CANONICAL LIVING OVERVIEW GUIDE

Local Models, On-Device Inference, and Sovereign AI

A local-first AI guide covering model acquisition, licensing, hardware sizing, GGUF, runtimes, accelerators, offline operation, privacy, sovereign controls, model lifecycle, evaluation, updates, and hybrid routing.

PERMANENT ID
MYTHOS-FG-AI-0003

PUBLICATION CLASS
FIELD GUIDE

VERSION / STATUS
1.0.0-candidate / CANDIDATE

PERMANENT PUBLICATION IDENTITY

Local Models, On-Device Inference, and Sovereign AI

Document ID
MYTHOS-FG-AI-0003

Version
1.0.0-candidate

Status
Candidate Focused Field Guide

Review date
2026-10-26

Publication position
This is a living field guide. It is designed for frequent revision while retaining a stable permanent identity. Candidate status means the content is ready for structured use and review but is not represented as an external certification or authority publication.

LEVEL	CLASS	CADENCE	EVIDENCE
L1-L4	FIELD GUIDE	QUARTERLY	SOURCE-BOUND
Foundational through frontier	Practical and educational	Interim updates as needed	Limitations remain visible

Reader orientation

Dimension	Engineering position
Primary domain	MYTHOS-AI - Artificial Intelligence & Agentic Systems
Secondary domain	MYTHOS-EMT; MYTHOS-SEC; MYTHOS-CLD
Audience	AI platform engineers, desktop developers, security teams, sovereign-infrastructure architects, and local-model practitioners.
Prerequisites	AI model fundamentals, operating systems, hardware, software packaging, security, and inference concepts.
Publisher / owner	Mythos Systems / Aaron Stovall
Public classification	Public technical guidance

Expected outcomes

Select local models and runtimes according to capability, hardware, licensing, privacy, and operational needs.

Size CPU, GPU, NPU, memory, storage, context, and throughput realistically.

Operate models offline with verified artifacts, controlled updates, and reproducible configuration.

Design sovereign and hybrid AI architectures with explicit data, identity, egress, and fallback boundaries.

Evaluate local systems on representative tasks, resource limits, privacy, and recovery.

How to use this guide

Read the guide as a progressive system rather than a disconnected encyclopedia. Foundational material establishes the mental model; practitioner sections convert it into repeatable implementation and operations; advanced sections expose architecture, scale, security, and failure behavior; frontier sections describe technologies whose evidence or maturity is still evolving.

Suggested reading path
New practitioners should follow the chapters in order and complete the laboratories. Experienced engineers may begin with the architecture, failure, validation, or frontier sections, then use the related-document map to deepen a specific topic.

Learning and assurance model

Level	Reader outcome
L1 - Foundational	Terminology, first principles, component roles, packet or state flow, and simple working examples.
L2 - Practitioner	Implementation, configuration, automation, testing, troubleshooting, and operational evidence.
L3 - Advanced	Architecture, scale, concurrency, resilience, threat models, performance, and complex migrations.
L4 - Frontier	Emerging systems, research directions, technical uncertainty, readiness gates, and adoption signals.

Contents

CHAPTER 01

Local and sovereign AI architecture

Define why inference is local and which trust boundaries change.

Chapter operating position
Define why inference is local and which trust boundaries change.

1.1 Local, edge, private, and sovereign modes

Local, edge, private, and sovereign modes must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Differentiate on-device, workstation, site, private cloud, air-gapped, national, and hybrid execution. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for local, edge, private, and sovereign modes.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating local, edge, private, and sovereign modes as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for local, edge, private, and sovereign modes.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

1.2 Data and egress boundaries

Data and egress boundaries must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Control prompts, files, embeddings, telemetry, model calls, updates, licenses, and support data. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for data and egress boundaries.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating data and egress boundaries as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for data and egress boundaries.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

1.3 Identity and policy

Identity and policy must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Authorize users, applications, models, tools, data, and administrative actions locally. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for identity and policy.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating identity and policy as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for identity and policy.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

1.4 Availability and survivability

Availability and survivability must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Plan offline operation, dependency removal, degraded modes, backups, and recovery. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for availability and survivability.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating availability and survivability as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for availability and survivability.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 01 laboratory and review

Practical laboratory
Create a sovereign AI architecture for a sensitive workflow and prove which information can and cannot leave each boundary.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore verifiable sovereign AI appliances and federated local model fabrics.

Review gate
Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 02

Model selection, licensing, and artifact trust

Choose a model as a governed software and data artifact.

Chapter operating position
Choose a model as a governed software and data artifact.

2.1 Capability and task fit

Capability and task fit must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Compare reasoning, coding, extraction, retrieval, multilingual, multimodal, tool use, and structured output. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for capability and task fit.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating capability and task fit as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for capability and task fit.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

2.2 Licensing and acceptable use

Licensing and acceptable use must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Review model license, weights, derivatives, redistribution, commercial use, dataset obligations, and policy. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for licensing and acceptable use.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating licensing and acceptable use as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for licensing and acceptable use.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

2.3 Artifact formats and metadata

Artifact formats and metadata must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Understand checkpoints, safetensors, GGUF, ONNX, tokenizers, chat templates, adapters, and configuration. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for artifact formats and metadata.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating artifact formats and metadata as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for artifact formats and metadata.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

2.4 Integrity and provenance

Integrity and provenance must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Verify source, hashes, signatures, conversion, quantization, model card, tests, and trusted distribution. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for integrity and provenance.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating integrity and provenance as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for integrity and provenance.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 02 laboratory and review

Practical laboratory
Qualify two local models through license, artifact, tokenizer, conversion, capability, safety, and provenance review.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore signed model bills of materials and transparency logs for weight artifacts.

Review gate
Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 03

Hardware sizing and execution backends

Map model architecture and precision to real device resources.

Chapter operating position
Map model architecture and precision to real device resources.

3.1 Memory budgeting

Memory budgeting must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Estimate weights, KV cache, activations, runtime buffers, context, batch, multimodal encoders, and fragmentation. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for memory budgeting.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating memory budgeting as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for memory budgeting.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

3.2 CPU, GPU, NPU, and accelerator paths

CPU, GPU, NPU, and accelerator paths must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Compare cores, vector units, memory bandwidth, VRAM, unified memory, device transfer, and supported operators. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for cpu, gpu, npu, and accelerator paths.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

Treating cpu, gpu, npu, and accelerator paths as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for cpu, gpu, npu, and accelerator paths.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

3.3 Backend selection

Backend selection must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Evaluate llama.cpp, ONNX Runtime, MLX, vendor runtimes, containers, and native application integration. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for backend selection.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating backend selection as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for backend selection.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

3.4 Thermal, power, and sustained performance

Thermal, power, and sustained performance must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Measure clocks, throttling, fan, battery, heat, energy, and long-running stability. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for thermal, power, and sustained performance.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating thermal, power, and sustained performance as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for thermal, power, and sustained performance.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 03 laboratory and review

Practical laboratory
Create a hardware-fit estimator and validate predicted memory and throughput against two precisions and context lengths.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore disaggregated memory, chiplets, photonic interconnects, and adaptive CPU/GPU/NPU execution.

Review gate
Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 04

Quantization and local optimization

Reduce footprint while measuring quality and platform effects.

Chapter operating position
Reduce footprint while measuring quality and platform effects.

4.1 Weight precision and formats

Weight precision and formats must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Compare FP16, BF16, FP8, INT8, INT4, mixed precision, group size, scales, and zero points. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for weight precision and formats.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating weight precision and formats as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for weight precision and formats.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

4.2 Post-training and calibration

Post-training and calibration must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Use representative calibration, clipping, outlier treatment, error metrics, and task evaluation. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for post-training and calibration.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating post-training and calibration as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for post-training and calibration.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

4.3 KV-cache and activation optimization

KV-cache and activation optimization must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Adjust cache precision, context, batching, offload, flash attention, and memory layout. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for kv-cache and activation optimization.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating kv-cache and activation optimization as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for kv-cache and activation optimization.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

4.4 Quality and performance verification

Quality and performance verification must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Measure task accuracy, structured output, perplexity limits, latency, throughput, memory, and power. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for quality and performance verification.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating quality and performance verification as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for quality and performance verification.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 04 laboratory and review

Practical laboratory
Quantize a small model or simulate quantization and compare memory, latency, and task quality across precisions.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore adaptive per-layer quantization and runtime quality-aware precision.

Review gate
Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 05

Runtime, session, and application integration

Embed local inference into reliable desktop, mobile, and edge applications.

Chapter operating position
Embed local inference into reliable desktop, mobile, and edge applications.

5.1 Model loading and lifecycle

Model loading and lifecycle must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Manage discovery, validation, loading, warmup, unload, updates, rollback, and concurrent access. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for model loading and lifecycle.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating model loading and lifecycle as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for model loading and lifecycle.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

5.2 Prompt and chat templates

Prompt and chat templates must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Control system messages, roles, tokens, special tokens, tool schemas, stops, and model-specific formats. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for prompt and chat templates.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating prompt and chat templates as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for prompt and chat templates.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

5.3 Streaming and cancellation

Streaming and cancellation must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Deliver tokens, progress, backpressure, user cancellation, timeouts, and cleanup. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for streaming and cancellation.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating streaming and cancellation as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for streaming and cancellation.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

5.4 Isolation and resource governance

Isolation and resource governance must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Limit files, network, memory, CPU, GPU, tools, processes, and tenant sharing. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for isolation and resource governance.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating isolation and resource governance as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for isolation and resource governance.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 05 laboratory and review

Practical laboratory

Build a local inference session manager with model validation, streaming, cancellation, resource limits, and safe unload.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore OS-native AI runtimes and capability-isolated local agent containers.

Review gate

Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 06

Offline knowledge, tools, and updates

Operate useful AI systems without assuming permanent cloud services.

Chapter operating position
Operate useful AI systems without assuming permanent cloud services.

6.1 Local retrieval and memory

Local retrieval and memory must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Store documents, indexes, embeddings, metadata, citations, and user-controlled memory locally. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for local retrieval and memory.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating local retrieval and memory as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for local retrieval and memory.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

6.2 Local tools and automation

Local tools and automation must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Expose filesystem, search, code, databases, devices, and workflows through least-privilege interfaces. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for local tools and automation.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating local tools and automation as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for local tools and automation.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

6.3 Update distribution

Update distribution must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Use signed model, runtime, index, policy, and tool updates with staged rollout and rollback. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for update distribution.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating update distribution as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for update distribution.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

6.4 Air-gap transfer and scanning

Air-gap transfer and scanning must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Control removable media, manifests, hashes, malware scanning, quarantine, approval, and import evidence. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for air-gap transfer and scanning.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating air-gap transfer and scanning as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for air-gap transfer and scanning.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 06 laboratory and review

Practical laboratory

Design an offline update bundle containing model, runtime, index, policy, and tool changes with complete verification and rollback.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore delay-tolerant synchronization and content-addressed sovereign AI distribution.

Review gate

Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 07

Privacy, security, and operational assurance

Prove that local execution delivers the intended trust benefit.

Chapter operating position
Prove that local execution delivers the intended trust benefit.

7.1 Threat modeling

Threat modeling must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Assess local users, malware, administrators, model files, prompt injection, tools, memory, telemetry, and physical access. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for threat modeling.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating threat modeling as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for threat modeling.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

7.2 Data protection

Data protection must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Use OS isolation, encryption, access control, secure deletion, logging, retention, and user-visible controls. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for data protection.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating data protection as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for data protection.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

7.3 Model and tool security

Model and tool security must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Test malicious models, parsers, templates, plugins, prompts, indirect injection, and unsafe actions. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for model and tool security.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating model and tool security as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for model and tool security.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

7.4 Incident and recovery

Incident and recovery must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Revoke models or tools, preserve evidence, restore trusted artifacts, rotate secrets, and validate state. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for incident and recovery.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating incident and recovery as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for incident and recovery.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 07 laboratory and review

Practical laboratory

Red-team a local AI application for artifact tampering, prompt injection, tool misuse, memory leakage, and telemetry egress.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore attested local inference and encrypted model execution.

Review gate

Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

CHAPTER 08

Evaluation, routing, and hybrid architecture

Decide when local, private, or cloud models should handle work.

Chapter operating position
Decide when local, private, or cloud models should handle work.

8.1 Representative task evaluation

Representative task evaluation must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Measure capability, safety, privacy, latency, throughput, context, reliability, and structured output. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for representative task evaluation.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating representative task evaluation as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for representative task evaluation.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

8.2 Hardware-aware model routing

Hardware-aware model routing must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Route by task, model fit, memory, current load, deadline, energy, and privacy. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for hardware-aware model routing.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Implementation sketch

```
local_route:
  privacy_class: restricted
  preferred_runtime: llama.cpp
  artifact:
    format: gguf
    sha256: required
    license_review: required
  limits:
    ram_gib: 32
    context_tokens: 32768
    max_parallel: 2
  fallback:
    remote_allowed: false
  telemetry:
    content_logging: disabled
```

Failure patterns

- Treating hardware-aware model routing as a model capability claim disconnected from complete system behavior.
- Using one benchmark, model judge, or successful demonstration as proof of production readiness.
- Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.
- Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.
- Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.
- Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for hardware-aware model routing.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

8.3 Hybrid escalation

Hybrid escalation must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Define when data may be minimized, summarized, redacted, or approved for stronger remote models. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

- Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for hybrid escalation.
- Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.
- Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.
- Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.
- Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.
- Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

- Treating hybrid escalation as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for hybrid escalation.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

8.4 Lifecycle and retirement

Lifecycle and retirement must be treated as an observable AI-system mechanism rather than a model label, benchmark claim, prompt trick, or framework feature. Monitor model quality, vulnerabilities, licensing, runtime compatibility, replacements, and archived evidence. The implementation should name authoritative inputs, data and model versions, identities, state transitions, permissions, resource limits, telemetry, failure behavior, uncertainty, rollback, and acceptance criteria.

This guide traces the subject through the artifact, license, hardware, runtime, context, inference, tool or knowledge boundary, telemetry, update, and recovery chain. A fluent response, high aggregate score, successful tool call, or green service dashboard is not sufficient proof. Intended behavior must be reconciled with hidden tests, source evidence, negative and adversarial cases, latency and resource measurements, user-visible results, and residual limitations.

Implementation begins with a small, inspectable baseline and explicit invariants. Introduce one model, dataset, retrieval source, tool, memory tier, optimization, or permission at a time; preserve before-state evidence; test expected and prohibited behavior; and retain a reversible path. Generated plans and model judgments remain subject to deterministic validation and human authority.

Troubleshooting locates the first divergence from the expected contract or state machine. Account for tokenizer and template mismatch, stale data, context loss, model nondeterminism, caching, quantization, scheduling, permissions, prompt injection, missing telemetry, partial tool effects, and distribution shift. Completion requires reproducibility, causal evidence, validated recovery, and clear disclosure of uncertainty.

Engineering practices

Define the objective, authority, data, model, owner, constraints, risks, and acceptance criteria for lifecycle and retirement.

Separate model behavior from retrieval, tools, memory, policy, interface, and human oversight.

Version prompts, schemas, data, model artifacts, runtime configuration, evaluations, and release evidence.

Test nominal, negative, adversarial, long-context, degraded, overloaded, recovery, and rollback conditions.

Retain source, trajectory, tool, citation, metric, profile, decision, and user-outcome evidence.

Prefer least privilege, typed contracts, bounded budgets, staged rollout, abstention, and reversibility.

Failure patterns

Treating lifecycle and retirement as a model capability claim disconnected from complete system behavior.

Using one benchmark, model judge, or successful demonstration as proof of production readiness.

Ignoring tokenizer, prompt template, source authority, permissions, cache, hardware, or version boundaries.

Allowing tools, plugins, retrieval content, or memory to redefine trusted instructions.

Optimizing latency, tokens, or cost without measuring quality, safety, and tail failures.

Declaring completion without hidden tests, evidence, residual limitations, rollback, and monitoring.

Validation and evidence

Method	Expected evidence
Examine	Requirements, authority, model, data, prompts, tools, memory, schemas, versions, and assumptions for lifecycle and retirement.
Observe	Requests, selected context, trajectories, tool calls, model outputs, citations, caches, telemetry, resources, and user outcomes.
Test	Expected, prohibited, adversarial, malformed, long-context, overloaded, failed, recovery, and rollback cases.
Measure	Task quality, calibration, grounding, safety, latency, throughput, memory, tokens, cost, energy, and operator effort.
Reconcile	Intended system behavior against evaluated model, data, runtime, tools, sources, deployment, and residual risk.

Chapter 08 laboratory and review

Practical laboratory

Create and test a local-first router with privacy classes, capability thresholds, resource limits, fallback, and user approval.

Laboratory evidence package

- Architecture or topology showing boundaries, identities, dependencies, and authoritative inputs.
- Versioned configuration or code with explanatory comments and reproducible setup instructions.
- Nominal-path validation plus at least two injected failure or boundary conditions.
- Structured logs, traces, metrics, packet captures, test output, or other appropriate evidence.
- A concise decision record covering findings, limitations, residual risks, and next actions.

Frontier extension

Explore distributed personal model ensembles and privacy-preserving collaborative inference.

Review gate

Advance only when another engineer can reproduce the result, explain the architecture, identify the failure boundary, and distinguish measured evidence from assumptions or projections.

Integrated learning roadmap

The guide is intended to be revisited. First establish fluency, then build a small implementation, operate it under failure, automate repeatable work, and finally evaluate emerging techniques against a stable baseline. Progress is demonstrated through evidence rather than reading completion alone.

Stage	Demonstrated capability
Foundation	Explain the system from first principles and trace its critical path without relying on product terminology.
Build	Implement the smallest representative system with explicit contracts, identities, telemetry, and tests.
Operate	Diagnose injected failures, recover safely, and preserve evidence of the causal chain.
Automate	Convert a proven manual workflow into bounded, idempotent, observable automation.
Architect	Design for scale, resilience, security, interoperability, and lifecycle change.
Explore	Evaluate frontier techniques using stated hypotheses, limitations, and adoption gates.

Source authority register

Source policy
Official standards, specifications, and project documentation remain with their issuing organizations. Mythos Systems provides original engineering interpretation, synthesis, implementation guidance, and relationship mapping. Source verification dates do not guarantee that an authority has not changed since review.

01. Hugging Face - GGUF Documentation
Role: GGUF format and model distribution context.
Verified: 2026-07-26
Official source: https://huggingface.co/docs/hub/gguf

02. ggml-org - llama.cpp
Role: Local inference, GGUF, quantization, backends, server, and tooling.
Verified: 2026-07-26
Official source: https://github.com/ggml-org/llama.cpp

03. Microsoft - ONNX Runtime Generate API
Role: On-device and cross-platform generative-AI runtime.
Verified: 2026-07-26
Official source: https://onnxruntime.ai/docs/genai/

04. Apple - MLX
Role: Apple-silicon array and machine-learning framework.
Verified: 2026-07-26
Official source: https://github.com/ml-explore/mlx

05. vLLM Project - Quantization
Role: Supported quantization families and deployment considerations.
Verified: 2026-07-26
Official source: https://docs.vllm.ai/en/latest/features/quantization/

06. NIST - NIST AI 600-1 - Generative Artificial Intelligence Profile
Role: Generative-AI risk profile and recommended actions.
Verified: 2026-07-26
Official source: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

07. OWASP GenAI Security Project - OWASP Top 10 for LLM and GenAI Applications 2025
Role: Risk and mitigation baseline for generative-AI applications.
Verified: 2026-07-26
Official source: https://genai.owasp.org/llm-top-10/

Authority application and refresh triggers

Authority class	Required library action
Protocols and specifications	Recheck interoperability, security considerations, defaults, and migration guidance when a referenced specification changes.
Security and assurance frameworks	Update control mappings, evidence expectations, risk language, and assessment procedures after material framework revision.
Languages, runtimes, and projects	Re-run examples, tests, deprecation checks, and operational recommendations against supported releases.
Benchmarks and evaluations	Re-execute claims when workloads, methods, models, compilers, drivers, hardware, or scoring rules change.
Operational evidence	Incorporate validated incidents, failure tests, reader corrections, and measured implementation outcomes into the next revision.

Limitations, freshness, and revision control

Boundary	Statement
Publication limitation	This guide synthesizes broad engineering practice. It does not replace product documentation, legal advice, safety certification, or environment-specific design review.
Evidence limitation	Not every pattern is benchmarked in every environment. Examples illustrate engineering methods and must be validated against actual versions, workloads, hardware, policy, and failure conditions.
Freshness limitation	Vendor behavior, software libraries, AI models, security threats, and emerging standards can change quickly. Volatile claims require source revalidation before operational adoption.
Adoption limitation	Frontier content is not an implicit adoption recommendation. Adoption requires defined value, qualification gates, controlled proof, rollback, ownership, and residual-risk acceptance.
Status boundary	Candidate Focused Field Guide describes the publication, not certification of a product, implementation, engineer, or organization.

Revision record

Version	Revision statement
1.0.0-candidate	Initial candidate living-guide edition. Published 2026-07-26; scheduled review 2026-10-26.

Search and relationship metadata

Dimension	Engineering position
Primary domain	MYTHOS-AI - Artificial Intelligence & Agentic Systems
Secondary domain	MYTHOS-EMT; MYTHOS-SEC; MYTHOS-CLD
Publication class	Field Guide / Canonical Living Overview Guide
Skill levels	L1 Foundational; L2 Practitioner; L3 Advanced; L4 Frontier
Tags	local models, on-device AI, sovereign AI, GGUF, llama.cpp, ONNX Runtime, MLX, offline AI, quantization, hybrid routing
Canonical route	/library/field-guides/mythos-fg-ai-0003/

Living publication
Corrections, expanded laboratories, improved diagrams, authority changes, and new engineering evidence should revise this permanent publication rather than create disconnected replacements.