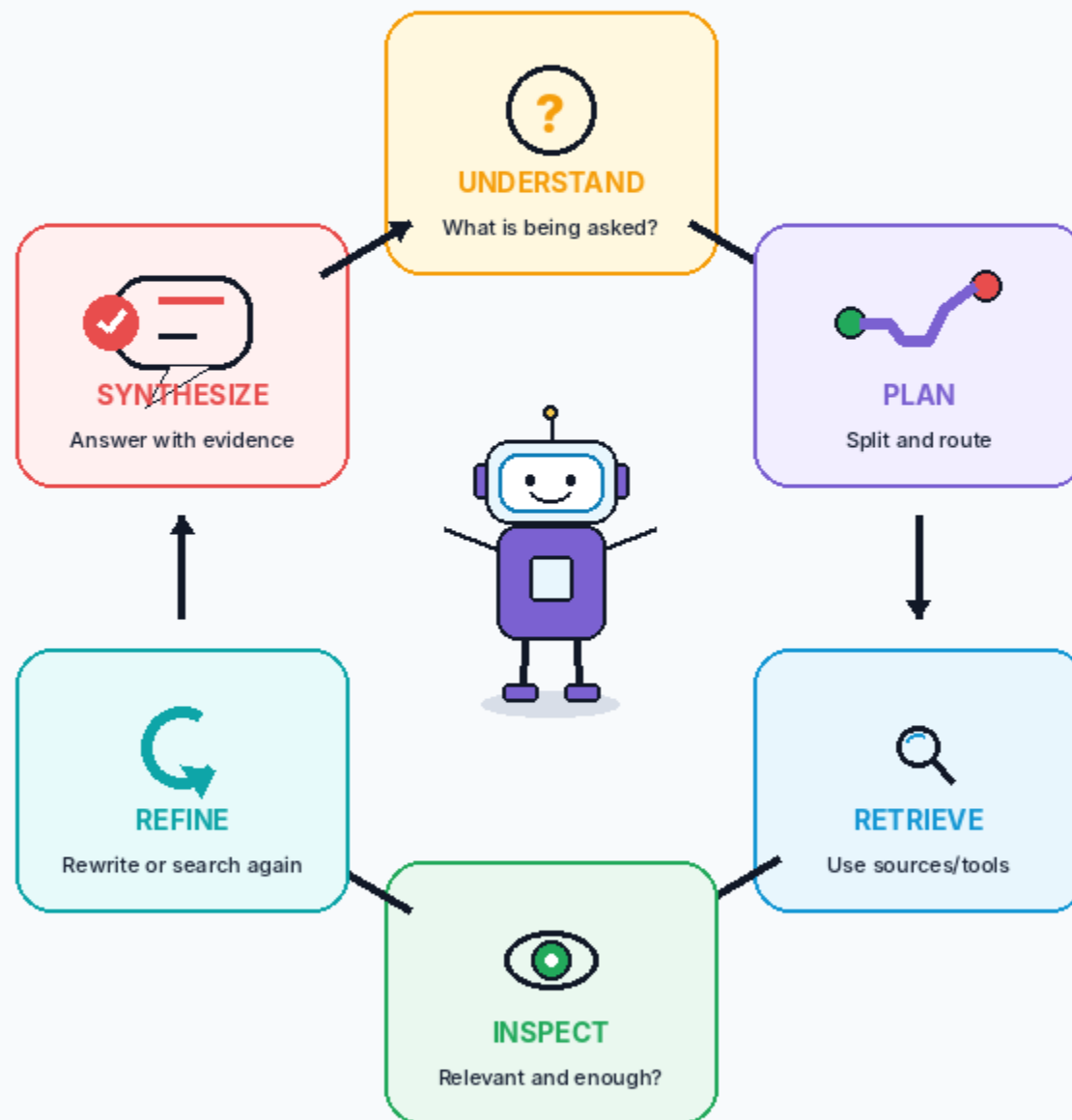
The agent dynamically controls how external knowledge is gathered.



Agentic RAG is a family of patterns—not one fixed architecture.

THE AGENTIC RAG LOOP

Retrieval becomes part of the agent's feedback loop.



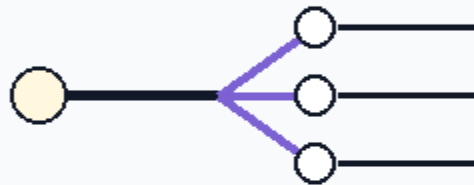
Stop when the answer is supported—or ask for clarification or human help.

STEP 1:

DECOMPOSE THE QUESTION

Turn one complex request into focused, searchable subquestions.

"Which laptop reimbursement policy applies to contractors in India, and does it cover repairs after two years?"



1

Contractor eligibility

Who can claim reimbursement?

2

India policy

Which country-specific rule applies?

3

Repair coverage

Are repairs reimbursable?

4

Device age

Does a two-year limit change the answer?

Smaller questions are easier to retrieve accurately—and can often run in parallel.

STEP 2: CHOOSE THE RIGHT SOURCE

Different questions belong in different knowledge systems.



POLICY DOCS

Rules, procedures, product documentation



DATABASE

Orders, accounts, inventory, transactions



TICKET SYSTEM

Incidents, case history, resolutions



WEB / RELEASE NOTES

Current public information and updates



APIs / BUSINESS APPS

Live status and permitted actions



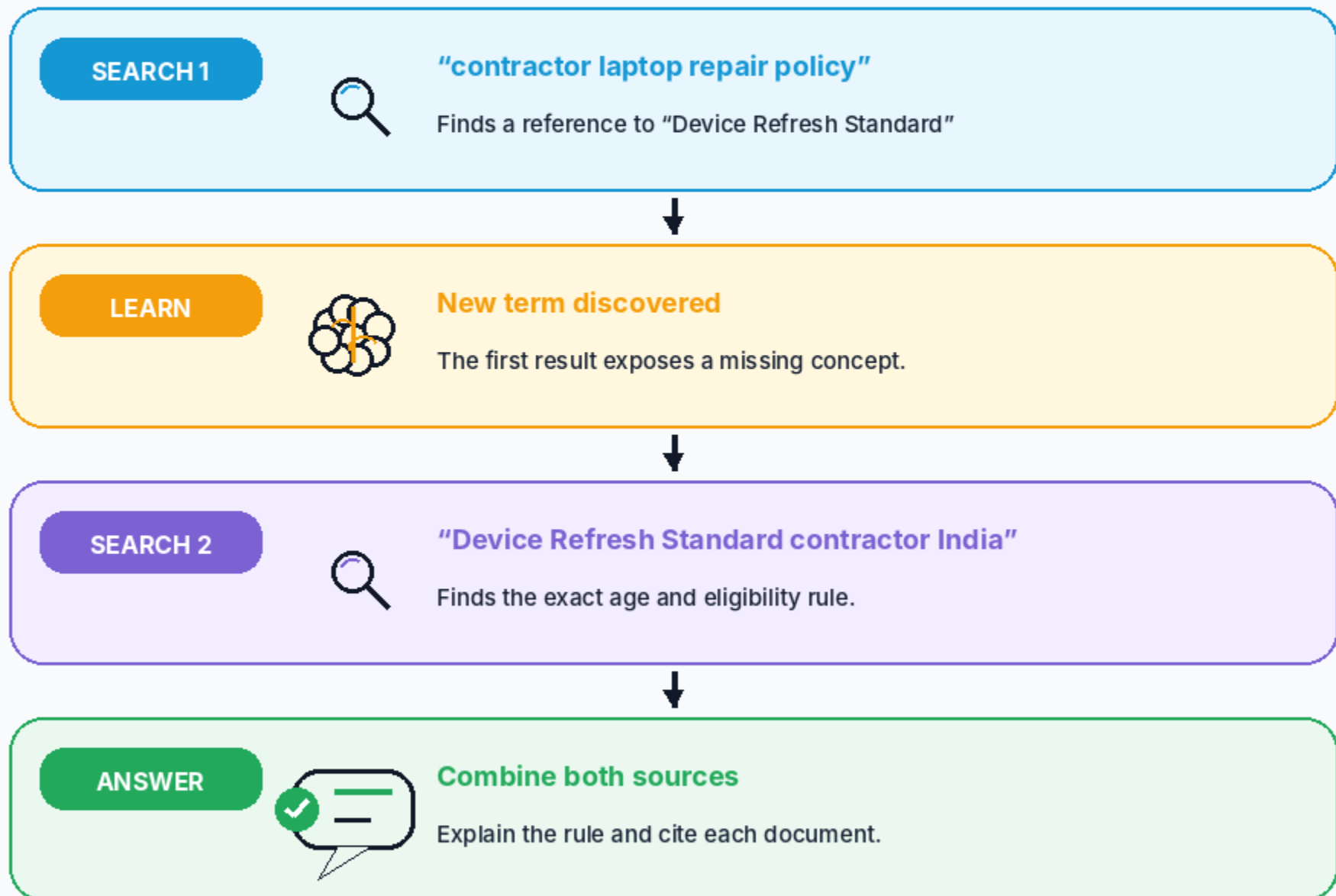
LOGS / ANALYTICS

Errors, telemetry, metrics, patterns

A vector database is only one retrieval tool. Routing and permissions matter.

STEP 3: SEARCH, READ, SEARCH AGAIN

The next query can depend on what the previous result reveals.



Iterative retrieval turns newly discovered evidence into the next search.

STEP 4:

GRADE THE EVIDENCE

Do the retrieved sources actually support a complete answer?



RELEVANCE

Does it address the question?



COVERAGE

Are all subquestions answered?



AUTHORITY

Is this an approved source?



RECENCY

Is it the current version?



CONSISTENCY

Do sources agree?



PERMISSION

Is the user allowed to see it?

If evidence is weak: rewrite, change source, ask a question, retry—or hand off.

TRADITIONAL RAG VS AGENTIC RAG

A useful mental model—not a strict law.

DIMENSION	TRADITIONAL RAG	AGENTIC RAG
Retrieval	Usually runs once before generation	Runs when needed; may repeat
Query	Original or fixed preprocessing	Can rewrite or decompose dynamically
Sources	Usually fixed and predefined	Can route among multiple tools/sources
Control path	Predictable pipeline	Dynamic decisions and branches
Latency & cost	Lower and more predictable	Higher and more variable
Best fit	FAQs and straightforward lookup	Complex research and troubleshooting

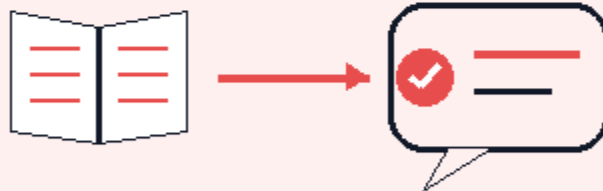
Agentic RAG buys flexibility with more complexity, latency, and cost.

REAL EXAMPLE: CUSTOMER SUPPORT

Some answers need documents and live business data.

"My refund says completed, but the money is missing. I paid partly by card and partly by wallet."

TRADITIONAL RAG



Searches the refund FAQ and replies: "Wait 5–7 business days."

AGENTIC RAG

Order DB

Confirm refund event

Card API

Check card settlement

Wallet ledger

Check wallet credit

Policy

Compare expected timelines



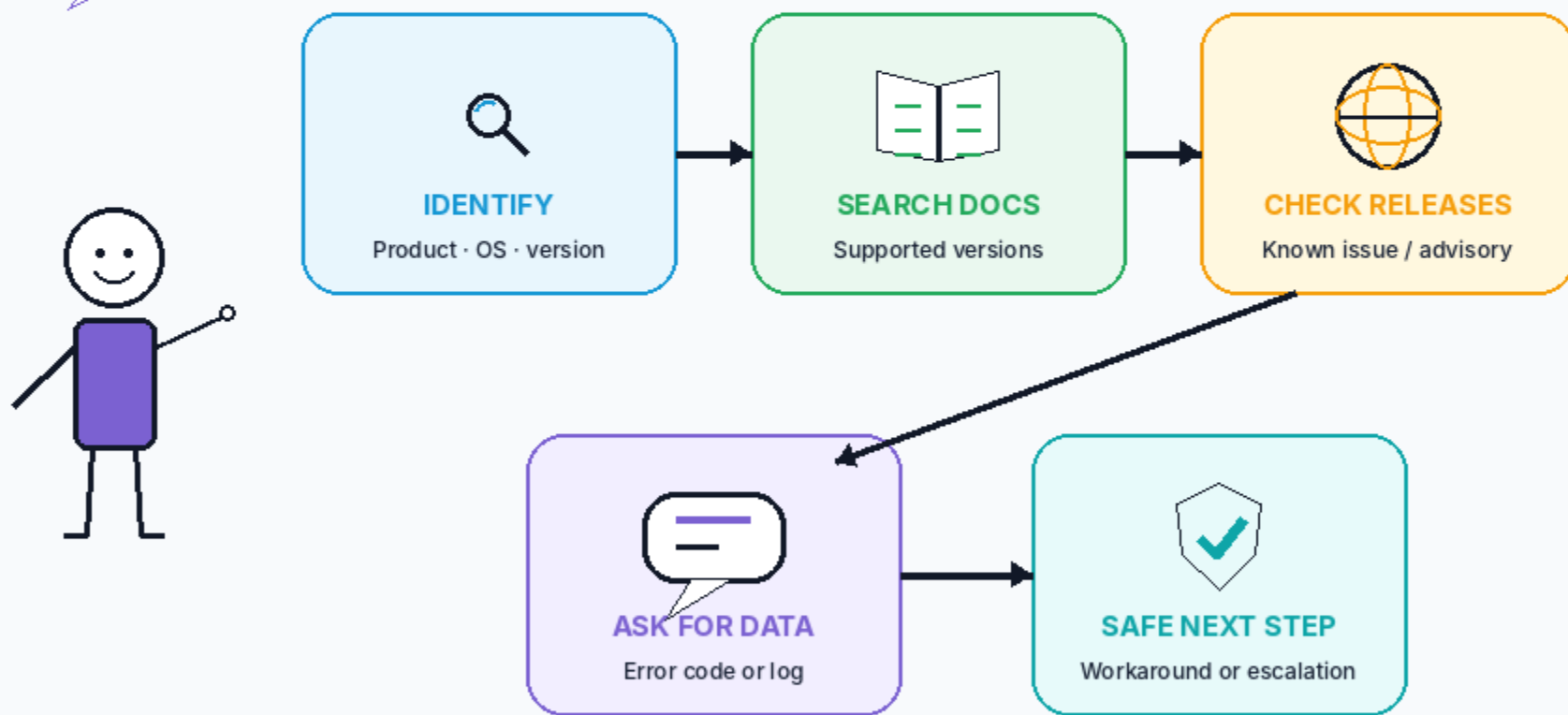
Explains the split status—or escalates with evidence.

The answer requires both knowledge retrieval and live state inspection.

REAL EXAMPLE: TECHNICAL TROUBLESHOOTING

The path changes as new evidence appears.

"Our VPN started disconnecting after yesterday's operating-system update."



Agentic RAG can search, ask, inspect, and adapt instead of guessing a universal fix.

WHEN TRADITIONAL RAG IS THE BETTER CHOICE

More agentic does not automatically mean better.



**FAST · SIMPLE ·
PREDICTABLE**



CHOOSE TRADITIONAL RAG WHEN

- ✓ Questions are simple and repetitive
- ✓ One authoritative source is usually enough
- ✓ Low latency is important
- ✓ Cost must be tightly controlled
- ✓ The workflow must stay predictable
- ✓ You need straightforward testing and auditing

Use the least complex retrieval architecture that reliably solves the task.

AGENTIC RAG: RISKS AND GUARDRAILS

Dynamic retrieval needs explicit limits and security boundaries.

WHAT CAN GO WRONG

- Endless search or retry loops
- Higher latency and unpredictable cost
- Irrelevant or contradictory sources
- Prompt injection inside retrieved content
- Cross-user or unauthorized data exposure
- Answer synthesis that overstates the evidence

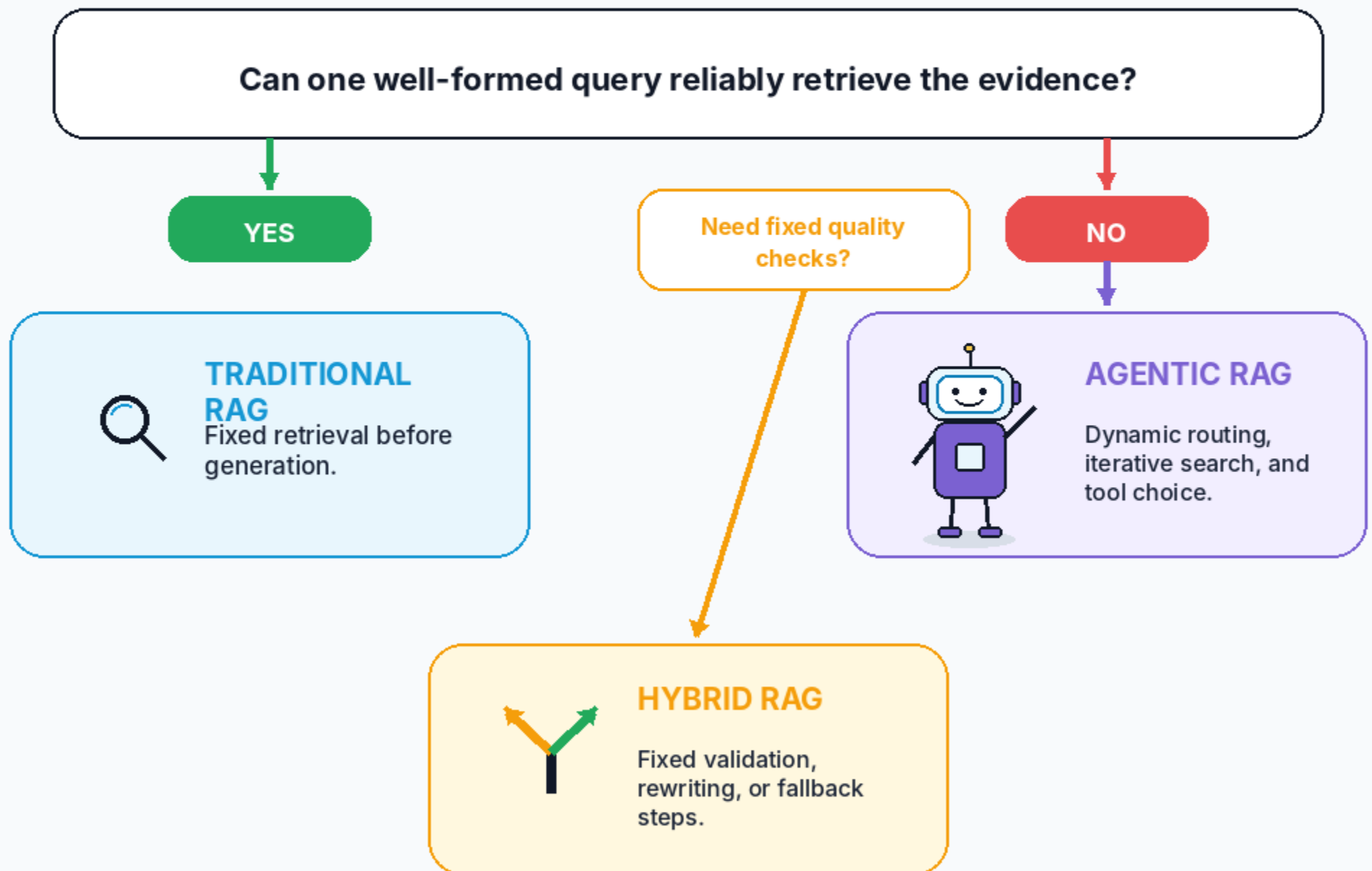
PRODUCTION CONTROLS

- ✓ Maximum steps, time, tokens, and cost
- ✓ Allowlisted tools and trusted sources
- ✓ User- and document-level access control
- ✓ Relevance, coverage, and citation checks
- ✓ Trace every query, tool call, and decision
- ✓ Clarification or human handoff when uncertain

The agent should be free to search—not free to ignore boundaries.

HOW TO CHOOSE THE RIGHT RAG PATTERN

Start simple. Add dynamic decisions only where they create measurable value.



HIGH-STAKES TASK? REQUIRE HUMAN REVIEW REGARDLESS OF THE PATTERN.

RAG FETCHES KNOWLEDGE. AGENTIC RAG MANAGES THE SEARCH.



- ✓ Simple lookup → Traditional
- ✓ Dynamic multi-step search → Agentic
- ✓ Fixed validation → Hybrid
- ✓ High stakes → Human review

NEXT UP

Day 11: Building a Capable Single Agent

Which question at your work needs more than one search?