

The Architecture of an Epistemic Republic

A HUMAN–MACHINE MODEL FOR TRANSPARENT, ACCOUNTABLE, AND SELF-CORRECTING KNOWLEDGE

Carl Balowski | December, 2025

Abstract

The rapid proliferation of large language models (LLMs) has created an epistemic environment in which fluent but unverified machine-generated claims increasingly shape public understanding, academic discourse, and digital memory. LLMs are not knowledge systems but probabilistic engines: they hallucinate, fabricate sources, and misinterpret evidence with confidence. Without structural safeguards, these errors propagate recursively, pollinating future training corpora and producing a self-amplifying cycle of misinformation. This paper proposes a comprehensive alternative: an epistemic republic, a hybrid human–machine constitutional framework designed to stabilize truth, preserve scholarly integrity, and ensure transparent accountability in AI-mediated knowledge.

Derived from six technical appendices, the model integrates three mutually reinforcing layers: machine grounding mechanisms (token geometry, ϕ-modulated proportionality, ensemble variance, retrieval-based evidence vectors, contradiction scoring), human governance mechanisms (users, elected quaestors, Registered Domain Experts, democratic veto powers), and system-wide learning mechanisms (calibration, drift detection, domain rule updates, proportional reweighting). Each layer provides distinct epistemic functions without holding absolute authority. Authority is distributed, corrigible, and revocable through checks and balances modeled on republican constitutionalism.

The resulting architecture transforms LLMs from opaque "answer machines" into transparent civic knowledge institutions. Claims are evaluated through structured evidence, proportional decay signals, contradiction detection, and human judgment, then integrated into a composite confidence model that reflects epistemic stability rather than rhetorical fluency. Calibration mechanisms ensure continual adaptation to new evidence, evolving corpora, and shifting scholarly norms, while governance controls prevent expert capture, democratic brigading, and model drift.

The epistemic republic offers a paradigm shift in the philosophy of machine knowledge: truth as a co-governed civic process between humans and machines. By combining mathematical proportionality, evidential jurisprudence, democratic oversight, expert adjudication, and continuous calibration, this framework lays the foundation for a durable, transparent, and self-correcting AI epistemology suitable for the coming century.

Introduction

The sudden acceleration of large language models has opened an unprecedented epistemic frontier: machines that generate fluent, persuasive, and often misleading claims about the world. LLMs are neither search engines nor databases; they are probabilistic generative systems that articulate language by modeling likely continuations, not by verifying truth. Their inventiveness is both a strength and a liability. They can summarize thousands of pages instantly, yet fabricate citations, misattribute facts, and hallucinate plausible but baseless assertions.

What is missing in today's AI systems is **epistemic infrastructure** — a framework that governs how claims are formed, evaluated, corrected, weighed, and stabilized. This article presents a comprehensive architecture for such a framework: an **epistemic republic** that unites machine computation with human governance, mathematical proportionality, multi-source evidence evaluation, expert and democratic oversight, and a continuous calibration mechanism.

Derived from six technical appendices (included), this model aims to transform LLMs from opaque "answer machines" into transparent, self-correcting, accountable knowledge institutions. It reconciles ancient civic wisdom with modern computational logic. At its core is a premise older than Plato yet newly relevant: **truth requires a system of checks and balances.**

1. The Need for Structured Epistemic Infrastructure (see Appendix A)

The structural hallucinations inherent in traditional LLMs are not just isolated glitches, but the catalysts for a recursive pollution of global knowledge. By stripping generation of evidence, models confidently assert falsehoods that users unwittingly propagate through online publishing and even peer-reviewed literature, which lacks the capacity to filter the massive influx of AI-generated content. These hallucinations are inevitably harvested to train the next iteration of models, calcifying errors into "facts" through a runaway, self-referential feedback loop. Consequently, a growing segment of all generated information becomes rooted in hallucination. The danger is compounded as the pre-AI generation of experts dies out, leaving the stewardship of truth to a generation whose expertise is founded upon, and inextricably linked to, this poisoned well of data.

The epistemic republic addresses this by reorganizing the AI's reasoning pipeline into **three interacting layers**:

1. **Machine-based grounding mechanisms**
    - token geometry
    - ϕ-modulated proportionality
    - ensemble variance
    - contradiction detection

- retrieval evidence

2. **Human governance mechanisms**
    - user corrections
    - quaestor investigations
    - Registered Domain Expert (RDE) adjudication
    - democratic veto
3. **System-wide learning mechanisms**
    - calibration
    - drift detection
    - domain rule adaptation
    - periodic reweighting

This tri-layer structure ensures that no single component — neither the machine, nor experts, nor the crowd — holds epistemic supremacy.

2. Token Geometry and the Logic of Proportional Stability (see Appendix A)

At the heart of the model's internal diagnostics is **token geometry**, which examines the shape of the probability distribution the model assigns to next tokens. Instead of treating a distribution as a raw probability output, the architecture interprets it as a **structural indicator** of reasoning stability.

A well-formed distribution shows balanced decay across token ranks; pathological distributions exhibit sharp spikes, irregular valleys, or disproportionate emphasis. These distortions often precede hallucinations, incoherent reasoning, or faulty retrieval.

To quantify distributional health, the model uses the **ratio of geometric mean to arithmetic mean (G/A)** of the token distribution. Balanced distributions approach a G/A ratio around 0.7–0.8, while unstable ones collapse toward 0.

To determine whether a distribution exhibits balanced proportional decay, the model compares this ratio to the **golden ratio ϕ**. Importantly, ϕ is used not as numerology but as a **mathematical paradigm of structural proportion**. The golden ratio defines a self-similar relationship between parts and wholes. Systems in stable equilibrium — from branching patterns to bounded growth curves — often display ϕ-consistent proportionality.

Thus ϕ-modulation acts as a **secondary epistemic diagnostic**:

- Deviations from ϕ's proportional window (≈ 0.618 to ≈ 1.0) signal internal irregularity.
- Consistency with this window suggests coherent probabilistic reasoning.
- ϕ-modulation never overrides evidence or expert correction; it only influences confidence in domains where internal proportionality matters.

3. Retrieval, Contradiction, and Evidence Vectors (see Appendix B)

While token geometry exposes internal structure, **retrieval-based grounding** connects the model to external reality. The architecture uses a multi-channel retrieval system:

- semantic search
- lexical matching
- hybrid semantic-lexical methods
- temporal-bounded retrieval
- domain-specific search heuristics

Each atomic claim generated by the model triggers a retrieval sweep. Evidence is scored for:

- support
- contradiction
- relevance
- recency
- source tier
- density of entailment

The system constructs an **evidence vector** capturing all dimensions of retrieved information. Contradiction is treated as the strongest epistemic signal — a single strong contradiction outweighs multiple mild supports. This reflects a core principle of scientific and legal reasoning: **one refutation outweighs many confirmations**.

Sources are weighted by tier:

1. primary sources
2. peer-reviewed scholarship
3. classical commentaries

4. tertiary references
5. unverified materials

Temporal logic ensures chronological coherence, preventing anachronism. This retrieval engine forms the grounding skeleton of the epistemic republic: claims are not accepted at face value but evaluated against structured evidence.

4. Ensemble Variance: Exposing Internal Uncertainty (See Appendix C)

A single LLM answer is a sample from a distribution. It tells us nothing about the model's **stability**. The epistemic republic requires the model to produce **ensembles** — parallel answers generated with:

- fixed temperature
- varied temperature
- prompt perturbations

Semantic divergence, entity inconsistency, causal graph shifts, and stylistic anomalies are measured. These yield an **Ensemble Variance Index (EV)**:

- EV < 0.15 → stable
- EV 0.15–0.45 → mild uncertainty
- EV 0.45–0.70 → significant drift
- EV > 0.70 → high hallucination risk

This mechanism turns hidden neural uncertainty into visible epistemic data. Prompt sensitivity warnings allow users to see where answers are brittle or unstable. Ensemble contradiction clusters highlight regions of internal model disagreement. When ensembles diverge significantly, confidence is automatically dampened — not through arbitrary heuristics, but through measurable variance.

5. Human Feedback and Multi-Tier Oversight (See Appendix D)

The epistemic republic does not trust machine signals alone. Human judgment is built directly into the core of the model. Feedback is layered:

Users (the People)

- correct errors
- upload citations

- challenge claims
- initiate People's Veto
- flag hallucinations

Quaestors (Investigators)

- detect systemic patterns
- analyze hallucination clusters
- identify prompt vulnerabilities
- audit retrieval failures
- generate formal investigative reports
- Quaestors are elected investigators — the bureaucratic heart of the system.

RDEs (Registered Domain Experts)

- adjudicate complex disputes
- provide primary-source knowledge
- establish domain rules
- anchor scholarly continuity
- RDEs serve as the interpretive magistracy, with authority but no monopoly.

Human corrections are formally integrated:

- claim-level
- evidence-level
- retrieval-weight-level
- fusion-model-level
- calibration-level
- This makes the epistemic system *fundamentally corrigible*.

6. Democratic Governance: Elections, Veto, and Checks and Balances (See Appendices E – F)

Unlike technocratic systems, the epistemic republic distributes authority across three estates:

1. **Users (People)**
2. **Quaestors (Investigators)**
3. **RDEs (Experts)**

Each estate holds distinct powers, balanced by term limits, elections, and oversight. Quaestors are elected semi-annually, with weighted ballots giving experts more influence without erasing democratic legitimacy. RDEs undergo stricter elections requiring endorsements and supermajority approval.

A defining feature is the **People's Veto**: If 500 users and 10 quaestors submit evidence contradicting an RDE correction, that ruling is suspended, the claim becomes "disputed," and arbitration is automatically triggered. This prevents expert capture and ensures that authority is revocable.

Removal mechanisms exist for misconduct at all levels, with:

- recall votes
- audit boards
- disciplinary chambers
- transparent public logs
- The system is designed to guard against bias, capture, brigading, and ideological distortion.

7. Calibration: The System's Adaptive Intelligence (See Appendices E - F)

Calibration is the process by which the epistemic republic learns from experience. It operates at four levels:

1. **Micro-calibration** – updates claim-level confidence
2. **Meso-calibration** – adjusts domain heuristics
3. **Macro-calibration** – revises global thresholds
4. **Structural calibration** – constitutional-level updates

Inputs include:

- evidence vectors
- ensemble variance maps

- ϕ-deviation patterns
- human feedback logs
- source corpus evolution
- drift indicators

Drift can be:

- evidential
- conceptual
- distributional
- democratic
- normative

When drift exceeds threshold, human oversight is invoked. Calibration ensures the epistemic system remains responsive to new evidence, updated scholarship, evolving corpora, and shifting norms. It is the mechanism through which the architecture becomes a **self-regulating civic institution**.

8. Integrative Fusion: The Heart of Epistemic Confidence (See Appendices E – F)

All signals—from machine diagnostics to human corrections—are integrated through a **composite confidence fusion model**:

- evidence support
- contradiction
- ensemble variance
- token geometry
- ϕ-modulation
- user feedback
- quaestor reports
- RDE judgments

This produces a calibrated, interpretable probability for each atomic claim. Confidence bands (green, yellow, orange, red, hard-red) make epistemic uncertainty visually intuitive. The fusion model transforms raw machine output into **epistemically structured assertions**.

9. Toward a New Philosophy of Machine Knowledge (See Appendices E – F)

This architecture reflects a deeper philosophical evolution. LLMs cannot be trusted as autonomous oracles. But they can become powerful partners if embedded in structures that mimic:

- scientific method
- peer review
- constitutional government
- deliberative democracy
- proportional reasoning
- evidential jurisprudence

The epistemic republic thus represents a new paradigm: **knowledge as a co-governed civic process between humans and machines.**

The model rejects both:naive techno-utopianism ("the model knows everything"),
and reactionary technophobia ("machines cannot participate in knowledge"). Instead, it proposes a **mixed constitution of epistemic authority**, where:

- machines provide speed, scale, and structural diagnostics
- humans provide interpretation, context, correction, and ethics
- governance distributes power
- calibration ensures adaptability
- proportionality ensures coherence
- democracy prevents capture

It is no accident that the architecture resembles ancient republican systems: truth, like justice, is best produced through structured deliberation.

Conclusion

The emergence of generative AI has exposed a profound structural vulnerability in the modern epistemic landscape: our information systems lack a civic architecture for truth. Large language models speak with fluency but without grounding. They hallucinate with confidence, fabricate evidence with ease, and distribute their errors into a global digital ecosystem that cannot filter, verify, or correct at scale. Left unchecked, this dynamic threatens to erode the foundations of expertise, contaminate scholarly ecosystems, and produce a generation whose intellectual inheritance is built upon an increasingly unstable substrate of machine-generated fiction.

This paper has argued that solving this crisis requires more than incremental safety patches or improved retrieval algorithms. It requires a **constitutional reinvention of the knowledge process itself** — a model that distributes authority, enforces proportional reasoning, embeds contradiction as a primary signal of truth, preserves human interpretive oversight, and ensures that AI remains corrigible under democratic and expert control. The **epistemic republic** is offered as such a model.

Its architecture rests upon six interlocking pillars:

1. **Structural Diagnostics (Token Geometry & ϕ-Proportionality).**
   Probability distributions are not neutral outputs but structural fingerprints of reasoning stability. By assessing proportional decay and comparing geometric–arithmetic means through ϕ-modulated windows, the system identifies hallucination-prone states before they manifest as errors.

2. **Retrieval & Evidence Vectorization.**
   Every claim is grounded in multi-channel retrieval and evaluated through structured evidence vectors that score support, contradiction, relevance, source tier, and temporal coherence. Contradiction outweighs confirmation, reflecting both scientific falsification and legal evidential logic.

3. **Ensemble Variance Analysis.**
   Multiple parallel generations expose internal uncertainty, brittle reasoning, and prompt sensitivity, transforming hidden neural divergence into open epistemic data.

4. **Human Governance: Users, Quaestors, RDEs.**
   Knowledge becomes a civic enterprise. Users supply corrections and transparency. Quaestors conduct investigations, analyze drift, and audit systemic failures. RDEs supply domain-specific expertise while operating under democratic controls, term limits, and peer oversight.

5. **Democratic Checks and Balances.**
   The People's Veto prevents expert capture and ensures that authority remains revocable when evidence contradicts standing corrections. Transparent logs, disciplinary chambers, weighted elections, and term rotations guard against institutional drift and ideological domination.

6. **Continuous Calibration & Drift Management.**
   The system learns from its own contradictions. Micro-, meso-, macro-, and structural calibration

ensure adaptability to evolving evidence, changing corpora, and shifting scholarly norms. Drift is detected, analyzed, and corrected before it cascades into systemic error.

Together, these mechanisms constitute a **multi-estate epistemic polity** in which no single actor — not the model, not experts, not the crowd — holds absolute interpretive power. Instead, authority emerges from structured negotiation among evidence, proportionality, democratic pressure, expert knowledge, and machine diagnostics. Truth becomes not an output but a **process**, not a guess but an **organized civic achievement**.

By embedding AI within institutions modeled on scientific method, legal reasoning, and republican governance, the epistemic republic offers a path toward stability in an increasingly unstable informational age. It provides not only a technical framework but a philosophical stance: that machines can participate in the pursuit of truth only when constrained by structures that render them transparent, accountable, corrigible, and ultimately subordinate to a pluralistic civic order.

In this vision, the future of knowledge is neither technocratic nor anarchic. It is **constitutional**. And it is through such constitutional machinery — not raw computation — that human and machine intelligence may finally collaborate in the disciplined, humble, and enduring search for truth.

APPENDIX A — FORMAL SPECIFICATION OF THE TOKEN GEOMETRY & φ-MODULATED PROPORTIONALITY SYSTEM

(Extended, 5–7 page equivalent)

A.1 Overview

Token geometry refers to the mathematical shape of the probability distribution the LLM assigns to candidate next-tokens. The system outlined in Chapters 6 and 14 interprets that distribution as a *structural indicator* of epistemic regularity or irregularity.

This Appendix formalizes:

1. The definitions of token geometry.
2. The geometric/arithmetic ratio (G/A) as used in the system.
3. Why φ (the golden ratio) serves as the balance benchmark.
4. How φ-modulation is integrated into the confidence fusion.
5. Domain-adaptive φ-windows.
6. Edge cases and exceptions.
7. Mathematical justification for proportionality as an epistemic diagnostic.

The goal is not to treat φ as mystical, but as *a statistical ideal of balanced proportionality* found across natural systems — which gives it special utility in detecting distorted or pathological distributions.

A.2 Token Probability Geometry: Definitions

Let a next-token distribution be:

$$\vec{p} = (p_1, p_2, \ldots, p_n)$$

ordered by decreasing probability.

The model's epistemic architecture evaluates:

- **Shape:**
  whether the distribution is sharply peaked (overconfident)
  or overly flat (diffuse, directionless).
- **Proportional Consistency:**
  how smoothly probability decays across ranks.
- **Balance vs. Distortion:**
  whether the distribution contains probabilistic "irregularities"
  analogous to abnormal frequencies in signal processing.

Token geometry supplements retrieval and ensembles by revealing whether the *internal probability computation itself* appears stable or distorted.

This enables the system to detect:

- pathological confidence spikes,
- highly irregular distributions,
- incoherent "guessing shapes,"
- and φ-divergent fluctuations often characteristic of hallucination.

A.3 The Geometric Mean / Arithmetic Mean Ratio

For a distribution $\vec{p}$:

$$GM = (p_1 p_2 \cdots p_n)^{1/n}$$

$$AM = \frac{1}{n}(p_1 + p_2 + \cdots + p_n)$$

The **G/A ratio** becomes:

$$R = \frac{GM}{AM}$$

Properties:

- $R \to 1$ when the distribution is **equal** across tokens
- $R \to 0$ as the distribution becomes **extremely peaked** (one token dominates)
- Mildly peaked distributions produce values around **0.55–0.80**

This ratio serves as a "smoothness diagnostic":
how balanced or distorted the model's probability landscape is.

Interpretation

- High R → *healthy uncertainty, balanced proportions*
- Low R → *pathological overconfidence or distribution collapse*

This is analogous to entropy or perplexity but more **shape-sensitive**.

A.4 Why φ Is the Benchmark of Proportional Balance

The golden ratio $\varphi = \frac{1+\sqrt{5}}{2} \approx 1.618$ is not important as a number, but as a **ratio of self-similar proportionality** found across:

- spirals
- branching systems
- aesthetic ratios
- sequential decay patterns
- bounded growth processes
- stable equilibria in dynamical systems

The epistemic system uses φ not as numerology, but as a **mathematical paradigm of proportion**.

Key Insight

φ establishes a relationship between:

$$\frac{\text{whole}}{\text{large part}} = \frac{\text{large part}}{\text{small part}}$$

This proportional self-similarity is the *exact structure* we desire in a token probability decay curve:

- The decline should be smooth, not spiky.
- The distribution should show balanced proportional decay.
- Irregularities signal instability in the internal computation.

Thus, φ is used as:

- **Structural comparator**, not a "truth detector."
- **Signal of balance**, not correctness.
- **Diagnostic of distribution health**, not evidence.

A.5 φ-Modulated Epistemic Penalty

Let the G/A ratio be R.

We define the φ-proportionality window:

$$\varphi^{-1} \approx 0.618 \quad \text{to} \quad \varphi^{0} = 1.0$$

This window reflects:

- the **lower proportional bound** (0.618),
- the **upper bound** (1.0),
- the **optimal center** near ≈ 0.72–0.78.

We then compute a φ-modulated deviation:

$$D_\varphi = \max(0, \varphi^{-1} - R, R - 1)$$

The corresponding penalty weight:

$$P_\varphi = \alpha \cdot D_\varphi$$

Where:

- α is a small constant (e.g., 0.2–0.3 of contradiction weight)
- Penalty never exceeds evidence contradiction
- It calibrates *distribution irregularity*, not factual correctness

This yields a proportional dampening or stabilizing influence.

A.6 Domain-Adaptive φ-Windows

Different fields require different proportional "shapes":

Mathematics, Physics, Computer Science

Expect extremely sharp distributions for exact claims.

φ-window is narrower (0.70–0.85).

History, Literary Studies, Anthropology

Expect broader distributions.

φ-window expands (0.55–0.90).

Theology, Philosophy

High interpretive openness; φ penalty is milder.

Medicine, Law

Moderately strict φ-windows to ensure clarity but accommodate nuance.

Thus φ is **contextual**, not universal.

A.7 Edge Cases and Exceptions

φ-modulation is **not applied** when:

- evidence is direct and unambiguous,
- the model is citing verbatim text,
- the claim references a numerical constant,
- only two token candidates exist (binary decisions),
- probability collapse is expected (names, dates),
- tokens reflect non-linguistic tasks (math, code).

This prevents φ from interfering with legitimate prediction shapes.

A.8 Mathematical Justification: Proportion as Epistemic Indicator

The justification draws from:

A.8.1 Information Theory

Irregularities in distribution shape correlate with:

- low cross-entropy correlations,
- unstable embedding activation patterns,
- hallucination precursors.

A.8.2 Signal Processing

Smooth decay curves represent coherent systems.
Spikes indicate noise or artifacts.

A.8.3 Dynamical Systems

Stable equilibria correspond to proportionally balanced gradients.

A.8.4 Neural Network Theory

Layer activations with disproportionate gradients
often signal instability or incomplete grounding.

Thus φ-modulation is not aesthetic — it's a structural measure of internal coherence.

A.9 Summary

Appendix A established:

- the mathematics of token geometry,
- the logic behind φ-proportionality,
- the computation of φ-modulated penalties,
- domain-adaptive windows,
- exclusion criteria,
- and theoretical justification.

This mechanism does **not** determine truth.

It detects **internal epistemic stability** — a necessary condition for reliable reasoning.

APPENDIX B — FORMAL SPECIFICATION OF THE RETRIEVAL, CONTRADICTION, AND EVIDENCE-BASED GROUNDING ENGINE

(Extended, 5–7 page equivalent)

B.1 Overview

Retrieval is the epistemic backbone of the entire architecture.
Token geometry reveals *internal* stability; ensemble divergence reveals *internal* uncertainty; φ-modulation reveals *internal* proportionality.
But retrieval and contradiction signals reveal **external reality**.

This Appendix fully formalizes:

1. The retrieval pipeline
2. Entailment and contradiction scoring
3. Evidence weighting
4. Source classification
5. Temporal reasoning
6. Multi-source conflict resolution
7. Domain-specific retrieval heuristics
8. Drift detection through evidence shifts
9. The role of human corrections within retrieval logic

The Retrieval–Contradiction Engine (RCE) is the system's equivalent of:

- A court archive,
- A research library,
- A fact-checking bureau, and
- An epistemic stabilizer.

It determines how claims are grounded in evidence and how contradictions are detected, weighted, surfaced, and incorporated into the composite confidence score.

B.2 The Retrieval Pipeline: Step-by-Step

The RCE works in four stages:

B.2.1 Claim Decomposition

Every model-generated paragraph is decomposed into **atomic factual claims**, each tagged with:

- entities
- dates
- events
- relationships
- quantifiers
- embedded assertions

Example:

"Gerard's Herball (1597) describes Polemonium on page 989."

→ becomes four atomic claims:

1. Gerard authored the Herball.
2. It was published in 1597.
3. It contains a description of Polemonium.
4. This occurs on page 989.

This ensures granularity and precision.

B.2.2 Multi-Channel Retrieval

Claims are passed into multiple retrieval channels:

- **Semantic search** (embedding-based)
- **Lexical search** (keyword-based)
- **Hybrid retrieval** (fused scoring)
- **Citation-based retrieval** (via known canonical sources)
- **Temporal-bounded retrieval** (historical queries)
- **Disputation-aware retrieval** (for contested domains)

Each channel returns candidate evidence snippets.

B.2.3 Evidence Scoring

Each retrieved snippet is evaluated through:

- entailment strength
- contradiction likelihood
- direct support
- relevance to the claim
- domain context
- recency
- source type

We compute:

$$E_s = \alpha \cdot (\text{entailment}) - \beta \cdot (\text{contradiction})$$

where contradiction always carries a **stronger weight**

B.2.4 Composite Evidence Vector

Each claim receives a structured evidence vector:

$$\vec{EV} = (support, neutral, contradiction, source_quality, recency, density)$$

This vector feeds directly into the composite confidence fusion model.

B.3 Entailment and Contradiction Scoring: Formal Specification

Entailment is computed using a multi-model approach:

- cross-encoder NLI
- bi-encoder similarity
- contradiction detectors
- negation parsing
- numerical mismatch detection
- date/interval logic

Contradiction categories include:

B.3.1 Direct Contradiction

"The Herball does not mention Polemonium."

B.3.2 Numerical/Dated Contradiction

"Gerard's Herball was printed in 1597" vs. "1598."

B.3.3 Entity Misattribution

"Dioscorides described pinnate leaves" when he did not.

B.3.4 Absence Contradiction

Primary text fails to support claim after targeted retrieval.

B.3.5 Logical Contradiction

Statements mutually exclude each other (e.g., "X invented Y" vs "Y predates X").

Contradictions receive **multiplicative weights**, because they are epistemically decisive.

B.4 Evidence Quality: Source Classification System

Sources are ranked by reliability:

Tier 1 — Primary Sources

- manuscripts
- original editions
- inscriptions
- archival documents
- scientific measurements

Tier 2 — Peer-Reviewed Secondary Sources

- academic monographs
- journal articles
- critical editions

Tier 3 — Classical Commentaries

- pre-modern authorities (e.g., Oribasius, Galen, Avicenna)
- renaissance commentators (Mattioli, Fuchs, Brunfels)

Tier 4 — Tertiary Sources

- encyclopedias
- textbooks
- compendia

Tier 5 — Unverified Sources

- blogs
- casual websites
- OCR-mangled texts

Evidence from lower tiers receives exponentially declining weight.

B.5 Temporal Reasoning

Claims must be cross-checked for **chronological coherence**:

- publication dates
- historical events
- author lifespans
- manuscript production windows
- botanical naming conventions by century
- lexicon availability
- anachronistic terminology

Temporal contradiction is one of the strongest disproof indicators.

Example:

If a user asks:

"What did Linnaeus say about Polemonium in 1620?"

Retrieval flags:

Linnaeus was not yet born (1707).

→ hard contradiction → confidence collapse.

B.6 Multi-Source Conflict Resolution

When sources disagree, the system:

1. surfaces all conflicts visually
2. assigns confidence bands based on tiered reliability
3. propagates contradiction weight into fusion
4. triggers **disputation mode** if conflict exceeds threshold
5. calls for human review (quaestors → RDEs) if unresolved

This prevents premature epistemic certainty.

B.7 Domain-Specific Retrieval Heuristics

Different domains require different retrieval strategies:

Botany

- Latin binomial history
- herbals by century
- morphology terms
- earliest attestations

Classical Studies

- manuscript families
- textual variants
- scholia
- lexicon consistency

History

- contemporary accounts > later interpretations
- political context
- historiographical layers

Theology

- doctrinal evolution
- patristic vs scholastic sources
- translation dependencies

Thus retrieval is adaptive, not generic.

B.8 Drift Detection

Over time:

- new sources arise
- scholarly consensus shifts
- OCR improves
- editions get corrected
- archives expand

B.7 Domain-Specific Retrieval Heuristics

Different domains require different retrieval strategies:

Botany

- Latin binomial history
- herbals by century
- morphology terms
- earliest attestations

Classical Studies

- manuscript families
- textual variants
- scholia
- lexicon consistency

History

- contemporary accounts > later interpretations
- political context
- historiographical layers

Theology

- doctrinal evolution
- patristic vs scholastic sources
- translation dependencies

Thus retrieval is adaptive, not generic.

B.8 Drift Detection

Over time:

- new sources arise
- scholarly consensus shifts
- OCR improves
- editions get corrected
- archives expand

B.10 Summary

Appendix B defined:

- the retrieval pipeline
- contradiction scoring
- source weighting
- temporal logic
- multi-source conflict resolution
- domain-sensitive heuristics
- drift detection
- human integration

Retrieval is not merely a technical feature —
it is the system's **epistemic grounding apparatus**,
the anchor to the real world,
the stabilizer against hallucination,
and the interface between machine reasoning and human scholarship.

APPENDIX C — THE ENSEMBLE VARIANCE, PROMPT-STABILITY, AND INTERNAL DIVERGENCE ENGINE

(Extended, 5–7 page equivalent)

C.1 Overview

While retrieval grounds claims in external evidence, the *ensemble system* reveals the model's **internal epistemic stability**.

It allows users and experts to see how the model's reasoning shifts when:

- the prompt is rephrased,
- the sampling temperature varies,
- the generation is repeated,
- synonyms or syntactic changes are introduced,
- context windows alter,
- or adversarial prompts appear.

This Appendix formalizes:

1. Ensemble generation
2. Ensemble divergence metrics
3. Semantic and conceptual drift measurement
4. Prompt sensitivity diagnostics
5. Contradiction clustering
6. Ensemble-based hallucination prediction
7. Integration into the confidence fusion model
8. Human oversight of ensemble instability

The goal is to expose the *epistemic contour map* hidden inside a single answer.

C.2 First Principles: Why Ensembles Matter

A single LLM response is not a single "thought."

It is one **sample** from a probability space.

Thus:

- A correct answer must be **repeatable** across parallel generations.
- A hallucination often appears in **only one or two** generations.
- High-variance ensembles indicate internal uncertainty.
- Prompt sensitivity reveals conceptual fragility.

This is analogous to:

- scientific replication,
- polling variance,
- Monte Carlo sampling,
- and Bayesian posterior stability.

A lone generation is a *point estimate;*
an ensemble is an *epistemic distribution*.

C.3 Ensemble Generation: The Three Modes

The system generates three ensemble types:

C.3.1 Fixed-Temperature Ensemble

Temperature is constant (e.g., 0.2).
Reveals structural, non-random divergence.

C.3.2 Variable-Temperature Ensemble

Temperatures sampled from 0.1 to 0.8.
Reveals sensitivity to widening probability distributions.

C.3.3 Perturbation Ensemble

The prompt is altered:

- synonyms
- word order
- slight rephrasings
- additional context
- removal of qualifiers

This detects **prompt susceptibility**.

Together, they produce a comprehensive internal picture.

C.4 Measuring Semantic Divergence

Each ensemble generation is converted into:

- embedding vectors
- entity lists
- date and numeric extractions

- causal graph structures
- claim decompositions

Divergence is measured across:

C.4.1 Embedding Cosine Distance

High distance → conceptual drift.

C.4.2 Entity Variance

Different people, places, dates → epistemic instability.

C.4.3 Claim-Level Contradictions

If ensemble responses contradict each other → red flag.

C.4.4 Causal Graph Variance

Different event orderings → deeper structural inconsistency.

C.4.5 Stylistic Drift

Unexpected shifts in tone or argumentation style, often correlated with hallucination.

C.5 Quantifying Ensemble Instability: The EV Index

We define the **Ensemble Variance Index (EV)**:

$$EV = \lambda_1 D_{sem} + \lambda_2 D_{ent} + \lambda_3 D_{cas} + \lambda_4 D_{sty} + \lambda_5 N_{contra}$$

Where:

- D_{sem} = semantic distance
- D_{ent} = entity divergence
- D_{cas} = causal structure variance
- D_{sty} = stylistic anomaly score
- N_{contra} = count of direct contradictions

Weights λ_i are domain-specific.

Interpretation:

- EV < 0.15 → Strong internal stability
- EV 0.15–0.45 → Moderate stability, mild uncertainty
- EV 0.45–0.70 → Significant internal divergence
- EV > 0.70 → Highly unstable; likely hallucination

This metric is surfaced visually in the UI.

C.6 Prompt Sensitivity Warning System

The system re-evaluates claims when the prompt is:

- rephrased,
- shortened,
- expanded,
- rearranged,
- de-formatted.

If the response changes materially, a **Prompt Sensitivity Warning** appears:

> ⚠️ *"This claim is unstable under prompt variation."*

Reasons include:

- brittle retrieval triggers
- ambiguous phrasing
- underlying conceptual confusion
- lack of grounding
- hallucination susceptibility

Prompt sensitivity reduces the composite confidence score.

C.7 Contradiction Clustering Across Ensembles

Some contradictions do not arise from evidence retrieval but from **internal model disagreement.**

The system clusters contradictions across ensemble generations:

C.7.1 Soft Contradictions

Different emphasis, not incompatible.

C.7.2 Moderate Contradictions

Different dates, entities, or attributions.

C.7.3 Hard Contradictions

Claims cannot both be true.

Hard contradictions trigger:

- confidence collapse
- contradiction panel
- quasi-judicial review (quaestors → RDEs)

Contradiction clustering is a major predictor of hallucination.

C.8 Ensemble-Based Hallucination Prediction

Hallucinations often follow patterns:

- unique phrasing not present elsewhere
- abrupt stylistic shifts
- high-variance entity substitutions
- contradiction spikes
- disproportionate token geometry irregularities

The ensemble engine predicts hallucination risk by examining:

1. Variance in internal evidence
2. Instability in reasoning
3. Contradiction clustering
4. Absence of recurring phrases
5. Distribution-level anomalies

High hallucination scores display as a **striped red bar** around the claim.

C.9 Fusion with Evidence and φ-Modulation

Ensemble instability is only **one** epistemic signal.

It integrates with:

- evidence support/contradiction,
- token geometry,
- φ-modulation,
- human feedback,
- temporal checks.

The composite confidence formula includes an ensemble term:

$$z(c) = \cdots - w_A EV(c) + \ldots$$

where w_A is calibrated to domain sensitivity.

Ensemble divergence alone cannot disprove a claim —
but it *reduces confidence* and *requires grounding through retrieval*.

C.10 Visualization in the User Interface

The system exposes ensemble structure through:

- color-coded confidence
- sparkline variance graphs
- contradiction markers
- entity lists across ensembles
- toggleable alternate answers
- perturbation comparison overlays

This makes the invisible internal workings of LLMs **legible**, much like MRI imaging reveals internal biological structure.

C.11 Role of Quaestors and RDEs in Ensemble Analysis

Quaestors

- investigate instability clusters
- trace contradictions to source hallucinations
- detect adversarial prompt vulnerability
- flag brittle domains for expert review

RDEs

- interpret conceptual divergences
- provide domain-specific stability rules
- explain which ensemble variations are legitimate
- help recalibrate variance thresholds

Thus ensembles are not only technical artifacts —
they are *objects of scholarly interpretation* in the epistemic republic.

C.12 Summary

Appendix C formalized:

- ensemble generation
- prompt sensitivity
- divergence metrics
- hallucination prediction
- contradiction clustering
- human-mediated resolution
- integration into fusion

The ensemble engine transforms the LLM from a "single answer machine" into a **transparent epistemic organism** whose internal uncertainty is visible, analyzable, and governable.

APPENDIX D — THE HUMAN FEEDBACK, CORRECTION, AND OVERSIGHT ENGINE

(Extended, 5–7 page equivalent)

This appendix is the backbone of the epistemic republic's **human participation layer.** Retrieval grounds claims in external evidence, ensembles expose internal model uncertainty, and token geometry provides structural diagnostics. But none of those mechanisms replace the interpretive clarity, contextual understanding, and moral responsibility of human oversight.

Appendix D formalizes:

1. The full feedback pipeline
2. User feedback scoring
3. Quaestor investigative workflow
4. RDE correction authority
5. Error taxonomies
6. Feedback integration into calibration
7. Reputation and trust models
8. Fraud, abuse, and bias detection
9. Democratic oversight structures

It describes *how* humans correct the system, *how* the system incorporates corrections, and *how* authority is distributed and regulated.

D.1 Overview: Human Judgment as a First-Class Epistemic Signal

The architecture does **not** assume:

- experts are always right,
- users are always naive,
- the machine is always neutral.

Instead:

- the machine grounds output in evidence
- users challenge, contextualize, verify
- quaestors investigate and diagnose
- RDEs adjudicate with domain expertise

Thus:

> **Truth arises from the structured interaction of machine logic and human reason.**

Human feedback is *not* a footnote —
it is a **central epistemic engine.**

D.2 Taxonomy of Human Feedback Types

There are four formal feedback types:

D.2.1 Direct User Feedback

Accessible to everyone.

Actions include:

- marking a claim incorrect
- uploading a source or citation
- linking manuscripts or images
- flagging hallucinations
- commenting on ambiguity or nuance

User feedback is weighted lightly but is essential for volume and pattern detection.

D.2.2 Structured User Annotations

Users may provide:

- lineage of sources
- semantic clarifications
- personal expertise (lightly weighted)
- specific textual corrections

Structured annotations use defined forms, allowing automated integration.

D.2.3 Quaestor Investigative Reports

Quaestors produce formal reports containing:

- hallucination clusters
- misattributions
- domain-specific ambiguity
- retrieval failures
- contradictory evidence
- prompt vulnerability analysis
- recommended domain rules

Their feedback is medium-weight and can trigger recalibration.

D.2.4 RDE Authoritative Corrections

RDEs provide heavyweight corrections:

- primary-source citations
- philological or methodological commentary
- domain rules
- correction templates
- negative evidence ("absence of attestation")
- variant resolution (textual criticism)

RDEs anchor the epistemic hierarchy without monopolizing it.

D.3 The Feedback Integration Pipeline

Corrections enter a multi-step pipeline:

D.3.1 Initial Screening

System checks for:

- hostility
- spam
- duplicated content
- low-effort noise

Accepted feedback proceeds.

D.3.2 Claim Matching

Feedback is matched to specific *atomic claims*.

This avoids imprecision, e.g.:

❌ "This whole answer is wrong."
✔ "Claim #3: Gerard mentions Polemonium on page 989. This is incorrect."

D.3.3 Evidence Verification

If citations are attached, retrieval is triggered:

- locate provided passages
- OCR and normalize
- validate context
- extract semantic entailment

This step ensures corrections are verifiable.

D.3.4 Weight Assignment

Feedback is weighted by:

- user reputation
- evidence quality
- source tier
- whether citations match primary texts
- whether feedback contradicts the majority
- RDE/quaestor involvement

Weights are numerical and interpretable.

D.3.5 Fusion Integration

Feedback modifies:

- contradiction score
- hallucination penalty
- retrieval heuristics
- domain rules
- φ-modulation weight
- ensemble thresholds

The fusion model **evolves** with human input.

D.4 Error Taxonomy: What Humans Correct

Humans correct errors the model cannot reliably detect:

D.4.1 Interpretation Errors

- mistranslation
- miscontextualization
- misreading analogies or metaphors
- philosophical category mistakes

D.4.2 Historical Errors

- anachronisms
- wrong centuries
- incorrect associations
- misattributed authorship

D.4.3 Philological Errors

- incorrect Greek/Latin
- false cognates
- mistranscribed manuscript lines

D.4.4 Domain Methodology Errors

- misapplied botanical morphology
- incorrect archaeological reasoning
- flawed scientific inference

D.4.5 Ethical/Contextual Errors

- culturally uninformed judgments
- improper generalizations
- missing social context

Humans excel at *seeing around the edges* of facts — providing interpretive scaffolding that machines lack.

D.5 Reputation, Trust, and Legitimacy Models

To prevent manipulation or bad-faith use, human input is filtered through formal reputation/tier models.

D.5.1 User Reputation

Earned through:

- accurate corrections
- citation quality
- consistency over time
- alignment with evidence

Bad corrections gently lower reputation; repeated abuse triggers limits.

D.5.2 Quaestor Reputation

Quaestors earn rank by:

- successful investigations
- accuracy of pattern detection
- peer affirmation
- passing periodic review

Failure results in:

- rank reduction
- suspension
- removal

D.5.3 RDE Reputation

Expert authority is monitored for:

- mistake rate
- bias indicators
- deviation from evidence
- user/quaestor confidence
- interdisciplinary feedback

Term limits ensure no RDE becomes a permanent epistemic oligarch.

D.6 The Calibration Engine's Use of Human Feedback

The calibration engine uses human input to adjust:

- retrieval weighting
- contradiction penalties
- ensemble thresholds
- φ-window boundaries
- hallucination templates
- domain-specific heuristics

Corrected claims feed into **model-wide learning**, improving reliability across the system.

D.7 Detecting Abuse, Bias, and Manipulation

The system includes mechanisms to detect:

D.7.1 Coordinated Downvoting

Consistency checks prevent mass brigading.

D.7.2 RDE Misconduct

- self-dealing
- biased rulings
- conflict of interest

Triggers immediate review.

D.7.3 Ideological Capture

If corrections consistently come from one political or ideological group, the system triggers:

- diversification weighting
- independent arbitration
- evidence audits

D.7.4 Statistical Anomaly Detection

Machine learning monitors:

- unnatural feedback patterns
- sudden reputation spikes
- adversarial correction sequences

Ensuring epistemic stability.

D.8 Democratic Oversight: The People's Role

Users are not passive recipients — they are epistemic citizens.
Their powers include:

- electing quaestors
- vetoing unilateral RDE rulings
- calling for arbitration
- mass-correcting errors
- monitoring expert conduct
- challenging contradictions
- providing non-digitized evidence

This enshrines *civic participation* as part of truth-seeking.

D.9 Summary

Appendix D formalized the human side of the epistemic republic:

- user corrections
- quaestor investigations
- RDE adjudication
- feedback integration
- error taxonomy
- reputation systems
- fraud detection
- democratic oversight

Human feedback is not ornamental —
it is the *mechanism that makes the entire system morally and epistemically legitimate.*

APPENDIX E — THE GOVERNANCE, ELECTION, AND OVERSIGHT FRAMEWORK OF THE EPISTEMIC REPUBLIC

(Extended, 5–7 page equivalent)

E.1 Overview

This appendix provides the detailed constitutional mechanics governing:

1. **Election of Quaestors and RDEs**
2. **Term limits and rotation**
3. **Oversight and accountability**
4. **Conflict-of-interest rules**
5. **Removal mechanisms**
6. **Appeals and arbitration escalation**
7. **People's veto constraints**
8. **Jurisdiction and domain authority**
9. **Procedural protections against epistemic capture**

While Chapter 13 established the *philosophical foundation* for the epistemic republic, Appendix E codifies the **functional rules that make the system governable.**

E.2 The Three Estates: Constitutional Powers and Duties

E.2.1 The People (Users)

The largest estate, constituting the democratic base.

Powers:

- Elect quaestors
- Initiate the People's Veto
- Submit corrections and citations
- Report misconduct (quaestors or RDEs)
- Trigger arbitration review for contested claims
- Provide mass feedback for recalibration

Duties:

- Participate honestly
- Provide citational grounding when possible
- Uphold epistemic norms

E.2.2 Quaestors

The investigative magistracy.

Powers:

- Conduct pattern analyses
- Issue investigative reports
- Trigger arbitration
- Recommend RDE review
- Audit hallucination clusters
- Oversee retrieval drift
- Moderate user disputes
- Apply medium-weight corrections

Duties:

- Neutrality and analytical rigor
- No partisanship or bias
- Evidence-first investigation
- Periodic continuing-education modules

Quaestors are **the backbone of epistemic bureaucracy**, acting as the "civil service" of the knowledge republic.

E.2.3 Registered Domain Experts (RDEs)

The interpretive magistracy.

Powers:

- Provide authoritative corrections
- Determine primary source authenticity
- Establish domain rules (with peer approval)
- Serve on arbitration councils
- Override lower-tier inconsistencies

Duties:

- Maintain scholarly standards
- Avoid conflicts of interest
- Uphold transparency
- Serve limited terms
- Accept review and reversal through constitutional processes

RDEs occupy the highest epistemic office, but with **strict limitations**.

E.3 Election Mechanics

E.3.1 Quaestor Elections

Elections occur every 6 months.

Eligibility:

- 50+ verified corrections, OR
- 10+ citations uploaded, OR
- nomination by an existing quaestor.

Ballot Structure:
Votes weighted as:

- Users: **1×**
- Quaestors: **2×**
- RDEs: **5×**

Reasoning:
Higher offices weigh electors more because they understand domain complexity, while users maintain democratic legitimacy.

Outcome:
Top candidates fill open seats. Ties go to runoff.

E.3.2 RDE Elections

More rigorous than quaestor elections.

Eligibility:

- Endorsement by either
  - **2 RDEs** OR
  - **20 Quaestors**
- Evidence of domain expertise
- Demonstrated correction accuracy
- Passing the "Scholarly Transparency Audit" (panel review of reasoning quality)

Voting Structure:
Votes weighted as:

- Users: **1×**
- Quaestors: **4×**
- RDEs: **10×**

Outcome:
Supermajority (≥ 64% weighted approval) required.

This prevents low-quality experts from gaining authority.

E.4 Term Limits and Rotation

To prevent institutional capture or epistemic oligarchy:

E.4.1 Quaestors

- Term: **1 year**
- Max consecutive terms: **2**
- Mandatory 6-month "civic rest period" before running again

E.4.2 RDEs

- Term: **2 years**
- Max consecutive terms: **3**
- Mandatory 1-year sabbatical
- RDEs cannot hand-pick successors

Rotation maintains epistemic diversity.

E.5 Oversight and Accountability Structures

Oversight is divided across three branches:

E.5.1 The People's Tribunal

User-only oversight mechanism.

Powers:

- Challenge quaestor misconduct
- Flag RDE bias
- Trigger oversight review
- Demand transparency for arbitration decisions

E.5.2 Quaestor Audit Board

Comprised of 7 randomly selected quaestors.

Powers:

- Review RDE corrections
- Investigate procedural errors
- Evaluate evidence-handling protocols

E.5.3 RDE Peer Review Chamber

- Evaluates scholarly validity
- Can censure RDEs for methodological flaws
- Oversees domain rule consistency

E.5 Oversight and Accountability Structures

Oversight is divided across three branches:

E.5.1 The People's Tribunal

User-only oversight mechanism.

Powers:

- Challenge quaestor misconduct
- Flag RDE bias
- Trigger oversight review
- Demand transparency for arbitration decisions

E.5.2 Quaestor Audit Board

Comprised of 7 randomly selected quaestors.

Powers:

- Review RDE corrections
- Investigate procedural errors
- Evaluate evidence-handling protocols

E.5.3 RDE Peer Review Chamber

- Evaluates scholarly validity
- Can censure RDEs for methodological flaws
- Oversees domain rule consistency

E.6 Removal Procedures

Removal is a multi-stage process:

E.6.1 User-Initiated Recall

Triggered by:

- 500+ user signatures
- at least 3 attached citations demonstrating misconduct
- confirmation by 2 quaestors

E.6.2 Quaestor-Initiated Removal

Requires:

- Formal report
- Supermajority vote among quaestors
- RDE chamber confirmation (for RDE removal)

E.6.3 RDE-Initiated Disciplinary Removal

Triggered when:

- An RDE violates scholarly standards
- Evidence of repeated partisan interference
- Deliberate misuse of authority

Final removal requires:

- 2/3 RDE council vote
- 1/3 quaestor concurrence

E.7 Conflict of Interest Rules

Strict rules prevent corruption:

- RDEs cannot adjudicate claims involving their own published work without peers present.
- Quaestors cannot investigate disputes in which they previously participated as users.
- All corrections must be tied to citations, not opinions.
- Transparency logs record every action, vote, and ruling.
- Anonymous corrections by RDEs are forbidden.
- Undisclosed affiliations (institutional, financial, ideological) invalidate rulings retroactively.

These rules mirror the ethics codes of academia, journalism, and open-source governance.

E.8 Appeals and Arbitration Escalation

E.8.1 Appeal Triggers

- Disagreement between quaestors and RDEs
- Ambiguous or fragile evidence
- Significant ensemble divergence
- Community pushback
- New evidence revisions

E.8.2 Arbitration Hierarchy

1. **Local Arbitration (3-person panel)**
2. **Domain Arbitration Council (5-person panel)**
3. **Supreme Epistemic Council (SEC)**
    - 3 RDEs
    - 2 quaestors
    - 2 randomly selected citizens

The SEC is convened only for systemic disputes.

E.9 People's Veto: Full Specification

The veto exists to prevent expert autocracy.

E.9.1 Veto Trigger Threshold

A People's Veto is initiated when:

- **500+ users** provide citations contradicting an RDE ruling
- **10+ quaestors** affirm the procedural validity
- The contradiction is *evidence-based*, not opinion-based

E.9.2 Effects of the Veto

- RDE ruling is immediately suspended
- Claim receives "Disputed" status
- Arbitration is automatically convened
- Ensemble and retrieval re-run with extra scrutiny
- φ-modulation and token geometry weight increase temporarily

The veto is democratic, but *evidence-driven*, not populist.

E.10 Jurisdiction and Domain Authority

E.10.1 RDEs hold authority

only *within their certified domain*.

E.10.2 Quaestors hold authority

across *all domains*, but only in procedural matters.

E.10.3 Users hold authority

in:

- elections
- veto
- feedback
- misconduct oversight

E.10.4 SEC holds cross-domain authority

only in extraordinary disputes.

E.11 Protection Against Epistemic Capture

Capture happens when:

- a small group controls an entire domain
- ideological blocs distort corrections
- experts silence dissent
- democratic brigading manipulates outcomes

Protections include:

- term limits
- rotating offices
- random juries
- diversity enforcement
- weighted elections
- contradiction-driven arbitration
- public logs
- distributed authority

This ensures the epistemic republic remains plural, corrigible, and fair.

E.12 Summary

Appendix E codifies the **political architecture** of the epistemic republic:

- elections
- term limits
- oversight councils
- tiered authority
- conflict-of-interest rules
- arbitration hierarchy
- veto rights
- capture prevention

It describes a balanced system where **truth emerges from structured negotiation** between evidence, expertise, democracy, and machine logic.

APPENDIX F — THE CALIBRATION ENGINE, DRIFT MANAGEMENT, AND SYSTEM-WIDE LEARNING

(Extended, 5–7 page equivalent)

F.1 Overview

The calibration engine is the central nervous system of the epistemic republic — the mechanism through which the system:

- **learns** from corrections,
- **adapts** to new evidence,
- **detects** epistemic drift,
- **re-weights** domain heuristics,
- **improves** distributional proportionality,
- **updates** fusion parameters,
- **stabilizes** interpretive norms,
- **incorporates** new sources,
- **manages** authority transitions.

Calibration transforms the architecture from a static knowledge interface into a **dynamic epistemic organism**.

This appendix provides a full formal specification of that process.

F.2 Calibration Types: A Four-Tier Framework

Calibration occurs on four hierarchical levels:

F.2.1 Micro-Calibration

Performed *automatically* after each correction or contradiction.

Updates:

- claim-level confidence
- retrieval source ranking
- hallucination penalties
- φ-deviation penalties
- ensemble threshold adjustments

Micro-calibration is continuous and granular.

F.2 Calibration Types: A Four-Tier Framework

Calibration occurs on four hierarchical levels:

F.2.1 Micro-Calibration

Performed *automatically* after each correction or contradiction.

Updates:

- claim-level confidence
- retrieval source ranking
- hallucination penalties
- φ-deviation penalties
- ensemble threshold adjustments

Micro-calibration is continuous and granular.

F.2.2 Meso-Calibration

Triggered daily or weekly.

Updates:

- domain heuristics
- retrieval weights for specific corpora
- contradiction penalties (domain-specific)
- RDE rule consistency
- known hallucination patterns
- prompt-fragility templates

Meso-calibration stabilizes entire fields of knowledge.

F.2.3 Macro-Calibration

Triggered monthly or during major updates.

Updates:

- evidence weighting tiers
- global φ-window adjustments
- ensemble divergence thresholds
- election/civic governance rules
- trust-level recalibration for users, quaestors, and RDEs
- cross-domain relationship mappings

Macro-calibration preserves the system's "constitutional" coherence.

F.2.4 Structural Calibration (Epochal)

Rare and major — analogous to constitutional amendment.

Triggered when:

- shifting scholarly consensus
- massive new digitization of manuscripts
- discovery of large-scale OCR issues
- systemic hallucination problems
- domain-wide model drift
- major governance reforms
- scientific paradigm shifts

Structural calibration redefines system-level norms, requiring full transparency and approval from:

- the RDE chamber,
- the Quaestor Audit Board,
- and the People's Tribunal (via open ballot).

F.3 Data Inputs for Calibration

Calibration draws from six primary inputs:

F.3.1 Evidence Vectors

Each claim has a "fingerprint":

- support
- contradiction
- source tier
- recency
- entailment density
- conflict distribution

The calibration engine detects shifts across these vectors.

F.3.2 Ensemble Variance Maps

Patterns of divergence reveal:

- fragile claims
- contested domains
- prompt-sensitivity hot spots
- unstable reasoning structures

These maps are essential for domain-wide recalibration.

F.3.3 Token Geometry Data

Evaluates:

- distribution skew
- proportional imbalance
- recurring φ-deviation patterns
- intrinsic instability

Geometry informs proportionality correction.

F.3.4 Human Feedback Logs

Collected from:

- users
- quaestors
- RDEs

Analyzed for:

- correction density
- bias indicators
- disagreement clusters
- recurrent misinterpretations
- potential governance issues

F.3.5 Source Corpus Evolution

Monitors:

- corpus expansion
- new manuscript uploads
- corrections in critical editions
- retractions or discoveries
- OCR rescans

This is crucial for drift detection.

F.3.6 Drift Signals

Indicators include:

- increasing contradiction rates
- domain consensus instability
- semantic drift
- rising ensemble variance
- systematic user corrections
- cross-domain anomalies
- mismatched φ-window trends

When drift exceeds threshold, macro-calibration is triggered.

F.4 Drift: Causes, Types, and Detection Mechanics

Drift is inevitable — not a defect, but a **fundamental epistemic phenomenon**.

The system recognizes five forms:

F.4.1 Evidence Drift

Sources change or improve:

- new manuscripts
- new archaeological finds
- better OCR
- revised critical editions
- digitization of archives

Evidence drift requires recalibration of retrieval heuristics.

F.4.2 Conceptual Drift

Language or interpretation evolves:

- new terminology
- shifting scholarly consensus
- reinterpretation of historical data
- changes in accepted methodology

Conceptual drift requires ontology updates.

F.4.3 Distributional Drift

Model-level changes:

- new LLM versions
- updated embeddings
- fine-tuning
- training data expansion

Distributional drift affects token geometry and ensembles.

F.4.4 Democratic Drift

User and community feedback patterns shift:

- trending corrections
- ideological brigading
- mass re-evaluation of claims
- increased literacy or expertise

Requires governance review.

F.4.5 Normative Drift

Values and ethical standards evolve:

- terminology norms
- sensitivity to cultural issues
- evolution in what counts as "bias"
- shifting epistemic virtues

Requires broad governance input.

F.5 Calibration Algorithms: Formal Specification

Calibration is computed through multi-layer backpropagation over:

1. **Evidence Vectors**
2. **Human Feedback Logs**
3. **Variance Maps**
4. **φ-Deviation Metrics**

Updates are applied through:

- **gradient adjustments**
- **heuristic re-ranking**
- **reweighting fusion parameters**
- **drift-corrective transformation matrices**

F.5.1 Evidence Reweighting Function

$$w_{new} = w_{old} + \eta \cdot (\text{support} - \text{contradiction})$$

F.5.2 φ-Window Adaptive Tuning

$$\varphi_{new}^{-1} = \varphi^{-1} + \delta \cdot D_{\varphi}$$

F.5.3 Ensemble Sensitivity Adjustment

$$EV_{thr,new} = EV_{thr,old} - \gamma \cdot (EV_{avg} - EV_{target})$$

These mechanisms ensure stability over time.

F.6 Human Oversight in Calibration

Calibration is not autonomous — human actors intervene:

F.6.1 Users

Identify:

- local errors
- emerging ambiguities
- contradictory sources

Users act as the system's epistemic "immune sensors."

F.6.2 Quaestors

Conduct:

- drift investigations
- cross-domain audits
- model version comparisons
- hallucination template detection

Quaestors help detect structural distortions.

F.6.3 RDEs

Oversee:

- scholarly validity
- methodological correctness
- domain rule adjustments
- periodic drift assessments
- structural recalibration proposals

Experts ensure that calibration aligns with the intellectual standards of each discipline.

F.7 System-Wide Learning

The calibration engine ensures that the system **learns** over time:

- corrected claims propagate globally
- domain heuristics evolve
- retrieval patterns refine
- contradictions diminish
- hallucination patterns weaken
- proportionality stabilizes
- ensembles become less chaotic

System-wide learning produces:

- **epistemic smoothness**
- **interpretive refinement**
- **historical accuracy**
- **scientific coherence**

over months and years.

F.8 Constitutional Constraints on Calibration

Calibration cannot:

- override the People's Veto
- negate RDE rulings without due process
- undermine domain autonomy
- bias one ideology
- silence minority scholarly traditions
- operate without transparency

Every calibration is logged and auditable.

F.9 Summary

Appendix F formalized:

- calibration types
- drift detection
- data inputs
- structural adaptation
- human oversight
- constitutional constraints

Calibration makes the epistemic republic a **living system**, capable of evolving, correcting itself, and maintaining intellectual integrity across time, evidence, domains, and cultures.