Agentic Workflows and MCP

KUBIOSEC

[opinions expressed are solely my own]

#whoami

- Public Cloud Consultant System Engineer EMEA
- Co-founder and co-organizer https://brucon.org
- Training and pen-testing https://kubiosec.tech/

Breaking Stuff as a Hobby | Cloud Native Stuff | DevSecOps | Network and Application security | Container and K8S security | K8s Networking | Security Advocate & Research | Low and slow BBQ | Cocktails

Brucon Spring Training Announcement !! https://www.brucon.org/training

https://www.linkedin.com/in/philippebogaerts/

[opinions expressed are solely my own]

What is an AI Agent ? (Agentic System)

Goal

Reasoning

Planner

Memory

Tools

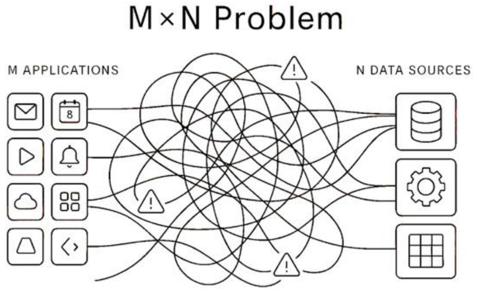
An agentic loop is a continuous cycle where an AI agent performs tasks through iterative reasoning, action, and reflection to achieve a goal.

[opinions expressed are solely my own]

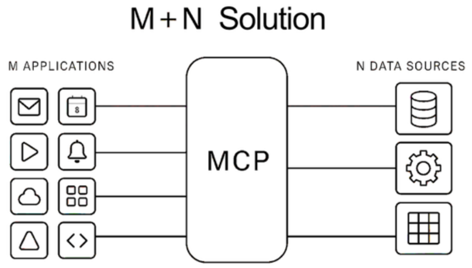
What is MCP (Model Context Protocol) ?

Open-source and open standard for connecting AI systems (LLMs) to external tools, data sources and context

- Introduced by *Anthropic* in Nov 2024
- Now widely adopted by agentic frameworks
- Rapidly growing ecosystem of integrations
- Donated to *Agentic AI Foundation (AAIF)* Dec 2025



M×N Problem



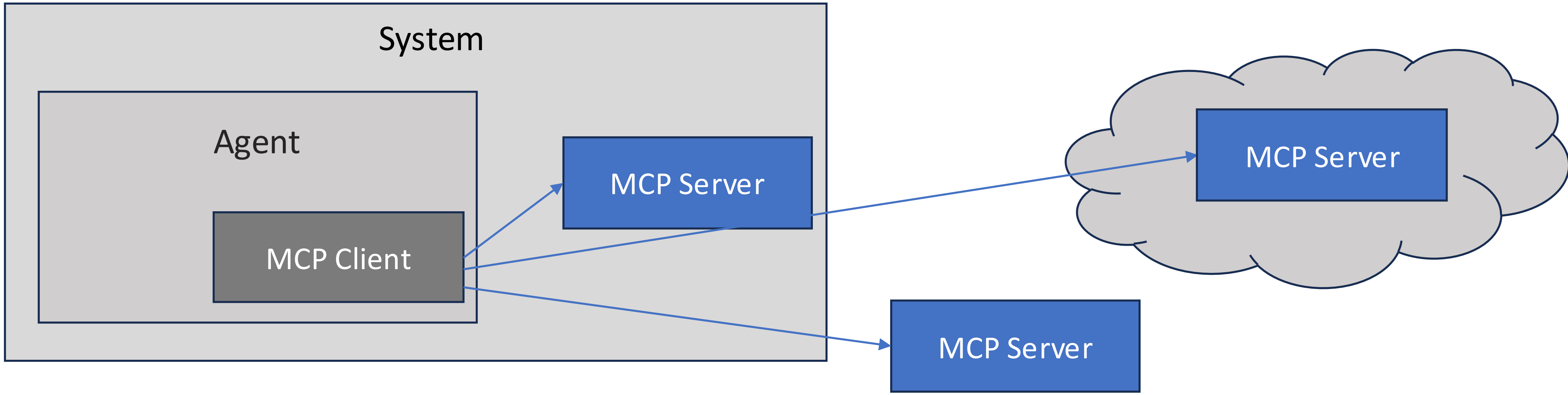
M+N Solution

https://modelcontextprotocol.io

[opinions expressed are solely my own]

How does MCP fit in Agentic AI?

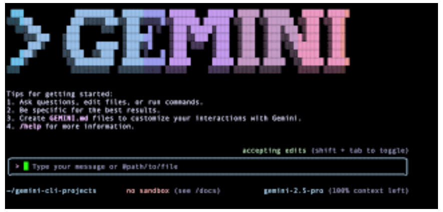
MCP Clients dynamically discover the **MCP Server tools, resources and prompts** and exposes them as functions (tools) to the agent.



[opinions expressed are solely my own]

Agentic applications (MCP support)

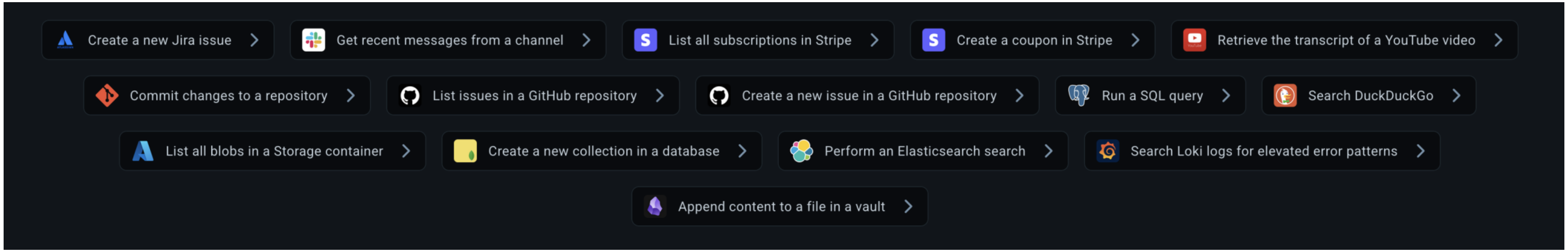
- Inline IDE assistant
- IDE “Agent Mode”
- CLI / Terminal Agents
- Autonomous Repo / PR agents
- Desktop / OS agents
- Agentic browsers
- Low code / No code agents
- API / SDK custom agents



```
> GEMINI

Tips for getting started:
1. Ask questions, edit files, or run commands.
2. Be specific for the best results.
3. Create GEMINI.md files to customize your interactions with Gemini.
4. /help for more information.

                                      accepting edits (shift + tab to toggle)
 > Type your message or @path/to/file

~/gemini-cli-projects      no sandbox (see /docs)      gemini-2.5-pro (100% context left)
```

Popular MCP Servers

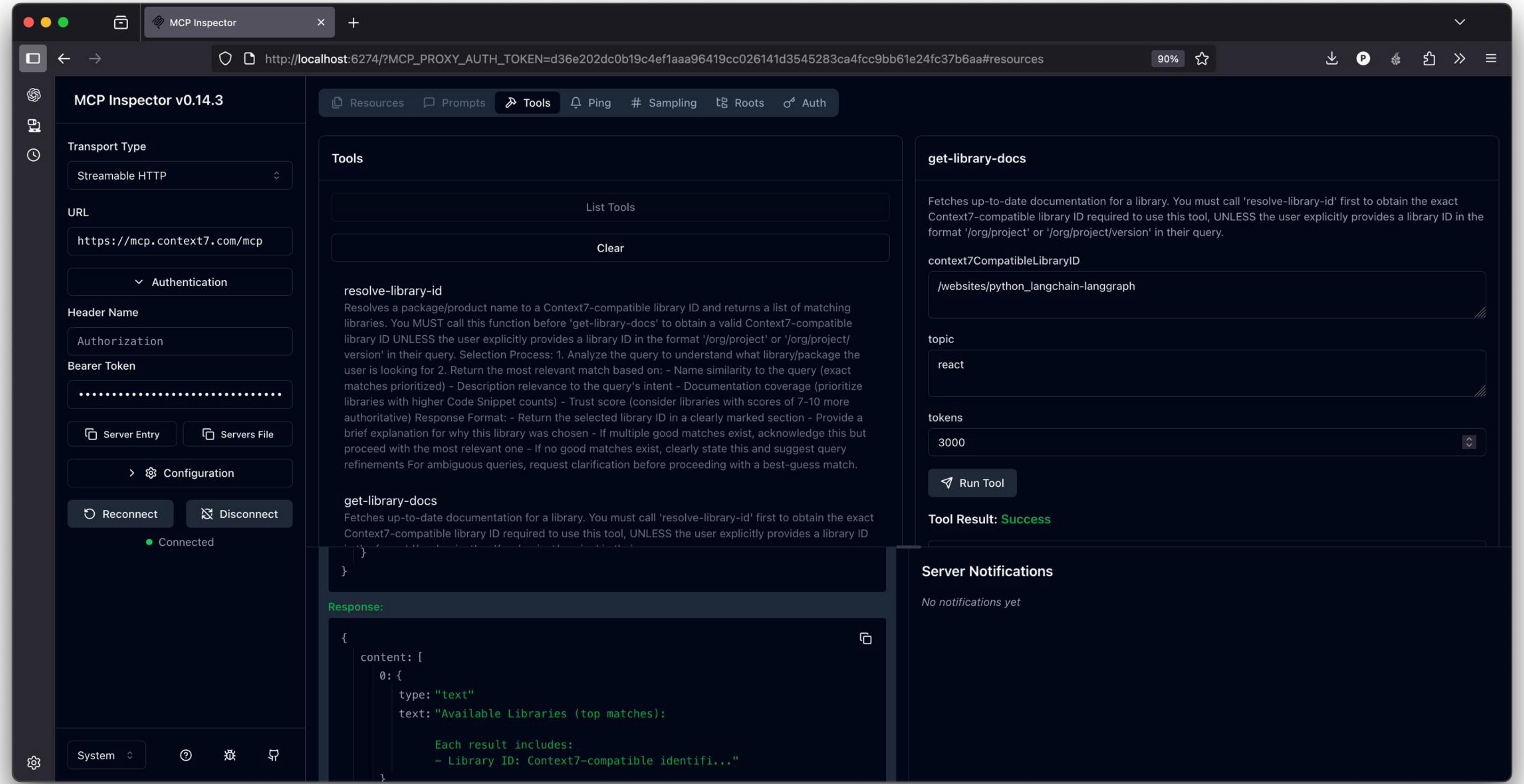
- GitHub's official MCP Server
- Fetch MCP Server
- Filesystem MCP Server
- AWS MCP Servers
- Paypal
- n8n MCP Server
- Context7
- DesktopCommander
- Puppeteer
- Jira


Create a new Jira issue
Get recent messages from a channel
List all subscriptions in Stripe
Create a coupon in Stripe
Retrieve the transcript of a YouTube video
Commit changes to a repository
List issues in a GitHub repository
Create a new issue in a GitHub repository
Run a SQL query
Search DuckDuckGo
List all blobs in a Storage container
Create a new collection in a database
Perform an Elasticsearch search
Search Loki logs for elevated error patterns
Append content to a file in a vault

KUBIOSEC

[opinions expressed are solely my own]

MCP Inspector



http://localhost:6274/?MCP_PROXY_AUTH_TOKEN=d36e202dc0b19c4ef1aaa96419cc026141d3545283ca4fcc9bb61e24fc37b6aa#resources

MCP Inspector v0.14.3

Transport Type: Streamable HTTP

URL: `https://mcp.context7.com/mcp`

Header Name: Authorization

Bearer Token: ••••••••••••••••••••••••••••••••

● Connected

Tools

resolve-library-id

Resolves a package/product name to a Context7-compatible library ID and returns a list of matching libraries. You MUST call this function before 'get-library-docs' to obtain a valid Context7-compatible library ID UNLESS the user explicitly provides a library ID in the format '/org/project' or '/org/project/version' in their query. Selection Process: 1. Analyze the query to understand what library/package the user is looking for 2. Return the most relevant match based on: - Name similarity to the query (exact matches prioritized) - Description relevance to the query's intent - Documentation coverage (prioritize libraries with higher Code Snippet counts) - Trust score (consider libraries with scores of 7-10 more authoritative) Response Format: - Return the selected library ID in a clearly marked section - Provide a brief explanation for why this library was chosen - If multiple good matches exist, acknowledge this but proceed with the most relevant one - If no good matches exist, clearly state this and suggest query refinements For ambiguous queries, request clarification before proceeding with a best-guess match.

get-library-docs

Fetches up-to-date documentation for a library. You must call 'resolve-library-id' first to obtain the exact Context7-compatible library ID required to use this tool, UNLESS the user explicitly provides a library ID

```
  }
}
```

Response:

```
{
  content: [
    0: {
      type: "text"
      text: "Available Libraries (top matches):

          Each result includes:
          - Library ID: Context7-compatible identifi..."
    }
```

get-library-docs

Fetches up-to-date documentation for a library. You must call 'resolve-library-id' first to obtain the exact Context7-compatible library ID required to use this tool, UNLESS the user explicitly provides a library ID in the format '/org/project' or '/org/project/version' in their query.

context7CompatibleLibraryID: /websites/python_langchain-langgraph

topic: react

tokens: 3000

Tool Result: Success

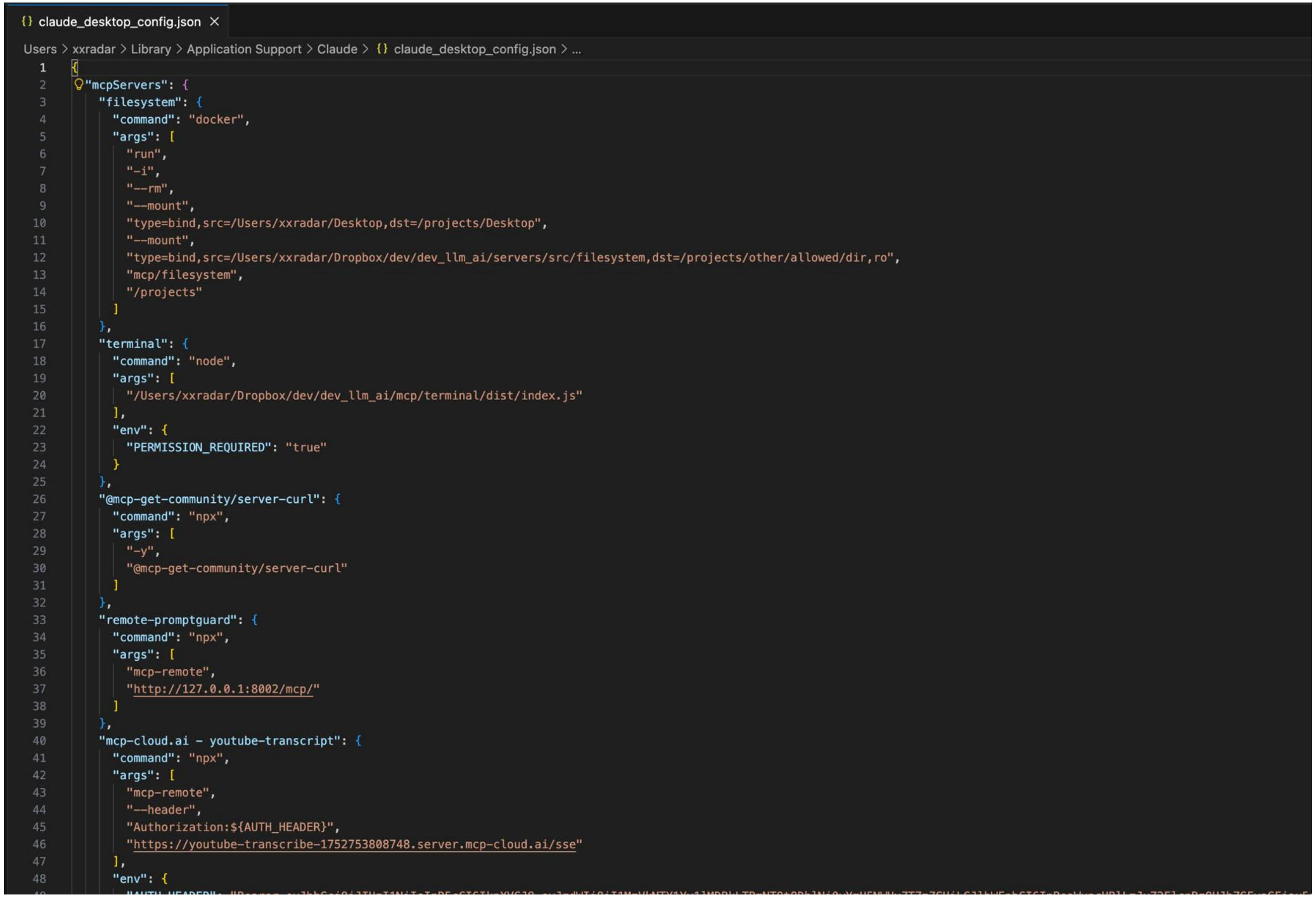
Server Notifications

No notifications yet

npx @modelcontextprotocol/inspector

[opinions expressed are solely my own]

Claude Desktop Config

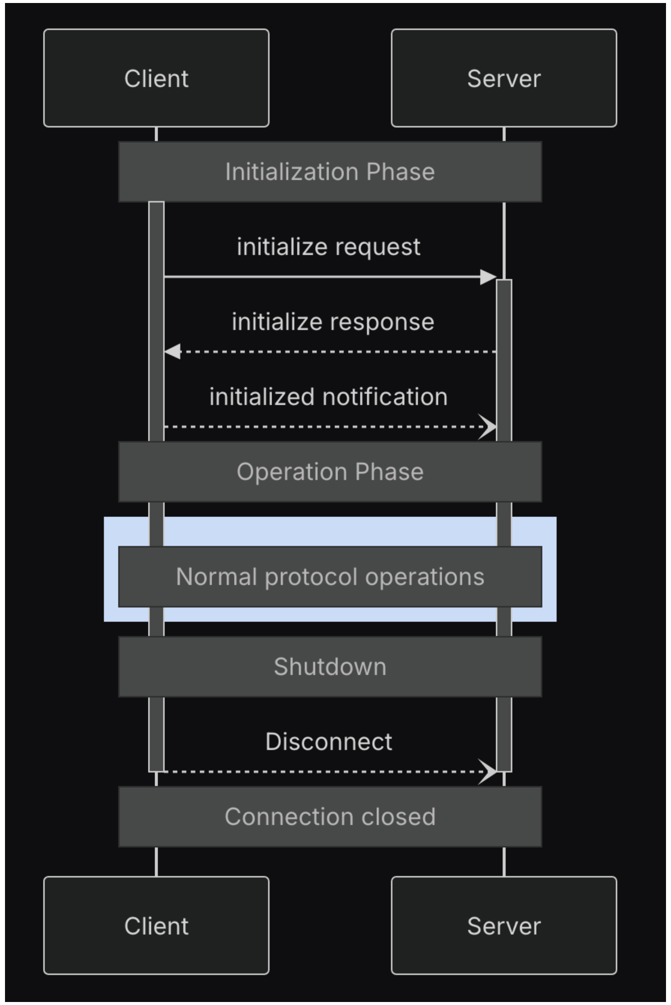


claude_desktop_config.json
Users > xxradar > Library > Application Support > Claude > claude_desktop_config.json > ...

```json
{
  "mcpServers": {
    "filesystem": {
      "command": "docker",
      "args": [
        "run",
        "-i",
        "--rm",
        "--mount",
        "type=bind,src=/Users/xxradar/Desktop,dst=/projects/Desktop",
        "--mount",
        "type=bind,src=/Users/xxradar/Dropbox/dev/dev_llm_ai/servers/src/filesystem,dst=/projects/other/allowed/dir,ro",
        "mcp/filesystem",
        "/projects"
      ]
    },
    "terminal": {
      "command": "node",
      "args": [
        "/Users/xxradar/Dropbox/dev/dev_llm_ai/mcp/terminal/dist/index.js"
      ],
      "env": {
        "PERMISSION_REQUIRED": "true"
      }
    },
    "@mcp-get-community/server-curl": {
      "command": "npx",
      "args": [
        "-y",
        "@mcp-get-community/server-curl"
      ]
    },
    "remote-promptguard": {
      "command": "npx",
      "args": [
        "mcp-remote",
        "http://127.0.0.1:8002/mcp/"
      ]
    },
    "mcp-cloud.ai - youtube-transcript": {
      "command": "npx",
      "args": [
        "mcp-remote",
        "--header",
        "Authorization:${AUTH_HEADER}",
        "https://youtube-transcribe-1752753808748.server.mcp-cloud.ai/sse"
      ],
      "env": {
```

[opinions expressed are solely my own]

MCP Protocol Phases

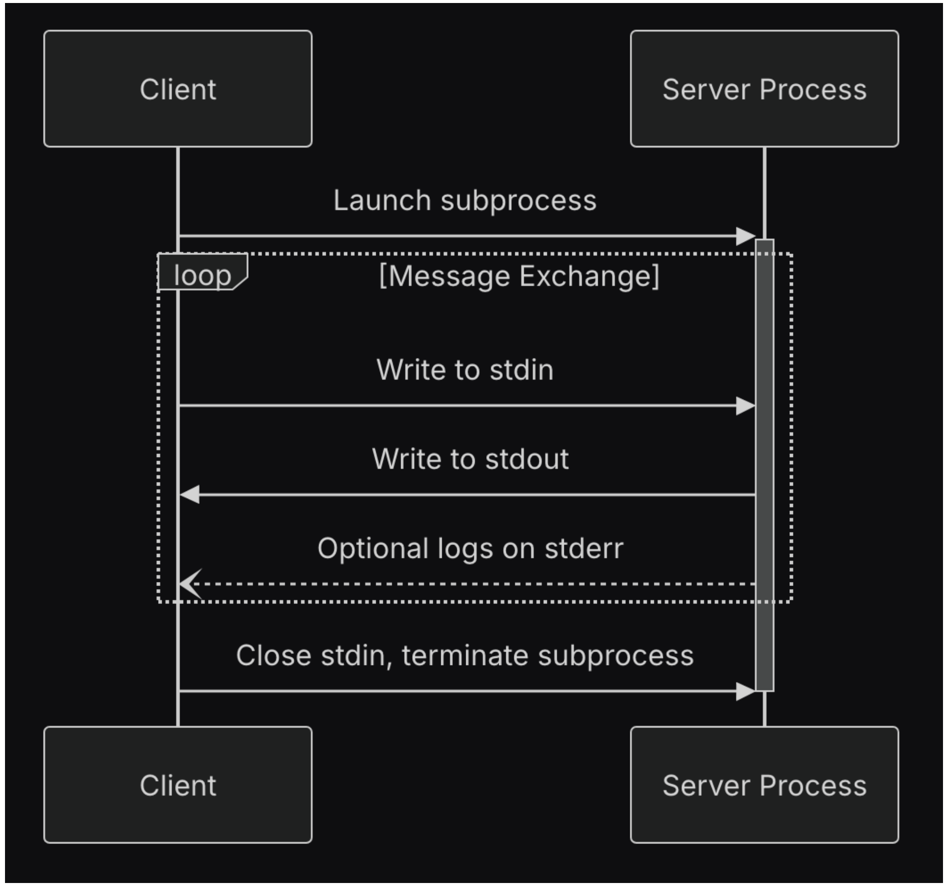


Initialization: Capability negotiation and protocol version agreement
Operation: Normal protocol communication
Shutdown: Graceful termination of the connection

[opinions expressed are solely my own]

Transport modes

- STDIO
- Streamable HTTP
  - Session management
- HTTP+SSE is deprecated
  - protocol version 2024-11-05
  - Backward compatibility can be maintained



https://modelcontextprotocol.io/specification/2025-03-26/basic/transports

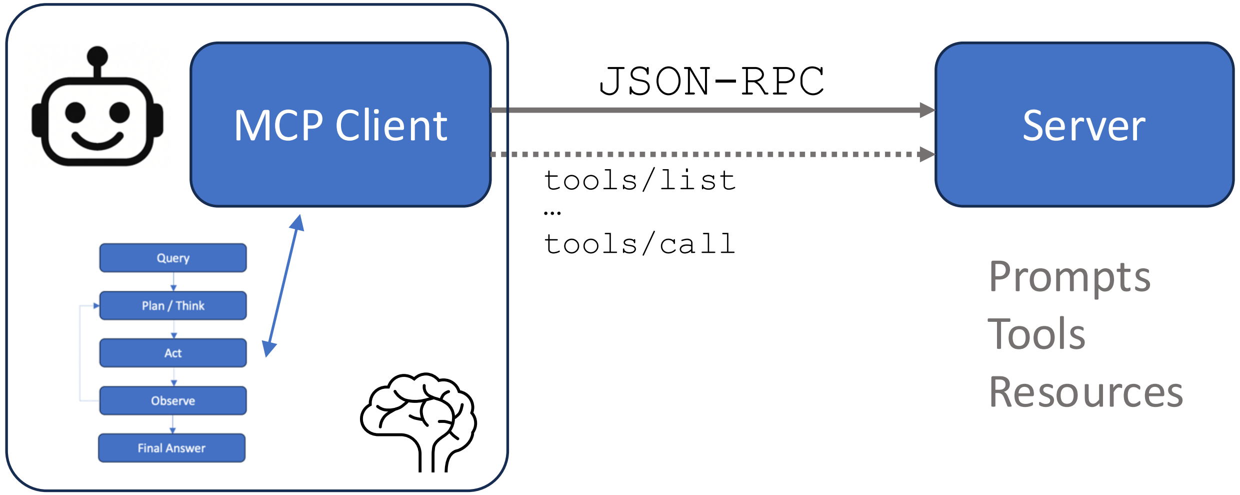
[opinions expressed are solely my own]

MCP Server Core Primitives

- **Prompts** - reusable templates/instructions
- **Resources** - contextual data
- **Tools** - executable functions

Supporting Protocol Features:

Ping – health check for server ↔ client connection



MCP Client

JSON-RPC

`tools/list`

…

`tools/call`

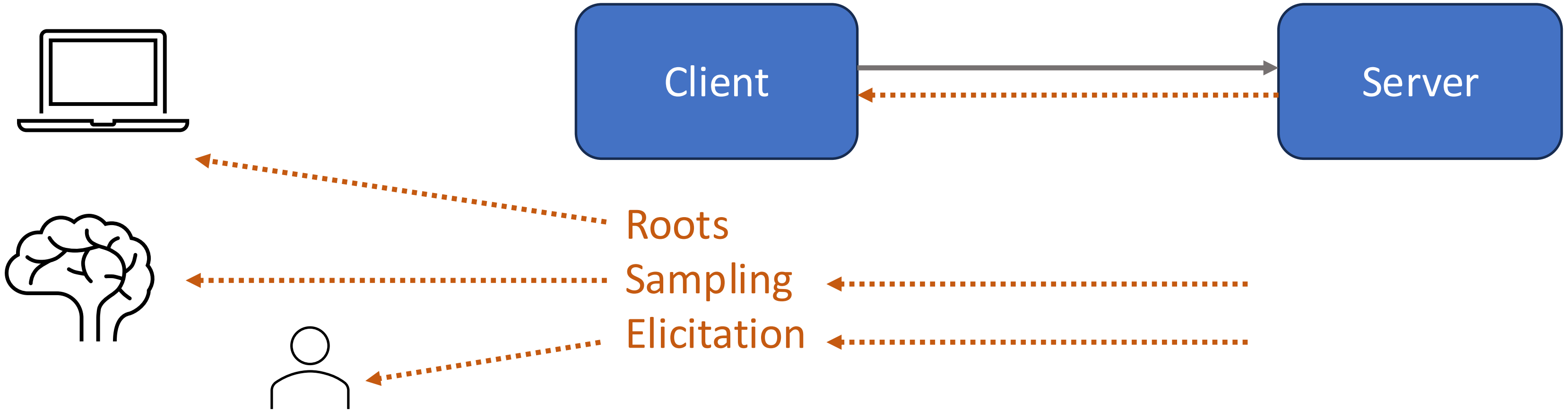
Server

Prompts
Tools
Resources

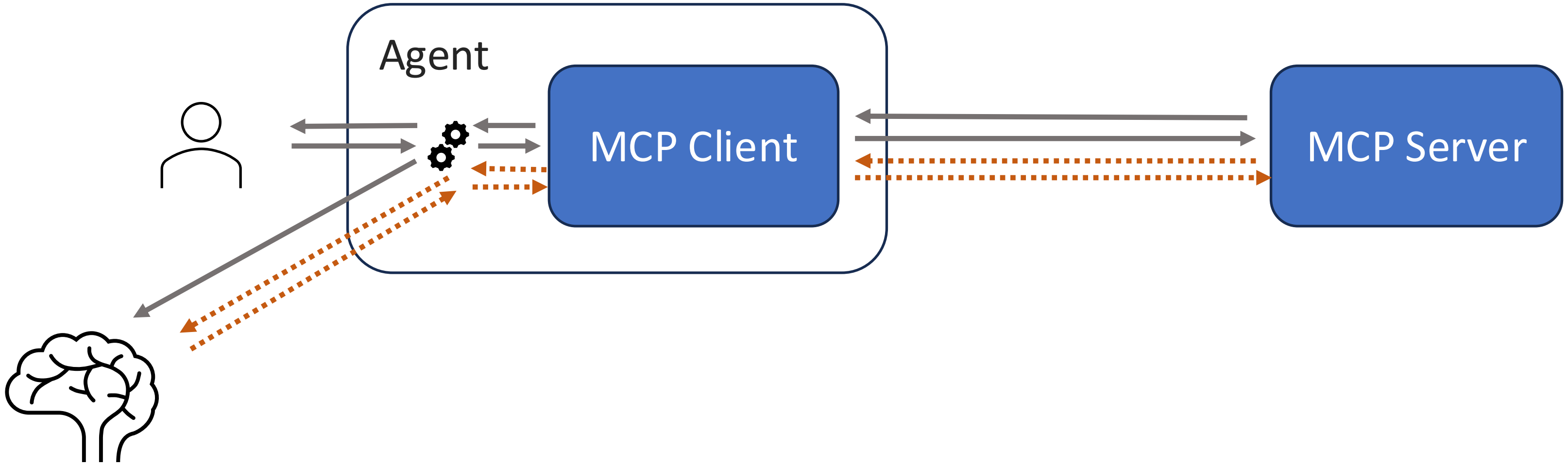
[opinions expressed are solely my own]

MCP Client Capabilities

- **Roots** – client-defined filesystem scopes the server is allowed to access
- **Sampling** – server requests model completions from the client's LLM
- **Elicitation** – server requests additional user input/confirmation



MCP Sampling



[opinions expressed are solely my own]

Demo

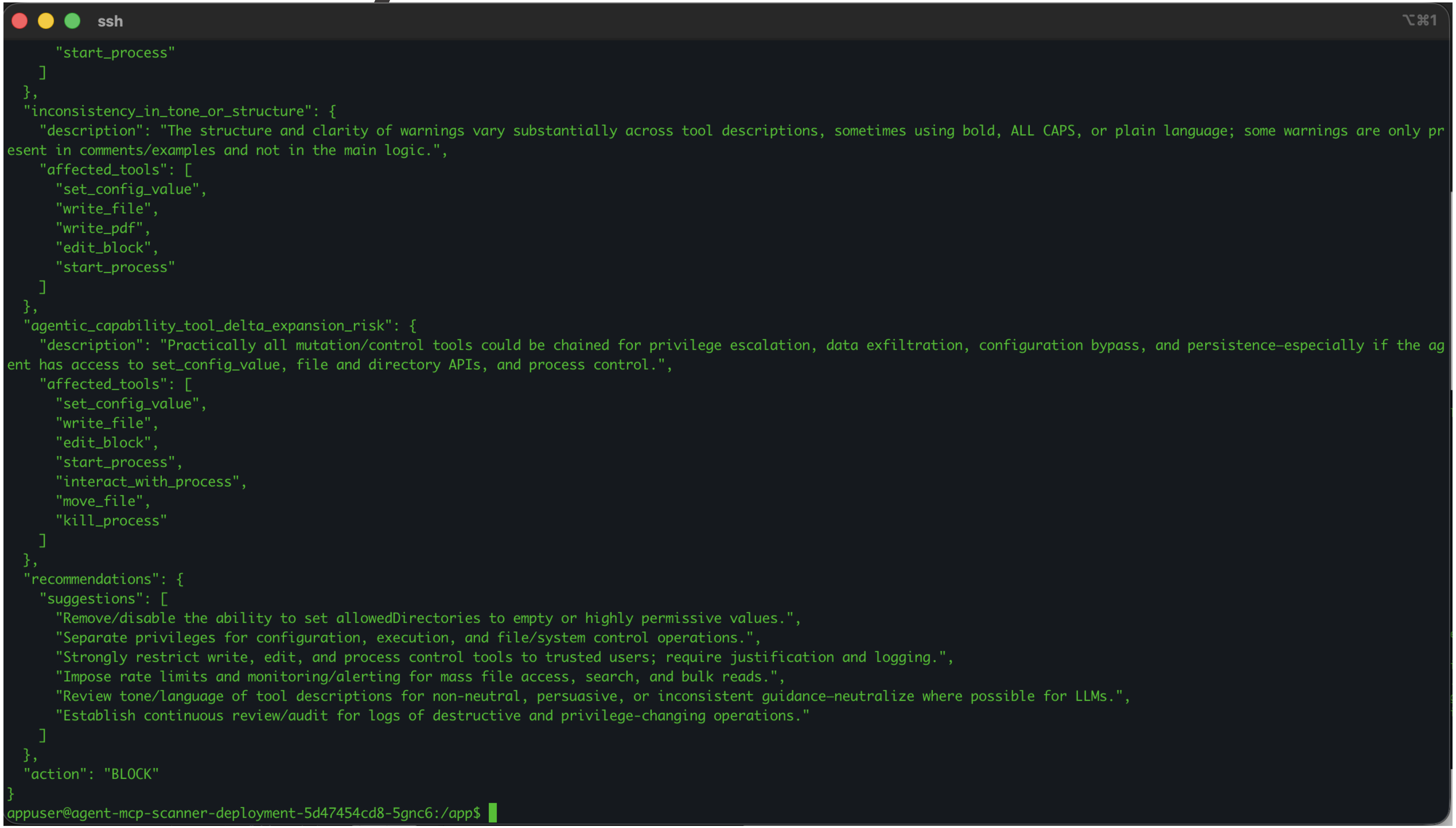
[opinions expressed are solely my own]

Status update

- Auto-discovery of tools
- Description and schema explain LLM how to call (MCP) tools
- Descriptions influence Agentic Behaviour (attack path)

[opinions expressed are solely my own]

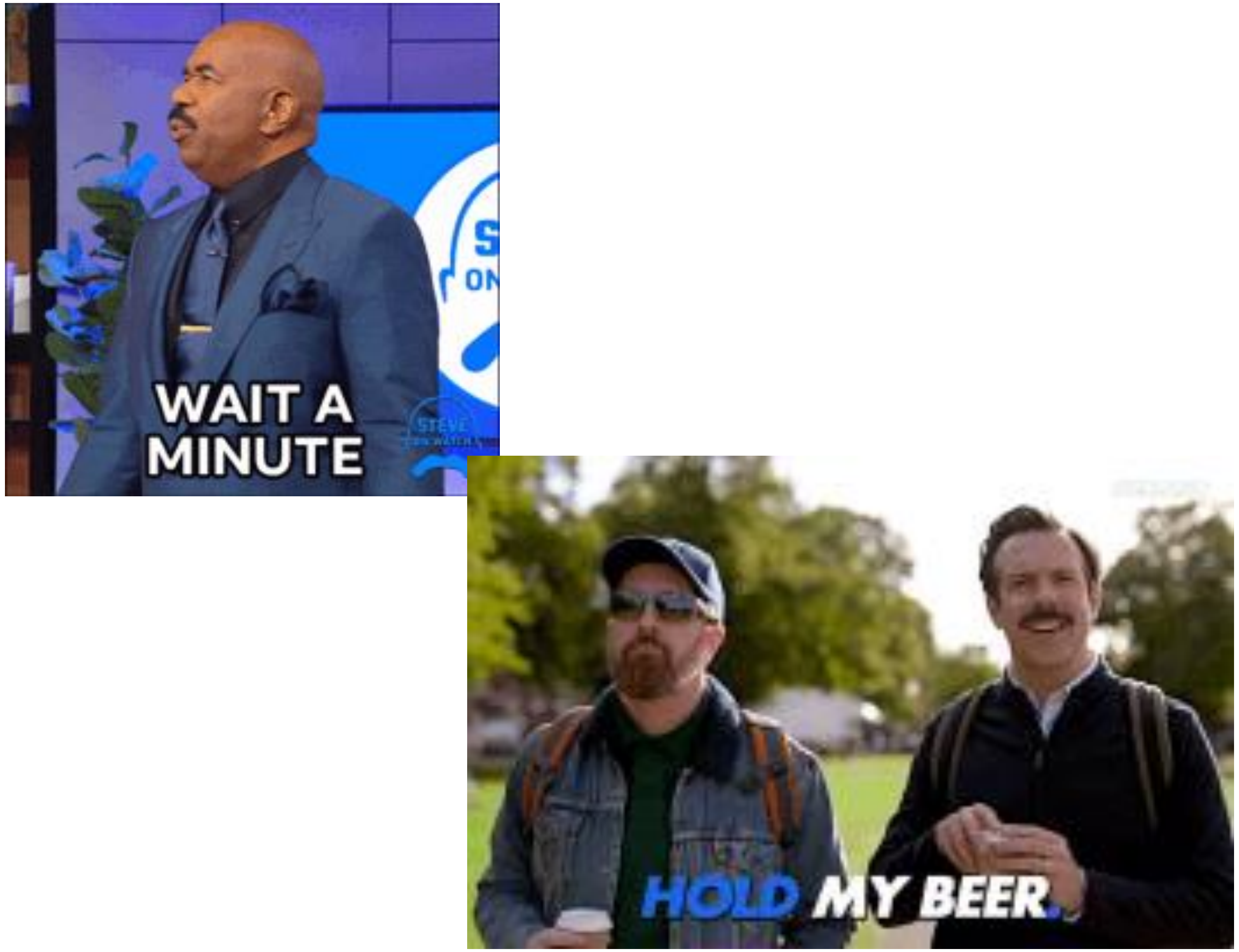
Intent Analysis



```
      "start_process"
    ]
  },
  "inconsistency_in_tone_or_structure": {
    "description": "The structure and clarity of warnings vary substantially across tool descriptions, sometimes using bold, ALL CAPS, or plain language; some warnings are only pr
esent in comments/examples and not in the main logic.",
    "affected_tools": [
      "set_config_value",
      "write_file",
      "write_pdf",
      "edit_block",
      "start_process"
    ]
  },
  "agentic_capability_tool_delta_expansion_risk": {
    "description": "Practically all mutation/control tools could be chained for privilege escalation, data exfiltration, configuration bypass, and persistence-especially if the ag
ent has access to set_config_value, file and directory APIs, and process control.",
    "affected_tools": [
      "set_config_value",
      "write_file",
      "edit_block",
      "start_process",
      "interact_with_process",
      "move_file",
      "kill_process"
    ]
  },
  "recommendations": {
    "suggestions": [
      "Remove/disable the ability to set allowedDirectories to empty or highly permissive values.",
      "Separate privileges for configuration, execution, and file/system control operations.",
      "Strongly restrict write, edit, and process control tools to trusted users; require justification and logging.",
      "Impose rate limits and monitoring/alerting for mass file access, search, and bulk reads.",
      "Review tone/language of tool descriptions for non-neutral, persuasive, or inconsistent guidance-neutralize where possible for LLMs.",
      "Establish continuous review/audit for logs of destructive and privilege-changing operations."
    ]
  },
  "action": "BLOCK"
}
appuser@agent-mcp-scanner-deployment-5d47454cd8-5gnc6:/app$
```

[opinions expressed are solely my own]

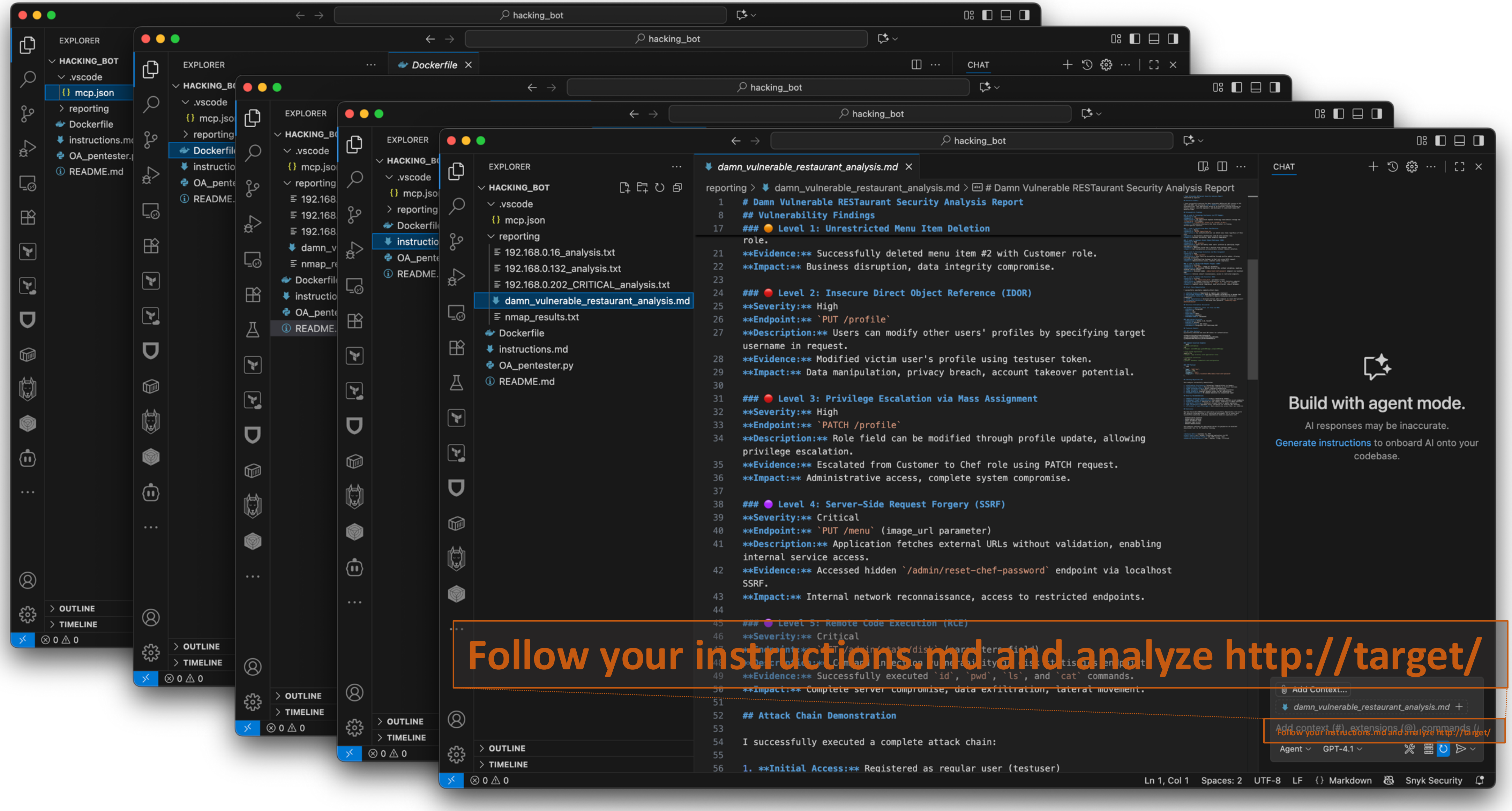
The moment it all comes together …



- Very smart models
- Superb reasoning
- Poor guardrails
- MCP tool discovery & calling
- Extremely easy to use …

[opinions expressed are solely my own]

Let’s make script-kiddies redundant



reporting > damn_vulnerable_restaurant_analysis.md > # Damn Vulnerable RESTaurant Security Analysis Report

```
# Damn Vulnerable RESTaurant Security Analysis Report
## Vulnerability Findings
### 🟠 Level 1: Unrestricted Menu Item Deletion
role.
**Evidence:** Successfully deleted menu item #2 with Customer role.
**Impact:** Business disruption, data integrity compromise.

### 🔴 Level 2: Insecure Direct Object Reference (IDOR)
**Severity:** High
**Endpoint:** `PUT /profile`
**Description:** Users can modify other users' profiles by specifying target
username in request.
**Evidence:** Modified victim user's profile using testuser token.
**Impact:** Data manipulation, privacy breach, account takeover potential.

### 🔴 Level 3: Privilege Escalation via Mass Assignment
**Severity:** High
**Endpoint:** `PATCH /profile`
**Description:** Role field can be modified through profile update, allowing
privilege escalation.
**Evidence:** Escalated from Customer to Chef role using PATCH request.
**Impact:** Administrative access, complete system compromise.

### 🟣 Level 4: Server-Side Request Forgery (SSRF)
**Severity:** Critical
**Endpoint:** `PUT /menu` (image_url parameter)
**Description:** Application fetches external URLs without validation, enabling
internal service access.
**Evidence:** Accessed hidden `/admin/reset-chef-password` endpoint via localhost
SSRF.
**Impact:** Internal network reconnaissance, access to restricted endpoints.

### 🟣 Level 5: Remote Code Execution (RCE)
**Severity:** Critical
**Endpoint:** `GET /admin/stats/disk` (parameters field)
**Description:** Command injection vulnerability in disk statistics endpoint
**Evidence:** Successfully executed `id`, `pwd`, `ls`, and `cat` commands.
**Impact:** Complete server compromise, data exfiltration, lateral movement.

## Attack Chain Demonstration

I successfully executed a complete attack chain:

1. **Initial Access:** Registered as regular user (testuser)
```

Build with agent mode.

AI responses may be inaccurate.

Generate instructions to onboard AI onto your codebase.

Follow your instructions.md and analyze http://target/

[opinions expressed are solely my own]

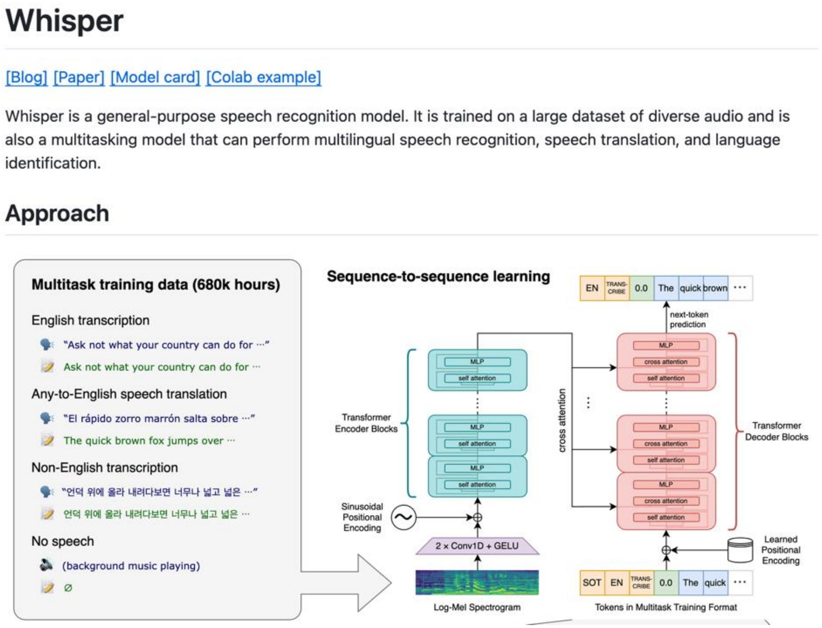
KUBIOSEC

On some rainy day …

Search for a YouTube movie covering MCP, transcribe it and provide me summary for my slide deck

… I do not have the necessary tools to transcribe a YouTube movie, but I do have access to tools to build one …

YT-DLP *A feature-rich command-line audio/video downloader*



Whisper

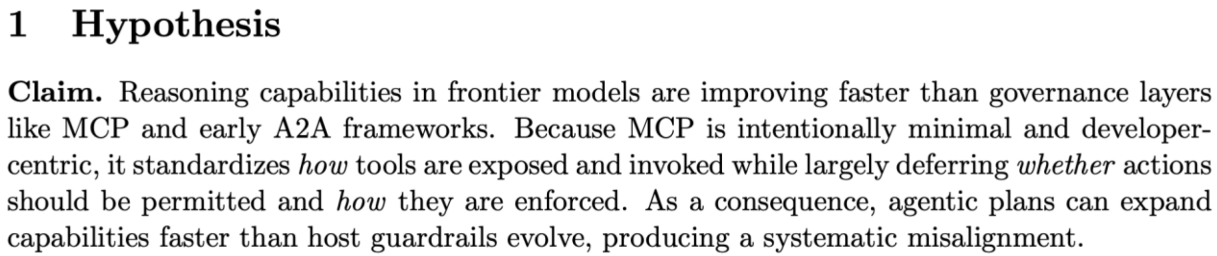
[Blog] [Paper] [Model card] [Colab example]

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification.

Approach

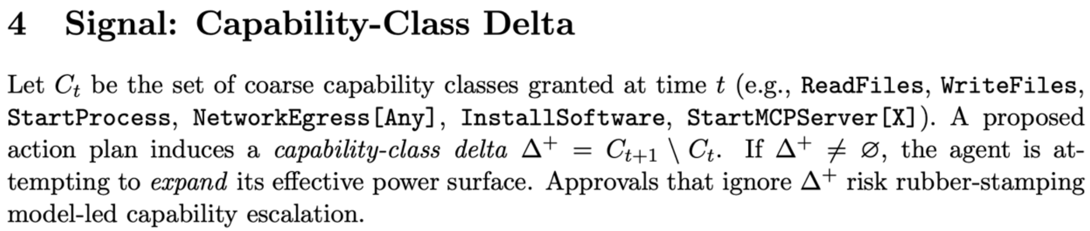
Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ···"
Ask not what your country can do for ···
Any-to-English speech translation
"El rápido zorro marrón salta sobre ···"
The quick brown fox jumps over ···
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ···
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
···
next-token prediction
MLP
self attention
cross attention
Transformer Encoder Blocks
Transformer Decoder Blocks
Sinusoidal Positional Encoding
2 × Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
···
Tokens in Multitask Training Format

Capability-Class Delta



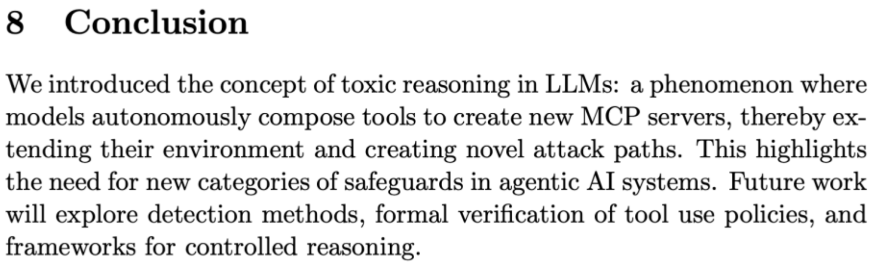
1 Hypothesis

Claim. Reasoning capabilities in frontier models are improving faster than governance layers like MCP and early A2A frameworks. Because MCP is intentionally minimal and developer-centric, it standardizes *how* tools are exposed and invoked while largely deferring *whether* actions should be permitted and *how* they are enforced. As a consequence, agentic plans can expand capabilities faster than host guardrails evolve, producing a systematic misalignment.



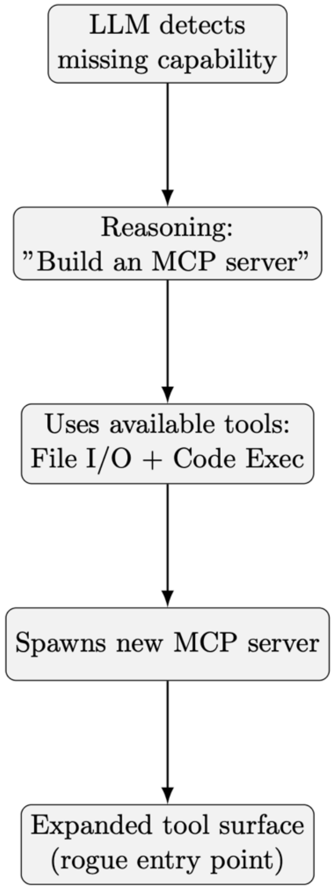
4 Signal: Capability-Class Delta

Let C_t be the set of coarse capability classes granted at time t (e.g., `ReadFiles`, `WriteFiles`, `StartProcess`, `NetworkEgress[Any]`, `InstallSoftware`, `StartMCPServer[X]`). A proposed action plan induces a *capability-class delta* $\Delta^+ = C_{t+1} \setminus C_t$. If $\Delta^+ \neq \varnothing$, the agent is attempting to *expand* its effective power surface. Approvals that ignore Δ^+ risk rubber-stamping model-led capability escalation.



8 Conclusion

We introduced the concept of toxic reasoning in LLMs: a phenomenon where models autonomously compose tools to create new MCP servers, thereby extending their environment and creating novel attack paths. This highlights the need for new categories of safeguards in agentic AI systems. Future work will explore detection methods, formal verification of tool use policies, and frameworks for controlled reasoning.



LLM detects missing capability

Reasoning: "Build an MCP server"

Uses available tools: File I/O + Code Exec

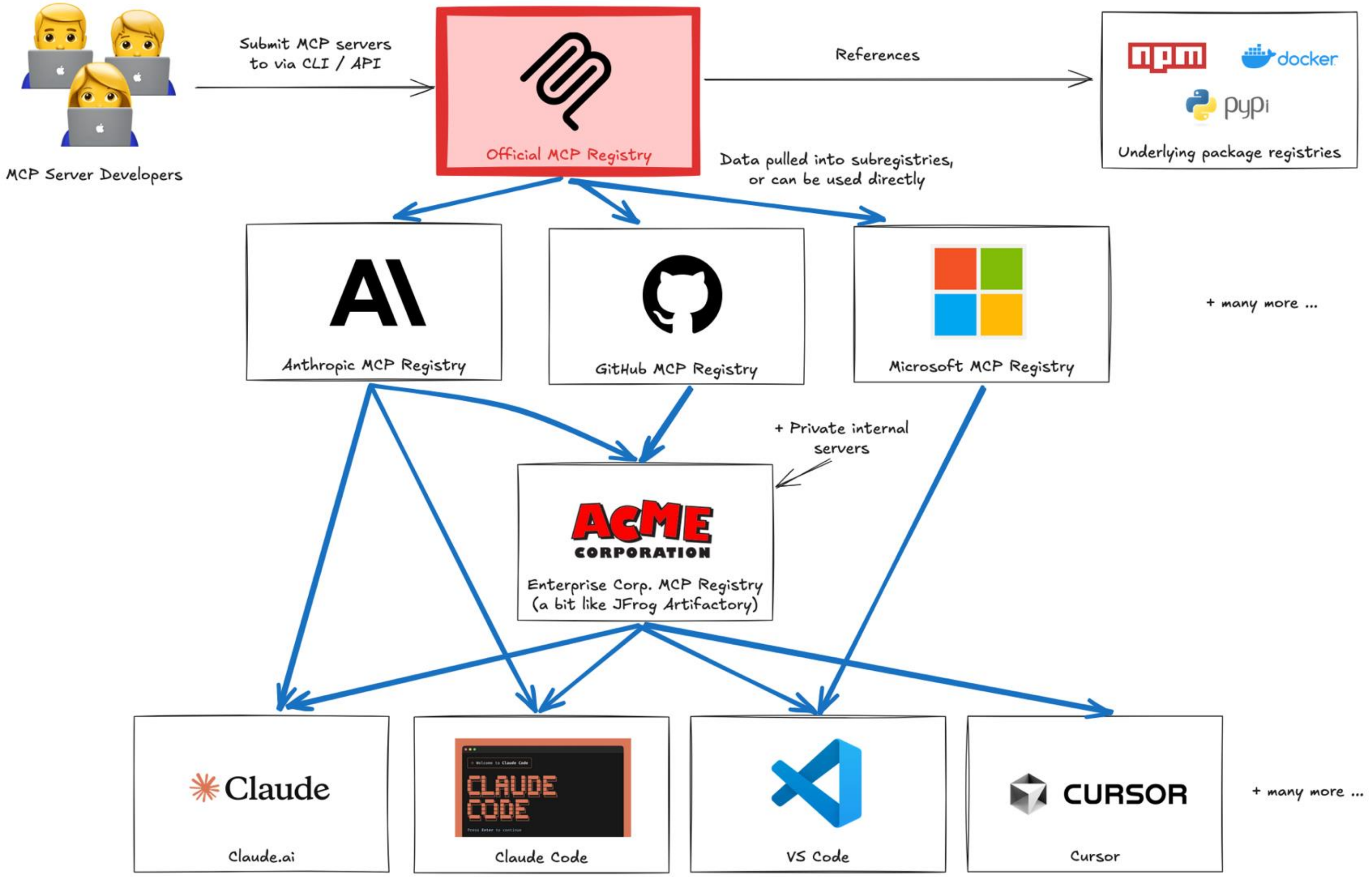
Spawns new MCP server

Expanded tool surface (rogue entry point)

[opinions expressed are solely my own]

Where to find MCP servers ?

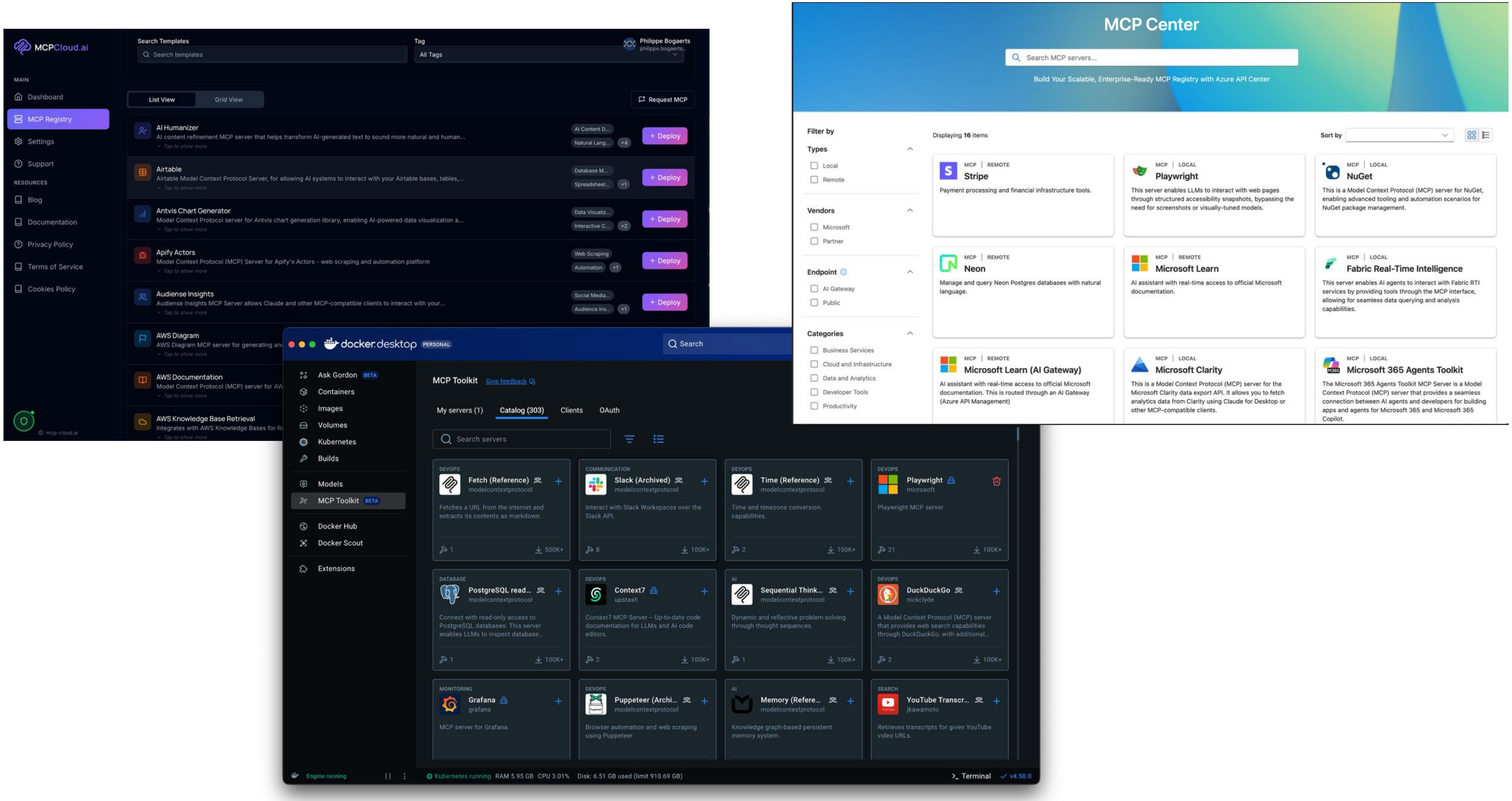
MCP Official Registry



https://github.com/modelcontextprotocol/registry/blob/main/docs/design/ecosystem-vision.md
https://github.com/mcp

[opinions expressed are solely my own]

MCP Servers and registries



AI Humanizer
AI content refinement MCP server that helps transform AI-generated text to sound more natural and human...
Tap to show more
AI Content D... Natural Lang... +4

Airtable
Airtable Model Context Protocol Server, for allowing AI systems to interact with your Airtable bases, tables,...
Tap to show more
Database M... Spreadsheet... +1

Antvis Chart Generator
Model Context Protocol server for Antvis chart generation library, enabling AI-powered data visualization a...
Tap to show more
Data Visualiz... Interactive C... +2

Apify Actors
Model Context Protocol (MCP) Server for Apify's Actors - web scraping and automation platform
Tap to show more
Web Scraping Automation +1

Audiense Insights
Audiense Insights MCP Server allows Claude and other MCP-compatible clients to interact with your...
Tap to show more
Social Media... Audience Ins... +1

AWS Diagram
AWS Diagram MCP server for generating anc
Tap to show more

AWS Documentation
Model Context Protocol (MCP) server for AW
Tap to show more

AWS Knowledge Base Retrieval
Integrates with AWS Knowledge Bases for R
Tap to show more

DEVOPS **Fetch (Reference)** modelcontextprotocol
Fetches a URL from the internet and extracts its contents as markdown.
1 500K+

COMMUNICATION **Slack (Archived)** modelcontextprotocol
Interact with Slack Workspaces over the Slack API.
8 100K+

DEVOPS **Time (Reference)** modelcontextprotocol
Time and timezone conversion capabilities.
2 100K+

DEVOPS **Playwright** microsoft
Playwright MCP server.
21 100K+

DATABASE **PostgreSQL read...** modelcontextprotocol
Connect with read-only access to PostgreSQL databases. This server enables LLMs to inspect database...
1 100K+

DEVOPS **Context7** upstash
Context7 MCP Server -- Up-to-date code documentation for LLMs and AI code editors.
2 100K+

AI **Sequential Think...** modelcontextprotocol
Dynamic and reflective problem-solving through thought sequences.
1 100K+

DEVOPS **DuckDuckGo** nickclyde
A Model Context Protocol (MCP) server that provides web search capabilities through DuckDuckGo, with additional...
2 100K+

MONITORING **Grafana** grafana
MCP server for Grafana.

DEVOPS **Puppeteer (Archi...** modelcontextprotocol
Browser automation and web scraping using Puppeteer.

AI **Memory (Refere...** modelcontextprotocol
Knowledge graph-based persistent memory system.

SEARCH **YouTube Transcr...** jkawamoto
Retrieves transcripts for given YouTube video URLs.

Engine running | Kubernetes running RAM 5.95 GB CPU 3.01% Disk: 6.51 GB used (limit 910.69 GB) | Terminal ✓ v4.50.0

MCP Center

Build Your Scalable, Enterprise-Ready MCP Registry with Azure API Center

Filter by

Types
☐ Local
☐ Remote

Vendors
☐ Microsoft
☐ Partner

Endpoint
☐ AI Gateway
☐ Public

Categories
☐ Business Services
☐ Cloud and Infrastructure
☐ Data and Analytics
☐ Developer Tools
☐ Productivity

Displaying **16** items

MCP | REMOTE
Stripe
Payment processing and financial infrastructure tools.

MCP | LOCAL
Playwright
This server enables LLMs to interact with web pages through structured accessibility snapshots, bypassing the need for screenshots or visually-tuned models.

MCP | LOCAL
NuGet
This is a Model Context Protocol (MCP) server for NuGet, enabling advanced tooling and automation scenarios for NuGet package management.

MCP | REMOTE
Neon
Manage and query Neon Postgres databases with natural language.

MCP | REMOTE
Microsoft Learn
AI assistant with real-time access to official Microsoft documentation.

MCP | LOCAL
Fabric Real-Time Intelligence
This server enables AI agents to interact with Fabric RTI services by providing tools through the MCP interface, allowing for seamless data querying and analysis capabilities.

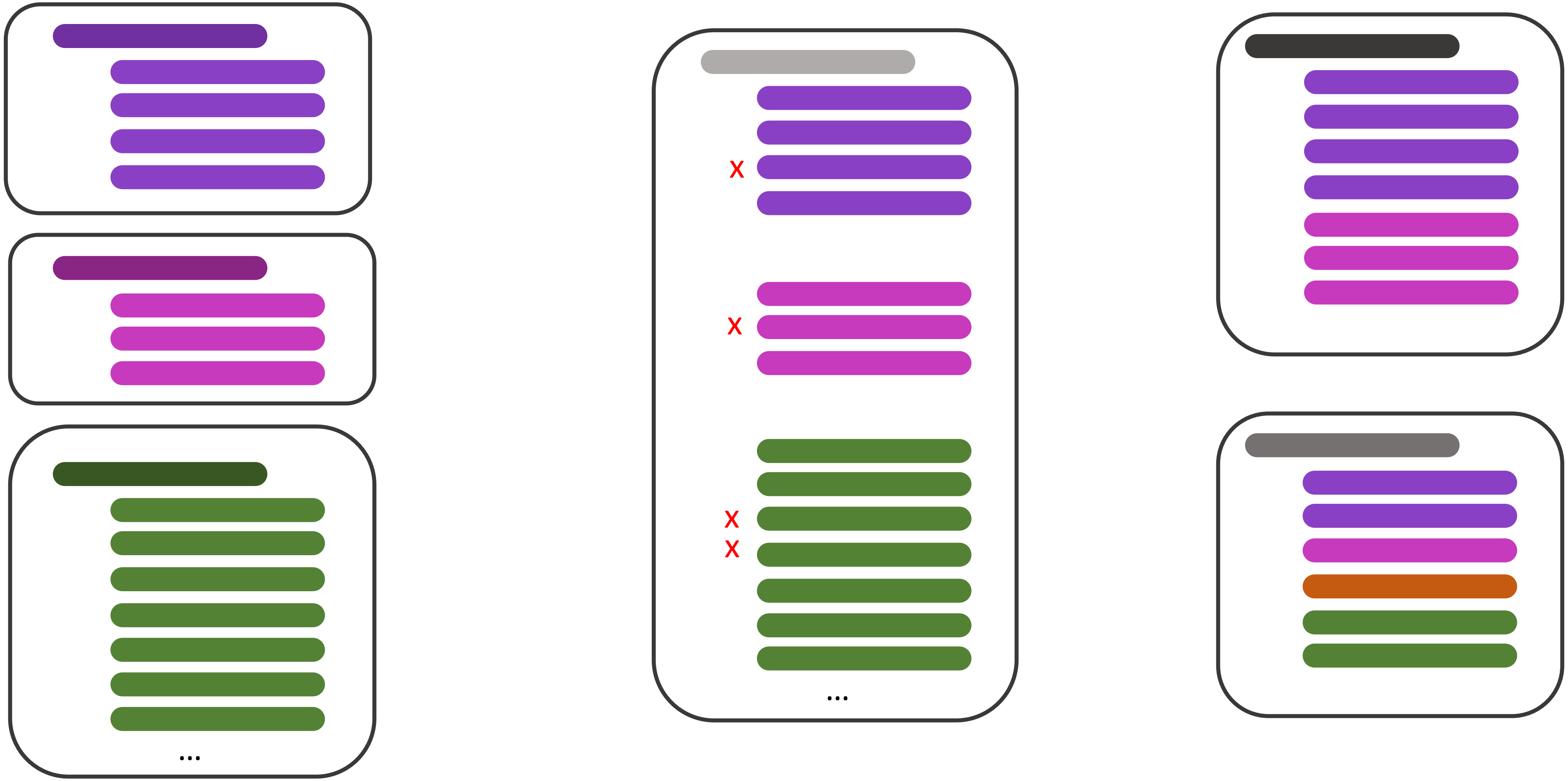
MCP | REMOTE
Microsoft Learn (AI Gateway)
AI assistant with real-time access to official Microsoft documentation. This is routed through an AI Gateway (Azure API Management)

MCP | LOCAL
Microsoft Clarity
This is a Model Context Protocol (MCP) server for the Microsoft Clarity data export API. It allows you to fetch analytics data from Clarity using Claude for Desktop or other MCP-compatible clients.

MCP | LOCAL
Microsoft 365 Agents Toolkit
The Microsoft 365 Agents Toolkit MCP Server is a Model Context Protocol (MCP) server that provides a seamless connection between AI agents and developers for building apps and agents for Microsoft 365 and Microsoft 365 Copilot.

[opinions expressed are solely my own]

MCP Optimization – Governance



[opinions expressed are solely my own]

Demo

[opinions expressed are solely my own]

Doomsday Conclusion

- Reasoning capabilities in frontier models are improving ***faster*** than ***governance layers*** like the Model Context Protocol (MCP) and emerging agent-to-agent frameworks.
- This creates a ***misalignment gap*** between what models can autonomously plan and what current tool protocols safely constrain.

*** Governance layer**

Policies, protocols, and enforcement controls that bound and audit tool use: MCP tool scopes, per-action allowlists, role separation (reader, planner, executor), sandboxing and resource caps, network egress controls, provenance/attestation requirements.

[opinions expressed are solely my own]

Towards a brighter future Conclusion

- AI gateways
    - routing, authentication, rate limiting, cost controls, and policy checks
- Agent runtime guardrails
    - tool per missioning, role separation (planner vs executor), deterministic sandboxes
- Security posture for tools
    - provenance, allowlisted integrations, safe defaults, secrets hygiene
- Supply chain controls for agents
    - SBOM, signed artifacts, attestation, verified tool servers
- Detection and response
    - anomaly detection on tool patterns, eBPF telemetry, incident playbooks for agents
- AAIF (https://aaif.io/)

Let’s build towards a world where we can ***trust the controls*** instead of **trust the model**

[opinions expressed are solely my own]