Moussaif Neo-AGI

Model Orchestration Reference

20-Model Free-Only Router | 8 HyperLoops | 5 Destruction Courts
12 HyperSkills | 7-Vessel Fugu Fusion | Daily 40-Loop Factory

AFRIPA Knowledge Vault 08_OBSIDIAN_VAULT

Active Providers & APIs

OpenRouter (Primary) [Configured]: URL: https://openrouter.ai/api/v1 | Models: moonshotai/kimi-k2.6 (default), free models for aux

NVIDIA (Fallback) [Configured]: URL: https://integrate.api.nvidia.com/v1 | Models: Nemotron-3-Ultra/Super/8B

ZenMux (Anthropic proxy) [Configured]: URL: https://zenmux.ai/api/anthropic | Models: Claude 3 Opus/Sonnet/Haiku/3.5-Sonnet

OpenAI-Codex (Image Gen) [Configured]: URL: Via Codex CLI auth | Models: gpt-image-2-high

Auxiliary Models (All via OpenRouter Free Tier)

- Vision: google/gemma-4-31b-it:free
- Compression: google/gemma-4-31b-it:free
- Web Extract: openai/gpt-oss-120b:free
- Title Generation: tencent/hy3:free
- Skills Hub: nvidia/nemotron-3-super-120b-a12b:free
- MCP: openai/gpt-oss-120b:free
- Approval: tencent/hy3:free

Smart Routing Config

- smart_model_routing.enabled: true
- smart_model_routing.cheap_model: tencent/hy3:free
- smart_model_routing.complex_model: nvidia/nemotron-3-ultra-550b-a55b:free
- smart_model_routing.threshold: 0.3
- provider_routing.sort: price
- provider_routing.data_collection: deny
- provider_routing.require_parameters: true

20-Model Router Roles (Free-Only via OpenRouter)

1 Supreme Architect (nvidia/nemotron-3-ultra): Hardest theory, strategy, final judgment | Budget: 8k

2 Deep Reasoner (nvidia/nemotron-3-super): Long causal chains, complex planning | Budget: 32k

3 Coding Agent (poolside/laguna-m-1): Repo edits, debugging, architecture | Budget: 8k

4 Fast Coder (poolside/laguna-xs): Quick patches, tool work | Budget: 4k

5 Research Synthesizer (tencent/hunyuan-3b): Web/files/papers, fact extraction | Budget: 16k

6 Critic Court (openai/gpt-oss-120b): Hallucination detection, weak logic | Budget: 4k

7 Compression Agent (openai/gpt-oss-20b): Context shrinking before expensive calls | Budget: 2k

8 Planner (google/gemma-4-26b): Task decomposition, dependency mapping | Budget: 4k

9 Executor (qwen/qwen3-coder): Deliverable production, code gen | Budget: 8k

10 Refiner (cohere/north-mini-code): Style, structure, formatting | Budget: 2k

11 Multilingual (google/gemma-4-26b): Arabic/French/English consistency | Budget: 2k

12 Visual Director (nvidia/nemotron-nano-omni): Image prompts, brand visuals | Budget: 1k

13 Video Director (nvidia/nemotron-nano-omni): Scripts, shots, motion prompts | Budget: 1k

14 SEO/Growth (tencent/hunyuan-3b): Keywords, funnels, growth loops | Budget: 2k

15 Outreach (openai/gpt-oss-20b): Cold email, DM, partnership pitches | Budget: 1k

16 Data Extractor (poolside/laguna-xs): Structured extraction from sources | Budget: 1k

17 Memory Curator (google/gemma-4-31b): Kernel extraction, reconstruction | Budget: 2k

18 Tool-Use Agent (poolside/laguna-m-1): API orchestration, browser, CLI | Budget: 4k

19 Low-Cost Draft (openai/gpt-oss-20b): First pass, exploration | Budget: 1k

20 Premium Final (nvidia/nemotron-3-ultra): Final synthesis, client-facing | Budget: 8k

Token Economy Rules

- 1. Classify every task: trivial | normal | complex | critical
- 2. Use cheapest sufficient model first
- 3. Compress context before expensive calls
- 4. Summarize long memory into task state
- 5. Draft with fast free models, critique with mid-tier, finalize with premium
- 6. For long tasks: create minimal task graph, execute in chunks, synthesize
- 7. Always remove redundancy, avoid verbose hidden planning
- 8. Output only useful artifacts

8 HyperLoops (Continuous Learning Engine)

1. **Kernel Extraction (MCD Extractor):** Input: Raw source (PDF, transcript, doc) -> Output: kernel-card.yaml with MCD statement | Validation: Regenerates 80%+ domain concepts
2. **Evidence Ranking (Researcher):** Input: Claims + sources -> Output: Epistemic tags on all claims | Validation: Zero ESTABLISHED without primary source
3. **Concept Disambiguation (Conceptualizer):** Input: Term cluster (e.g. content strategy/marketing/ops) -> Output: Concept map with IS_A, PART_OF, OVERLAPS... | Validation: No circular definitions, necessary/sufficient conditions
4. **Problem Decomposition (Planner):** Input: Complex operational goal -> Output: Task DAG with preconditions/inputs/outputs | Validation: Agent executes leaf tasks without human
5. **Destruction & Repair (5 Courts) (Critic (parallel)):** Input: Candidate artifact (HyperSkill, Loop, deliverable) -> Output: Pass/Fail per court + repair instructions | Validation: All 5 pass = version tag; any fail = repair re-test
6. **Tool Construction (Builder):** Input: Validated pattern from Loop 5 -> Output: Script/skill/workflow/API wrapper + tests | Validation: 10x speedup vs manual
7. **Agentic Execution (Executor + Critic loop):** Input: Task graph + tools from Loop 6 -> Output: Deliverable + execution trace + metrics | Validation: Output passes Destruction Court (Loop 5)
8. **Update & Version (Memory Curator):** Input: New source, failed court, tool improvement -> Output: Versioned artifacts + reconstruction memory | Validation: Regression test on prior passes

5 Destruction Courts (Pass All or Reject)

1. Hallucination Court: Any unsupported claims? Invented sources? | Fail: Yes = FAIL

2. Reductionism Court: Destroyed complexity? Confused abstraction levels? | Fail: Yes = FAIL

3. False Novelty Court: Actually new? Or renamed existing pattern? | Fail: Not new = FAIL

4. Weak Operationalization Court: Can an agent execute this without human intervention? | Fail: No = FAIL

5. Agentic Failure Court: Would an agent make better decisions than human author? | Fail: Yes = FAIL

Appeal: 3+ fails -> Supreme Architect (Nemotron-3-Ultra) reviews all courts + reconstruction memory -> Final verdict.

12 HyperSkills (Mandatory Per Capability)

- 1. Kernel Extraction
- 2. Concept Disambiguation
- 3. Evidence Ranking
- 4. Problem Decomposition
- 5. Mechanism Mapping
- 6. Hypothesis Generation
- 7. Destruction Testing
- 8. Tool Construction
- 9. Retrieval Optimization
- 10. Evaluation Design
- 11. Version Updating
- 12. Deployment Translation

Fugu Fusion Router (7-Vessel Parallel Fusion)

For high-stakes/complex tasks: fan out to 7 specialized vessels in parallel, then V1 Orchestrator synthesizes.

V1 Orchestrator (Nemotron-3-Ultra, temp=0.3): Synthesis, quality gate, version decision

V2 Researcher (Hunyuan-3B, temp=0.2): Evidence gathering, citations, fact-check

V3 Conceptualizer (Gemma-4-26B, temp=0.4): MCD extraction, concept maps, disambiguation

V4 Critic (GPT-OSS-120B, temp=0.7): Destruction Court simulation, failure modes, risks

V5 Builder (Laguna-M-1, temp=0.3): Code, tools, automation, API wrappers

V6 Skill Distiller (LFM-1.2B, temp=0.2): HyperSkill cards, training examples, distillation

V7 Retriever (Gemma-4-31B, temp=0.1): RAG chunking, GraphRAG edges, embedding optimization

Fusion Protocol: User Prompt -> V2..V7 parallel -> V1 synthesizes all 6 inputs -> Final Artifact + Destruction Court Log + Version Tag + Reconstruction Memory Update

Quick Load Commands

Full AGI Orchestrator:

```
hermes -s moussaif-agi-philosophy,moussaif-orchestrator,moussaif-codex-bridge,hermes-fattah-hyperloop-factory chat
```

Ebook Factory Pipeline:

```
hermes -s ebook-publishing-gumroad-kdp,moussaif-cover-design-system,moussaif-metadata-seo-engine,moussaif-pdf-render-qa chat
```

Faceless Content OS (Full Stack):

```
hermes -s faceless-content-os,faceless-script-hook-extractor,faceless-script-body-builder,faceless-youtube-shorts-optimizer,faceless-kling-video-director,faceless-thumbnail-designer-flux chat
```

Fusion Router (7-Vessel):

```
hermes -s fugu-fusion-router,moussaif-orchestrator chat
```

Daily HyperLoop Factory:

```
hermes -s hermes-fattah-hyperloop-factory,moussaif-agi-philosophy chat
```

Moussaif Vertical Specialist:

```
hermes -s moussaif-seo-content-refresh,moussaif-social-media-amplification,moussaif-email-sequence-builder chat
```

Key Files in AFRIPA Vault (08_OBSIDIAN_VAULT)

03_AUTONOMOUS_AGENCY/01_AGENT_ARCHITECTURE.md: Full 20-model spec, roles, token budgets

03_AUTONOMOUS_AGENCY/02_FUGU_FUSION_ROUTER.md: 7-vessel prompts, fusion protocol, quality metrics

03_AUTONOMOUS_AGENCY/03_HYPERLOOP_FACTORY.md: Daily 40-loop generation, 9 agents, registries

03_AUTONOMOUS_AGENCY/04_DESTRUCTION_COURTS.md: 5-court procedures, logs, appeal process

02_SKILLS/NEO_AGENCY_SKILLS_INDEX.md: All 300+ skills/KBs categorized with load commands

01_CORE_KB/07_KNOWLEDGE_ENGINEERING/01_KB_ENGINEERING_SOP.md: KB build pipeline, chunk schema, MCD/QMCD

04_PROJECTS/01_EBOOK_FACTORY_OS.md: Complete Gumroad/KDP pipeline (18 hyperskills + 50 GOAT)

06_TEMPLATES/PROJECT_TEMPLATES.md: 7 project templates (Ebook, Channel, Client, KB, Agent, Content, HyperLoop, Automation)

07_WORKFLOWS/CRON_JOBS.md: Daily/weekly/monthly/quarterly schedules + Hermes cronjob commands

11_QA_SYSTEMS/01_DESTRUCTION_COURTS.md: Project-level QA checklist, court log format, failure patterns

Epistemic Tags (Required on Every Claim)

ESTABLISHED: Verified data, API specs, primary sources

SUPPORTED: Multiple credible sources, needs context

PLAUSIBLE: Reasonable inference from evidence

THEORETICAL: Model-based, untested in this context

HYPOTHESIS: Testable theory, experiment designed

SPECULATIVE: Creative scenario, low evidence

UNKNOWN: Gap acknowledged honestly

NEEDS_VERIFICATION: Claim made, source pending