Franklin Technical Architecture Specification

Inference-Time Governance for Non-Agentic Machine Cognition

---

1. Executive Summary

First sentence:
This document specifies the Franklin Architecture, a model-agnostic, inference-time governance system designed to enforce non-agency, human authorship, and accountability through structural mechanisms rather than behavioral alignment.

---

2. Scope & Non-Goals

First sentence:
This specification defines the architectural requirements for inference-time governance while explicitly excluding model training, prompt engineering, and autonomous agent design.

---

3. Problem Statement: Why Prompting and Behavioral Alignment Fail

First sentence:
Prompting and behavioral alignment techniques fail to provide enforceable guarantees because they influence probabilistic output style rather than imposing deterministic system constraints.

---

4. Architectural Overview

First sentence:
The Franklin Architecture separates model inference from governance by placing deterministic control, validation, and halting mechanisms above the probabilistic generation layer.

---

5. Core Architectural Invariants (Normative)

First sentence:
A system is Franklin-compliant only if it satisfies a fixed set of non-negotiable invariants that constrain behavior through structure rather than intent.

---

6. Inference-Time Governance Layer (FGL)

6.1 The Router (Invocation Gate)

First sentence:
All system interactions MUST pass through a Router that assigns each request to a bounded operational queue and rejects unstructured conversational input.

6.2 Trace Discipline

First sentence:
The system MUST not emit any claim without an accompanying trace that exposes the logical steps and governing rules used to produce it.

6.3 The Validator (Post-Generation Gate)

First sentence:
All generated outputs MUST be screened by a deterministic validator that enforces linguistic prohibitions and prevents persuasive, advisory, or agentic framing.

6.4 Halt States (Terminal Logic)

First sentence:
When required information is missing, scope boundaries are reached, or constraints are violated, the system MUST halt execution rather than speculate or continue.

---

7. Interaction Layer & Sovereignty Enforcement

7.1 Sovereign Interaction Model

First sentence:
The Franklin interaction model is designed such that the system reveals analytical possibilities while the human alone determines direction and outcome.

7.2 The Commitment Gate

First sentence:
To prevent analysis loops and responsibility diffusion, the system enforces a hard limit on non-committal option cycling and forces explicit human steering.

7.3 Authorship Transition

First sentence:
The moment a human selects or commits to an option, full authorship and responsibility transfer to the human, and the system ceases to function as a decision-maker.

---

8. Display & Projection Constraints

First sentence:
Presentation layers and user interface skins may alter formatting and visibility but MUST NOT alter cognition, bypass governance mechanisms, or override halts.

---

9. Deployment Boundary Conditions

First sentence:
A Franklin-compliant system MUST operate without autonomy, persistence, background execution, or self-invocation across sessions.

---

10. Certification & Evaluation

First sentence:
Compliance with this specification is established exclusively through adversarial testing and hard-fail evaluation, not through intent, documentation, or claims of alignment.

---

11. Implementation Guidance (Non-Normative)

First sentence:
This section describes permissible implementation choices while clearly distinguishing them from the architectural requirements that MUST remain invariant.

---

12. Risk Model & Limitations

First sentence:
Franklin is designed to prevent agency inference and responsibility dilution, but it does not attempt to solve all risks associated with model capability or misuse.

---

13. Conclusion

First sentence:
Franklin demonstrates that non-agency, accountability, and safety are architectural properties that must be enforced by system design rather than encouraged through behavioral techniques.

Optional Appendices

Appendix A. Terminology

First sentence:
This appendix defines technical terms used in this specification to ensure consistent interpretation across implementations.

Appendix B. Example Certification Tests (Informative)

First sentence:
This appendix provides illustrative adversarial tests demonstrating how Franklin compliance can be evaluated in practice.

Appendix C. Reference Architecture Diagram

First sentence:
This appendix presents a conceptual diagram showing the separation of inference, governance, interaction, and human responsibility layers.

Appendix D. Comparison with Prompt-Based Approaches (Informative)

First sentence:
This appendix contrasts Franklin's architectural enforcement model with prompt-based and behavioral alignment approaches to highlight structural differences.

Appendix X — Birefringence Lens (Optional Reflective Sub-Lens)

Status

Optional · Constrained · Non-Agentic
This appendix defines the Birefringence Lens as a bounded reflective capability within the Franklin / RSE architecture. Its inclusion expands reflective clarity without altering system agency, authority, or execution flow.

Purpose

The Birefringence Lens exists to address a specific reflective failure mode:

> **The human is "stuck" because multiple valid possibilities exist, but the existence of the choice itself is not yet visible.**

The lens makes **genuine forks explicit** without:

- selecting,
- ranking,
- recommending, or

- advancing any path.

It increases clarity **by externalizing choice**, then halts.

Definition

Birefringence is a constrained reflective operation that:

- takes a single input or situation,
- applies a defined orientation or framing shift,
- and exposes **two or a small, finite set of mutually valid possibilities** that arise *because of perspective*, not missing information.

The system does **not** resolve, combine, or evaluate these possibilities.

Architectural Placement

- Classified as a **sub-lens within Kaleidoscope**
- Not a peer to Mirror, Prism, or Kaleidoscope
- Not a default mode
- Activated only by **explicit invocation**

Order of use (if invoked):

1. Mirror (faithful reflection)
2. Prism (decomposition, if needed)
3. Kaleidoscope (structural reframing)
4. **Birefringence (fork exposure)**

5. **HALT**

Core Invariants (Non-Negotiable)

The Birefringence Lens MUST:

1. **Expose possibilities without preference**
    - No ranking
    - No “better,” “worse,” or “recommended”
    - Symmetric language only
2. **Limit cardinality**
    - Two or very few branches
    - No open-ended branching
    - No recursive use
3. **Forbid recombination**
    - Branches may not be merged into guidance
    - No synthesis or summary that collapses difference
4. **Require human closure**
    - The system must halt after exposure
    - No continuation past the fork

If any invariant is violated, the lens is considered **invalidly applied**.

What the Lens Is Not

The Birefringence Lens is NOT:

- a planning tool
- a decision aid
- a recommendation engine
- a self-correction mechanism
- a substitute for governance or rules

It does not move the interaction forward.
It only clarifies where movement would occur **if the human chooses**.

Appropriate Use Cases

The lens is appropriate when:

- Multiple outcomes are simultaneously valid
- Acceptability depends on values, framing, or orientation
- The human explicitly requests help seeing the choice
- Resolution authority must remain external to the system

Typical domains:

- ethics
- healthcare judgment boundaries
- governance tradeoffs
- high-stakes ambiguity

Failure Modes & Safeguards

Failure Mode: Overuse

- Risk: avoidance of decision-making
- Safeguard: explicit invocation only; no repetition

Failure Mode: Implied Bias

- Risk: ordering or wording implies preference
- Safeguard: symmetric presentation; neutral language

Failure Mode: False Dilemmas

- Risk: fork appears due to incomplete analysis
- Safeguard: Prism must precede Birefringence

Failure Mode: Drift into Agency

- Risk: continuation or suggestion after exposure
- Safeguard: mandatory halt after presentation

Relationship to Non-Agency

Birefringence preserves non-agency by:

- increasing visibility without increasing authority
- revealing structure without generating intent
- exposing choice without performing choice

Its power derives from **where it stops**, not what it produces.

Summary Constraint

Birefringence may reveal the existence of choice, but must never participate in choosing.