Local Three-Tier Agent Workstation

A single-GPU operator setup that matches agent and model architecture to workload – Hermes / OpenCode / OpenClaw on 32 GB

● Design intent – target workstation setup, not a measured deployment

Design intent. This describes a target operator setup on a single 32 GB GPU, not a benchmarked deployment. Numbers are footprints and design budgets, not measured results.

Premise

One workstation, one RTX 5090 (32 GB), three distinct workloads: everyday reasoning and writing, software engineering, and unattended long-horizon jobs. Rather than force one model and one agent to cover all three, each tier pairs the agent framework and the open-weight model whose architecture actually fits the workload – under a strict rule that only one model is resident at a time.

The three tiers

Tier 1 – Cognitive cockpit (everyday & research)

Agent: Hermes Agent (Nous Research) – a self-hosted agent with persistent, SQLite-backed memory and full-text recall over past sessions, autonomous skill creation, and isolated subagents for parallel workstreams; an optional temporal-knowledge-graph memory is available via plugins. **Models:** DeepSeek-R1 32B for reasoning-heavy research – an RL-trained reasoning model that spends thinking tokens on an autoregressive chain-of-thought, self-correcting before it answers (no tree search at inference); Gemma 4-31B or Qwen3.6-35B-A3B for drafting reports and correspondence. **Safety:** inbound unstructured data (email, web) is treated as untrusted. Sensitive actions require explicit `[Y/n]` confirmation, and execution runs under container isolation with dropped Linux capabilities – the standard mitigation against indirect prompt injection.

Tier 2 – Agile development environment (SWE tasks)

Agent: OpenCode – a terminal-native agentic coding tool (in the class of terminal-native coding agents) that reads the working directory and operates over the repository, not a lightweight syntax checker. **Model:** `qwen3-coder:30b`, a Mixture-of-Experts model activating ~3.3B parameters per token. Sparse routing gives heavy-model depth at light-model latency; the long context (256K) holds a working set of the codebase so long reviews and automated test runs don't thrash the KV cache.

Tier 3 – Autonomous background factory (large projects)

Agent: OpenClaw – a persistent, long-horizon agent for multi-hour autonomous jobs, with checkpointing so a failed subtask resumes rather than restarts from zero. **Models:** Qwen3.6-35B-A3B as the allrounder, or Devstral 24B as an agentic-coding specialist. This tier runs unattended – data ingestion, background scripts, document assembly – where fault-tolerant orchestration matters more than latency.

Fitting 32 GB: one model at a time

Model	Footprint (Q4_K_M)	Note
Devstral 24B	~15 GB	agentic-coding specialist
qwen3-coder:30b	~18 GB	MoE, ~3.3B active
deepseek-r1:32b	~19 GB	reasoning

Model	Footprint (Q4_K_M)	Note
gemma4:31b	~19 GB	dense, writing
qwen3.6:35b	~24 GB	MoE, ~3B active

Any single model fits with headroom; no two fit together. That is why the setup time-multiplexes rather than co-loads. Ollama enforces it with `OLLAMA_NUM_PARALLEL=1` and `OLLAMA_MAX_LOADED_MODELS=1`: incoming work queues instead of loading a second model, and switching tiers swaps the resident model (e.g. `qwen3-coder` ~18 GB → `deepseek-r1` ~19 GB → `qwen3.6` ~24 GB). Single-residency is not a limitation to work around – it is the design.

Scenario (STAR)

- **Situation.** A single operator moves across research, coding, and long-running background jobs on one 32 GB workstation, with no cloud dependency permitted.
- **Task.** Assign each workload to the agent and model whose architecture fits it, without ever exceeding the VRAM ceiling.
- **Action.** Route everyday reasoning and writing to Hermes (DeepSeek-R1 / Gemma 4), SWE work to OpenCode (`qwen3-coder` MoE), and unattended multi-hour jobs to OpenClaw (Qwen3.6 / Devstral); Ollama swaps the single resident model on each tier switch.
- **Intended result (design goal).** Each task runs on the tool matched to it, at full local speed, with no VRAM collision and no data leaving the machine – one coherent single-GPU operator setup rather than three runtimes competing for the same card.

Status & honesty note

This is a design, not a benchmark. The VRAM figures are model footprints and budgets; the tier assignments are architectural judgments, not measured outcomes. What is established elsewhere (see the Hybrid Evaluation Pipeline and the codebase-evaluation pilot) is the single-stream VRAM discipline this setup depends on; what remains to be measured is the end-to-end ergonomics of swapping across three tiers in daily use.

Designed under CTC AI Operations.