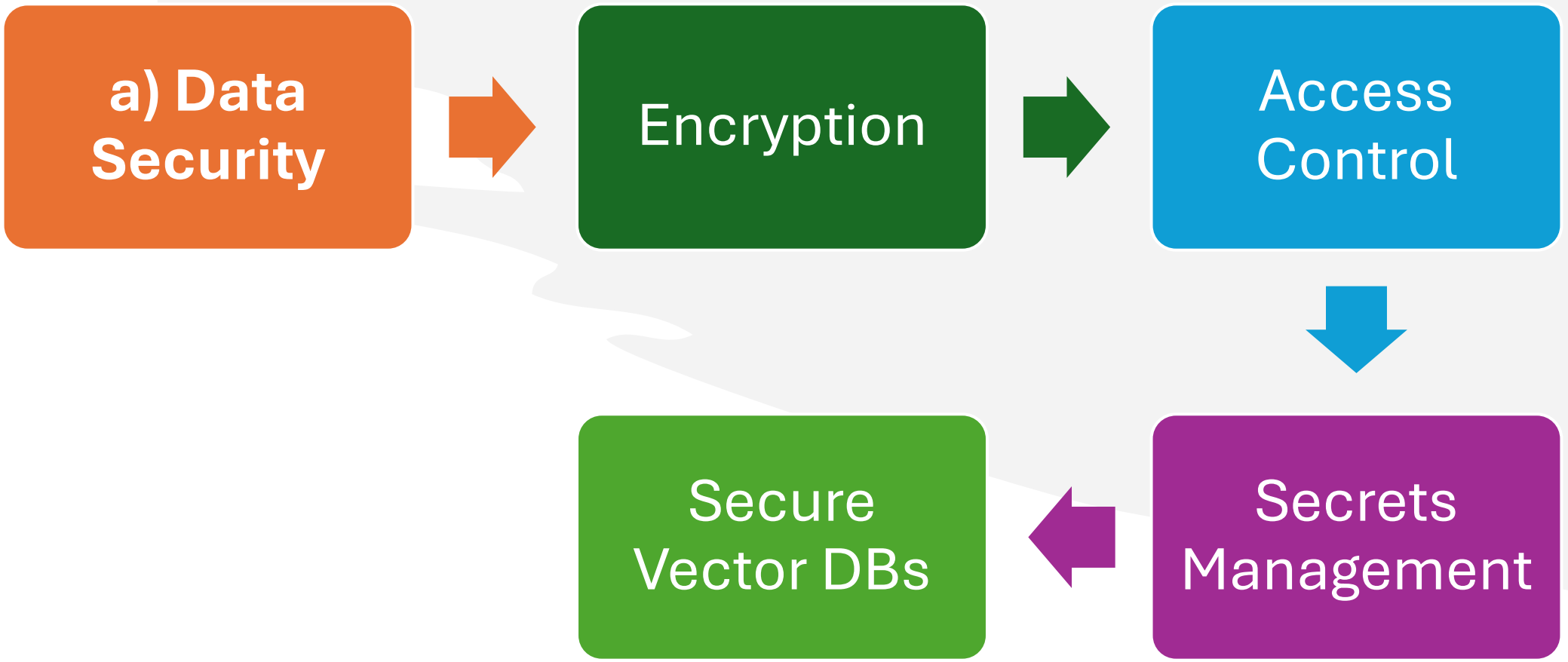
Security, Governance, MCP

Surendra Panpaliya

Security, Governance, Responsible AI

Security


a) Data Security
Encryption
Access Control
Secrets Management
Secure Vector DBs

Encryption

Use TLS in transit,

AES-256 at rest for embeddings,

chat logs, and documents.

Access Control:

Enforce **RBAC/ABAC**

(Role/Attribute-Based Access Control)

so only authorized users/agents see specific data.

Secrets Management

STORE API KEYS, TOKENS,

DB CREDENTIALS IN **VAULTS**

(HASHICORP VAULT,

AWS SECRETS MANAGER), NEVER IN CODE.

Secure Vector DBs

Using Milvus/Qdrant/pgvector,

Enable TLS + auth, and

restrict network access.

b) Application Security

- Input Sanitization
- Sandboxing:
- API Gateway

Input Sanitization

Guard against prompt injection

Malicious tool calls

Delete all records

Disguised in user input

Sandboxing

Run code-generation/execution in

Isolated containers

With resource limits.

API Gateway

Rate limiting + WAF rules

to block abuse and DOS attacks.

c) Monitoring & Incident Response

- Audit Logs
- Anomaly Detection
- Alerts

Audit Logs

Record queries,

model outputs, and

tool invocations.

Anomaly Detection

Flag unusual query volumes or

PII exposure attempts.

Alerts

INTEGRATE WITH SIEM

(SPLUNK, AZURE SENTINEL)

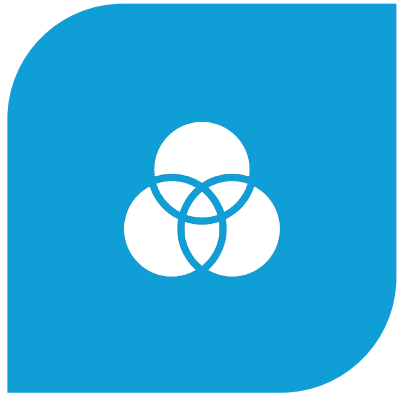
FOR REAL-TIME MONITORING.

2. Governance in GPT-5 Applications

a) Policy & Compliance

DATA RETENTION POLICIES

RIGHT TO BE FORGOTTEN



CROSS-BORDER DATA

Data Retention Policies

Define how long

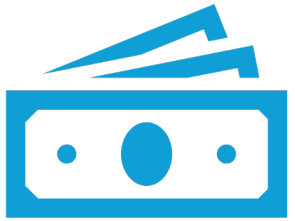
prompts/responses

are stored.

Right to be Forgotten

Implement deletion workflows

for user-specific embeddings



(GDPR/CCPA)

Cross-Border Data

Ensure embeddings

Logs stay in compliant

Regions (EU/India/US)

b) Lifecycle Governance

Model Registry

Prompt/Template Management

Change Control

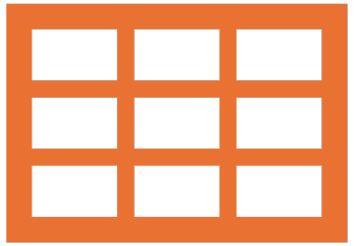
Model Registry

Track which

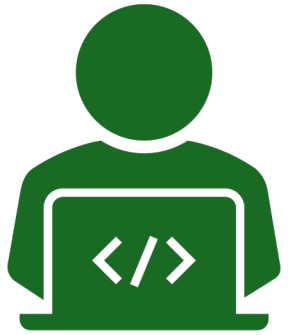
GPT-5 versions or

fine-tunes are in use

Prompt/Template Management



Centralize approved prompts,

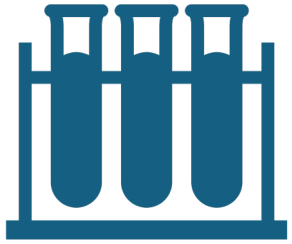


enforce version control (like code).

Change Control

Any update to prompts/tools

should go through review



test → approval.

Oversight Structures

Audit Trails

Full traceability — which model,

which embeddings,

which vector search, which answer.

KPIs

Track **accuracy, latency, cost,**

user satisfaction,

compliance incidents.

Fairness & Bias

Bias Audits

Balanced Training/Evaluation Data

Human Review Loops

Bias Audits

Regularly test outputs for

Demographic,

Geographic

Gender bias.

Balanced Training/Evaluation Data

Especially for fine-tuned GPT-5 models.

Human Review Loops

For sensitive domains (finance, healthcare, hiring).

Transparency

Explainability

Disclaimers

Confidence Scores

Explainability

Provide citations

to retrieved docs (RAG)

so users know

“where the answer came from.”

Disclaimers

Label AI-generated

output vs. human content.

Confidence Scores

Share retrieval confidence, not just polished text.

Accountability

- Human-in-the-Loop
- Incident Reporting
- Escalation Paths

Human-in-the-Loop

Approval workflows

for high-risk actions

(contracts, hiring, medical advice).

Incident Reporting

Allow users to flag

incorrect

unsafe output.

Escalation Paths

Define who in the organization is

accountable for AI decisions

Safety Guardrails

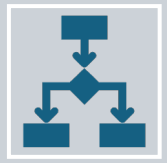
Toxicity Filters

Domain Guardrails

Evaluation Benchmarks

Toxicity Filters

Pre/post-process outputs



through moderation models.

Domain Guardrails

Restrict LLM to

Specific knowledge bases (via RAG),

disallow free hallucination.

Evaluation Benchmarks

Continuously test with red-team prompts

prompt injection,

jailbreaks,

policy violation tests

What is GenAI Governance?

Set of practices ensuring that

AI solutions comply with organizational policies,

legal frameworks, ethical standards,

and business objectives.

What is GenAI Evaluation?

Systematic assessment

to ensure AI models are accurate,

fair, reliable, robust, and

aligned with business requirements.

What is GenAI Evaluation?

Evaluation is a structured approach

to measure and improve the effectiveness,

accuracy, fairness, and safety

of Generative AI (GenAI) models.

Why Is GenAI Evaluation Essential at Walmart?

Ensuring Accuracy & Reliability

Mitigating Risks & Biases

Enhancing Customer Trust

Regulatory Compliance

Continuous Improvement

1️⃣ Ensuring Accuracy & Reliability

Confirms the AI-generated responses and

actions align with Walmart’s business requirements.

Maintains trust by providing consistently accurate results.

2️⃣ Mitigating Risks & Biases

Detects and reduces unwanted biases

that could harm Walmart's reputation.

Prevents incorrect decisions

that could lead to financial or operational risks

3️⃣ Enhancing Customer Trust

- Customers interact confidently
- with reliable, transparent AI solutions.
- Strengthens Walmart’s brand value
- by ensuring fairness and trustworthiness
- in automated interactions.

4️⃣ Regulatory Compliance

Ensures that Walmart’s AI solutions

comply with global regulations and

standards, avoiding legal issues.

5️⃣ Continuous Improvement

Provides insights into

model performance,

highlighting areas

for further development.

5️⃣ Continuous Improvement

Enables Walmart

to maintain competitive

advantage through adaptive and

improved GenAI solutions.

Key Metrics in GenAI Evaluation

Accuracy & Precision:

Correctness of AI responses.

Fairness & Bias:

AI decisions equitable across diverse user groups.

Key Metrics in GenAI Evaluation

Robustness:

Stability of AI under various conditions.

Safety & Ethics:

AI adherence to ethical guidelines and policies.

Potential Risks Without Effective Evaluation

- Misleading customer interactions.
- Financial losses from incorrect AI decisions.
- Reputational damage from
- biased or inappropriate outputs.
- Legal and compliance risks.

Practical Steps for Walmart GenAI Evaluators

Set clear, measurable criteria aligned with business objectives.

Implement standardized evaluation frameworks and tools.

Conduct regular audits and reviews.

Use feedback loops for continuous refinement

MCP Server using LangChain

Surendra Panpaliya
GKTCS Innovations
https://www.gktcs.com

Model Context Protocol (MCP)

Standardized framework

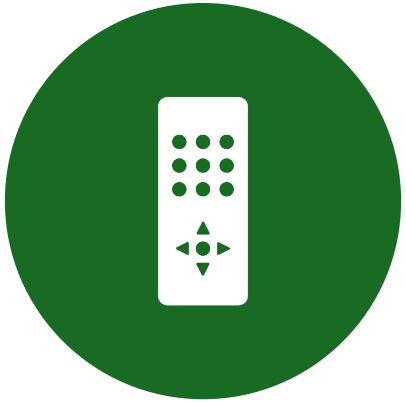
Allows **AI models**

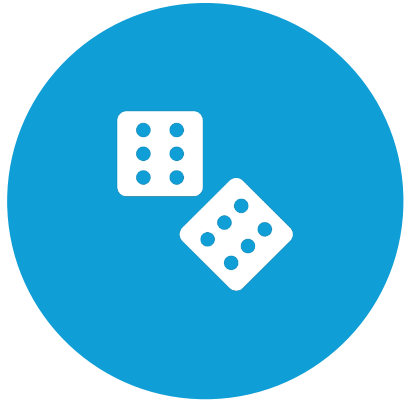
To **interact with real-world tools,**

APIs, and data sources

in a **safe, modular, and controlled way**.

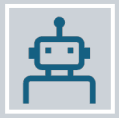
What is MCP (Model Context Protocol) Server?





LIKE A **TOOLBOX AND** **REMOTE CONTROL** FOR AI MODELS.

What is MCP (Model Context Protocol) Server?



Lets AI systems not just think and answer

but also **take real-world actions**

by calling APIs, tools, or databases

in a safe, modular way.

MCP Workflow

[AI Model / LLM]

↓

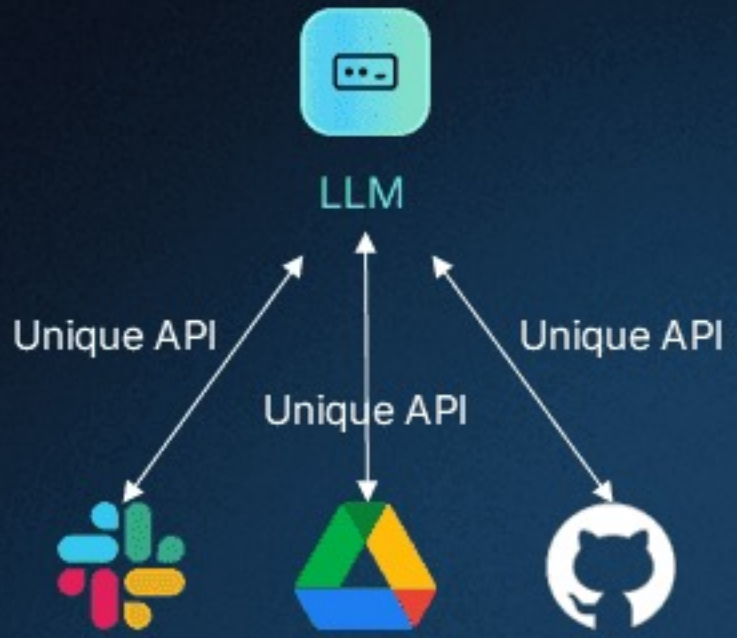
[MCP Protocol Layer]

↓

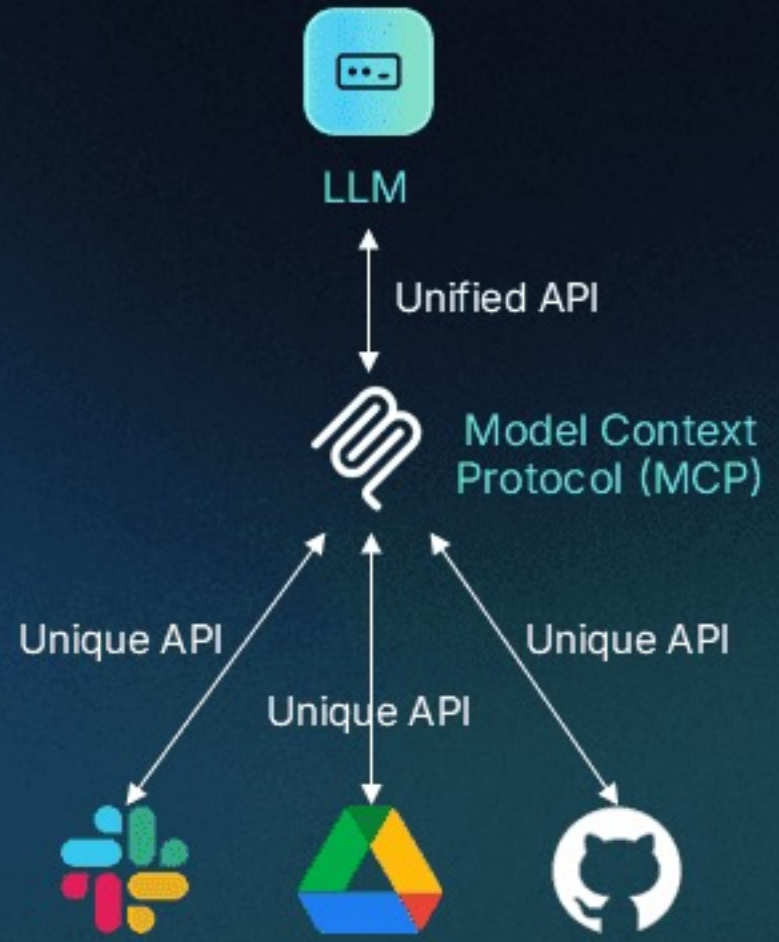
[Tools / APIs / Databases / Systems]

descope

Before MCP

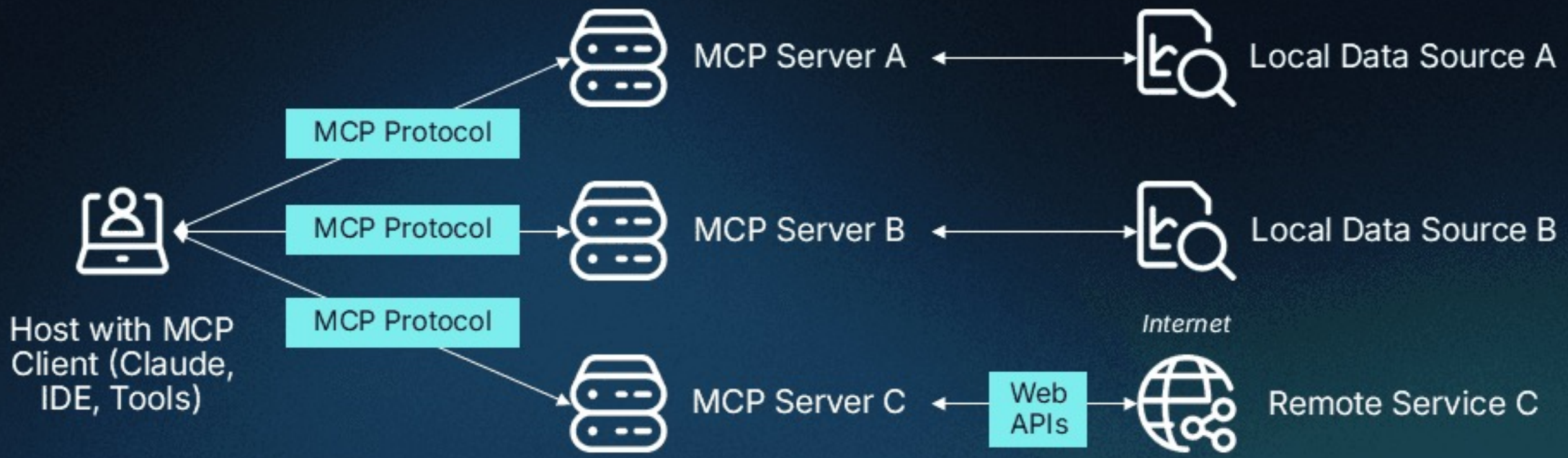


After MCP



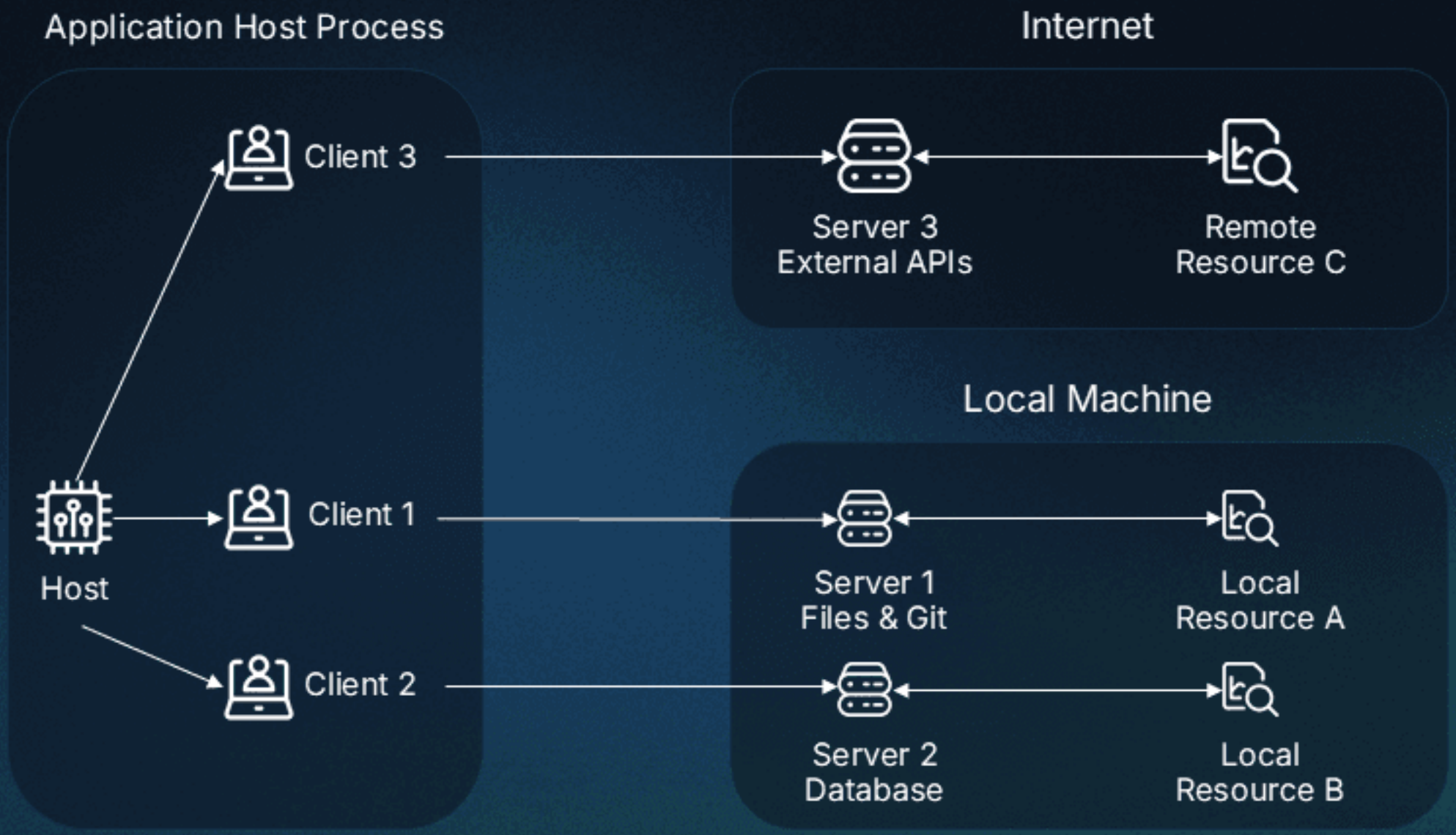
MCP Architecture

descope



MCP Core Components

descope



Why Do We Need MCP Server?

Large Language Models

Can **think**, **reason**, **write**, and **chat**.

But **cannot directly act**

Why Do We Need MCP Server?

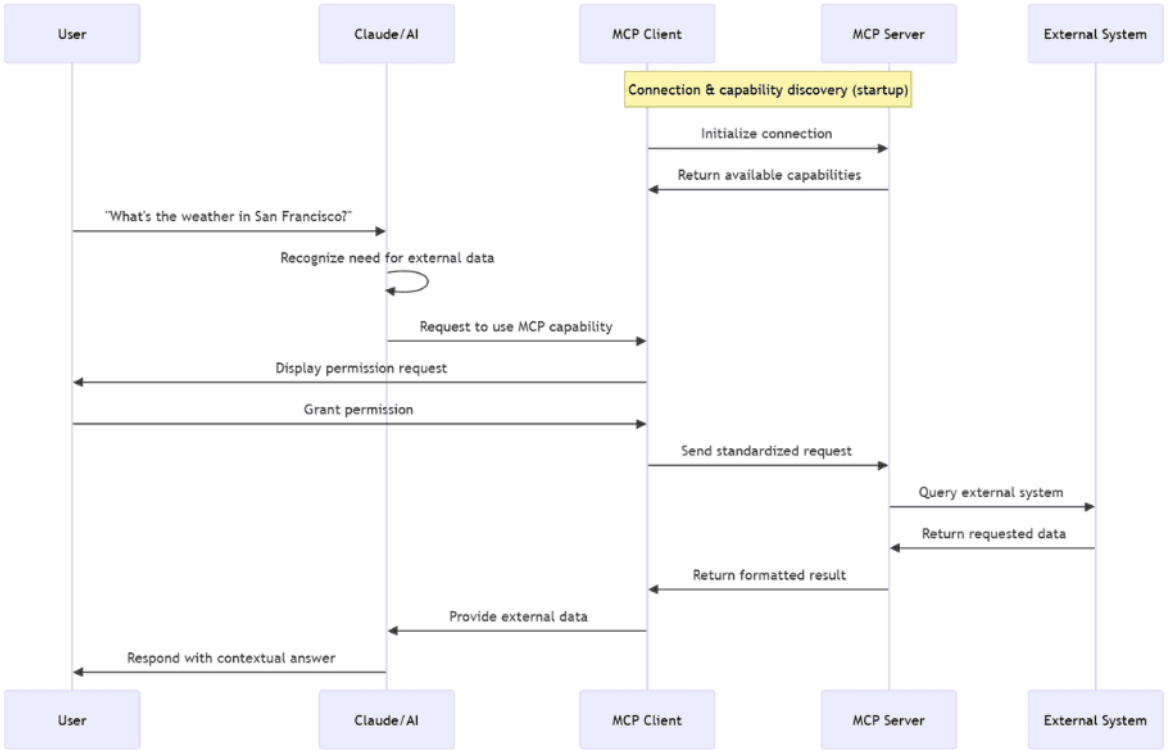
MCP Server bridges

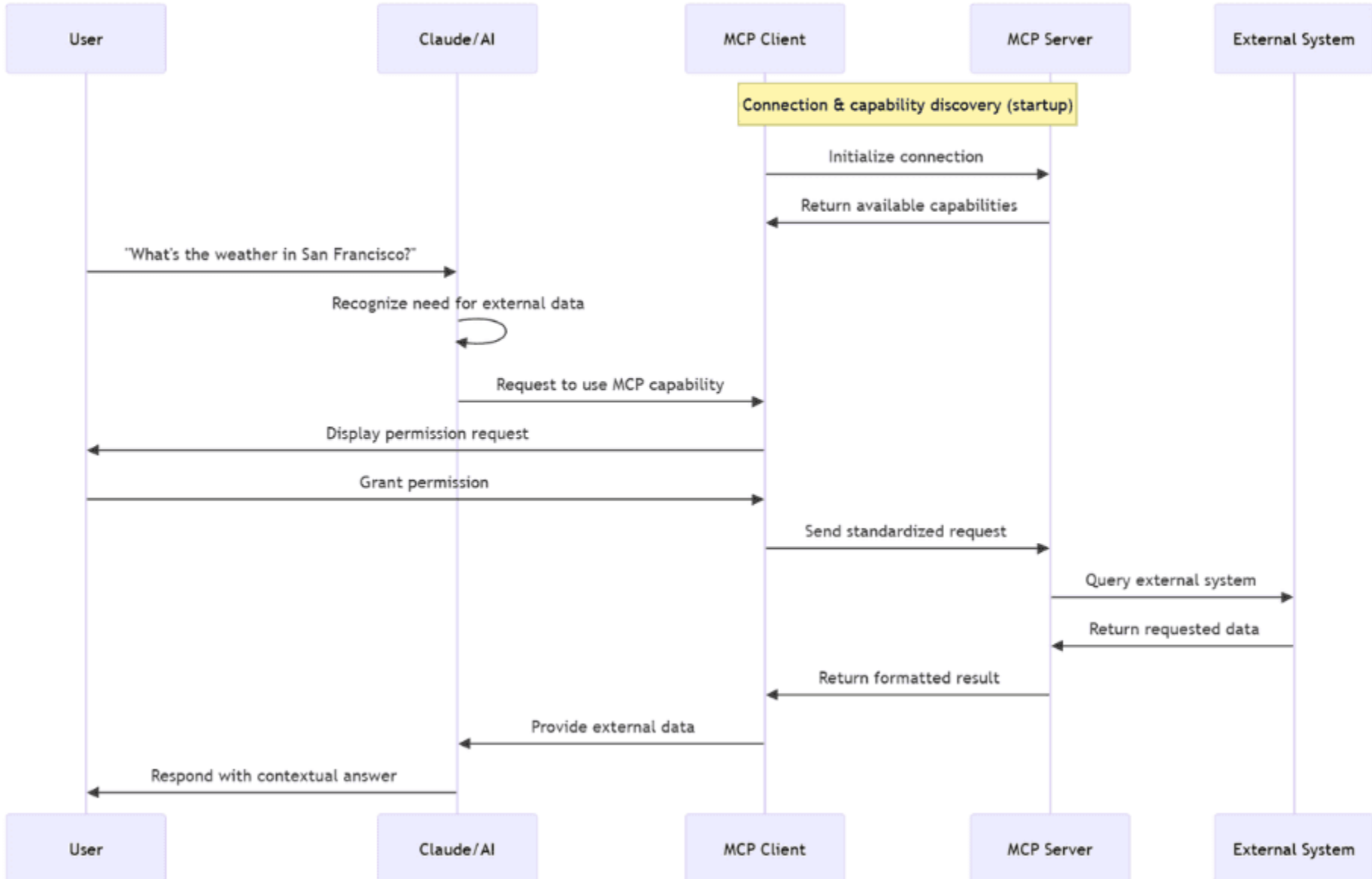
this gap by letting AI interact

with the real world

securely and efficiently.

Protocol handshake




User
Claude/AI
MCP Client
MCP Server
External System
Connection & capability discovery (startup)
Initialize connection
Return available capabilities
"What's the weather in San Francisco?"
Recognize need for external data
Request to use MCP capability
Display permission request
Grant permission
Send standardized request
Query external system
Return requested data
Return formatted result
Provide external data
Respond with contextual answer
User
Claude/AI
MCP Client
MCP Server
External System

Initial connection

When an MCP client

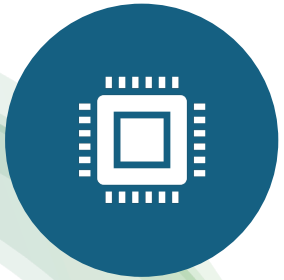
like Claude Desktop starts up

Connects to the configured MCP servers on your device.

Capability discovery

The client asks each server

"What capabilities do you offer?”



Each server responds with its

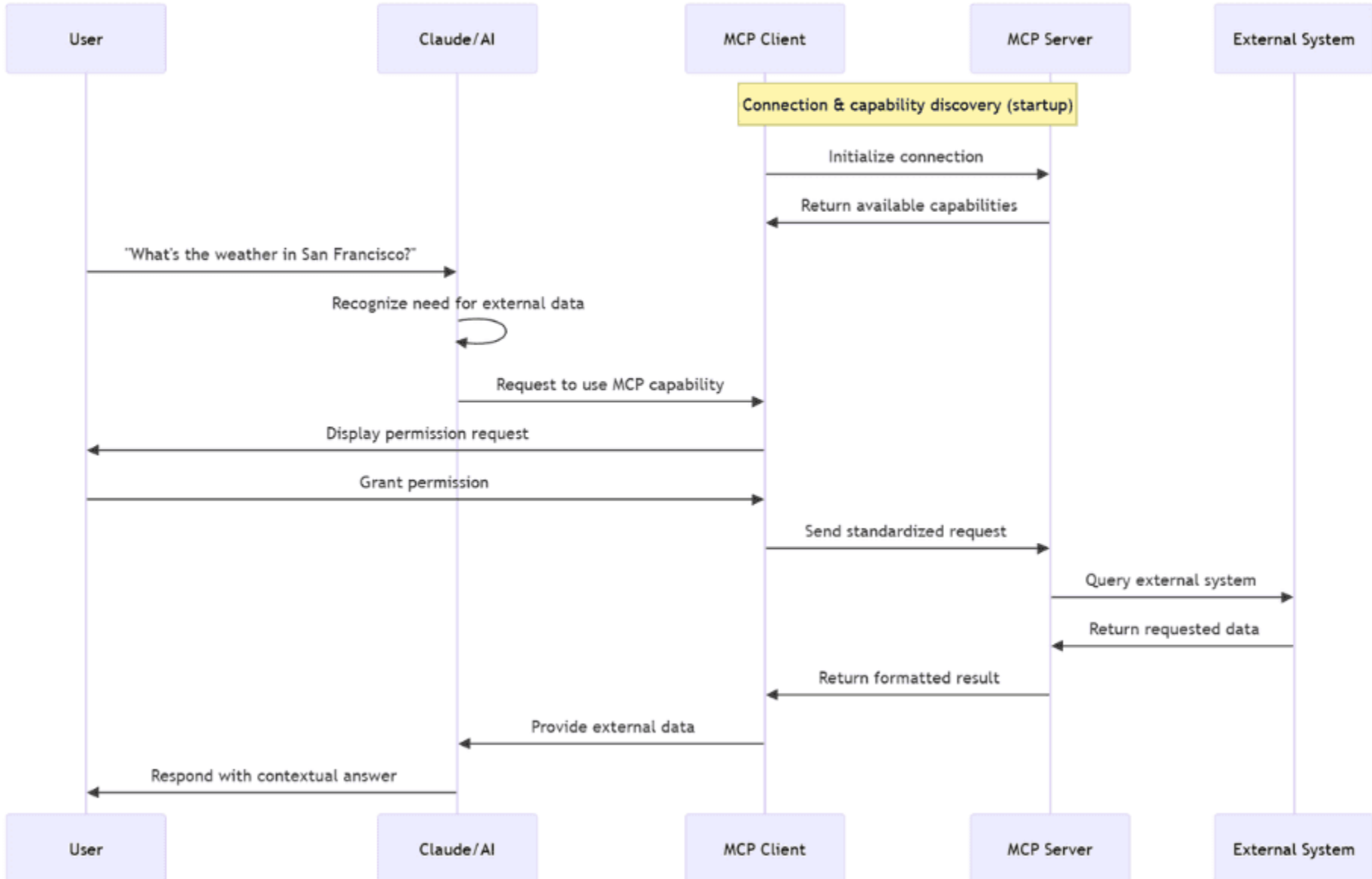
available tools, resources, and prompts.

Registration

The client registers these capabilities,

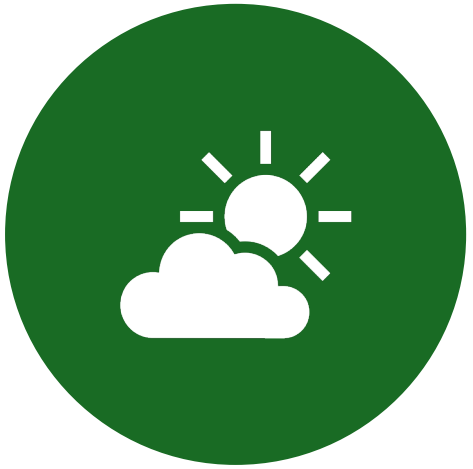
making them available for the AI

to use during your conversation.


User
Claude/AI
MCP Client
MCP Server
External System
Connection & capability discovery (startup)
Initialize connection
Return available capabilities
"What's the weather in San Francisco?"
Recognize need for external data
Request to use MCP capability
Display permission request
Grant permission
Send standardized request
Query external system
Return requested data
Return formatted result
Provide external data
Respond with contextual answer
User
Claude/AI
MCP Client
MCP Server
External System

From user request to external data

LET'S SAY YOU ASK CLAUDE



"WHAT'S THE WEATHER LIKE IN SAN FRANCISCO TODAY?"

Need recognition

CLAUDE ANALYZES YOUR QUESTION AND

RECOGNIZES IT NEEDS EXTERNAL,

REAL-TIME INFORMATION

THAT WASN'T IN ITS TRAINING DATA.

Tool or resource selection

Claude identifies that

it needs to use an MCP capability

to fulfill your request.

Permission request

The client displays a permission prompt

asking if you want to allow access

to the external tool or resource.

Information exchange

Once approved,

the client sends a request

to the appropriate MCP server

using the standardized protocol format.

External processing

The MCP server processes the request,

performing whatever action is needed

querying a weather service,

reading a file, or

accessing a database.

Result return

THE SERVER RETURNS

THE REQUESTED INFORMATION

TO THE CLIENT IN A STANDARDIZED FORMAT.

Context integration

Claude receives

this information and

incorporates it into

its understanding of the conversation.

Response generation

Claude generates a response

that includes the external information,

providing you with an answer

based on current data.

RAG vs Agentic AI vs MCP

Feature / Aspect	RAG	Agentic AI	MCP
Main focus	Retrieval + grounded generation	Planning, reasoning, tool orchestration	Standardized context & tool interface
Handles multi-step tasks	❌	✅	❌ (but can be used by agents)
Requires vector DB	Usually	Optional	Optional

RAG vs Agentic AI vs MCP

Feature / Aspect	RAG	Agentic AI	MCP
Tool/API integration	Minimal	Core feature	Yes, as standardized MCP tools
Interoperability	❌	❌ (custom per agent)	✅ cross-app/LLM
Example in Walmart	Return policy Q&A	Return eligibility + logistics	Serve policy DB & inventory API to any LLM client

References

https://medium.com/@HazlanRozaimi/generative-ai-reference-architecture-example-for-enterprises-260e95986da2

https://dr-arsanjani.medium.com/the-genai-reference-architecture-605929ab6b5a#cf5c

References

https://www.descope.com/learn/post/mcp

https://modelcontextprotocol.io/introduction

https://youtu.be/GQDHxlKJe_M

https://codingscape.com/blog/how-model-context-protocol-mcp-works-connect-ai-agents-to-tools

https://github.com/modelcontextprotocol

References

https://github.com/modelcontextprotocol/python-sdk?tab=readme-ov-file#mcp-python-sdk

https://claude.ai/public/artifacts/aed32faf-a9bc-43b8-8fd0-eb104a0cb261

https://claude.ai/public/artifacts/0a8124b7-3e44-4ba4-a159-b29669fcc799

Happy Learning!! Thanks for Your Patience ☺

Surendra Panpaliya

GKTCS Innovations