

MCP: The new “USB-C for AI” that’s bringing fierce rivals together

Other notable examples include servers that connect AI models to home automation systems, real-time weather data, e-commerce platforms, and music streaming services. Some implementations allow AI assistants to interact with gaming engines, 3D modeling software, and IoT devices.

What is “context” anyway?

To fully appreciate why a universal AI standard for external data sources is useful, you’ll need to understand what “context” means in the AI field.

With current AI model architecture, what an AI model “knows” about the world is baked into its neural network in a largely unchangeable form, placed there by an initial procedure called “pre-training,” which calculates statistical relationships between vast quantities of input data (“training data”—like books, articles, and images) and feeds it into the network as numerical values called “weights.” Later, a process called “fine-tuning” might adjust those weights to alter behavior (such as through reinforcement learning like RLHF) or provide examples of new concepts.

Typically, the training phase is very expensive computationally and happens either only once in the case of a base model, or infrequently with periodic model updates and fine-tunings. That means AI models only have internal neural network representations of events prior to a “cutoff date”

when the training dataset was finalized.

After that, the AI model is run in a kind of read-only mode called “inference,” where users feed inputs into the neural network to produce outputs, which are called “predictions.” They’re called predictions because the systems are tuned to predict the most likely next token (a chunk of data, such as portions of a word) in a user-provided sequence.

In the AI field, context is the user-provided sequence—all the data fed into an AI model that guides the model to produce a response output. This context includes the user’s input (the “prompt”), the running conversation history (in the case of chatbots), and any external information sources pulled into the conversation, including a “system prompt” that defines model behavior and “memory” systems that recall portions of past conversations. The limit on the amount of context a model can ingest at once is often called a “context window,” “context length, ” or “context limit,” depending on personal preference.

<https://arstechnica.com/information-technology/2025/04/mcp-the-new-usb-c-for-ai-thats-bringing-fierce-rivals-together/>