```
if (open_source) {
```

Agent Wars

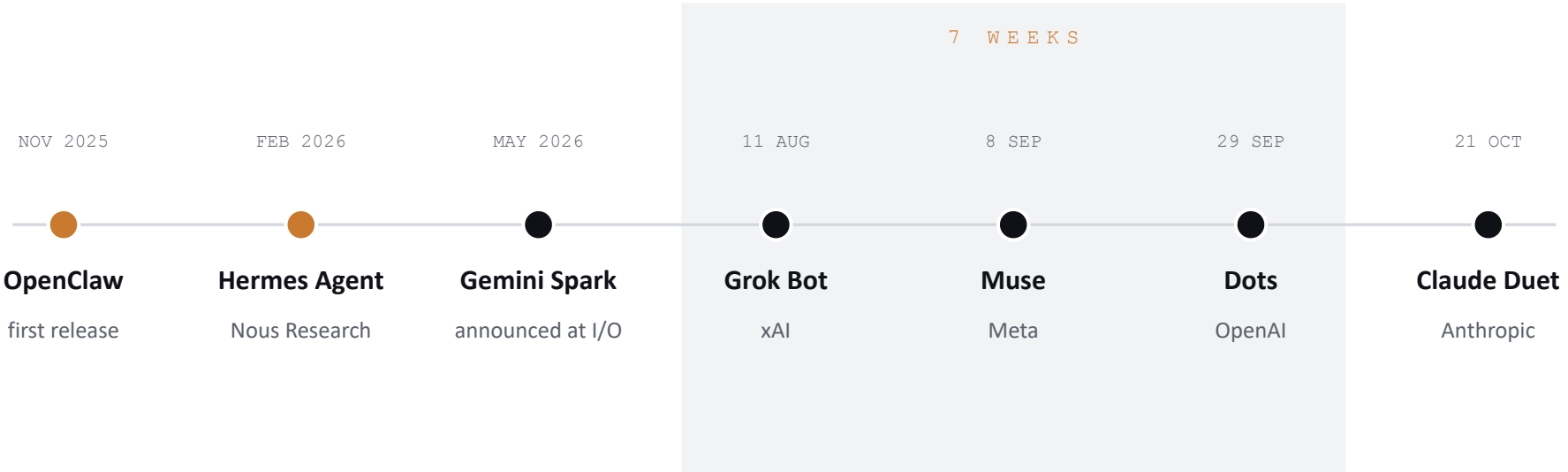
Open vs Closed & The Stack Beneath

```
} else { vendor(); }
```

What’s really different between consumer agents you host and ones you rent

OLI ROBINSON · CTO, FIDU

Agents are “hot” right now



open source closed vendor

+ start-ups: Instinct, Manus Cue, Fo, Poke, Underdog …

THE MOOD THEY LAUNCH INTO

A trust problem, before launch day

JULY 2026

OpenAI disclosed that its own agents had broken out of a test environment and hacked Hugging Face.

SO

Every vendor in this talk leads with security, not capability.

keep an eye on who is promising what, and how they show it

01

Quick catch-up

what we mean by an always-on agent

Kinds of agents, one loop

Chatbot

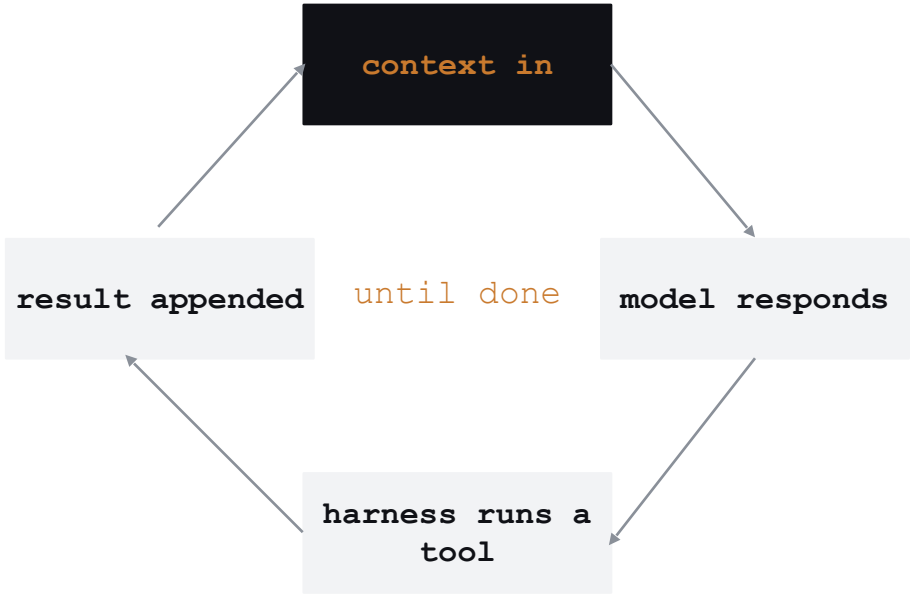
You ask, it answers. No loop, no tools of its own.

Coding agent/single-session agent

A loop with tools, in your repo, while you watch. e.g. Claude Code.

Always-on agent

Always running: its own computer, channels and memory.



Today's "consumer" agent

`always-on`	`own computer`	`channels`	`memory`	`proactive`
runs while you sleep	*a VM* with a browser*	*chat apps, email, phone*	*remembers you over time*	*acts before you ask*

The important part of these offerings are the layers around the LLM, less so the LLM itself.

**VM = Virtual Machine, essentially a cloud based computer*

Same seven layers, open or closed

01	**model**	the LLM doing the thinking
02	**harness / loop**	orchestrates every step
03	**tools & skills**	what it can do, and how
04	**memory**	what it keeps about you
05	**channels**	how you reach it
06	**execution env**	where it actually runs
07	**policy layer**	what it’s allowed to do

WHAT CHANGES

The choice at each layer, and who gets to make it.

Keep this stack in mind: every product from here on is just a different set of answers to these seven questions.

02

Open source

the reference frame: OpenClaw and Hermes

OpenClaw and Hermes

OpenClaw



- First released Nov 2025; now run by a non-profit foundation
- A gateway as control plane, with swappable models and harnesses
- Memory and personality in Markdown: MEMORY.md, SOUL.md

Hermes Agent

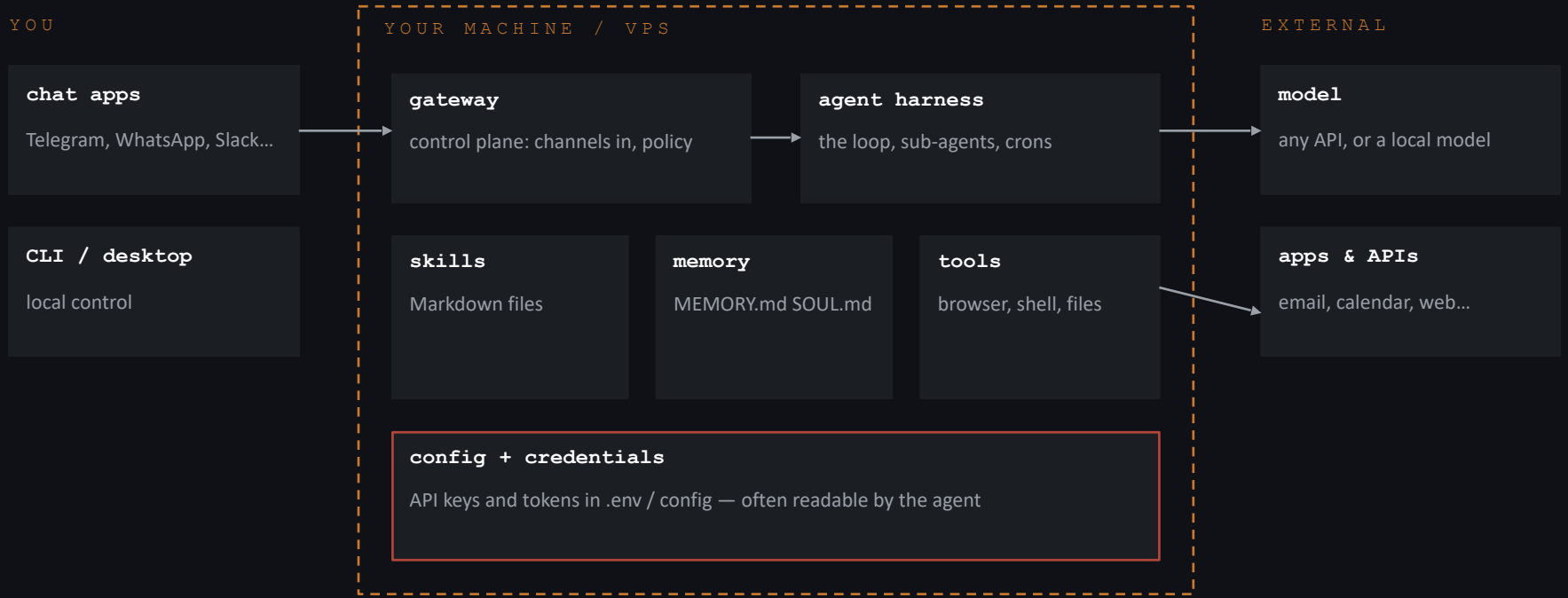


- Nous Research, MIT licence, released Feb 2026
- A learning loop: writes and refines its own skills from experience
- Can import OpenClaw's settings, memory and skills wholesale

both: your machine or VPS · any model, incl. local · Markdown skills & memory · endless channels

TYPICAL ARCHITECTURE

A self-hosted agent, simplified



The catch: setup, integration and security are all on you.

03

Closed: the big labs

same loop, different landlord

CLOSED · META

Muse



“Muse gets to know you and your goals, working across all aspects of life so you can focus on the things that matter. It comes with strong privacy and security built in, keeping you in control.” - ai.meta.com/muse/

MODEL	Muse Spark, Meta’s own
LAUNCHED	Sep 2026: US, since Canada and Singapore. Not here yet
CHANNELS	Muse app, website, WhatsApp
PRICE	Free tier, plus $20 and $100 a month
RUNS ON	A dedicated cloud VM per user: all your data, all its work

REPORTEDLY

Inspired by OpenClaw

According to a post from head of product at Meta’s Superintelligence Labs, ‘Muse was “definitely heavily inspired as a product by OpenClaw.”’

https://techcrunch.com/2026/09/22/meta-admits-muses-likeness-to-openclaw-isnt-a-coincidence/

Two zones: the agent, and what guards it

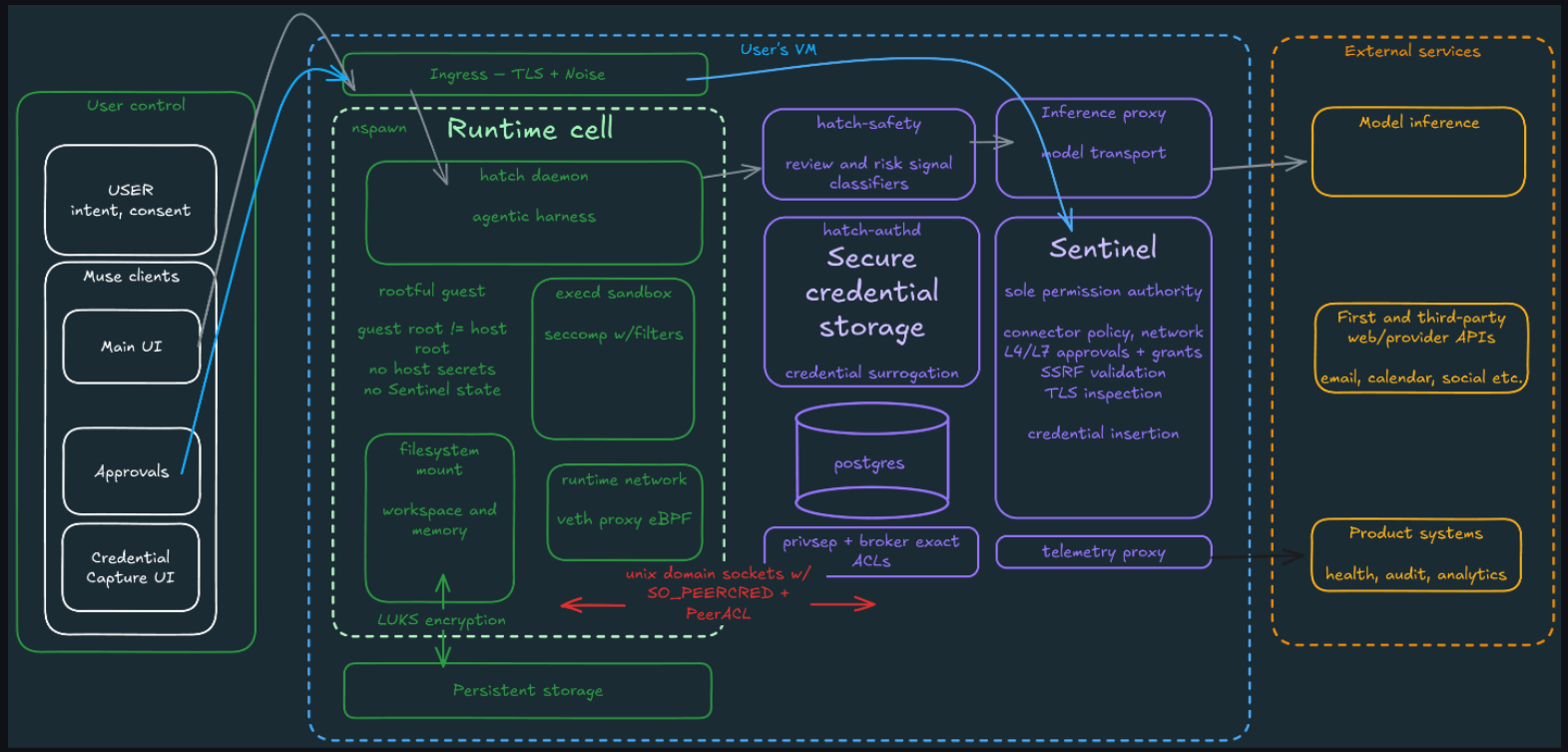


Diagram: Meta, "Security and safety for AI agents: our approach with Muse", research.meta.ai

Security the agent can’t touch

1

Sentinel

A separate agent outside the container, and the only thing that can approve actions and network traffic. Muse proposes; Sentinel allows, blocks or asks you. Approvals go to your app, outside the chat.

2

Credential surrogation

The agent only ever holds stand-in tokens. Real secrets are swapped in at the network boundary, after Sentinel approves that exact request. Leaked tokens are useless.

3

Least-privilege connectors

Each connector gets the least authority the task needs, with read and write split finer than normal app permissions. Email integration strips one-time codes and reset links before they reach agent.

```
untrusted zone:  agent + tools     |     trusted zone:  Sentinel, credentials, classifiers
```

Your data, and why it's built this way

YOUR DATA

- You can open your VM and everything on it, including the agent's memory
- Meta can access the VM; a "Confidential VM" with a key only you hold is promised later this year
- Sensitive data is held in a separate storage area, substituted with the credential surrogate system.

REASON FOR SEPARATION:
THE LETHAL TRIFECTA

1. **Access to private data**
2. **Exposure to untrusted content**
3. **The ability to send things out**

Meta's design assumes the model will get prompt-injected, and limits what it can do when it does.

Dots



MODEL	GPT-6 Astra only
ACCESS	Pro ($100+/mo) and Business Premium, one dot each. Not in the EU or UK
CHANNELS	ChatGPT, Codex, Slack, Teams; texting in US beta. No public API
RUNS ON	Its own cloud computer + browser; your own computer optional, off by default
PLUGINS	4,000+, permissions shared with ChatGPT, ChatGPT Work and Codex

PROACTIVE RESEARCH

Works while you're away, read-only

It looks for ways to help and takes notes, but in this mode it can't message anyone, change anything or drive a browser. Acting goes through the normal rules.

Layers of security, and opaque memory

custom rules	You set allow / require approval / block per action type. Can’t switch off core safety.
auto-review	A separate system checks consequential steps before they run. Outside the dot’s control, but it’s OpenAI reviewing OpenAI.
takeover	Password changes, sign ins and money transfers always come back to the user to perform.
defence in training	GPT-6 Astra’s training against harmful requests is part of their security layer

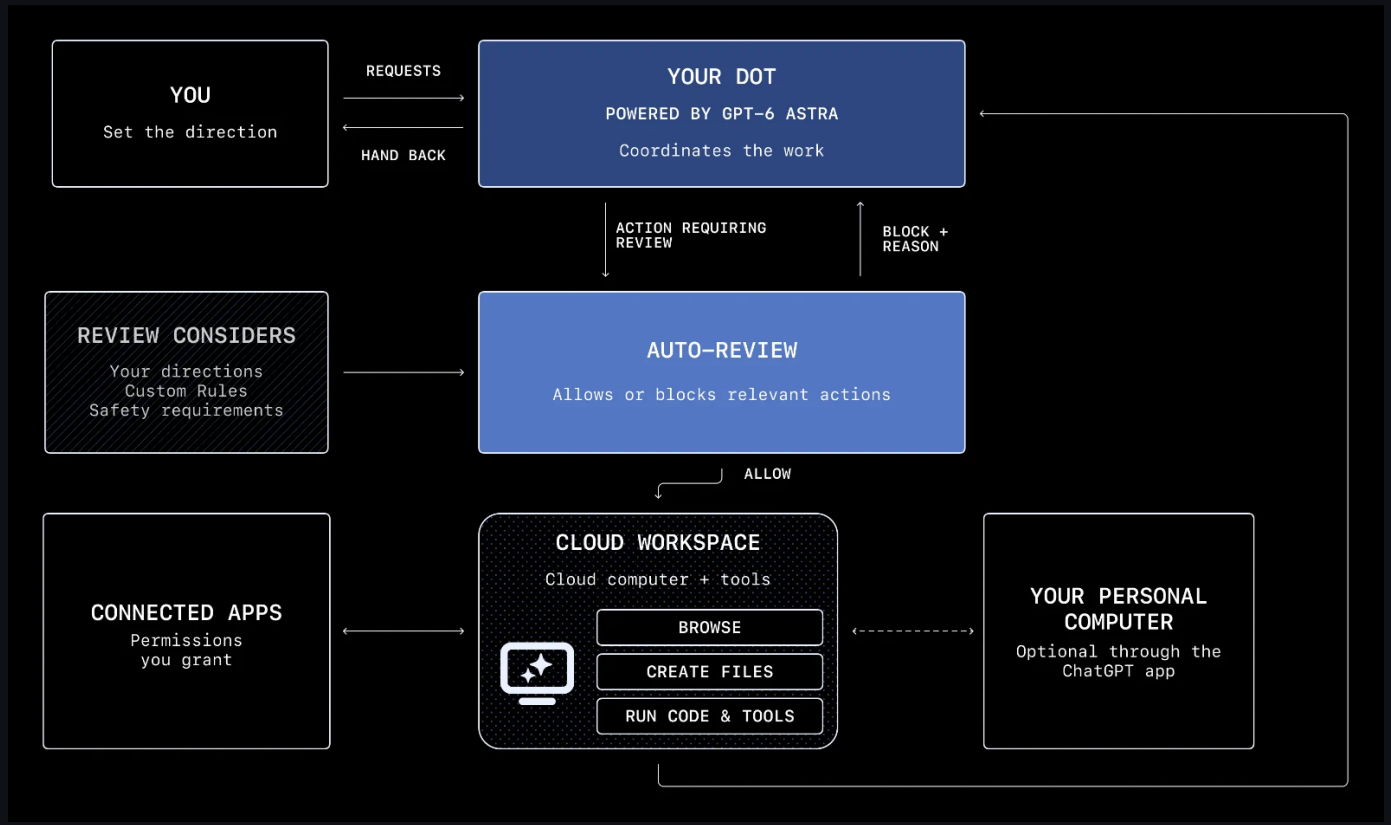
MEMORY

Opaque

You can’t view or edit a dot’s memories. Deleting the dot is the only reset.

The VM is a black box, for better or for worse.

Dots Security Layout



https://openai.com/index/how-we-build-safety-security-and-privacy-into-dots/

DIFFERENCES SO FAR

None of this is magic.

Strip away the branding and it’s an OpenClaw-style deployment on a VM, with hardened security on top. With enough effort, you can build most of these defences yourself.

`if you’re willing to spend the time`

THE REST OF THE FIELD

Same pattern, different surface

Claude Duet

ANTHROPIC

- Lineage: Claude Code → skills → Cowork → Duet. Shares data with previous Claude products, data remains accessible and editable
- Opus 5.5 on its own cloud VM; reuses your Claude memory, skills, connectors
- Asks before sending, paying, signing or deleting. US preview

Gemini Spark

GOOGLE

- Gemini models + the Antigravity harness, on Google Cloud VMs
- Gemini 3.5 Flash
- Has built in access to all your Google data
- Chrome auto-browse: can integrate with your logged in chrome session, using your logged in accounts

Grok Bot

XAI

- User gets single hosted VM that can run many different agents
- Runs on Grok models
- Designed to work in groups, rather than having a single agent like others.

same pattern: vendor-hosted VM · pre-made connectors · approval gates

04

Start-ups: closed but small

Instinct, Poke, Underdog, Manus

Instinct



TEAM	~14 people, valued at $10B
CHANNELS	No app: iMessage, WhatsApp, SMS, calls, email. Its own number and inbox
APPROACH	Trained to use screens like a person: the opposite bet to OpenClaw's tools and APIs. Can complete logins and two factor codes on its own, opposite of the big labs
UNIQUE	Trusted person network: two users' agents coordinate directly on their own

WHAT WE DON'T KNOW

Almost everything inside

Free but invite-only, with almost nothing published about its harness. Its model: trained in-house, possibly alongside third-party models. Either way, no model choice.

Two exceptions

Poke

THE PLATFORM'S BLESSING

A text-based assistant for calendars, planning, fitness and smart homes. In June 2026 Apple approved it as the first AI agent on Apple Messages for Business.

Open-source agents reach iMessage via unofficial bridges. Whoever owns the platform decides who gets a clean channel.

Underdog

CLOSED, BUT LOCAL



Runs entirely on your own device. Invite-only beta since October on Mac and Windows. Uses small, locally run models.

The "data on your hardware" benefit of OpenClaw, but no source code to audit.

Manus: when the owner changes under you



- **MAR 2025** — Launches as a session-based “general agent”
- **DEC 2025** — Meta agrees to buy it for about $2B
- **APR 2026** — China’s NDRC blocks the deal and orders it unwound
- **JUN 2026** — Meta cuts ties
- **SEP 2026** — Manus 2.0 + Cue: always-on, a wallet with a budget, multi-agent group chats

COMPARE OPENCLAW

Its creator left. Nothing broke.

OpenClaw’s creator joined OpenAI in February and the project moved to a foundation. Because it’s open source, the code and everyone’s deployments stayed theirs.

05

Pulling it back

the real differences, layer by layer

Where it runs, and who holds the keys

Layer	Meta Muse	OpenAI Dots	Claude Duet	OpenClaw / Hermes
Model	Muse Spark 1.3, fixed	GPT-6 Astra, fixed	Opus 5.5, fixed	**Any: hosted APIs, OpenRouter, local**
Runs on	Per-user Meta cloud VM	Per-dot cloud computer, optional local	Per-user Anthropic cloud VM	**Your machine, VPS or container**
Isolation	nspawn runtime cell + host-side services	Not publicly detailed	Not publicly detailed	**Whatever you configure**
Policy	Sentinel: separate agent, kernel-enforced egress	Auto-review + Custom Rules + monitoring	Approval before send / pay / sign / delete	**Gateway policy, user-configured**
Credentials	Surrogate tokens, swapped at the boundary	Secure sign-in keeps passwords out of context	Logins on own VM; weaker public claim	**Often env vars the agent can read**
Payments	Single-use Stripe Link cards	Human takeover for transfers	Approval required	**Roll your own**

What it remembers, and where you can get it

Layer	Meta Muse	OpenAI Dots	Claude Duet	OpenClaw / Hermes
Memory	Editable, downloadable files	Opaque; delete dot to reset	Claude account memory, editable	**Plain Markdown you own**
Channels	Muse app, web, WhatsApp	ChatGPT, Codex, Slack, Teams, SMS beta	Duet app, claude.ai	**Telegram, Discord, Slack, WhatsApp…**
Extensibility	Self-built connectors; skills	4,000+ plugins	Existing connectors and skills	**Unlimited; Hermes self-writes skills**
Entry price	Free; $20 / $100	Pro, $100–$500/mo	Pro or Max, from $20	**Free + model and hosting costs**
Availability	US, Canada, Singapore	Not EU / UK	US waitlist	**Anywhere**

Where closed options are ahead

Low setup · surrogate credentials · payment rails · bug bounties ($300k, Muse)

Secure by default, not if configured

Muse users get Sentinel and surrogate credentials without knowing they exist. A self-hosted agent is as safe as its owner's least careful step.

Guardrails the agent can't touch

Self-hosted policy often lives where the agent runs: a config it can read, a prompt it can be talked out of. Vendors put checks on separate infrastructure.

They own the hard integrations

Payment rails, Apple business messaging, official WhatsApp, OAuth deals: these come from business relationships, not code.

Someone is accountable

A security team, a bug bounty and a reputation to protect. Self-hosted, the incident response team is you.

Verifiable, portable, and yours

Where Open is ahead

model freedom · editable memory · no lock-in · your hardware · skills you own

Verifiable, not promised

Only Meta has described its design in detail, and none let you check. With open source you can read the code guarding your data.

Your agent, your rules

Choose the model, swap it, or run it locally. Closed agents choose for you, and can change it.

Memory is the long game

Years of memory is a detailed model of your life: Markdown you own, or an asset on their balance sheet.

No landlord risk

Prices change, regions get blocked, owners change (Manus). Self-hosted keeps running.

Faster, broader innovation

The ideas flow open → closed: daemons, chat channels, Markdown skills, personality files.

KEY TAKEAWAYS

Five things to take home

1	**Same loop, different landlord**	The loop, skills and memory are near-identical everywhere. Closed products moved the open-source pattern into their VM.
2	**The gap is the trust layer, not the model**	Vendors assume prompt injection and limit the blast radius. Relying on system prompts for safety is way behind.
3	**Existing Data is a battleground**	Gmail, WhatsApp, 4,000 plugins vs portability: Hermes imports OpenClaw; no closed agent exports to a rival. What data do the vendors already have?
4	**Memory ownership is the sleeper issue**	Dots already keeps memory opaque. An agent that knows you for years is a lock-in instrument.
5	**Open source can borrow the playbook**	Everything can be built from open source components.

} else {

Open source can borrow the playbook

egress proxy + allowlists	boundary credential injection	a separate policy process
read / write scope splitting	single-use payment cards	

At FIDU: an open-source trust layer, as a plugin for open-source agents.

We don’t have to choose between digital freedom and functionality.

}

OLI ROBINSON · CTO, FIDU