

Inferred Edge AI Update

May 2025

Top AI Industry News

- **Rapid Model Releases:** Major AI companies continue to release new models, spanning both proprietary platforms (Anthropic, Google, OpenAI) and open source offerings (Meta's Llama 4)
 - Latest models prioritize enhanced reasoning abilities
- **Model Context Protocol:** Anthropic launched MCP late last year as a standard protocol for AI models to access external data sources
 - Continued strong adoption across the AI developer community (~5K MCP Servers built)
- **Agent 2 Agent Protocol:** Google launched A2A on April 9 as a protocol for agents to collaborate with other agents
 - Remains in early stages of adoption with few production examples
- **AI Coding Expansion:** Major updates across AI development tools:
 - On May 6, OpenAI announced acquisition of Windsurf (AI-powered coding environment) for \$3B
 - Cursor (AI-powered coding environment) expanding MCP ecosystem, supporting 40+ external tools

Large Language Models: Latest Roster

Model	Parent	Release Date	License Type	Context Window*	Blended Price**	Initial Market Traction
Gemini 2.5 Pro Preview	Google	May 2025	Proprietary	2M	\$3.44	Multimodal analysis, particularly video
GPT-4.1	OpenAI	April 2025	Proprietary	1M	\$3.40	Document processing
o4-mini	OpenAI	April 2025	Proprietary	200K	\$1.93	High-volume processing in production
o3	OpenAI	April 2025	Proprietary	200K	\$17.50	Complex reasoning in specialized domains
Llama 4 Scout	Meta	April 2025	Open	10M	\$0.17	Edge computing
Llama 4 Maverick	Meta	April 2025	Open	1M	\$0.40	Multimodal applications
Gemini 2.5 Pro	Google	March 2025	Proprietary	1M	\$3.44	Long document processing/comprehension
GPT-4o	OpenAI	February 2025	Proprietary	128K	\$5.11	Real-time chat
Claude 3.7 Sonnet	Anthropic	February 2025	Proprietary	200K	\$6.00	Software engineering and coding
Claude 3.5 Sonnet	Anthropic	June 2024	Proprietary	200K	\$6.00	Coding and creative content generation

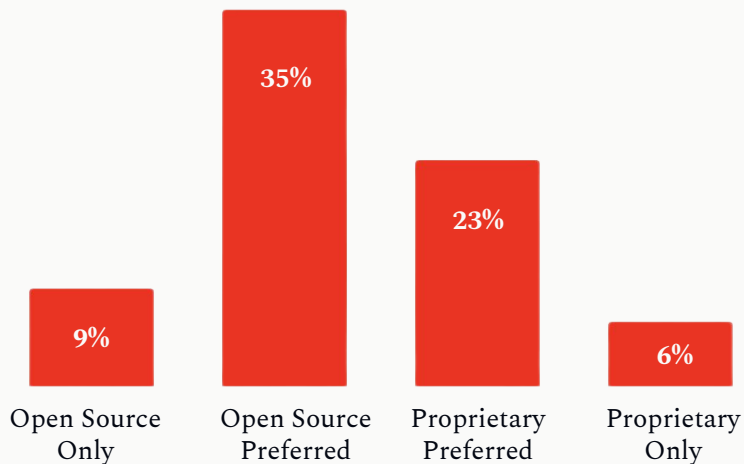
*maximum number of tokens processed in single interaction. **per 1M tokens; blended input/output costs.

Source: OpenAI, Anthropic, Google, Meta.

An Update on Open Source

As companies look to deploy Generative AI, they are increasingly using a combination of both open source (distributed with source code) and proprietary (productized) tooling across the entire tech stack:

Open Source vs. Proprietary Preferences



Preference Rationale

Open Source:

- **Cost (63% of respondents)**
- Ease of implementation (31%)
- Security, risk and system control (31%)

Proprietary:

- **Security, risk and system control (73%)**
- Better suited to org. use cases (34%)
- Resources available (30%)

Source: McKinsey, "Open Source Technology in the Age of AI," April 2025.

Inferred Edge Stack: Open Source vs. Proprietary

We believe a hybrid approach to tooling provides the best solution in terms of both quality and security.
Some examples of where we are using open source vs. proprietary in our tech stack:

Open Source

- Data validation and loading pipelines (Zod)
- Schema coercion & deterministic LLM result frameworks (BAML by Boundary)
- Databases technologies (Postgres, Kuzu, Weaviate)
- Deployment infrastructure (SST/Opencontrol)
- Security and Role Based Access Control (CASL)

External Proprietary

- Specialized databases (Neo4j)
- Foundation models (Anthropic, OpenAI)
- Cloud Providers (AWS)

Inferred Edge Proprietary

- Agentic frameworks and application business logic
- Domain-specific user experience and interface
- Domain-specific ontologies (knowledge graph schemas)
- Tracking and evaluation frameworks

Model Context Protocol (MCP): Overview

MCP is an open standard enabling bidirectional connections between AI models and external data sources.

- Developed by Anthropic to standardize how AI models access enterprise systems and tools
- Similar to APIs, MCP eliminates custom integration code for each new data source or tool that integrates with an AI model
 - A wide gap exists between MCP servers running locally on a laptop vs. enterprise deployments in the cloud. Most example servers use stdio (local-only protocol) instead of StreamableHTTP
- AI agents can now reason about and leverage specific enterprise tools to complete a task

Adoption Timeline

November 2024: Anthropic releases Model Context Protocol as an open standard

February 2025: >1K MCP Servers on GitHub

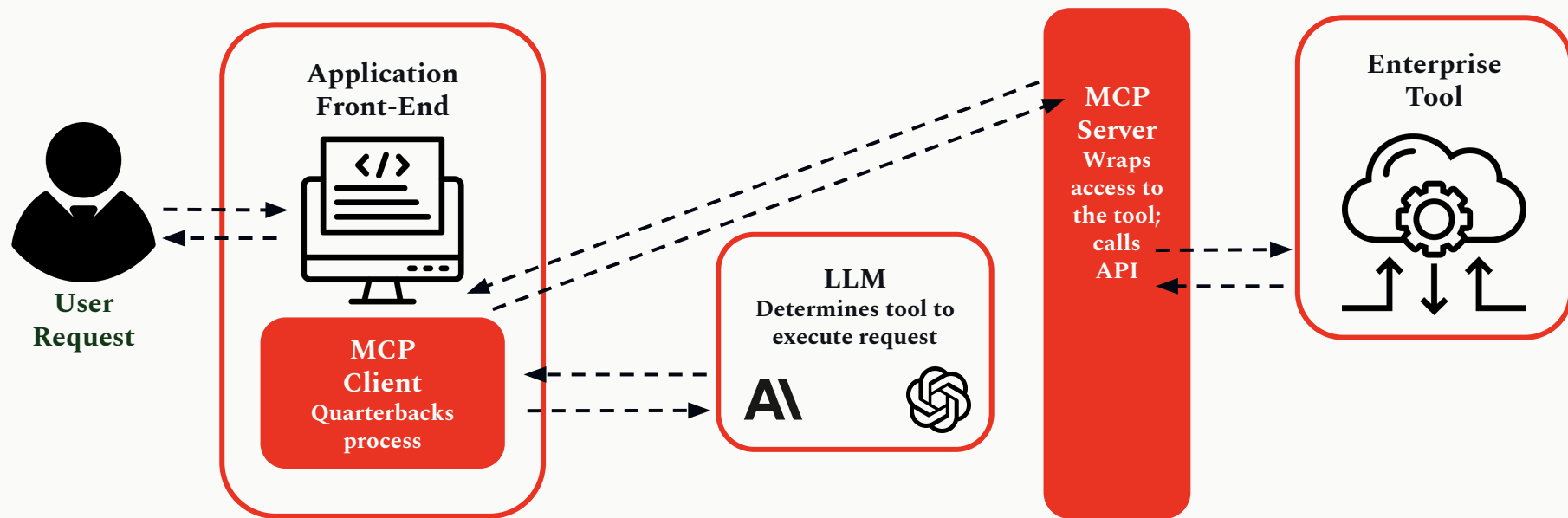
March 2025: Microsoft adds MCP support in Copilot Studio

March 2025: OpenAI introduces MCP support for its Agents SDK

April 2025: Google, Amazon + Azure adopt MCP

How MCP Works: Clients + Servers

MCP's architecture divides responsibilities between clients and servers, creating a standardized ecosystem that enables AI systems to work with enterprise tools.



MCP Servers are at Various Stages of Usability

Not all MCP Servers available on GitHub are “Official MCPs”, or MCPs written and released by the company itself.

Some non-official MCPs are recognized, while others may have no direct affiliation to the company at all.


Appropriate vetting and security are crucial.

Inferred Edge Vetted Servers


✓  **Neo4j:** graph database
read/write


✓  **Tavily:** web search

✓  **Github:** repository read/write


✓  **Kuzu:** graph database
read/write

✓  **RunReveal:** security log
review and analysis

 **FileSystem:** computer file
interaction


✓  **Notion:** project
management read/write

 **Google Calendar:**
read/write calendar events

✓  **Asana:** project
management read/write

Under Review

✓  **Weaviate:** vector database

✓  **Merge:** API integrations platform
Postgres: relational database

 read/write

Salesforce: CRM interaction

 **Slack:** workspace communication

 **Google Drive:** file read/write

 **Snowflake:** interaction with data
warehouse

 **Linear:** project management

 **Dropbox:** file storage

 **FRED:** Fed economic data



✓ Official MCP.

Recommended Next Steps

Data Readiness Assessment

- Audit existing data sources and structures for AI integration readiness
- Identify priority datasets for initial implementation
- Develop data governance frameworks to ensure compliance, quality and security

Use Case Prioritization

- Evaluate high-impact, low-complexity opportunities for initial deployment
- Conduct stakeholder workshops to identify pain points addressable by AI
- Define clear success metrics for pilot implementations

Implementation Roadmap

- Schedule a technical discovery session with Inferred Edge experts
- Develop a phased implementation plan with defined milestones