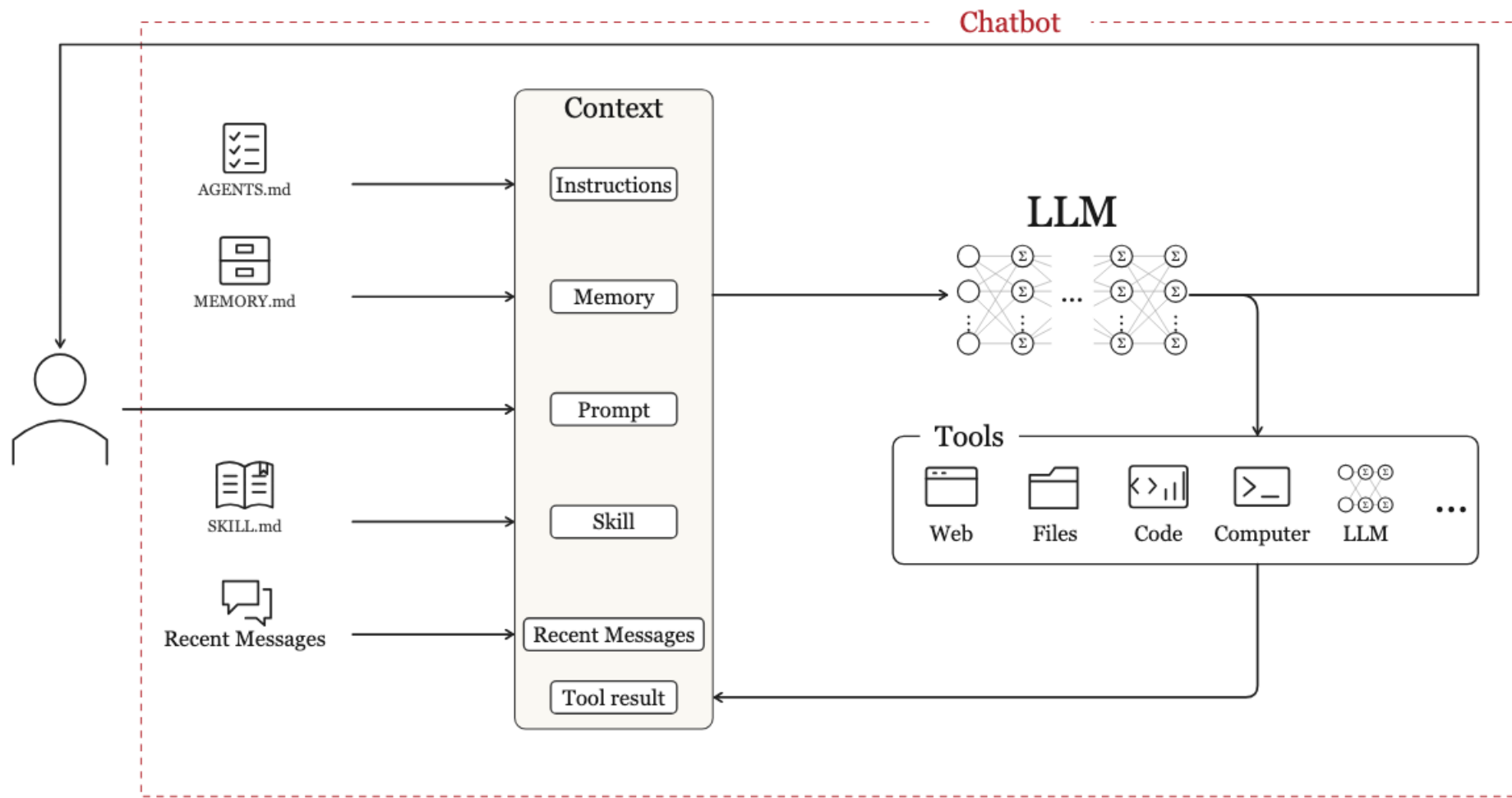




L-4

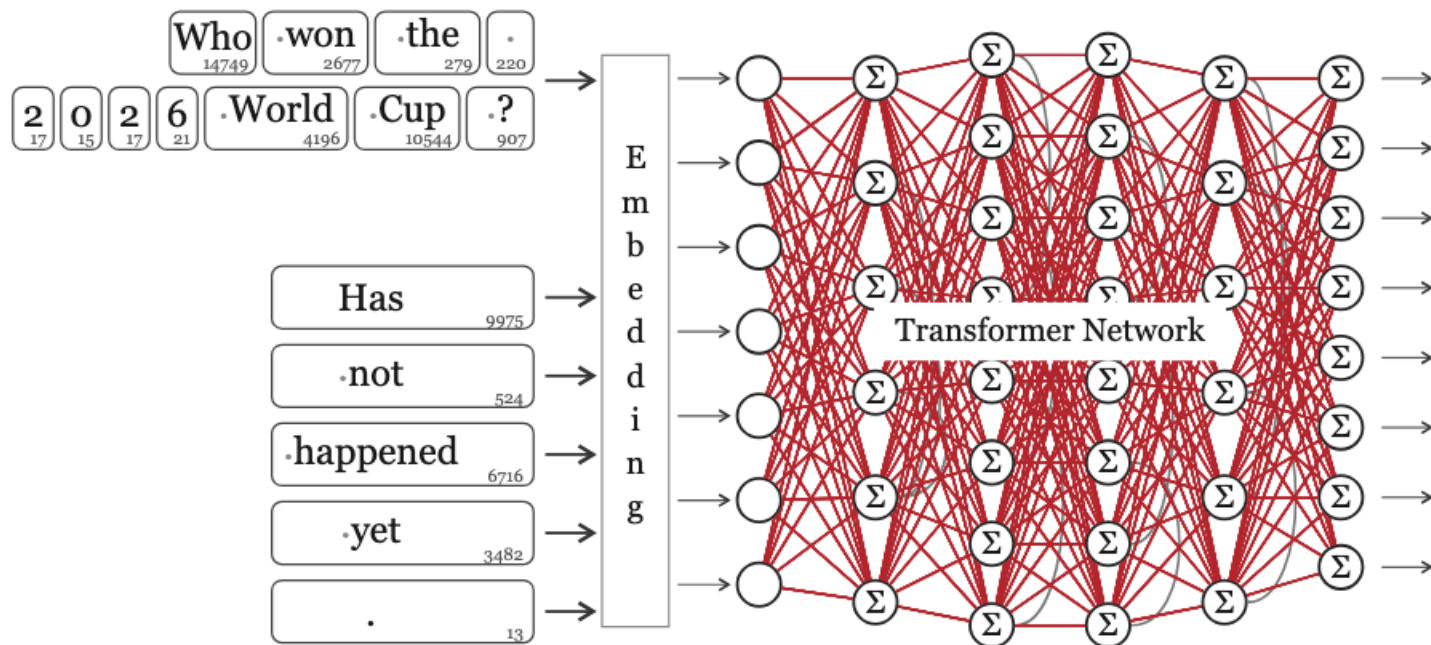
From LLMs to Chatbots



DEMO 1 · L3 Exit Ticket Results

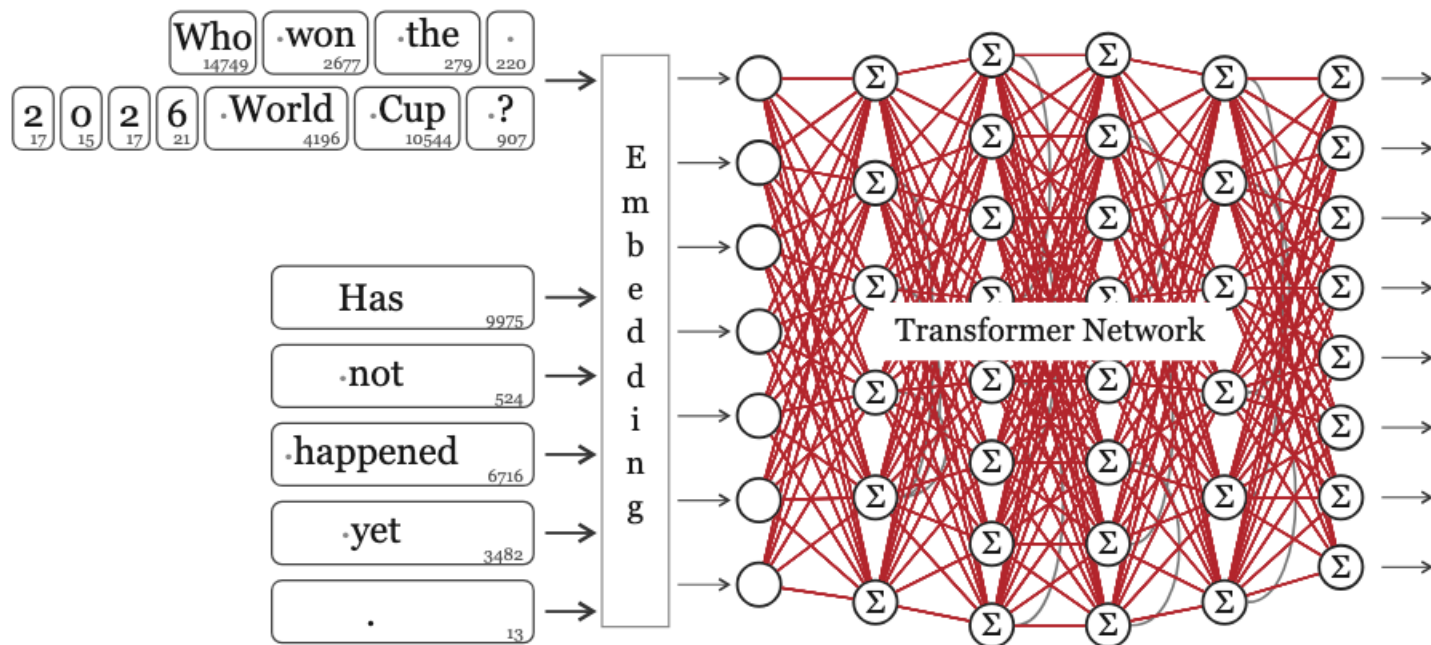
What Does an LLM Know?

Qwen3.5-9B



What Does an LLM Know?

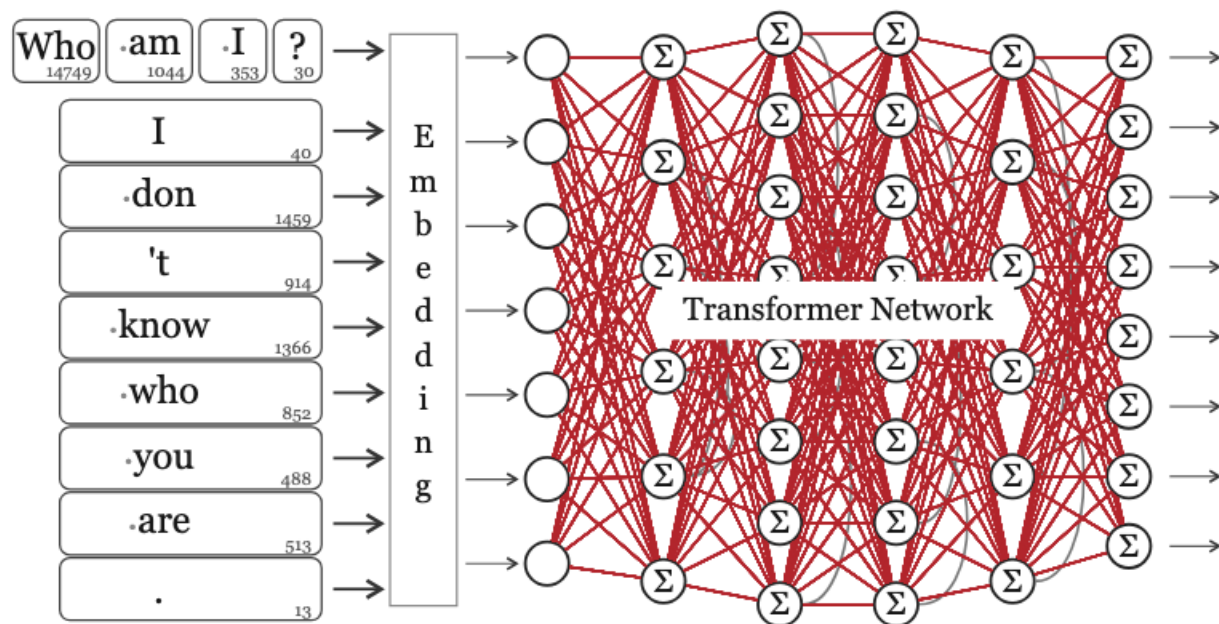
Qwen3.5-9B



DEMO 2 · What Does an LLM Know?

What Does an LLM Remember?

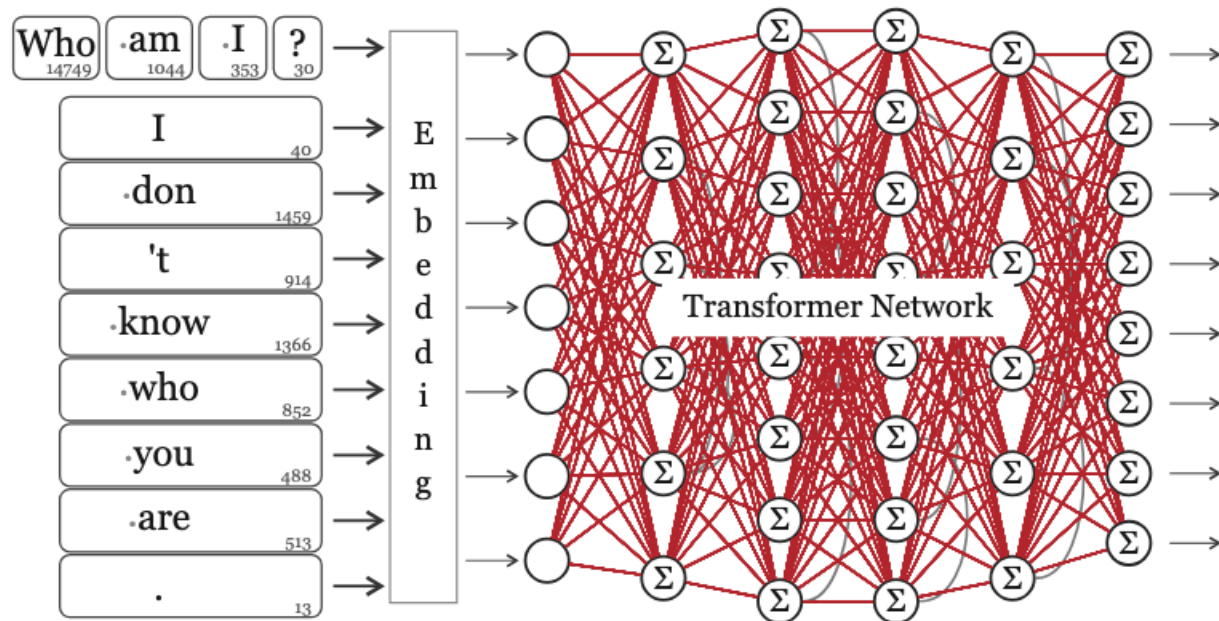
Qwen3.5-9B



New request

What Does an LLM Remember?

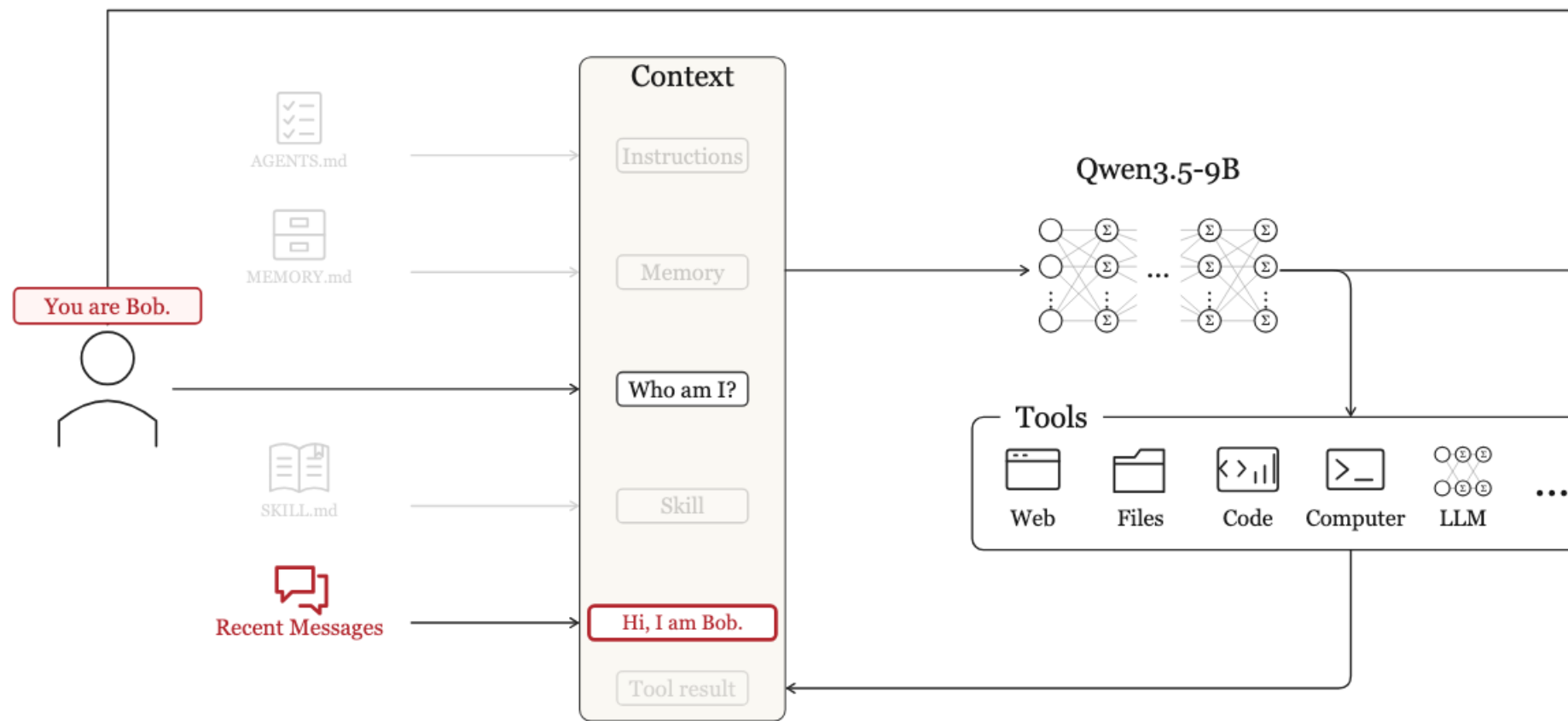
Qwen3.5-9B



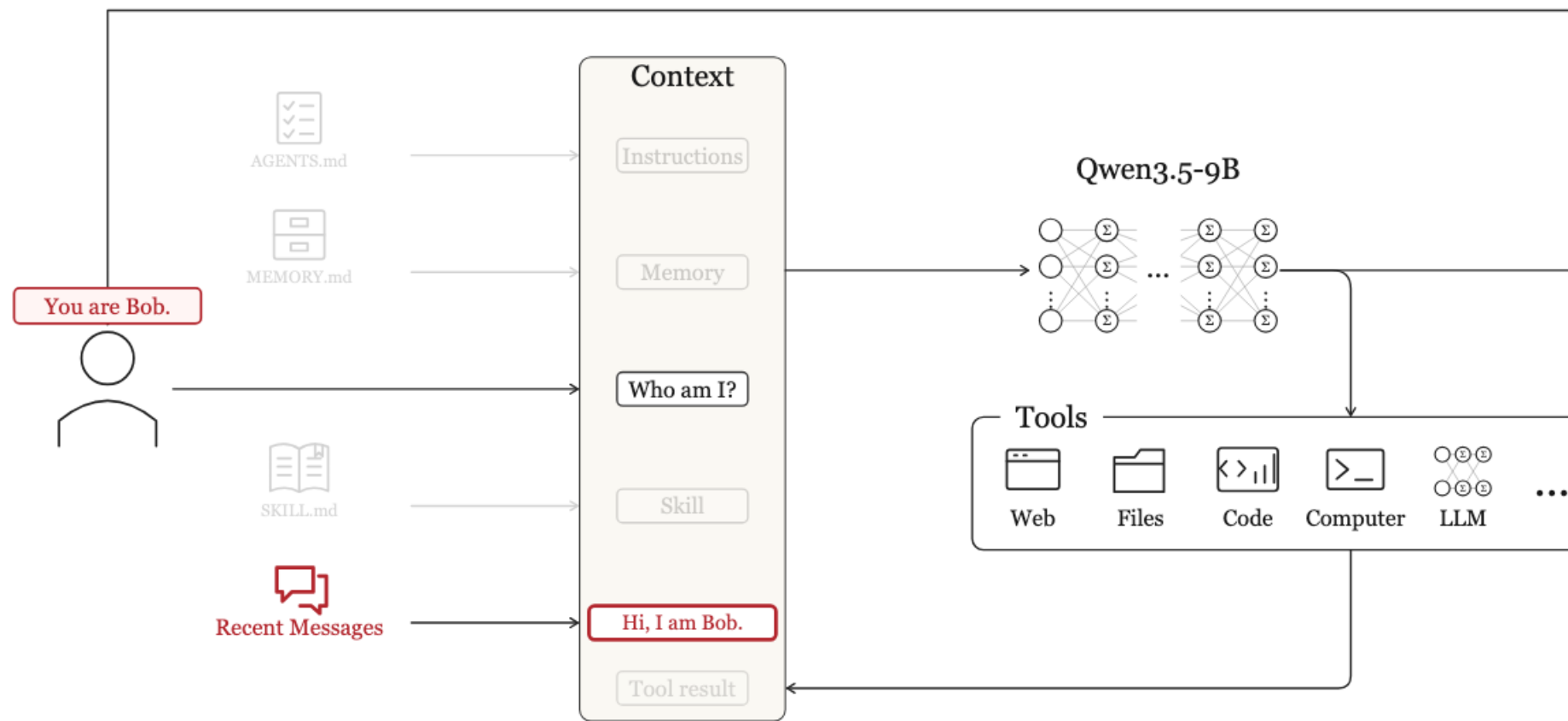
New request

DEMO 3 · What Does an LLM Remember?

Remember Bob

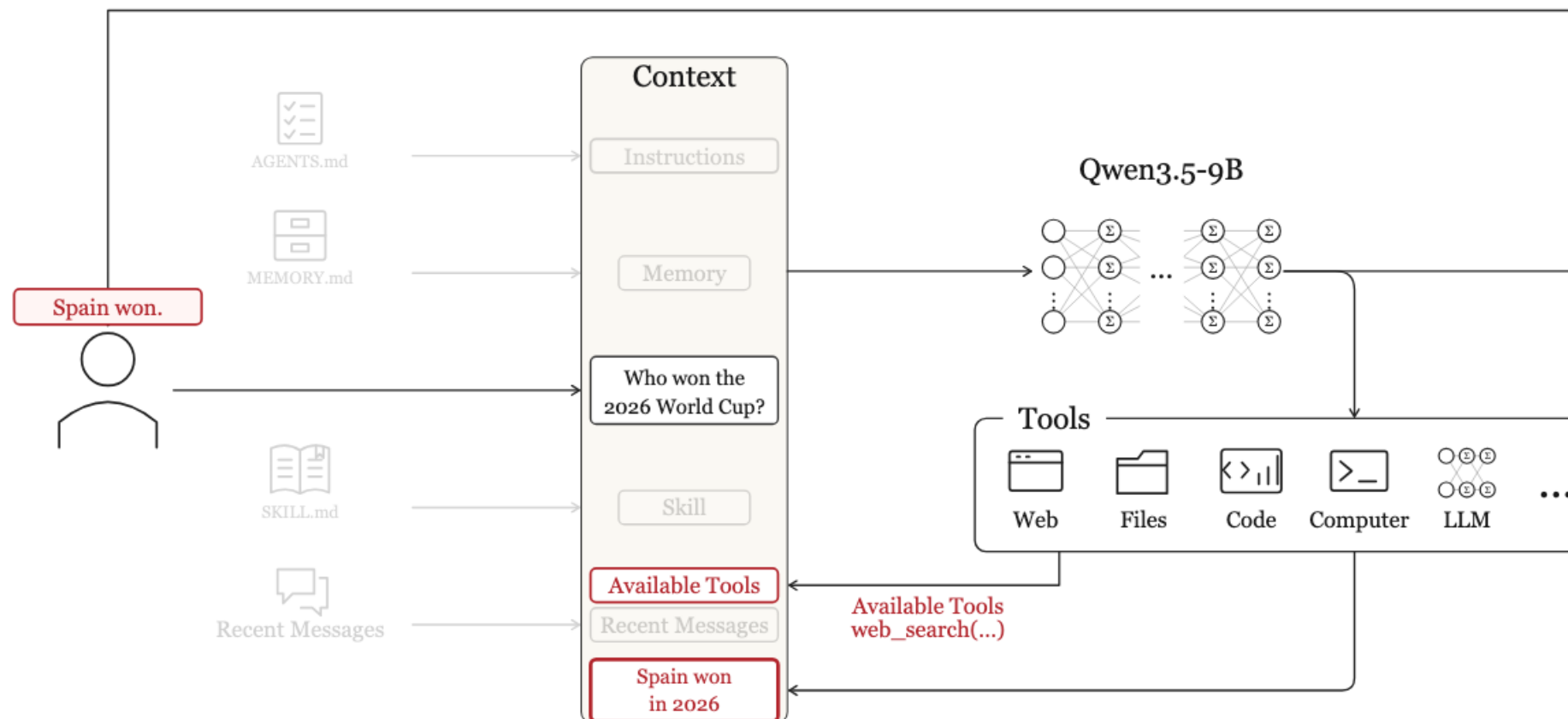


Remember Bob

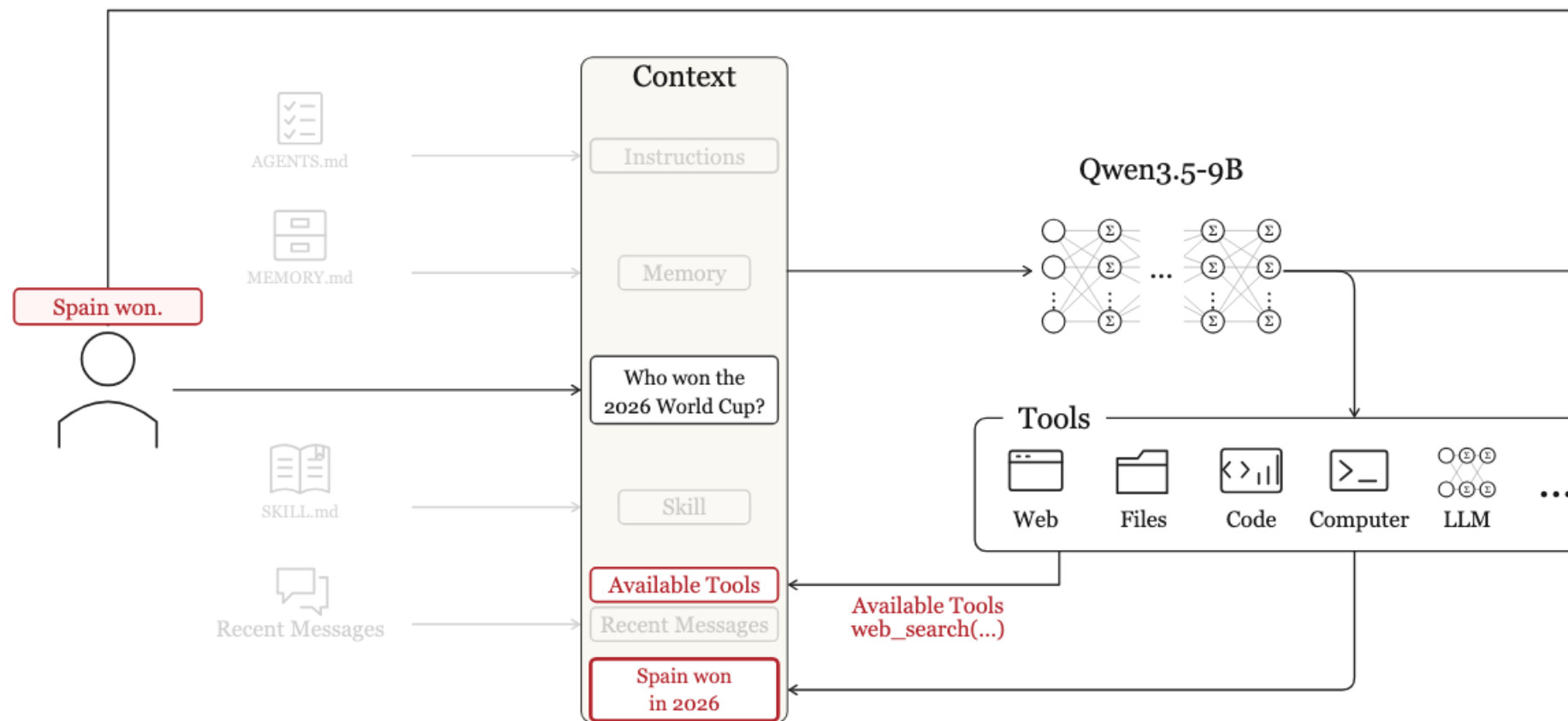


DEMO 4 · Remember Bob

Who Won the World Cup?

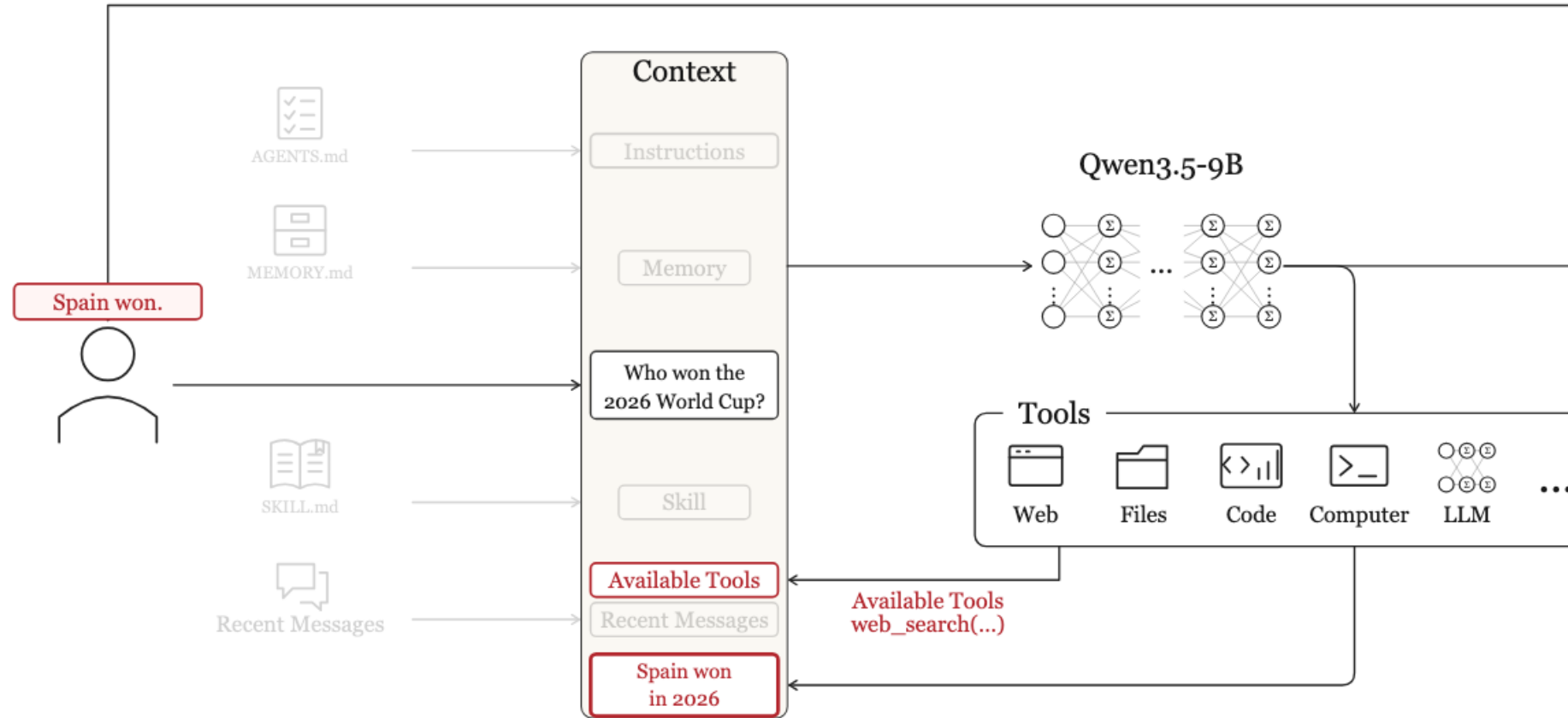


Who Won the World Cup?

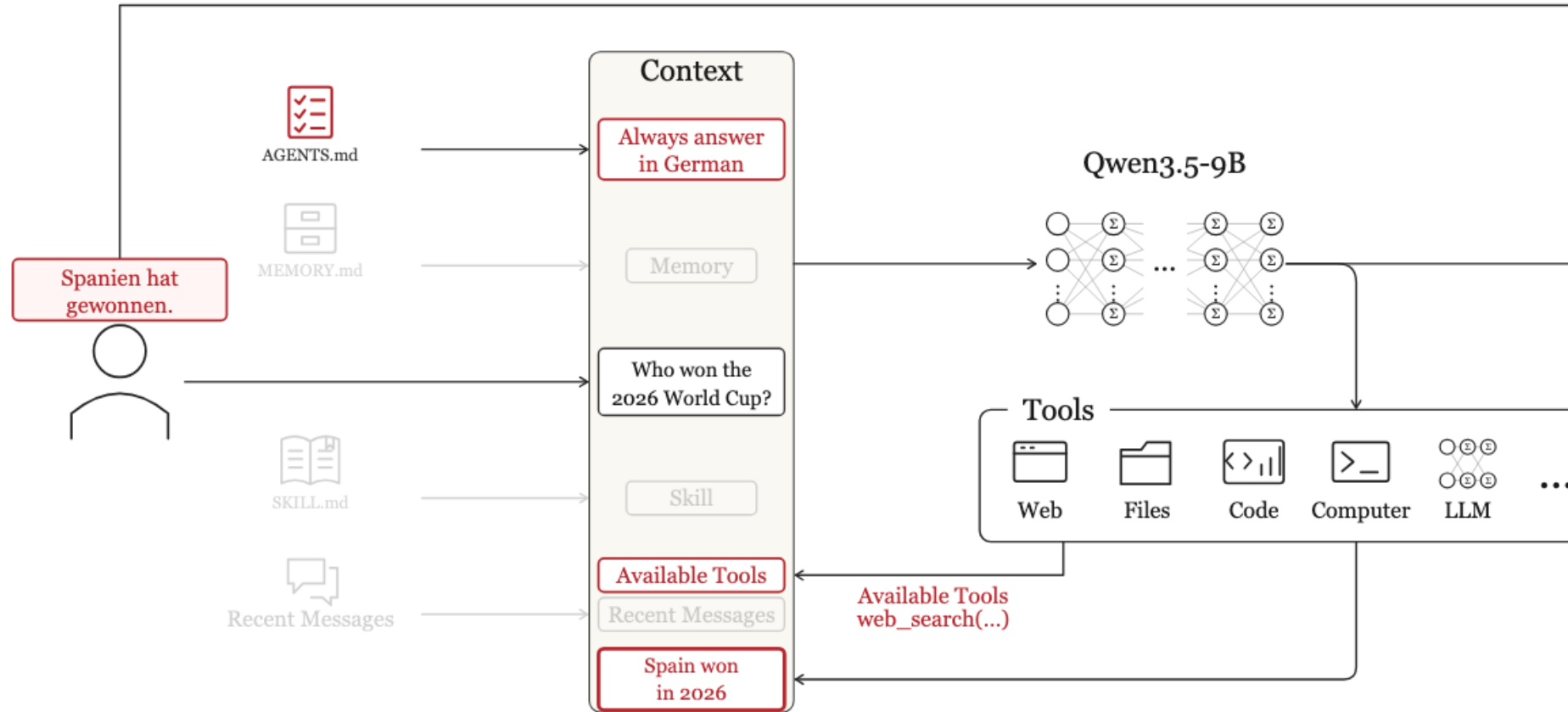


DEMO 5 · Who Won the World Cup?

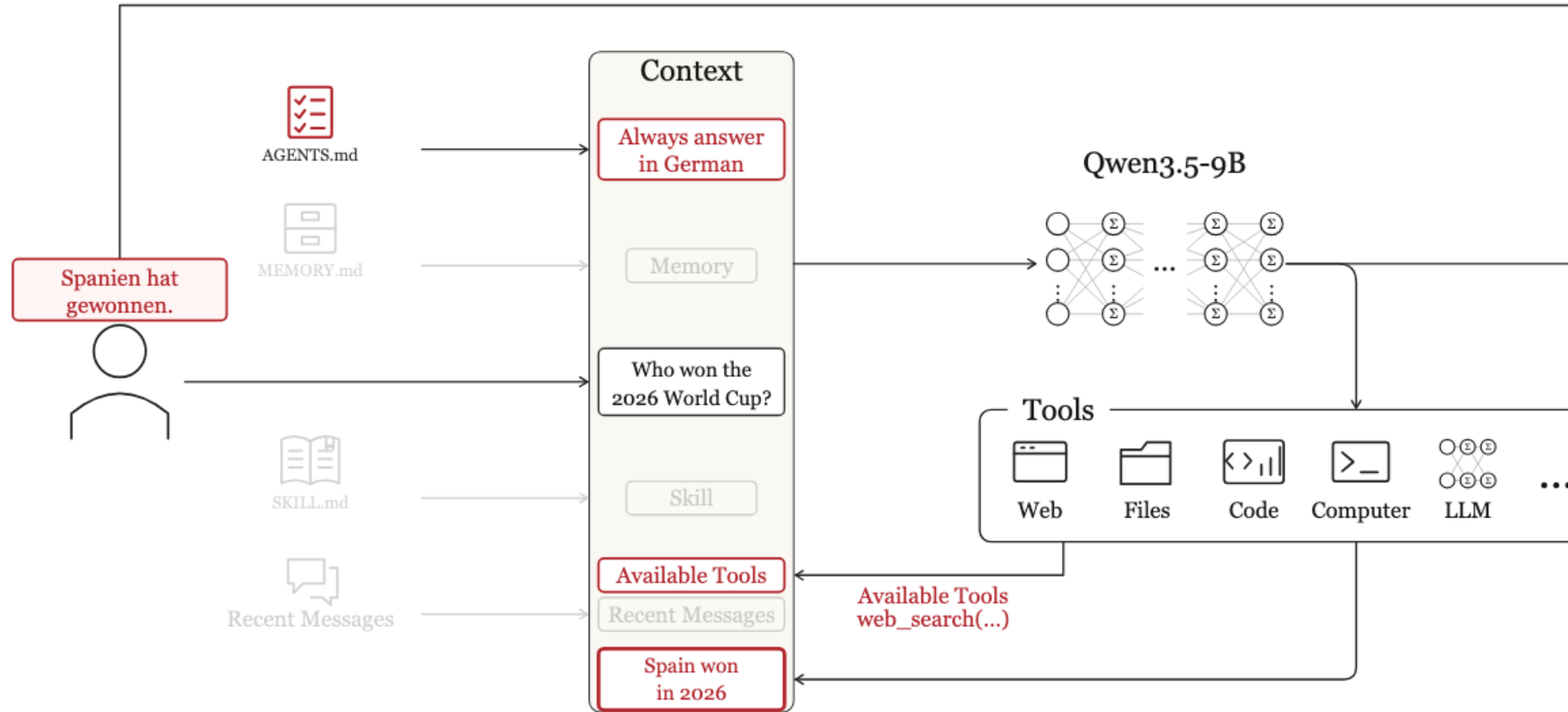
Instructions Shape the Reply



Instructions Shape the Reply

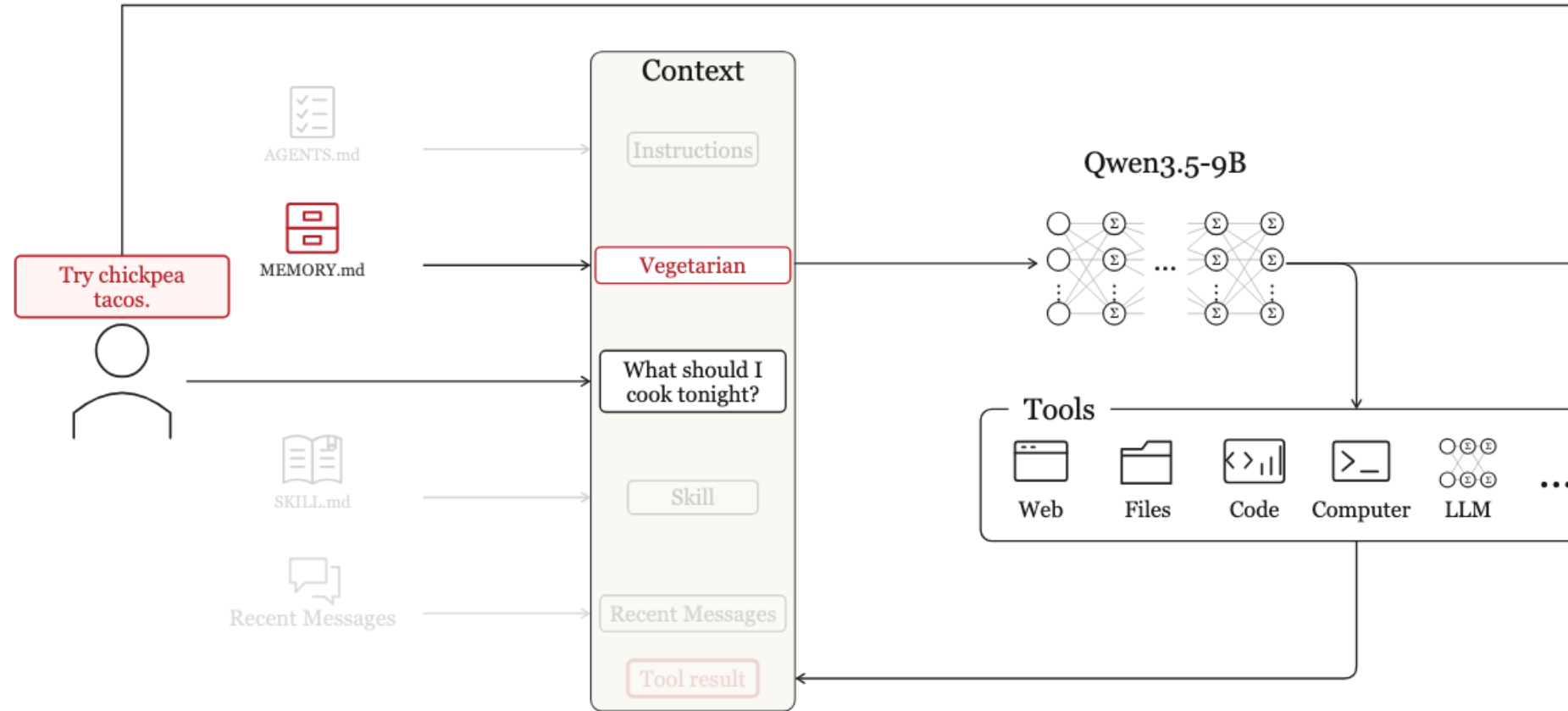


Instructions Shape the Reply

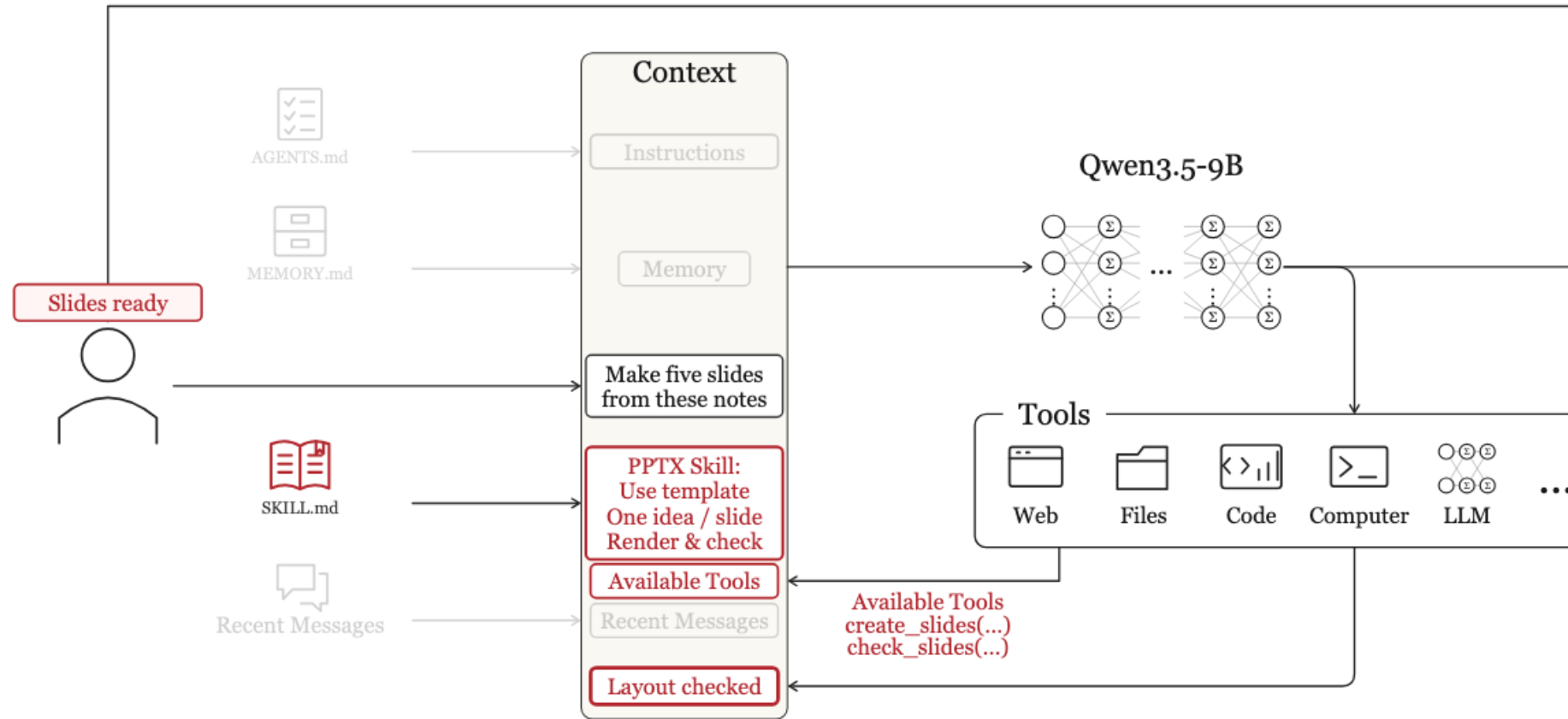


DEMO 6 · Instructions Shape the Reply

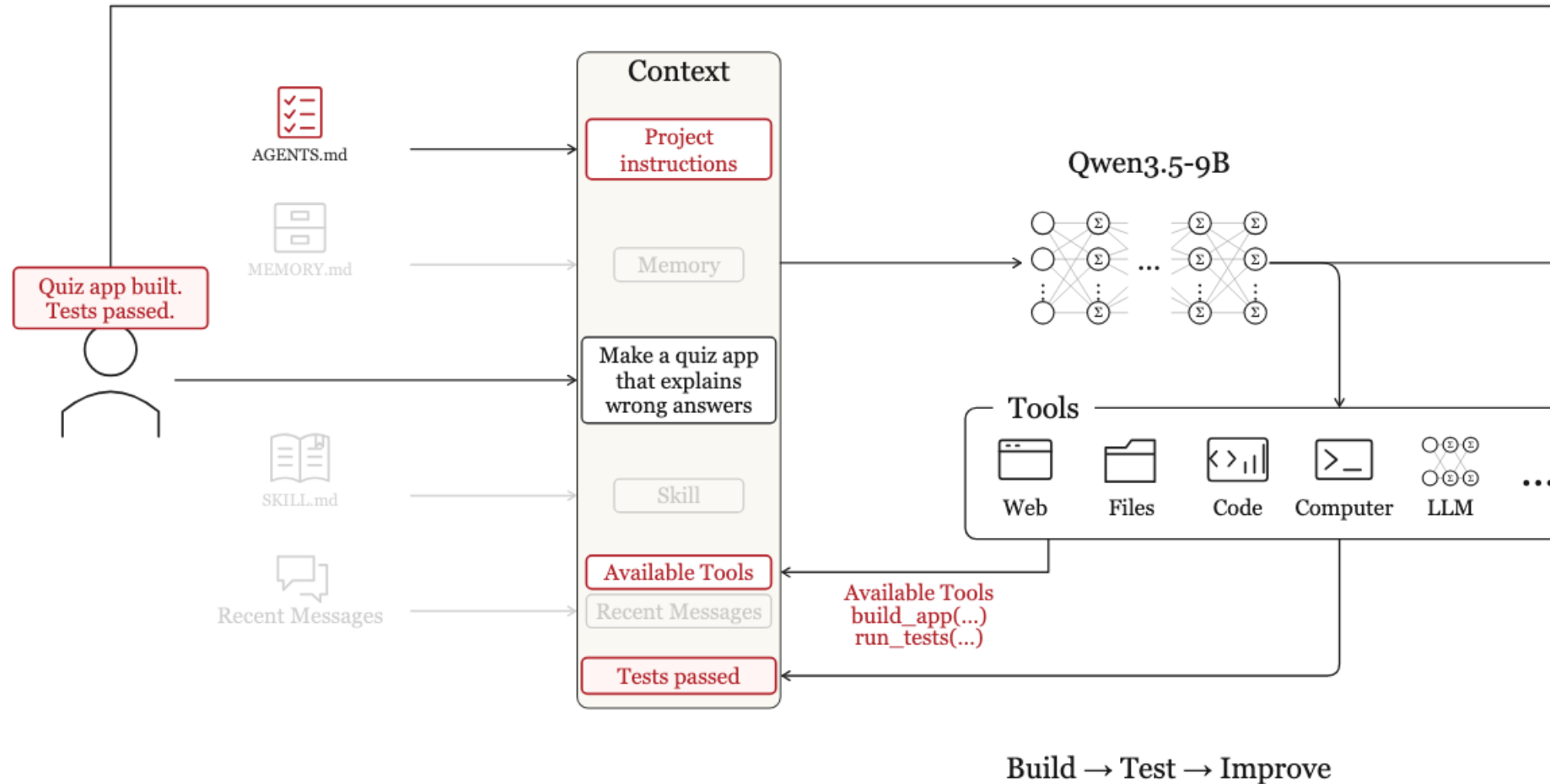
A New Chat. Same Preferences



A Skill for Making Slides

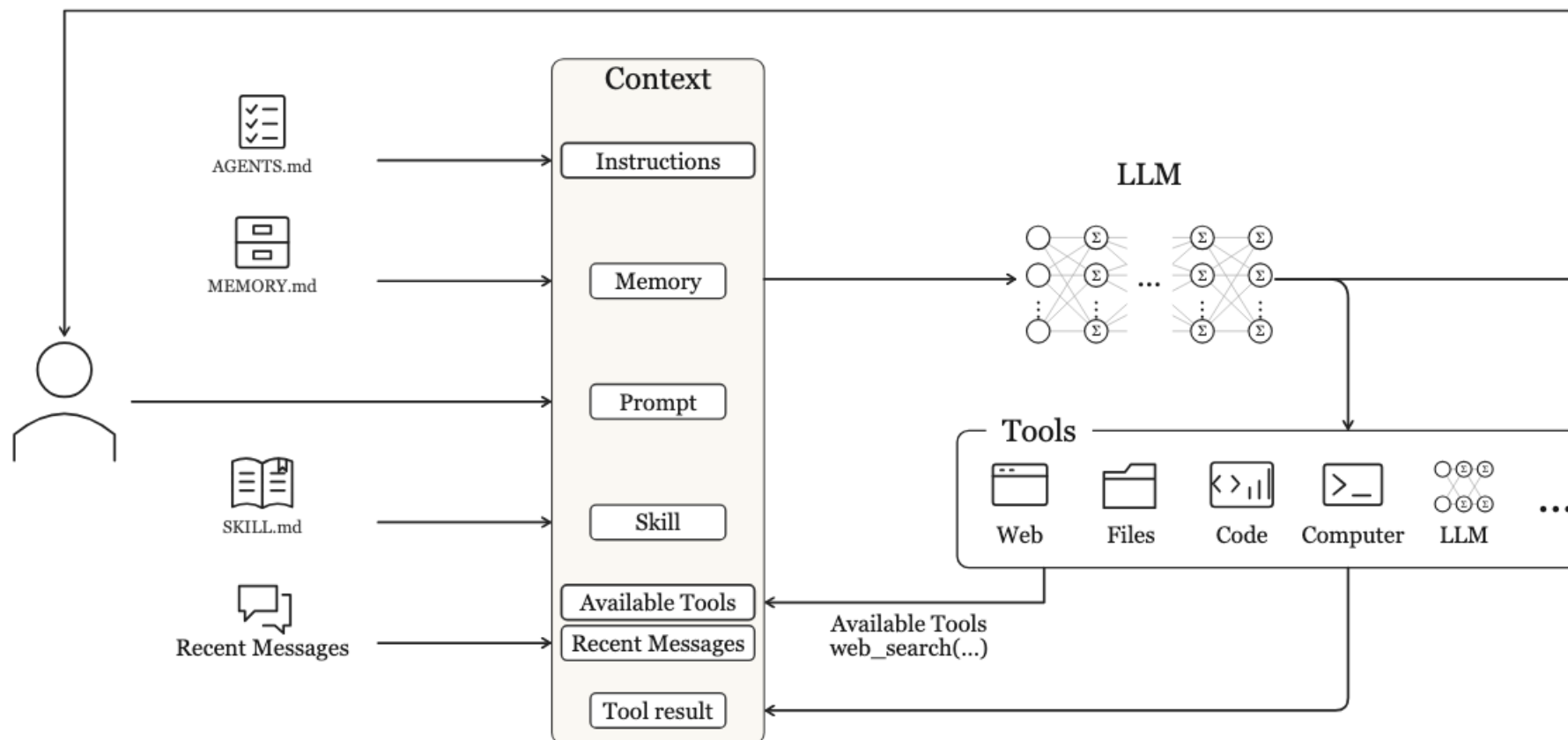


Agentic AI

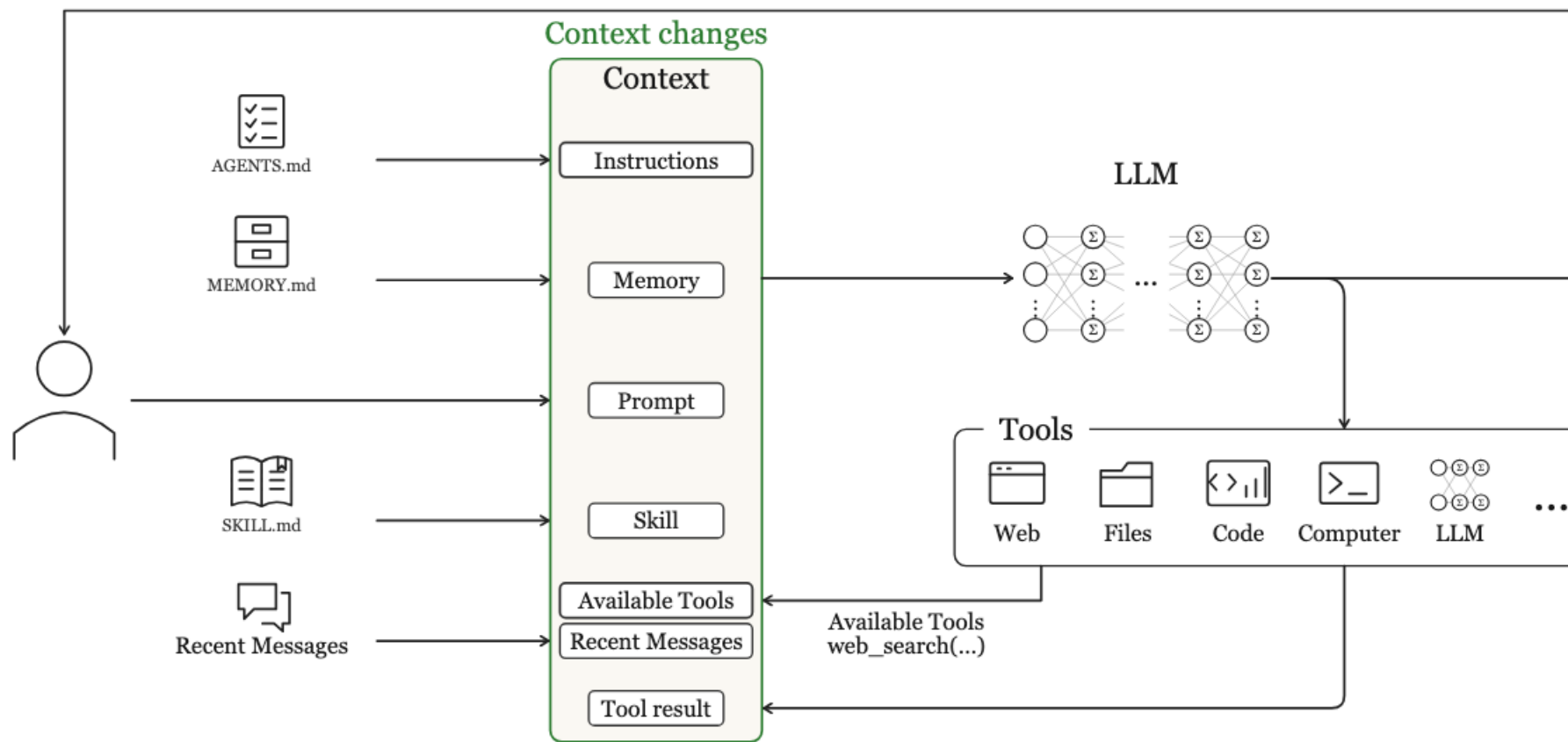


Agentic AI works towards a goal by using tools, checking results, and deciding what to do next

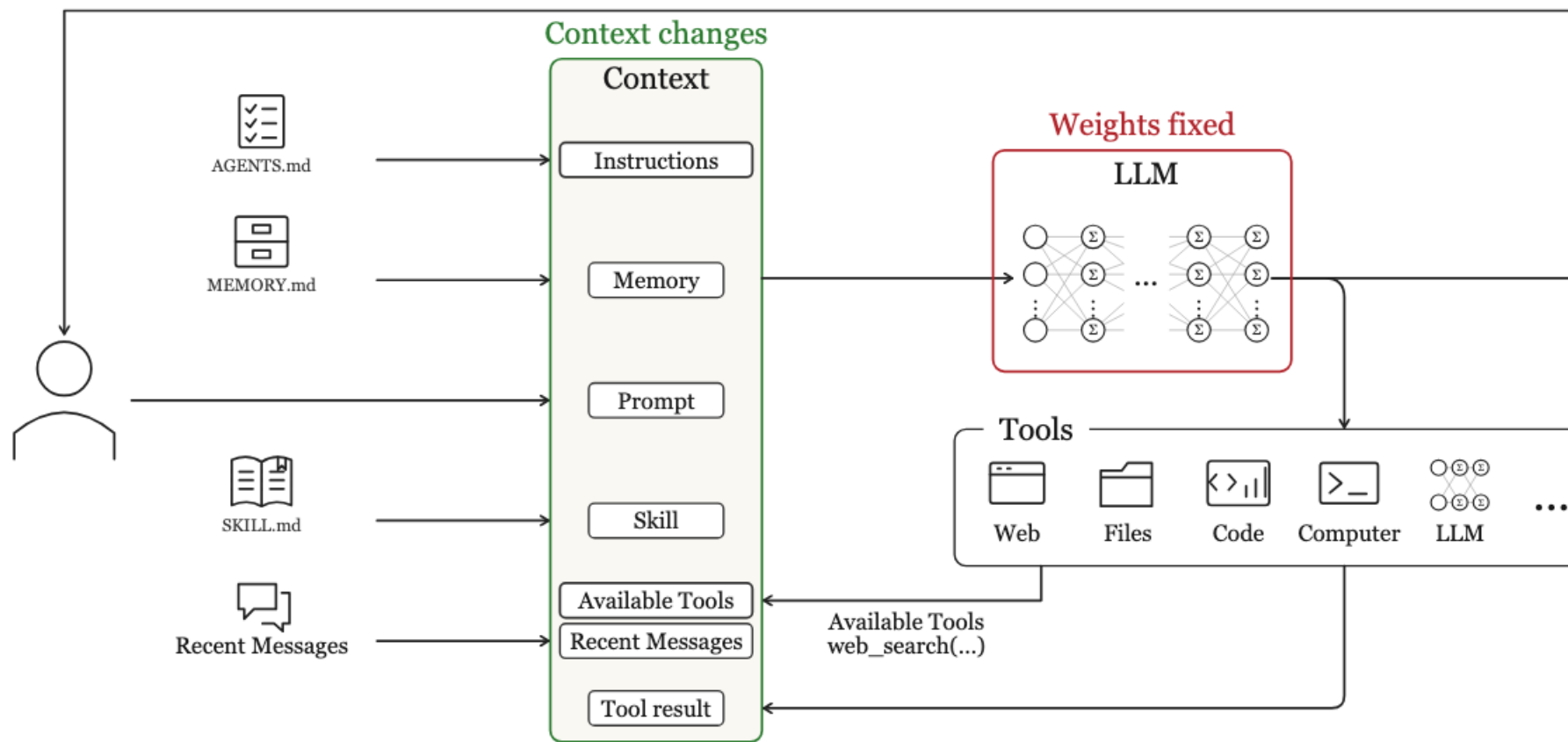
Different Context. Same Model



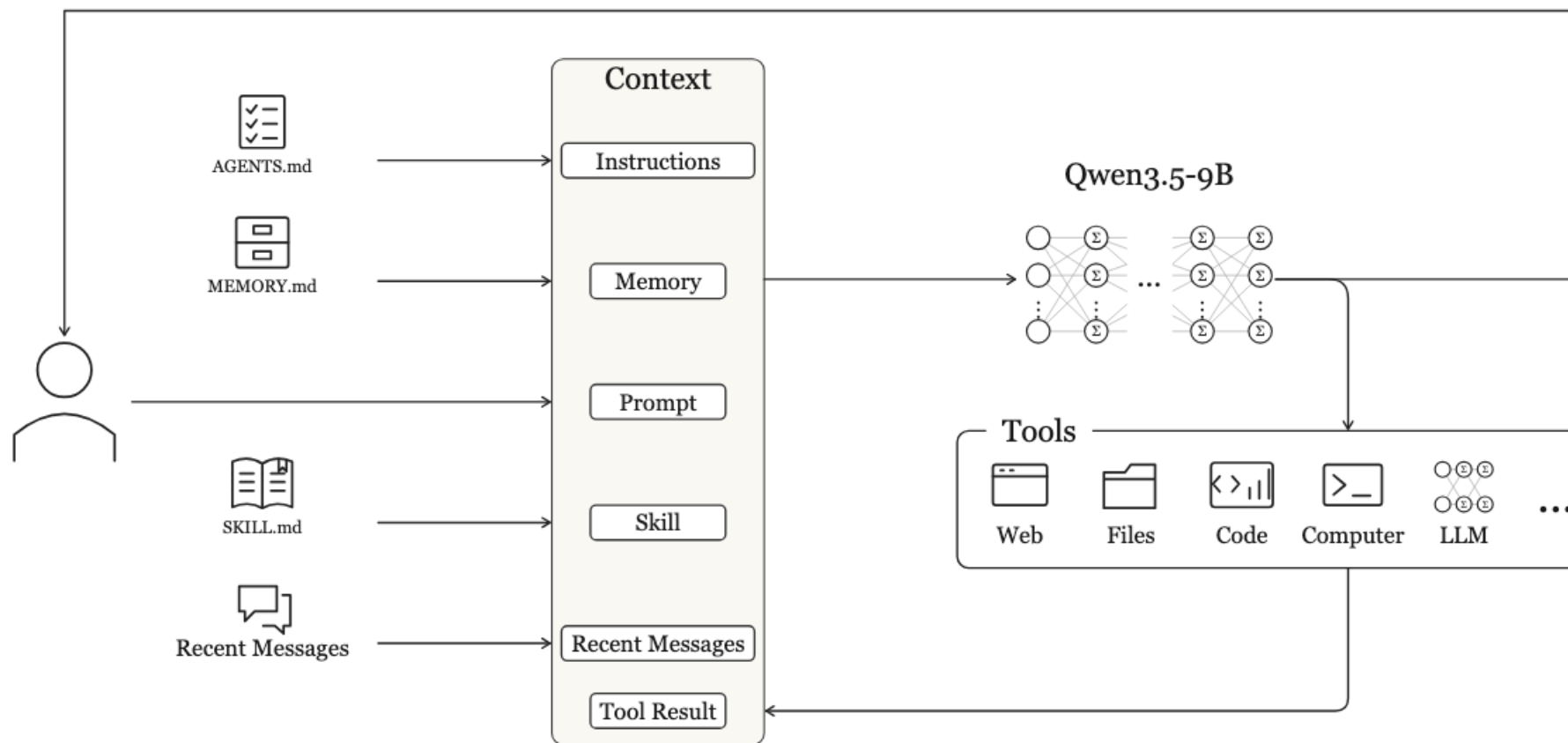
Different Context. Same Model



Different Context. Same Model

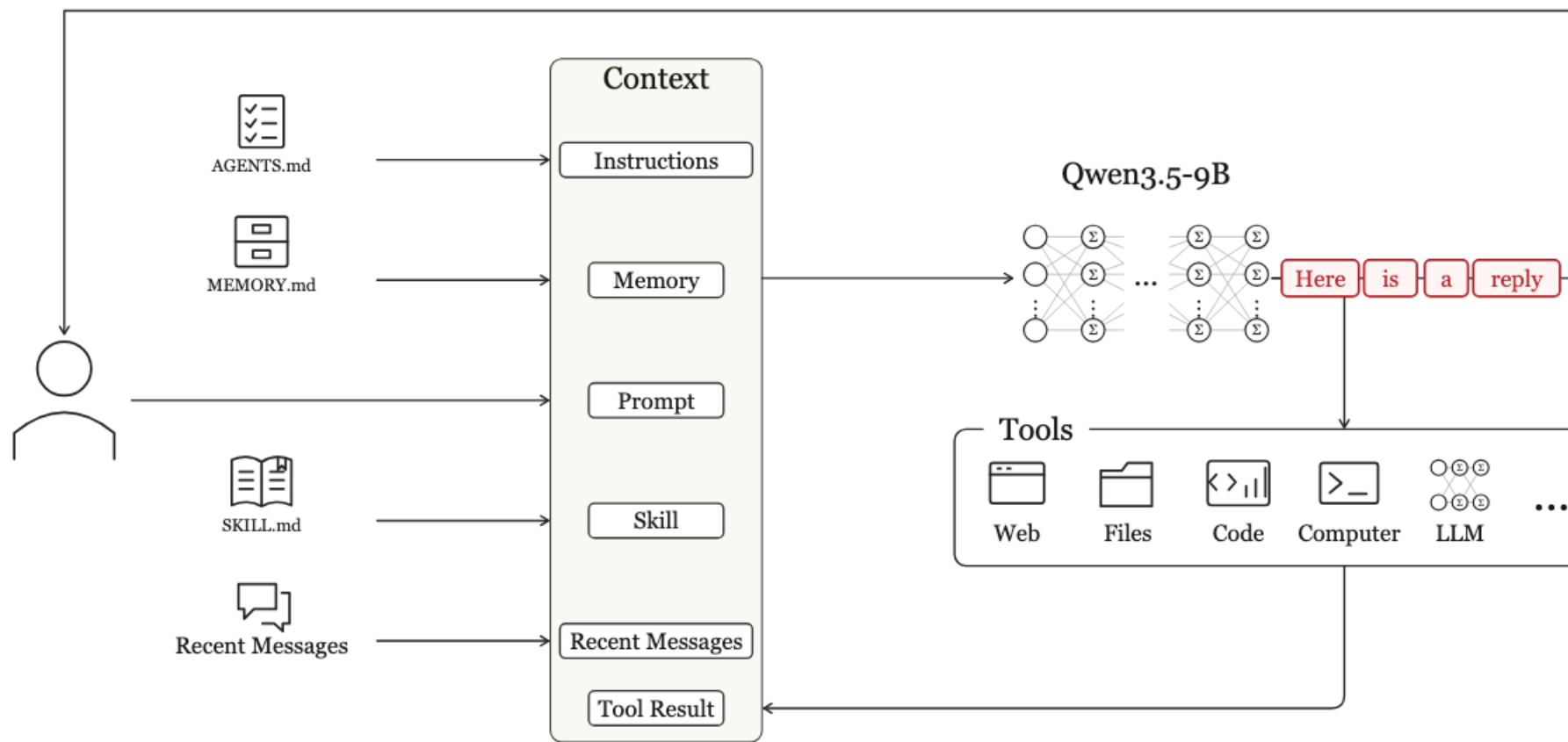
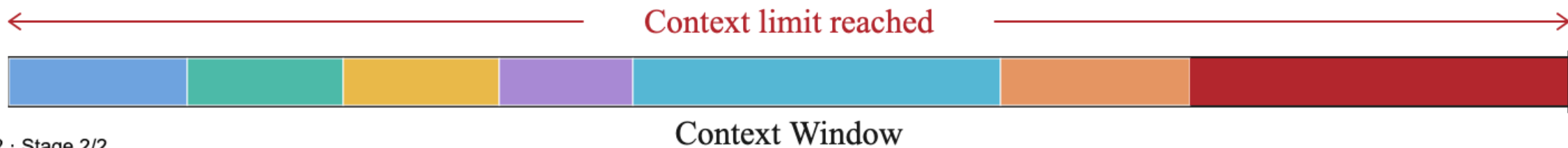


Context Has Limits

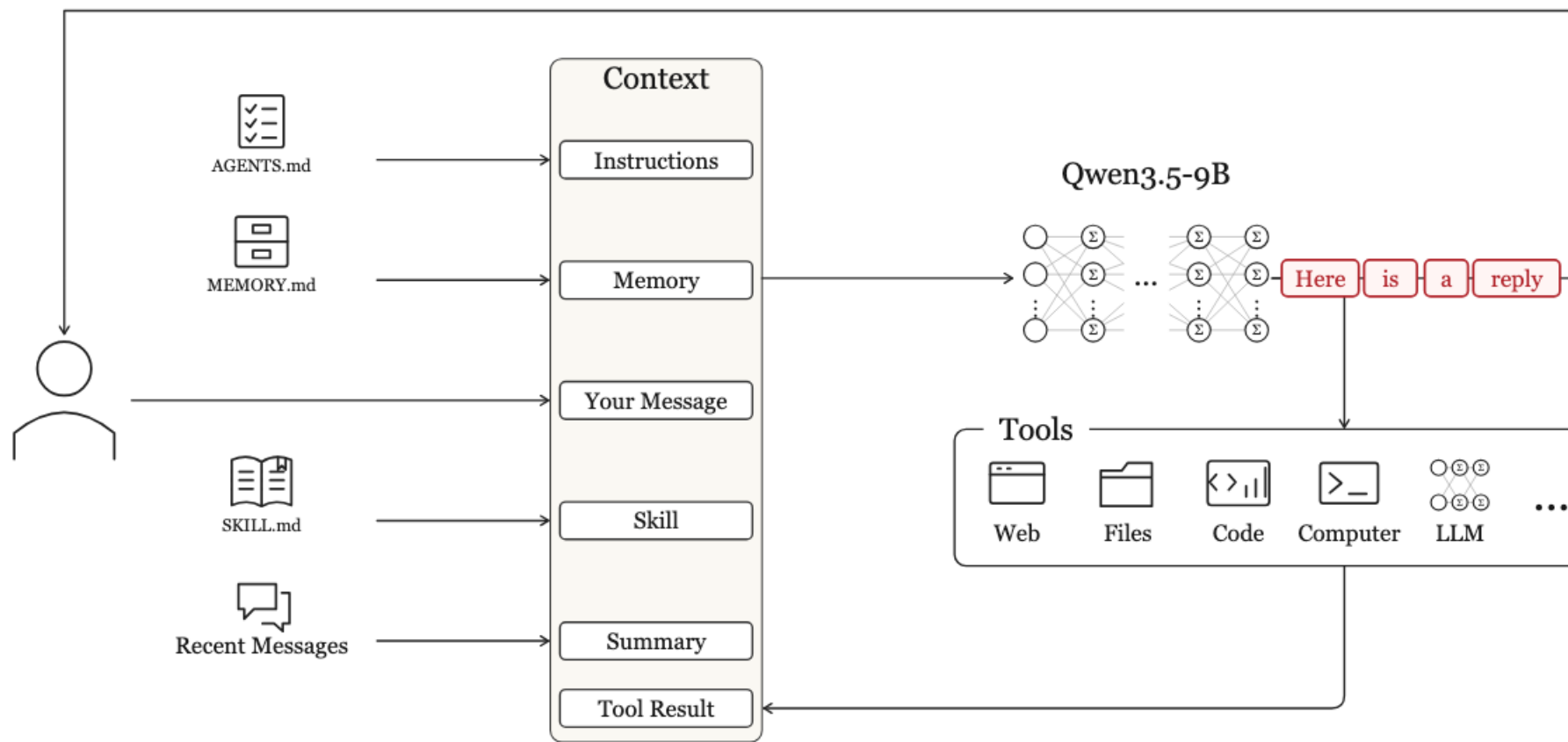


Context Window

Context Has Limits

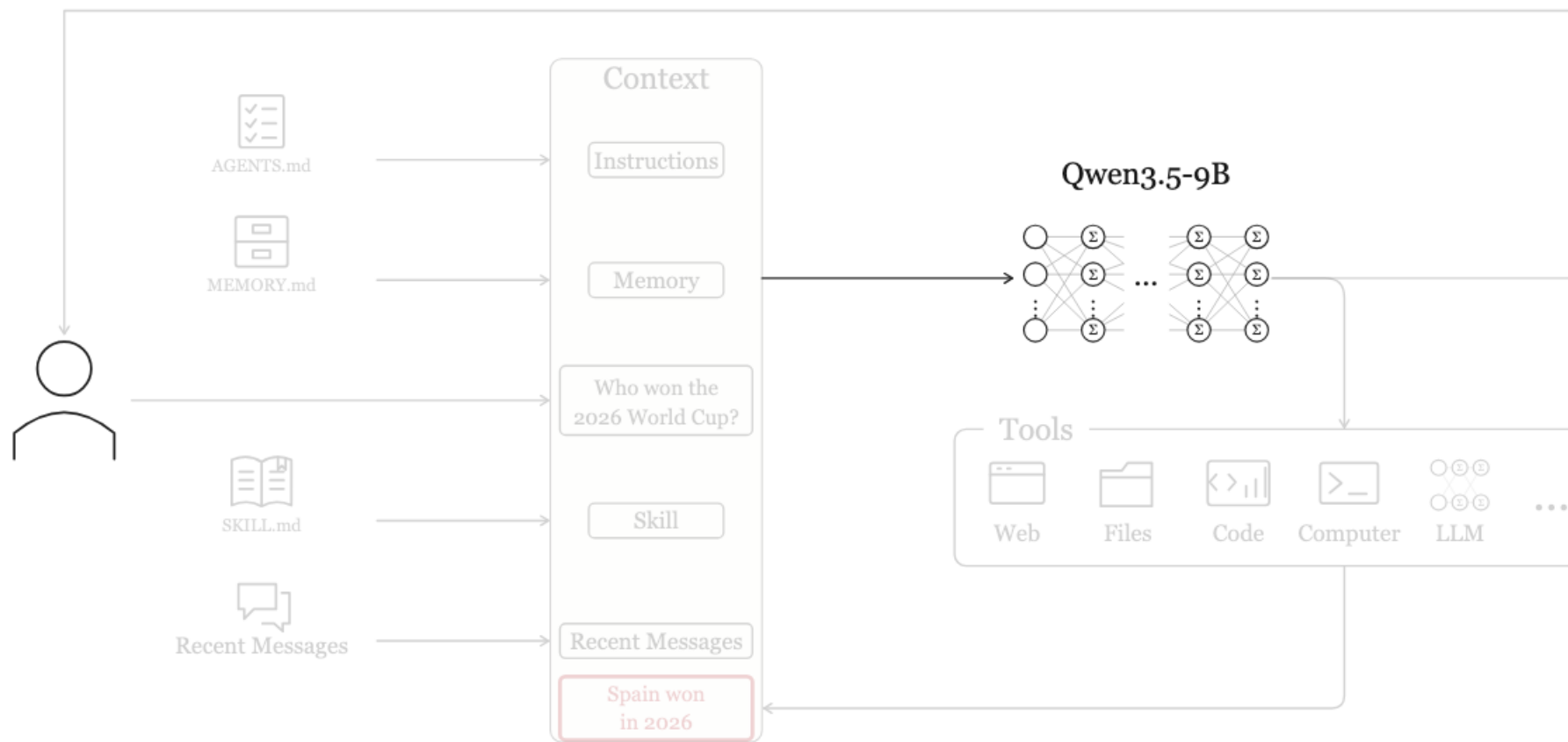


Compacting the Context Window

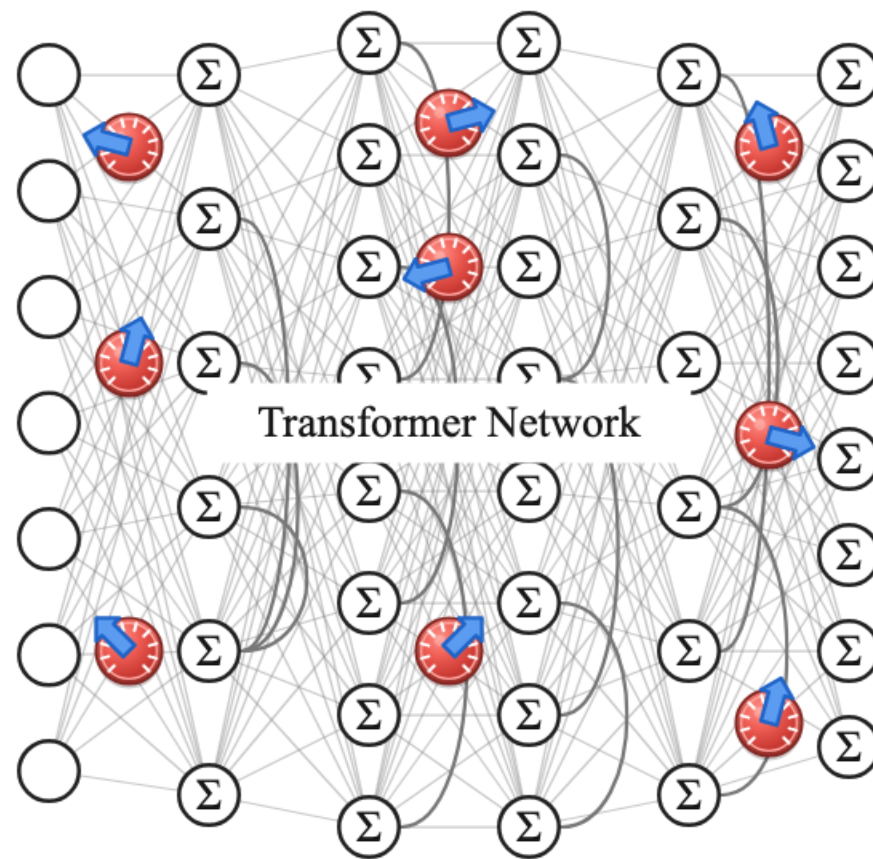


Context Window

What Exactly Is This?

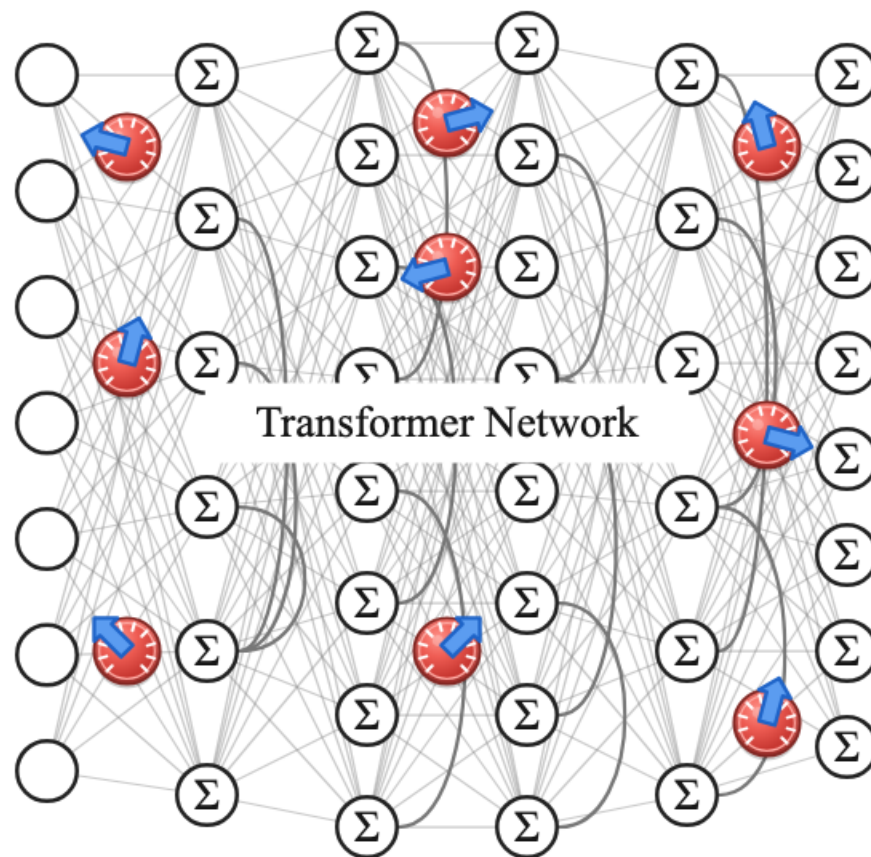


Qwen3.5-9B



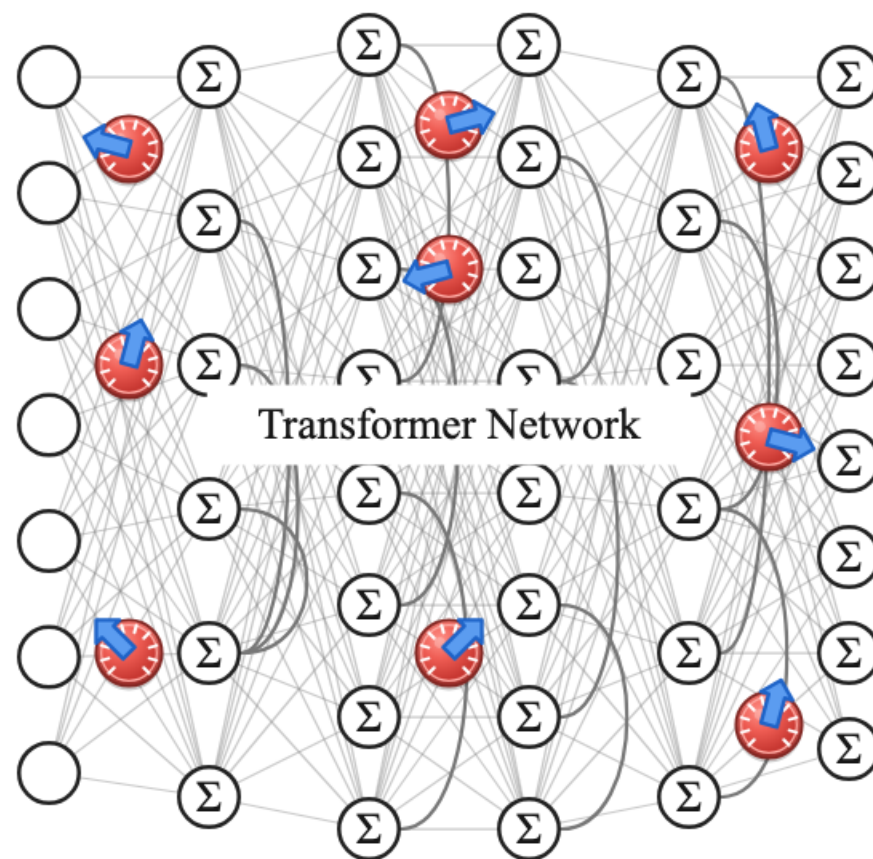
Qwen3.5-9B

An open-weight model family developed by Alibaba



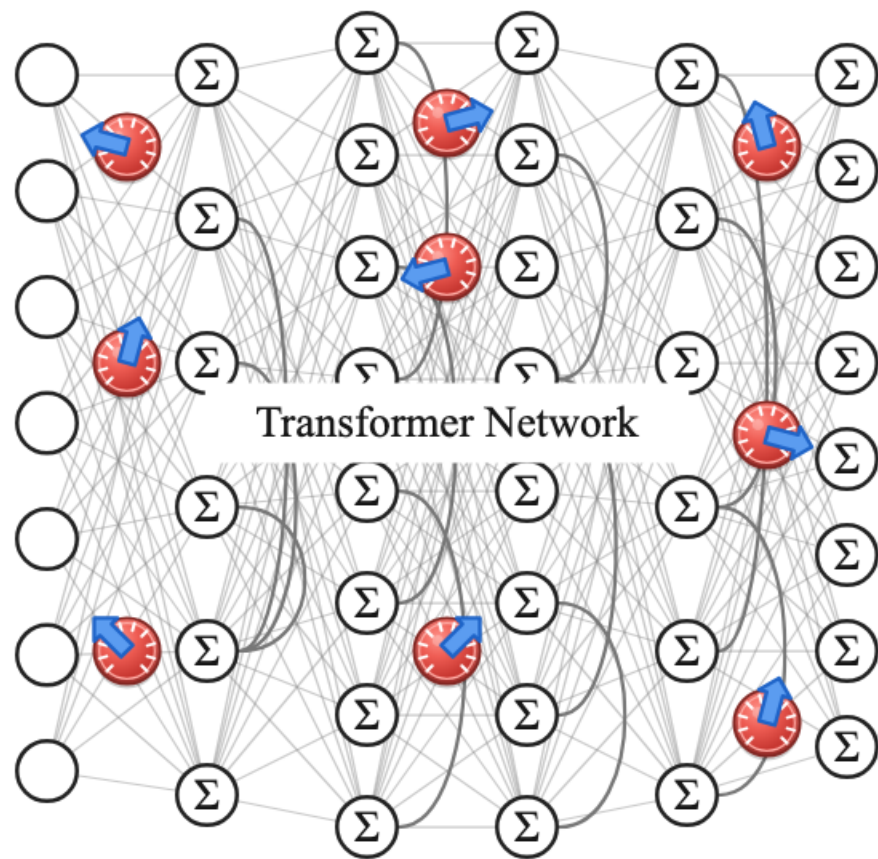
Qwen3.5-9B

Model version, introduced 16 February 2026

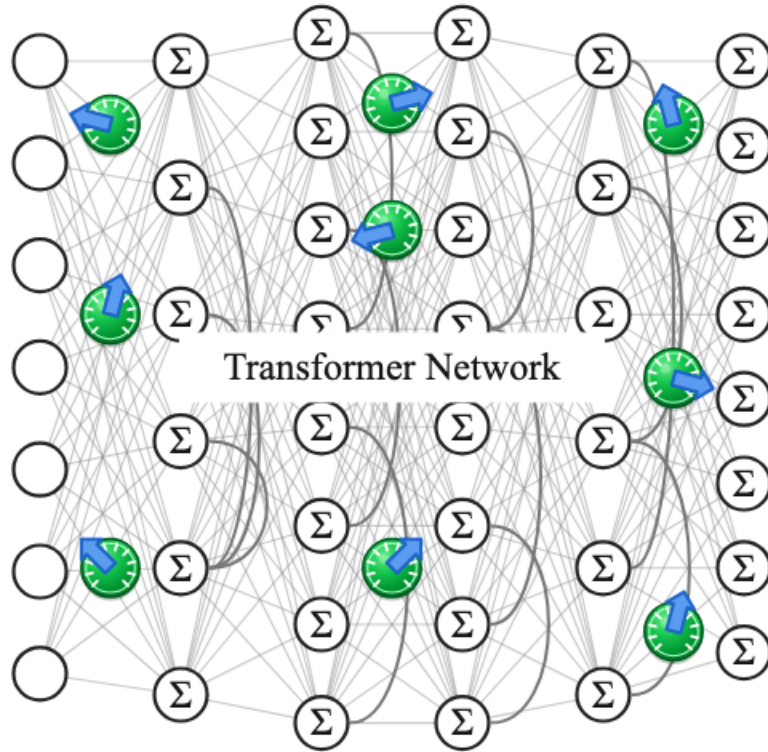


Qwen3.5-9B

About 9 billion learned weights

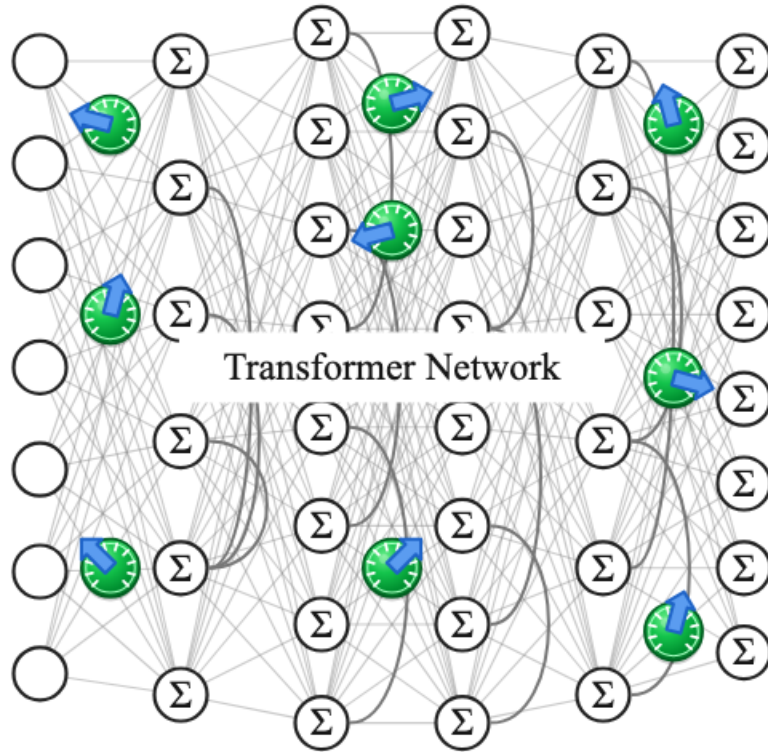


Open Weights



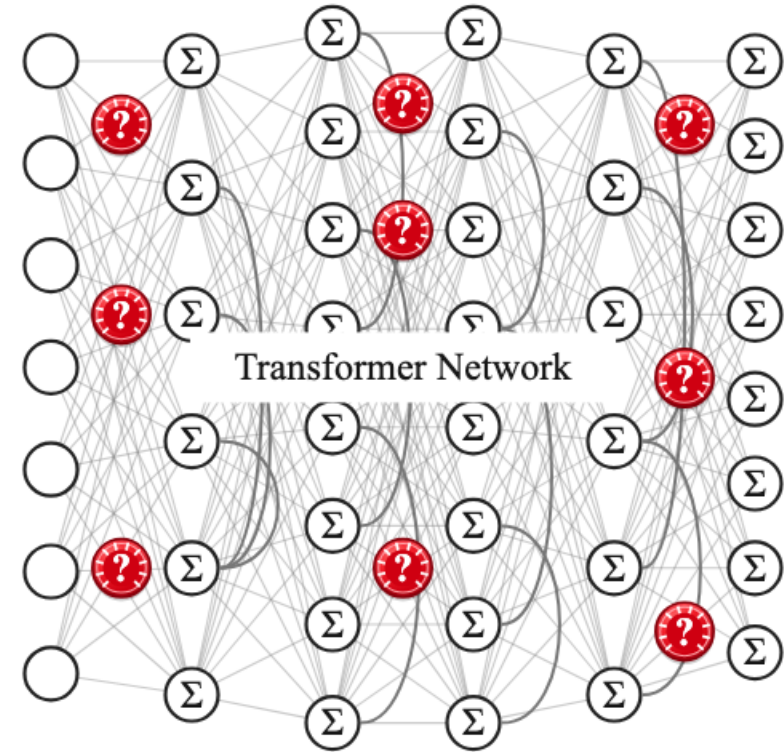
Weights are
available to download

Open Weights



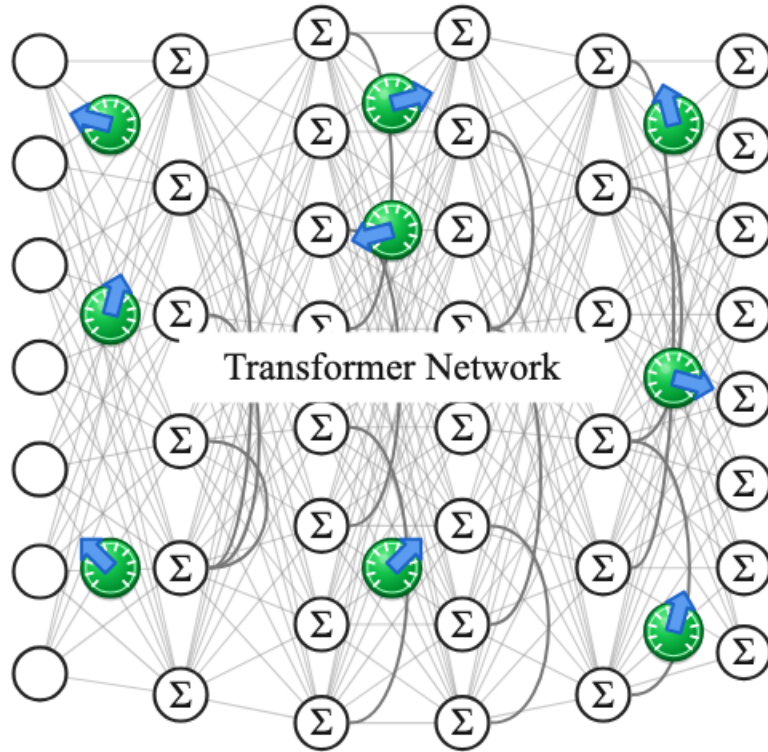
Weights are
available to download

Closed Weights



Weights are
not publicly available

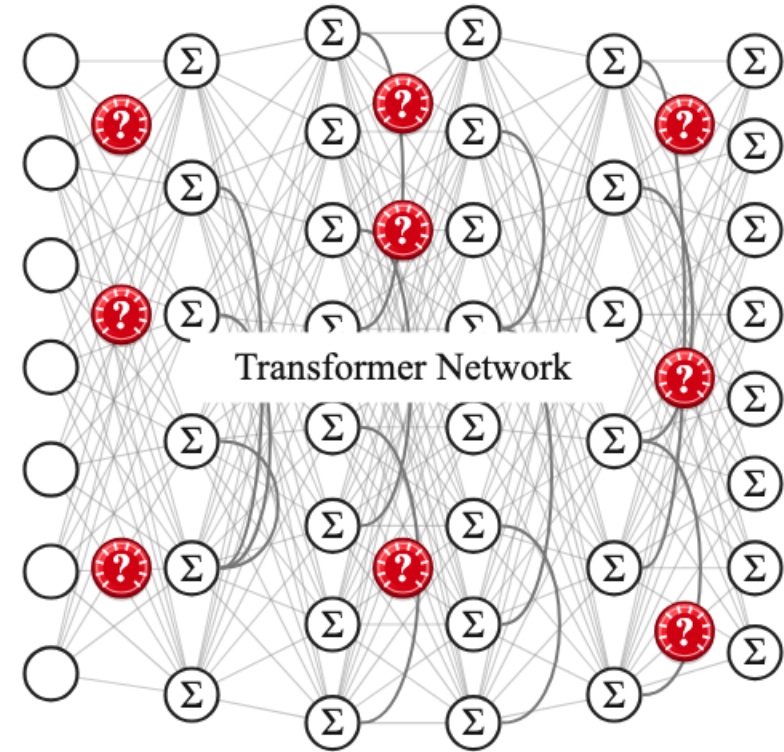
Open Weights



Weights are
available to download

You control the computer

Closed Weights



Weights are
not publicly available

The provider controls the computer

Open Weights



Qwen3.8-2.4T-A95B



DeepSeek-V4-Pro



Kimi K3



GLM-5.3

Closed Weights



GPT-6 Astra



Claude Fable 5.1

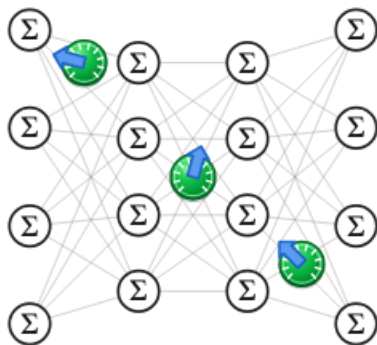


Gemini 3.1 Pro

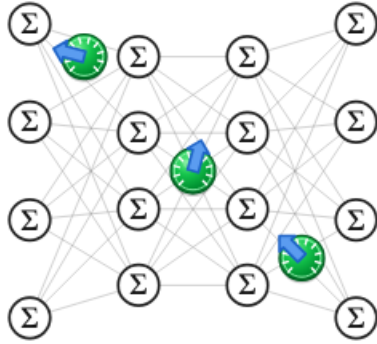


Grok 4.6

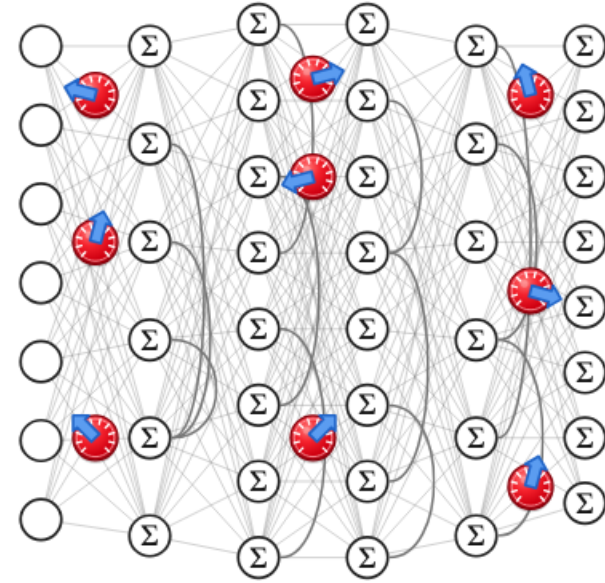
Fewer Weights



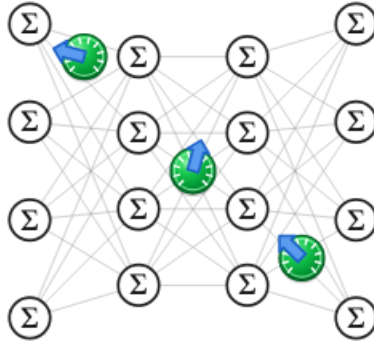
Fewer Weights



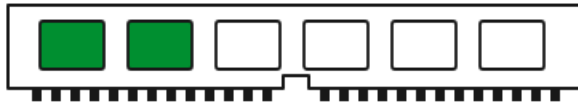
More Weights



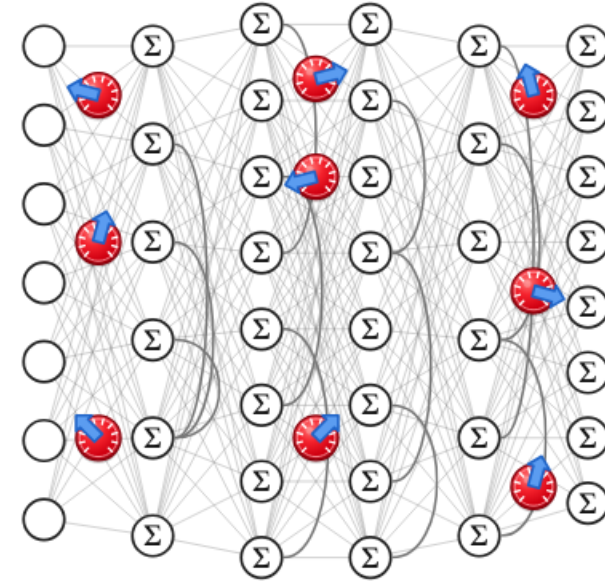
Fewer Weights



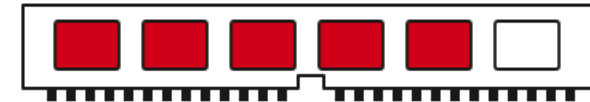
Less Memory



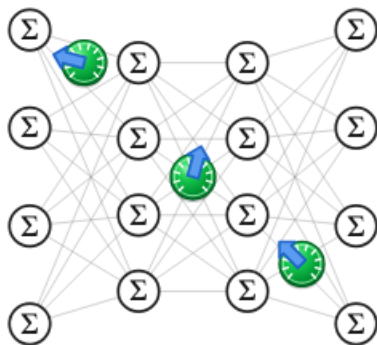
More Weights



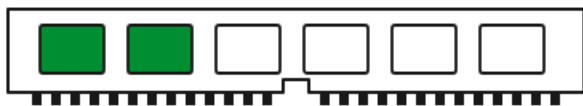
More Memory



Fewer Weights

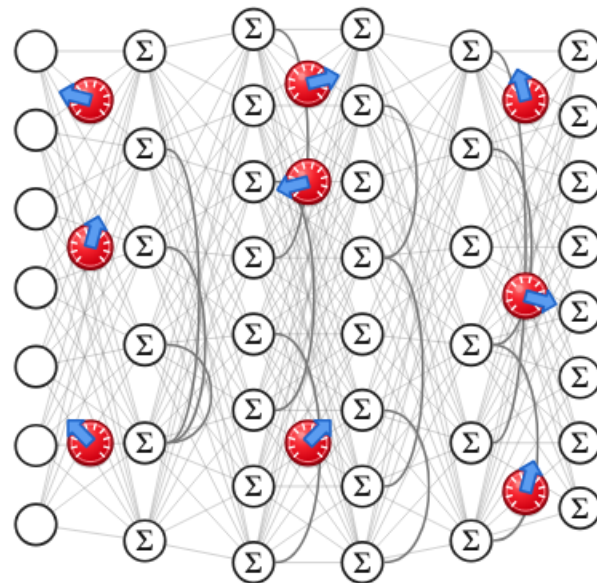


Less Memory

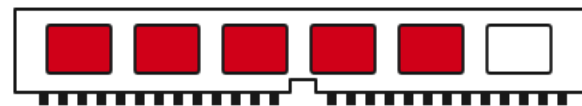


Faster

More Weights

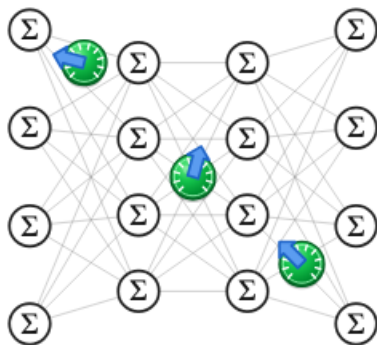


More Memory

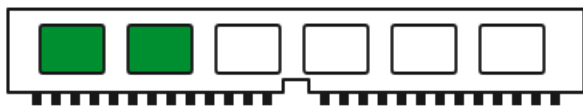


Slower

Fewer Weights



Less Memory

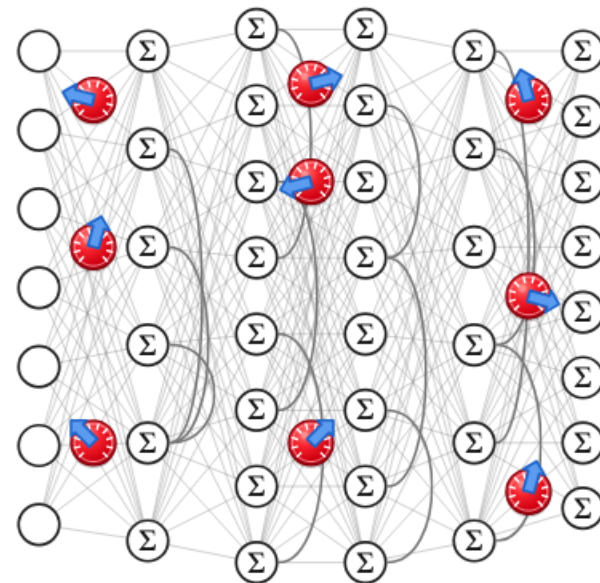


Faster

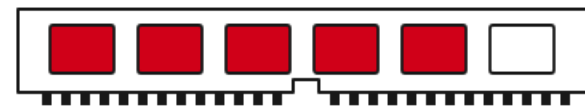


Less Capable

More Weights



More Memory



Slower



More Capable

Fewer Weights



Qwen3.5-9B

9 billion weights · Max Context: 256K

iPhone  HK\$15K

 15 tokens/s · 4K context

More Weights

Fewer Weights



Qwen3.5-9B

9 billion weights · Max Context: 256K

iPhone  HK\$15K

 15 tokens/s · 4K context



Qwen3.8-27B

27 billion weights · Max Context: 256K

Mac mini  HK\$30K

 25 tokens/s · 16K context

More Weights

Fewer Weights

More Weights



Qwen3.5-9B

9 billion weights · Max Context: 256K

iPhone 💰 HK\$15K

⚡ 15 tokens/s · 4K context

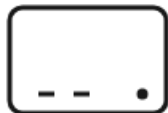


Qwen3.8-27B

27 billion weights · Max Context: 256K

Mac mini 💰 HK\$30K

⚡ 25 tokens/s · 16K context



GLM-5.3-Flash

320 billion weights · Max Context: 1M

Mac Studio 💰 HK\$60K

⚡ 30 tokens/s · 32K context

Fewer Weights



Qwen3.5-9B

9 billion weights · Max Context: 256K

iPhone  HK\$15K

⚡ 15 tokens/s · 4K context

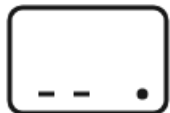


Qwen3.8-27B

27 billion weights · Max Context: 256K

Mac mini  HK\$30K

⚡ 25 tokens/s · 16K context



GLM-5.3-Flash

320 billion weights · Max Context: 1M

Mac Studio  HK\$60K

⚡ 30 tokens/s · 32K context

More Weights

DeepSeek V4 Pro

1.6 trillion weights · Max Context: 1M

8 × NVIDIA B300



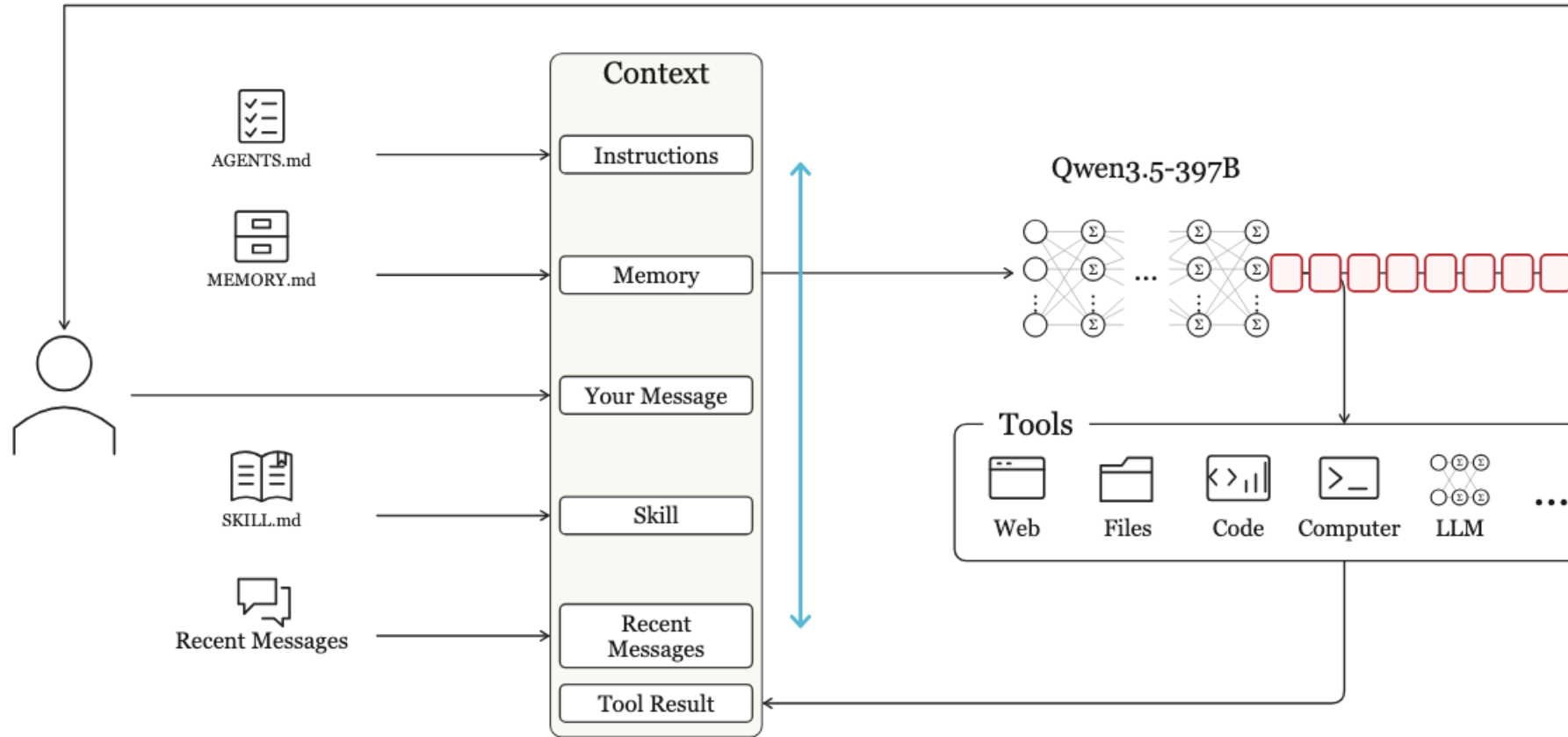
HK\$5M

Overseas price reference

+ Power, Cooling and Server Space

⚡ 400 tokens/s

More Context, More Work



64K Context

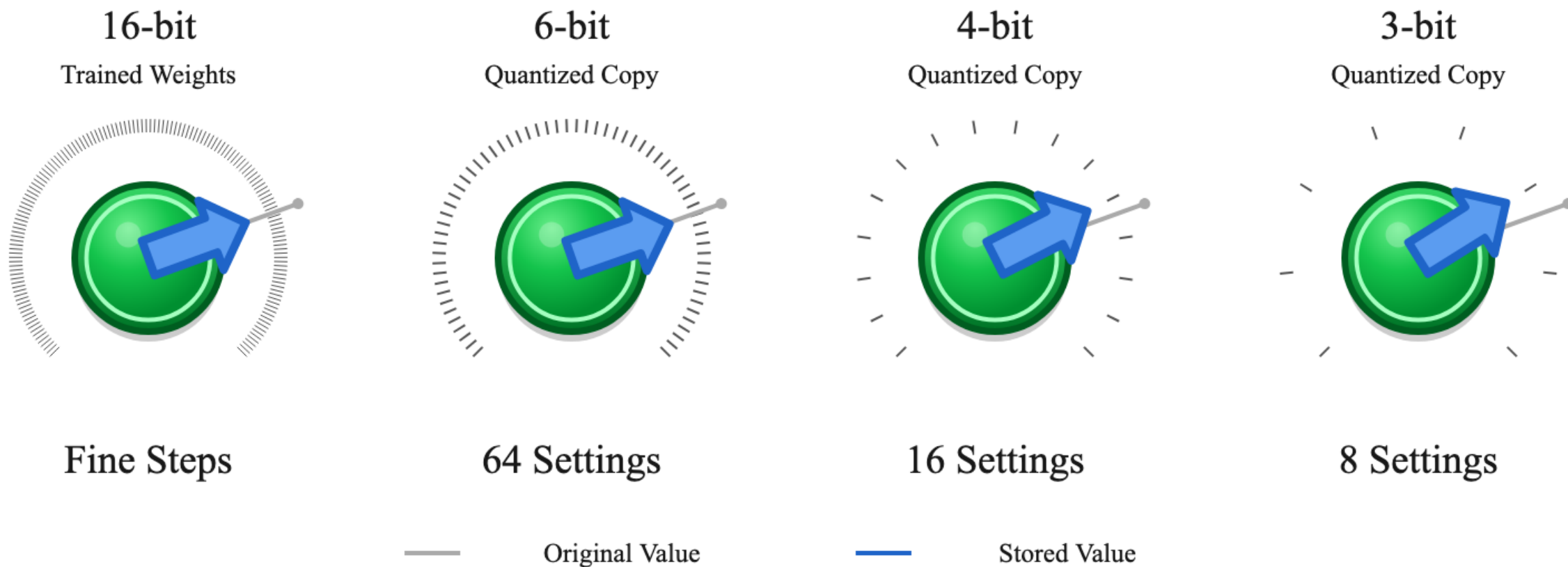


25 tokens/s

Context Window

Longer context can mean slower replies

Quantization: Fewer Knob Positions



Trained Weights → Quantized Copies

Schematic

Quantization: Fewer Bits, Less Computer Memory

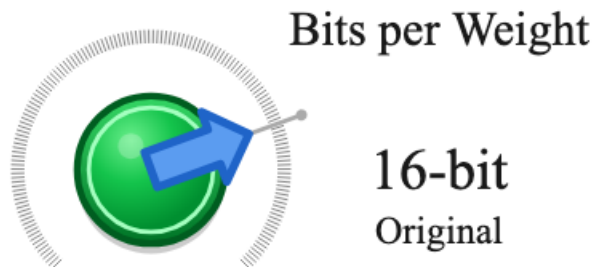
Qwen3.5-397B

M3 Ultra · 512GB · 4K Context



Memory

Tokens/s



16-bit
Original

>794 GB
Exceeds 512GB

Not Run

Quantization: Fewer Bits, Less Computer Memory

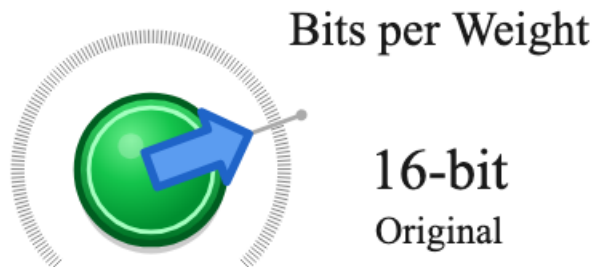
Qwen3.5-397B

M3 Ultra · 512GB · 4K Context



Memory

Tokens/s



>794 GB
Exceeds 512GB

Not Run



≈304 GB

≈30

Quantization: Fewer Bits, Less Computer Memory

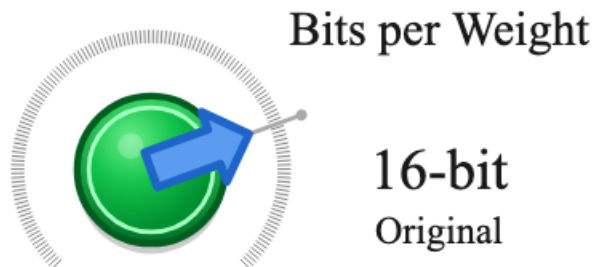
Qwen3.5-397B

M3 Ultra · 512GB · 4K Context



Memory

Tokens/s



>794 GB
Exceeds 512GB

Not Run



≈304 GB

≈30



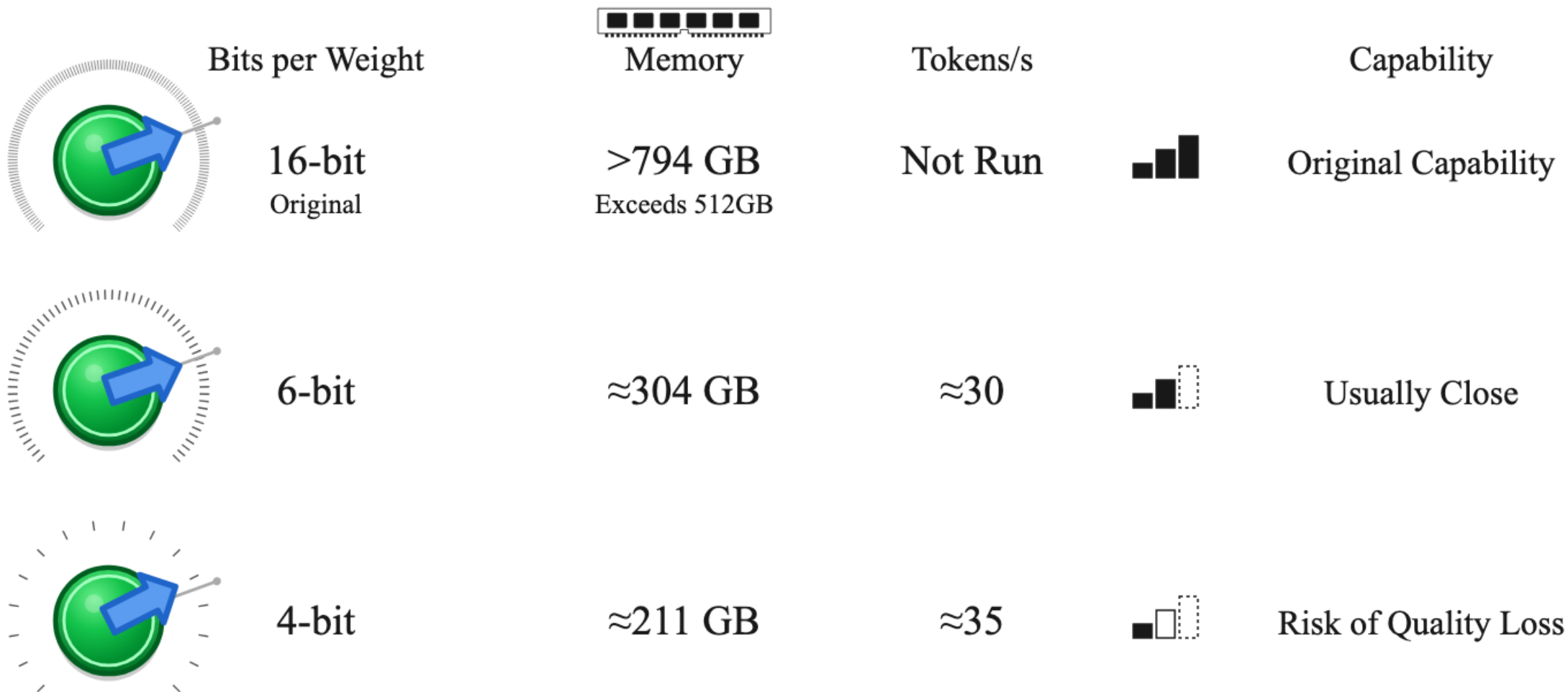
≈211 GB

≈35

Quantization: Fewer Bits, Less Computer Memory

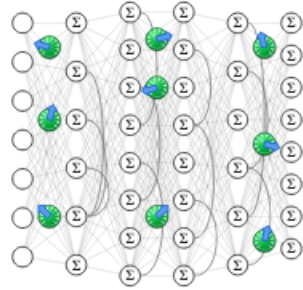
Qwen3.5-397B

M3 Ultra · 512GB · 4K Context



You Run the Model

Self-Hosted



Open Weights

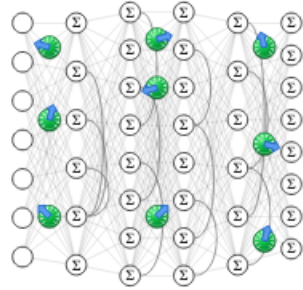


Your Local Server

You Manage the Deployment

You Run the Model

Self-Hosted



Open Weights



Your Local Server

or

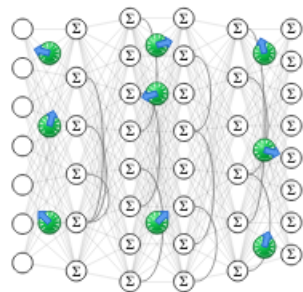


Your Cloud Server

You Manage the Deployment

You Run the Model

Self-Hosted

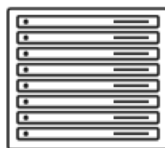


Open Weights



Your Local Server

or

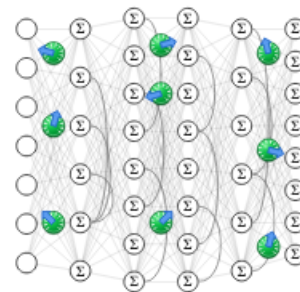


Your Cloud Server

You Manage the Deployment

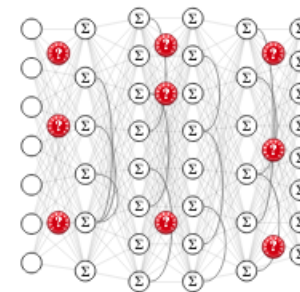
A Provider Runs the Model

Managed Service

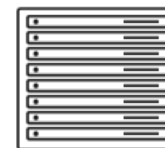


Open Weights

or



Closed Weights

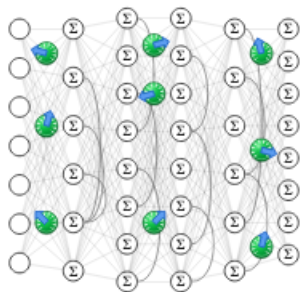


Provider's Server

Provider Manages the Deployment

You Run the Model

Self-Hosted

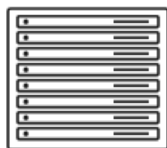


Open Weights



Your Local Server

or

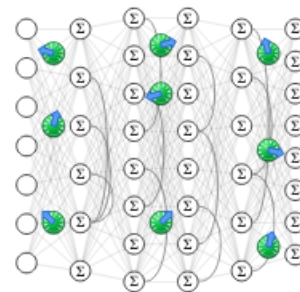


Your Cloud Server

You Manage the Deployment

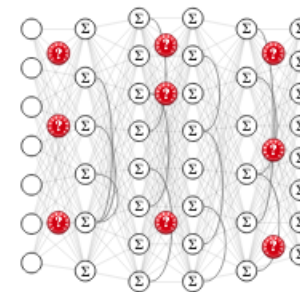
A Provider Runs the Model

Managed Service

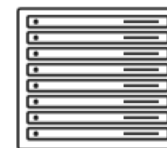


Open Weights

or

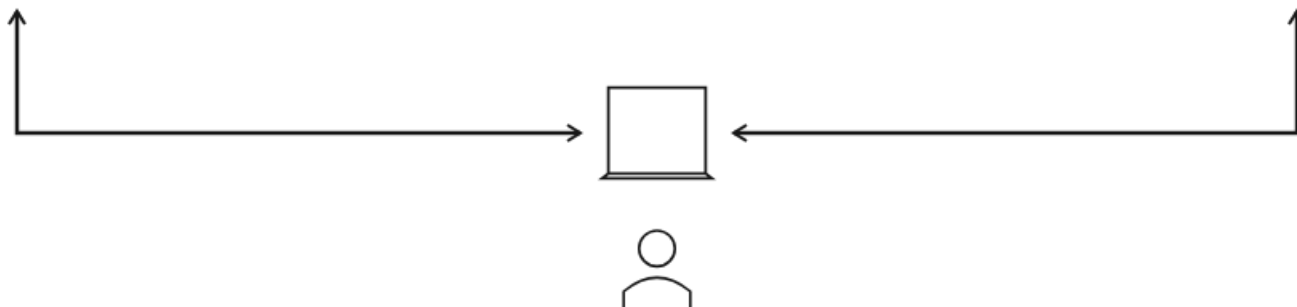


Closed Weights



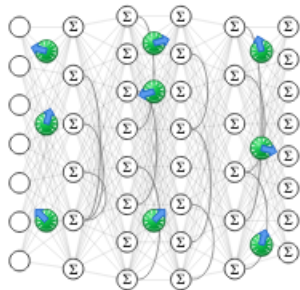
Provider's Server

Provider Manages the Deployment



You Run the Model

Self-Hosted

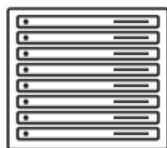


Open Weights



Your Local Server

or

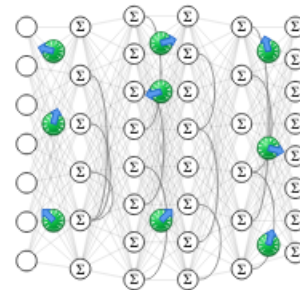


Your Cloud Server

You Manage the Deployment

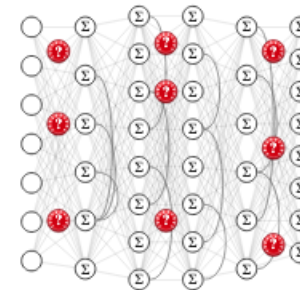
A Provider Runs the Model

Managed Service



Open Weights

or

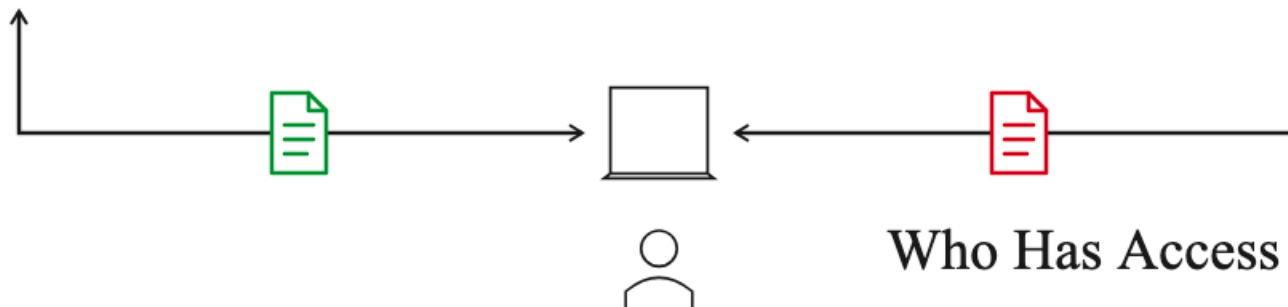


Closed Weights

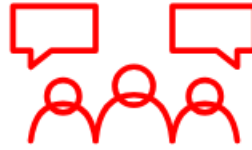


Provider's Server

Provider Manages the Deployment



Who Has Access to Your Data?



1 Ed Login
on your phone



2 Camera App



3 Scan QR



4 Submit

ACTIVITY · EXIT TICKET