Fabric — AI Operating System (cloud era · models are plugins, not dependencies · local + cloud)

Fabric Intelligence — the ONLY user-facing surface
Weave · Forge Studio · agents — users never pick models

↓

◈ Router — deterministic-first · fabric_router
classify intent → lane · min_cap floor · empty set → pool_exhausted

Seven capability lanes (logical roles — not one model each) · reason & code may burst to cloud

- **fast** — route · classify
- **reason** ☁ — + planning
- **code** ☁ — gen · bug-fix
- **review** — independent
- **vision** — art QA
- **plan** — architecture
- **embed** — retrieval

Model runtime — two coordinated seams

Local provider pool — fabric_providers · vLLM + Ollama
select_backend · lazy load + evict to fit RAM (128 GB) · keep-alive · least-in-flight

Cloud adapter — fabric_ai · Claude
failover · monthly budget · degrade to local

Model plugins — open-weight local (~1–2 resident) + frontier cloud, same plugin interface

- **Fast** — gemma3:12b
- **Reason** — qwen2.5:72b
- **Coder** — qwen3-coder-next
- **Vision** — qwen2.5vl:32b
- **Embed** 🔒 — nomic
- **Claude — reason · code · vision** — frontier · policy-gated · not downloaded

Cross-cutting rails — every request flows through these

- **Confidence Engine** — calibrated · learns
- **Knowledge + Memory** — dual-mode RAG · KB · history
- **Policy + 4-mode RBAC** — approval gates
- **Audit + Observability** — every call traced

Learning loops — cloud successes become local routing signals

- **Benchmark routing — fabric_bench** — measured latency + capability seeds cold routing
- **Learned routing — fabric_learn** — validated + approved outcomes bias future picks

Same interface, two deployments (one FORGE_PROFILE) →

- **Local — one workstation** — Mac · Ollama · JSONL RAG · private · $0
- **Hybrid — control plane + workers** — Postgres+pgvector · cloud routing · RBAC/SSO · vLLM

Principle: the AI layer is replaceable without changing Forge
Gateway → typed advisory result → Forge authority · routing and policy receipts → Canonical Event Spine

surface & governance rails · routing · local lanes · learning · local models & infra · cloud · ☁ = cloud-eligible lane