

How AI systems fit together

One model. Different roles around it.

LLM

THE MODEL

Generates responses from the context provided.
Tool-capable models can also propose actions.

Three decisions around the model

01 Knowledge & behaviour

RAG + fine-tuning

02 Control of execution

Workflows + agents

03 Integration

MCP + WebMCP

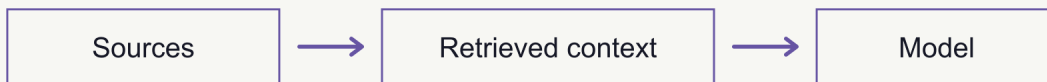
Choose the model using evidence from your own tasks.

Retrieve information. Adapt behaviour.

RAG and fine-tuning can work together.

RAG

Retrieval-augmented generation

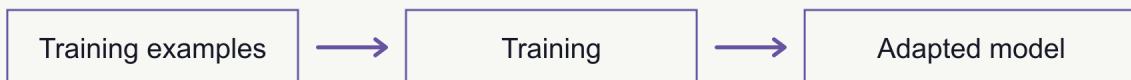


Use retrieved evidence to inform the response.

Useful when answers depend on current policies,
private documents or other changing sources.

Design for relevance, access and freshness.

Fine-tuning



Use additional training to shape task behaviour.

Useful for repeated patterns supported by good
examples and a measurable evaluation set.

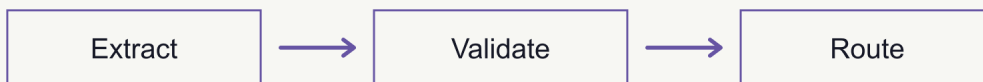
Evaluate improvements, regressions and maintenance.

Who chooses the next step?

The important distinction is control.

Workflow

The application defines the path.

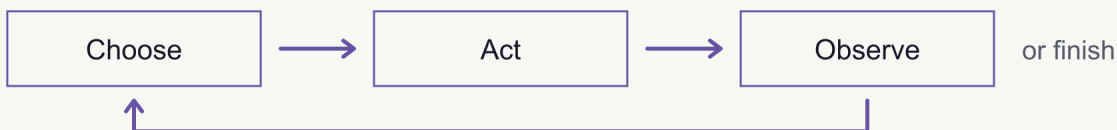


Use when the steps and decision rules are known.
Model calls can still happen inside the workflow.

Design exception paths and recovery.

Agent

The model helps direct the path as it learns more.



Use when the next action depends on the results.

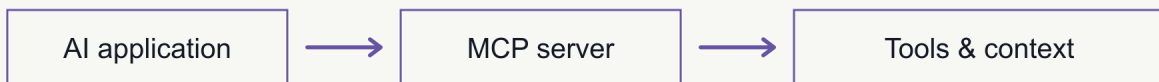
Bound permissions, stopping rules and cost.

Where does the action live?

Two interfaces with different integration surfaces.

MCP

Model Context Protocol

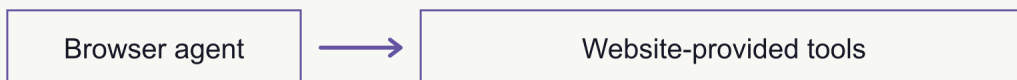


Standardize connections to tools and resources.
Useful for reusable service integrations across compatible AI applications.

Enforce authorization and service validation.

WebMCP

An emerging browser API



Expose structured website actions to browser agents.
Check browser support and validate every action.

WebMCP is distinct from an MCP server.

Choose for the job. Design for failure.

These are complementary roles, not intelligence levels.

01 What information must stay current?

Use retrieval where the answer depends on source evidence.

02 Is the execution path already known?

Keep defined processes explicit; use agents for adaptive work.

03 Which actions should the system have?

Expose only the tools and permissions the task requires.

Evaluate the whole system

Correctness · latency · cost · recoverability

REFERENCE NOTES

IBM: RAG and fine-tuning · Anthropic: agents and workflows

MCP specification · WebMCP draft · Reviewed 6 September 2026