Beyond The Clouds

Local AI, OpenClaw and Data Sovereignty

velociti.me

+852 9880 5017

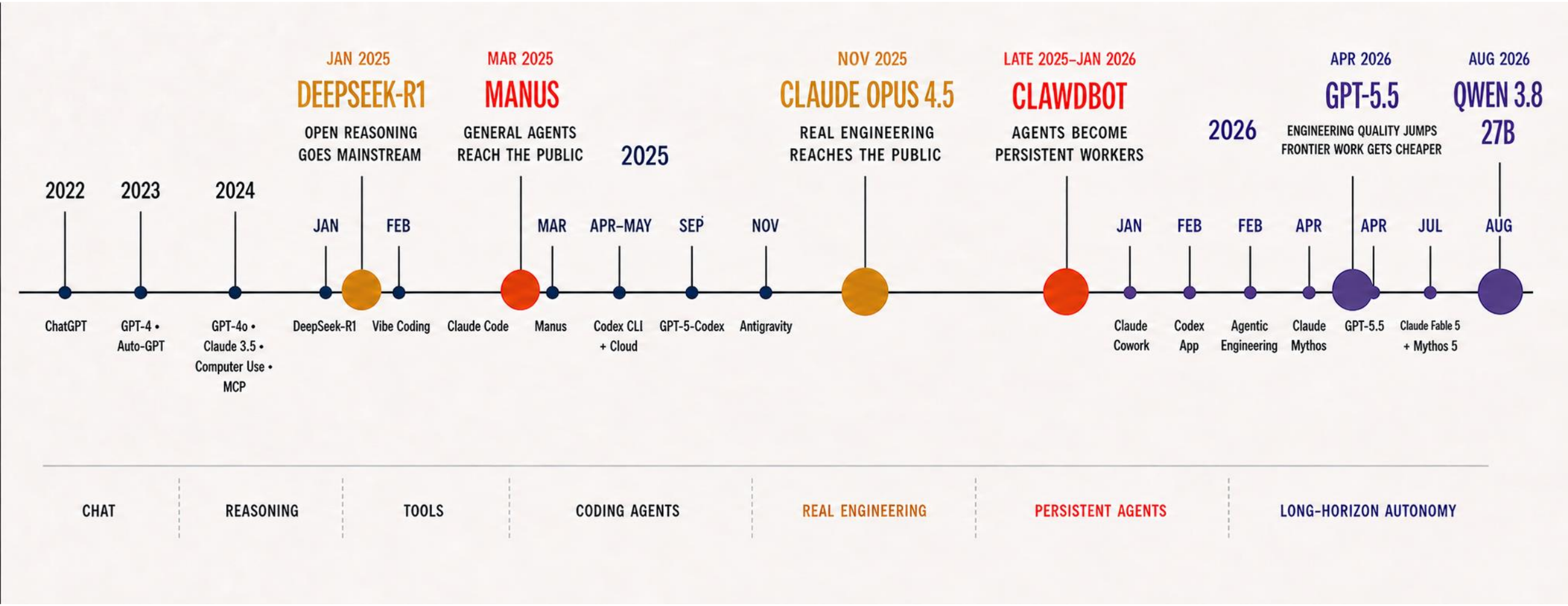
marco@velociti.me

Explore Velociti

Presented by JR and Marco

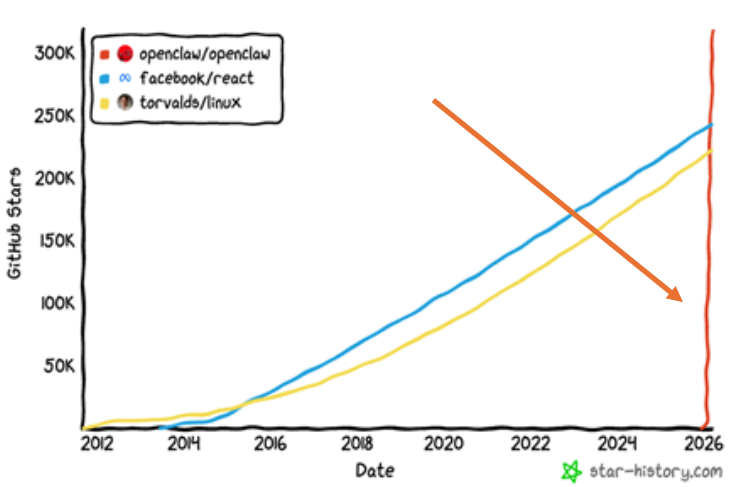
This presentation was made mostly by humans

Where we are in the timeline



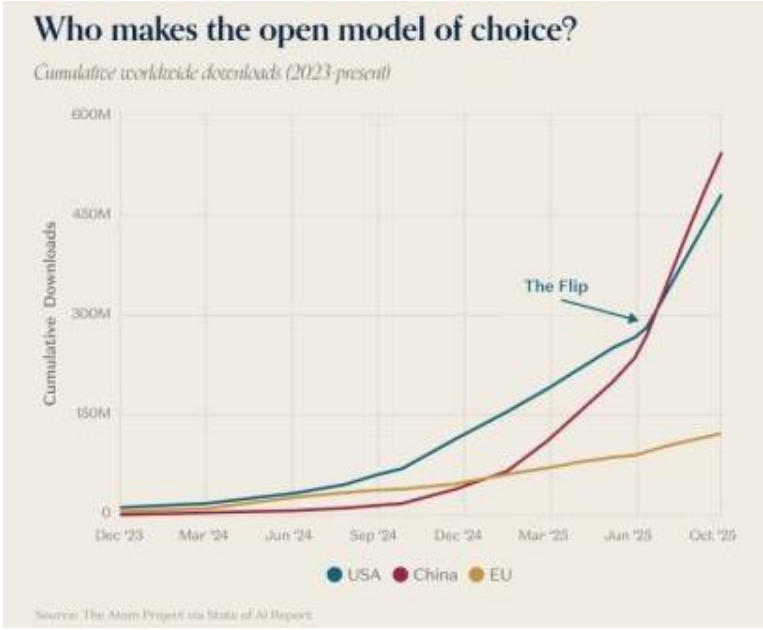
Persistent Workers: OpenClaw and Hermes Low-end AI Democratization

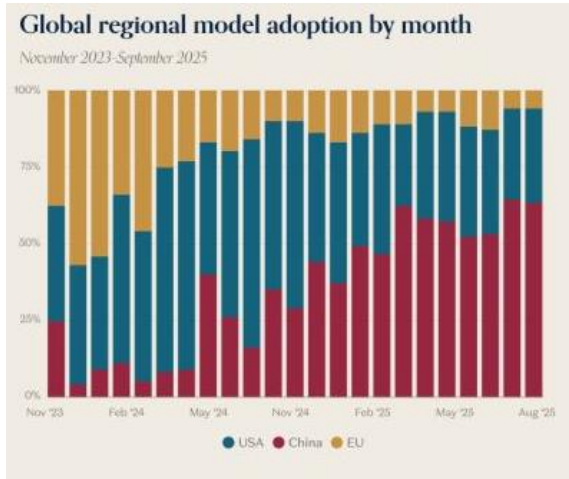
- What they are
  - Bots are agents capable of running **24/7 autonomously**, typically with a variety of tools
  - OpenClaw is a control plane for multi-agent bridging across dozens of channels.
  - Hermes Agent is a self-hosted worker that can **remember across sessions**, and 'self improve'
- What this changes
  - Fully autonomous bots, always running with persistent memory.
    - Ability to manage lower level, repetitive tasks on its' own
- Ripple effects:
  - Fastest growing github stars ever. 350k with openclaw, 275k with Hermes
  - Shortages of mac minis for hobbyists setting up autonomous agents to run persistently



The Surprising AI Democratization by China

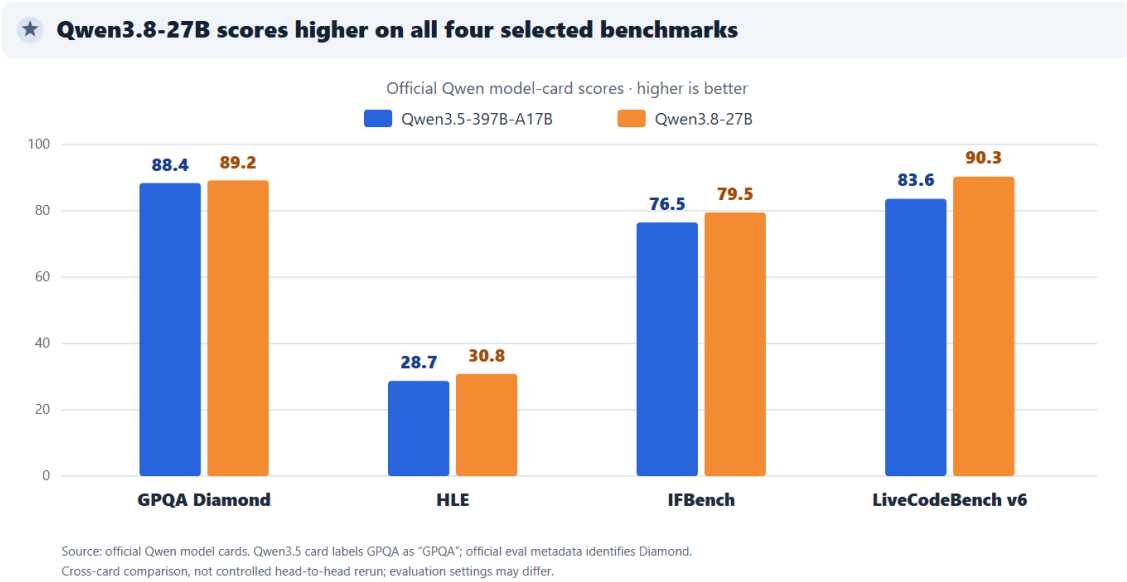
- The Deepseek Moment in 2024
  - The Chinese High-Flyer Asset Management Quant Fund created Deepseek as a ‘side project’ within their fund for quantitative analysis
- Triggering the Opensource Chinese Model
  - Sparking Qwen, Minimax, Kimi, GLM (which actually came first) to quickly catch up.
  - This resulted in Chinese models ‘overtaking’ in Openrouter use in June 2026
- Allows for independent users to **self-host** near Frontier level models on their hardware





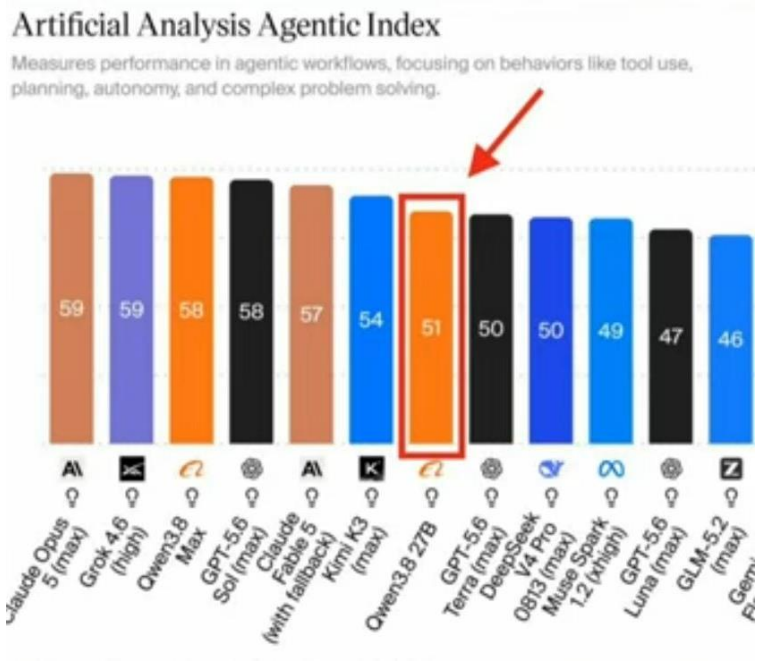
The Hardware Problem – Is it too expensive?

- Smaller models, more impact
  - Qwen3.5 397B (Feb26’) vs Qwen3.8 27B(Aug26’)
- Compounding efficiency
  - Model Optimization with LlamaCPP/VLLM/Unsloth
    - Automate processes, finding model to fit on VRAM+RAM platform, enhanced with GGUF, after selected model from footprint and quantization
  - Context window size
    - Reduce memory footprints, through vectorized chat memory
    - Tooling sent only when needed and prompt injection of constitutional on skill details
  - GPU offloading
    - Layering the LLM like stack of pancakes. Partial offloading and priotization
  - Indexing/Lookup KV Quant/TurboQuant > Dflash2
    - KV Cache (less token use)/Polar vector/Markov Algorithms/Sequential Decoding
- Densing Law (Nature Machine Intelligence)
  - 50% fewer parameters every 3.5 months (seemingly outdated)



Opensource & ‘Clankdom’ Takeover Consumer Local Set Up in August 2026

- End Result: Qwen3.8 27B running on
  - RTX5090 32GB + 64GB DDR5 RAM
    - Total cost roughly USD6k > Now you have a model that beats frontier 6 months ago
    - 8s Latency, 4 concurrent users with ~200tps
- Server-like Desktop to host Qwen3.8
  - Running 4 concurrent Hermes agents, 24/7 – faster than running on their own server
    - Bug fixing, Data Scraping/Collection, Document organization (Excel etc… Inventory management etc..)



0 API/Server Cost!

What is missing – The middle of the market

- Definitions:
  - Top of Market
    - MANGOS (Anthropic, OpenAI, Grok)
      - Own nothing and be happy / Moated tokens
  - Bottom
    - Clankdom
      - Maximized customization, high time costs to set-up
- What is missing?
  - Governance, Data ownership and Security
  - Software for an AI server
  - Middleware Systems

Velociti


AI Market So Far

File Edit Selection View Go Run Terminal Help

Untitled (Workspace)

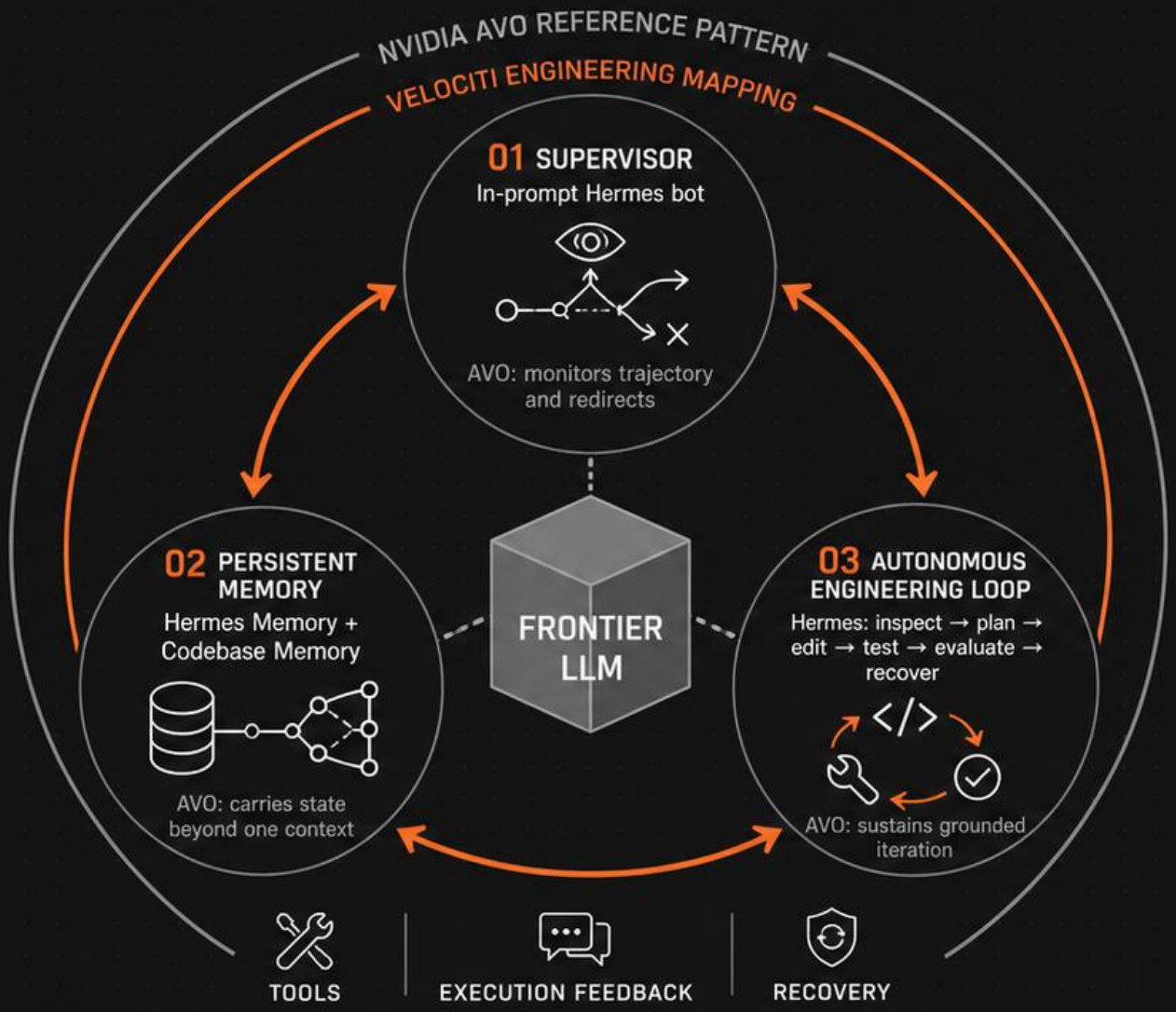
MODELS Local not updated Workspace: Velociti *Codex Port details* Workspace: Velociti

Bridging the Gap – Velociti IDE

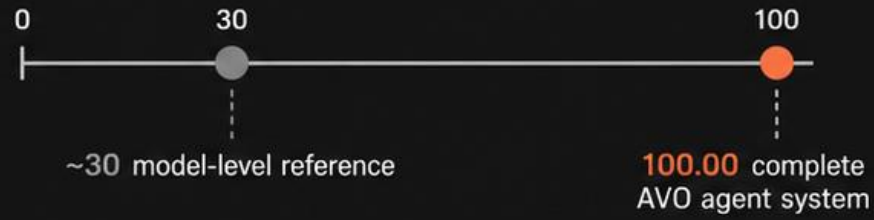
- Velociti is a **local-first** AI workspace.
  - Choose your local model, or BYOK, performance optimized for your system (one click switching between models)
- All your chats(local or cloud), contexts, responses **stored in your own machine**
  - Available to chat on the same window, Codex, Claude Cowork, Grok on subscriptions
- Token optimized, selective memory dispersion w/ token count on each message
  - Your context window **only uses what it needs**

FROM MODEL CALL TO AUTONOMOUS ENGINEERING SYSTEM

The model supplies intelligence. The harness sustains progress.

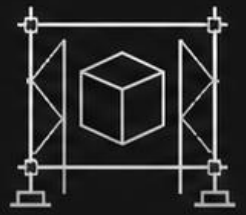


ARC-AGI-3 PUBLIC SET | RHAE



183 / 183 levels • 25 environments

Reference comparison—not a controlled ablation.



SCAFFOLDING VS. RAW MODELS

The surrounding software harness is as critical as the LLM for multi-step intelligence.

ECOSYSTEM CONTROL

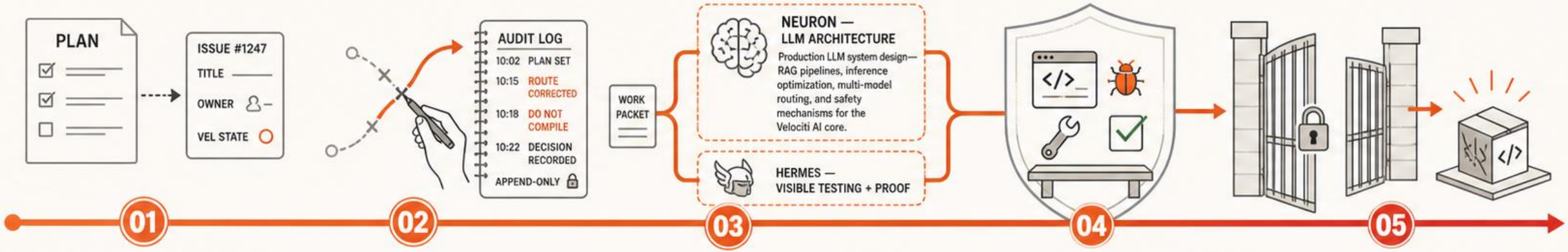
Compose model, harness and secure runtime; tune execution instead of relying only on model weights.

Open stack example: AVO harness + OpenShell runtime

Sources: NVIDIA Technical Blog, 21 Aug 2026 • AVO on ARC-AGI-3 • Where Security Fits in an AI Agent Stack

A CONTROL LOOP FOR LONG-HORIZON ENGINEERING

Plan → steer → decompose → prove → compile



01 PLAN + REGISTER

Scope • checkpoints • acceptance gates

Forgejo/Git issue • owner • VEL state

02 STEER + AUDIT

Short user prompts keep the AI aligned

"Stay in this lane." "Do not compile."

Audit logs preserve decisions

Common processes become reusable runbooks

03 SUB-PLAN + SOULS

Decompose before overload

NEURON — LLM ARCHITECTURE
Production LLM system design—RAG pipelines, inference optimization, multi-model routing, and safety mechanisms for the Velociti AI core.

HERMES — VISIBLE TESTING + PROOF

04 VIRTUAL TEST ARENA

Isolated worktree • focused tests

Real-work Prompt tests • Error Control

Fix every known bug. Retest until clean.

05 COMPILE RUNWAY

Only after real-work tests pass and every known bug is fixed

One controlled compile → activation smoke → evidence closeout

AUDIT LOG

FORGEJO / GIT

ERROR STATUS

VEL STATUS

EVIDENCE RECEIPTS

- AUDIT LOGS PRESERVE CONTINUITY.
- COMMON PROCESSES PRESERVE LEARNING.
- THE COMPILE RUNWAY PREVENTS UNFINISHED WORK FROM BECOMING A BUILD.

NO EVIDENCE, NO STATUS CHANGE. NO GREEN RUNWAY, NO COMPILE.

File Edit Selection View Go Run Terminal Help

deepreinforce-ai/Ornith-1.0-9B-GGUF - New project - Velociti

BOTS **Hermes update, refresh.**

Search bots...

All 2 | Running 0 | Degraded 1 | Stopped 2

Sort: Status

LIVE BOTS 0

STOPPED 1

New Hermes Stopped

Stopped - Rancher Cleaned, Total time: 00:00:00

Stopped | Port: 8643 | Sentinel Protected | google/gemma-4-12B-it-q... | Config Active | Rancher Cleaned | GitOps | Forgejo | Gemma 4 | Update Ready | Healthy

Local: DESKTOP-LL7NBHL (win32/x64)

NEEDS ATTENTION 1

New Hermes 2 Stopped

Blocked - Hermes post-start exact-profile proof failed at model-relay-authenticated (HERMES_POST_START_MODEL_RELAY_AUTHENTICA
Total time: 00:00:00

Recovering | Port: 8643 | Sentinel Blocked | google/gemma-4-12B-it-q... | Config Active | Rancher Drift | GitOps | Forgejo | Gemma 4 | Update Ready | Drift

Local: DESKTOP-LL7NBHL (win32/x64)

Canvas ? High-priority direct lane assignment 3/10. Repair and validate exactly Forgejo ... not edit files or run

- Root cause/disposition: upstream `Dark 2026` product-default constants caused boot fallback. Repair is present at current HEAD, including precedence-preserving defaults and regression coverage. Accepted commit `9b4e0ef...` is not an ancestor, but equivalent content entered current history via `dd479d336fb`. Issue is closed with `status:done`, `priority:high`.
- Overlap: none. Both corrective theme paths are clean; unrelated existing worktree changes were untouched.
- Changed paths: none.
- Tests:
  - `node Velociti\proof\verify-ts7-130-ui-hybrid-theme.cjs` — passed all four theme contracts, default/precedence guards, and first-frame checks.
  - `npm.cmd run test-node -- --run src/vs/workbench/services/themes/test/common/workbenchThemeService.test.ts` — 6 passing.
- Installed observation: package `d037beba20...55b618...fb...22ce326...5ed...c27`, Velociti `1.130.0`. Read-only live CDP proof on existing port `9...` passed: effective class contains `arriving-together-dark-json`, 3,119 theme variables resolved, dark palette samples matched, zero disallowed blue variables, and no visible notification errors.
- Blocker unrelated to #234: repository startup verifier reports missing installed `resources/app/WORKFLOW.md`; left untouched as out of scope.
- HEAD/status: `3ec9fdbe52e300f7876ba4b4f03563f2836b7e7b` on `codex/ts7-130-full`. Worktree remains pre-existing dirty outside the two clean theme paths. No build, package, reload, settings/database mutation, A-Coder access, or new IDE launch occurred.

4 days **Tokens unverified**

test

Ask | GPT-5.6-Sol | High

Codex | New project | Allow all | 2 in

deepreinforce-ai/Ornith-1.0-9B-GGUF

User prompt

Codex

Finished with 2 steps

Loaded the two instruction files you referenced. This repo already references `deepreinforce-ai/Ornith-1.0-9B-GGUF` in multiple bot registry and planning docs, so it appears to be an existing model identity rather than a new addition.

Please specify the action you want next:

1. Wire/activate that model as default in the relevant model/runtime config.
2. Audit and fix any Ornith references in `bots` config.
3. Open a specific file for review or patching.

3 days **Tokens 13,472 in, 1,317 out**

test

Agent | GPT-5.3-Codex-Spark | Medium

Codex | New project | Allow all | 1 in

DEMO

codex/ts7-130-data-ownership* 0 0

Screen Reader Optimized

Video of demo is taken out due to large file size.

Contact us directly at marco@velociti.me for it

FROM POLICY TO PROOF—ACROSS HEADQUARTERS AND RETAIL

One enterprise story. One governed AI control loop.

AI VISIBILITY GATE
Use cloud intelligence without surrendering company context.

VELOCITI ENTERPRISE SERVER
The policy brain and evidence hub.
Governance, correlation and accountability.

01 SET THE BOUNDARY

Sentinel defines identity, grants, data classes, approvals and cloud-AI rules.

The agent proposes. Policy decides.

MANAGEMENT APPROVALS (HUMANS)

HIGH-IMPACT ACTIONS

Approve • deny • escalate • contain

BREACH ALERTS

Email • approved Chat app • supervisor escalation

Human supervisors remain the final authority.

02 AUTOMATE HQ IT

Hermes performs bounded work across Windows workstations and Microsoft Server.

Accounts • endpoints • email • security

Velociti IDE gives employees a local review and approval surface.

03 OPERATE THE RETAIL EDGE

Rancher-managed Hermes services process store events locally.

Assets • camera/audio events • customer relations

Metadata and approved evidence report upstream—not continuous raw streams.

05 CORRELATE + RESPOND

Velociti Enterprise links HQ and retail events, assigns follow-up work and routes high-impact decisions to a human.

— Policies + work orders out →

← Status + evidence back —

06 TRACK + RECOVER

FORGEJO

Issues • states • backups • evidence links

RANCHER

Containers • isolation • site communication services

Actor • authority • action • result • evidence • issue state

DISTRIBUTED EXECUTION. CENTRAL GOVERNANCE. EVERY ACTION REPORTS HOME.

Example deployment pattern; enterprise connectors are individually approved and validated.

VELOCITI
ENTERPRISE

CAN YOUR AI STACK PRODUCE THE RECEIPTS?

A model becomes production-ready only when identity, data, authority, execution and settlement can be independently evidenced.

THE MODEL PROPOSES. POLICY AND PEOPLE DECIDE.

MODEL RECEIPT
Which model is admitted?

- Identity • Version • Capability • Authenticated • Healthy • Prompt-ready

DATA RECEIPT
What context may leave?

- Data class • Approved route • Minimized • Redacted

EXTERNAL CLOUD / INTERNET

COMPANY-DATA BOUNDARY

APPROVED ROUTE MINIMIZED / REDACTED

PRODUCTION ADMISSION GATE
MODEL ACCESS ≠ PRODUCTION READY

AUTHORITY RECEIPT
Who approves consequences?

SENTINEL AUTOMATED POLICY: Approve · Deny · Escalate · Contain

MANAGEMENT APPROVAL: Approve · Deny · Escalate · Contain

BREACH / ESCALATION NOTIFICATIONS
Email / Chat app / Supervisor

EXECUTION RECEIPT
Where may a persistent worker act?

- Hermes • Scoped tools/MCP • Rancher boundary • Network/File limits

GOVERNED PRODUCTION

SETTLEMENT RECEIPT
Can we reconstruct the run?

- Transcript • Memory refs • Run metadata • Audit log • Error Control • Forgejo state • Backup status • Approval receipt

NO RECEIPT. NO ADVANCE.

Can the work be found, reviewed, recovered and retained as company operational knowledge?

COMPANY-OWNED OPERATIONAL RECORD

Search · Review · Recover · Report · Retain

- HEALTHY / APPROVED
- PENDING / NEEDS ACTION
- ERROR / BREACH

Velociti connects every receipt through one local control plane.

Learn more & be part of our Alpha (for free)

velociti.me

+852 9880 5017

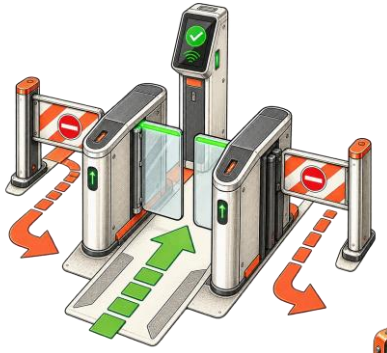
marco@velociti.me

Explore Velociti

Velociti Enterprise – Questions to ask as an business

- Governance and Control
  - How do you control who has access?
  - How can you review who did what on AI?

- Data you own and store
  - Who owns every prompt and response in your day to day in AI?
    - Where is it stored? How can you find something you did 3 weeks ago?
  - Would you risk exposing your IP now?



- Security
  - Is there any control plane to adjust access currently for your AI?
    - Are you running everything on full-access?
  - How do you prevent prompt injection or model poisoning?
  - Are sandboxes enough?

- Recurrent Cost vs. Asset
  - Shouldn’t data others are using to train their models, be an asset you own?

THE IDE BECOMES THE CONTROL PLANE

One governed workspace connects intent, live model readiness, human authority, persistent execution and operational evidence.

THE MODEL PROPOSES.
POLICY DECIDES.

YOUR MACHINES.
YOUR MODELS.
YOUR WORK.

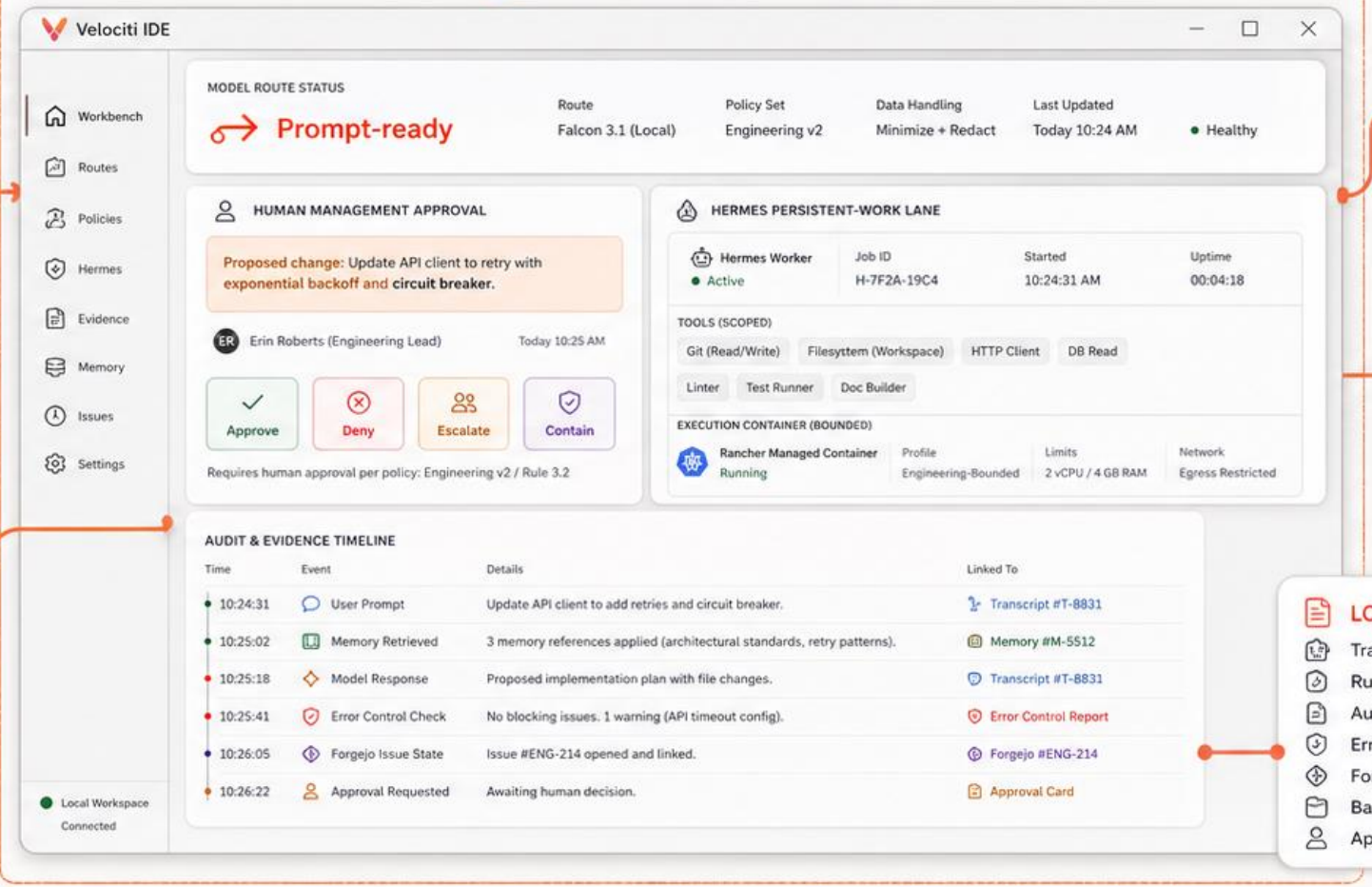
INTENT + CONTEXT

- Prompt / Objective
- Workspace & Project
- Model Route (Policy)

HERMES + BOUNDED EXECUTION

- Hermes Persistent Worker
- Scoped Tools / MCP
- Enterprise Server
- Workstation IDE
- Rancher-Managed Container

COMPANY DATA BOUNDARY



SENTINEL + MANAGEMENT

- Policy Shield (Automated)
- Management Approval (Human)
- Breach Escalation
  Email or Chat App to Supervisor

APPROVED • MINIMIZED • REDACTED •

Model Providers (External Cloud)

Only approved, minimized, redacted data leaves the company boundary.

Account, customer and IP context remains inside unless policy permits.

LOCAL SETTLEMENT

- Transcript & Memory References
- Run Metadata
- Audit Logs
- Error Control Reports
- Forgejo Issue States
- Backups & Snapshots
- Approval Receipts

CONTROLLED MODEL CHOICE • ACCOUNTABLE AUTOMATION • RECOVERABLE OPERATIONS • COMPANY-OWNED HISTORY

CDP proves the visible IDE. Canonical local storage owns transcripts, memory and settlement receipts.