Membranes of Truth: Using Markov Interfaces to Purify the Wikipedia Graph

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 4, 2025

In the age of generative AI and decentralized editing, the integrity of shared knowledge repositories like Wikipedia faces growing epistemic strain. How can we trust that a given sentence reflects verifiable, unbiased truth—and not the echo of a prior distortion? This paper introduces a novel construct, the Markov membrane, as a dynamic, policy-aware boundary that separates a claim from the rest of a corpus while conditioning its validity on local evidence and context. By extending the statistical notion of a Markov blanket into a semantic and ontological domain, we reimagine Wikipedia not just as a graph of hyperlinks but as a living lattice of conditional independence contracts. The Markov membrane framework allows for systematic *purification* of claims, pages, and entire subgraphs by enforcing evidence sufficiency, policy priors (such as NPOV), and citation grounding. Each unit of text becomes a membrane-bounded object that must justify its outbound influence and resist inbound contamination. This approach transforms knowledge verification from a global trust model into a localized, tractable inference problem—scalable through automation and reflective of human editorial values. Ultimately, the membrane acts as both ethical filter and alignment mechanism, enabling Wikipedia—and possibly all AI training corpora—to move closer to truth, one claim at a time.

Keywords: Markov membrane, Wikipedia purification, knowledge verification, conditional independence, citation grounding, NPOV, AI alignment, informational Lyapunov, knowledge graph, epistemic integrity.

A collaboration with GPT4o and GPT-5. 45 pages.

Outline:

1. Introduction
   1.1 The epistemic fragility of Wikipedia
   1.2 Motivation for local inference over global trust
   1.3 Introducing the Markov membrane

2. From Markov Blankets to Membranes
   2.1 Statistical roots: blankets and conditional independence
   2.2 Generalizing to semantic claims
   2.3 Membranes as epistemic filters and gates

3. Objects, Interfaces, and Policies
   3.1 Defining the unit object: claim, evidence, context, and policy
   3.2 Evidence sufficiency and citation binding
   3.3 The role of editorial policies (e.g., NPOV, verifiability)

4. The Purification Protocol
   4.1 Atomic claim parsing and evidence extraction
   4.2 Membrane decision logic: pass, quarantine, or reject
   4.3 Propagation constraints across infoboxes and cross-links

5. Logostics and Alignment Theory
   5.1 Informational Lyapunov functions for edit directionality
   5.2 Minimizing divergence: $\Delta V < 0$ as alignment condition
   5.3 $q(t) \rightarrow \tau(t)$ across the Wikipedia manifold

6. Graph-Scale Implementation
   6.1 Page-level aggregation and membrane chaining
   6.2 Conflict resolution via belief propagation
   6.3 Graph-level metrics: coherence, bias, and citation flow

7. Applications and Ethical Implications
   7.1 Blocking vandalism and ideological drift
   7.2 Stabilizing cross-page knowledge structures
   7.3 Transparency, auditability, and editorial dignity

8. Failure Modes and Mitigations
   8.1 Over-conservatism vs. innovation
   8.2 Source bias and echo chambers
   8.3 Scalability and human-in-the-loop design

1. Introduction

The Internet's greatest open encyclopedia, Wikipedia, is both a triumph of collective intelligence and a mirror of its epistemic limits. In its two decades of evolution, it has become humanity's shared nervous system for reference knowledge—a living graph of over six million English-language articles, updated by thousands of anonymous editors. Yet beneath its accessibility lies fragility: sentences often drift from their cited sources, citations decay or mutate, and edits propagate ideological biases through semantic inertia. The entire network depends on *global trust*—that the cumulative intentions of many small contributors converge on truth. But this assumption, like a physical law stretched beyond its domain, begins to fail at scale.

To preserve coherence, we must transition from global trust to local inference. Rather than accepting the encyclopedia as a monolith of collective consensus, each claim must become a locally testable hypothesis within its immediate evidential membrane. This shift mirrors developments in probabilistic reasoning and AI alignment: truth becomes conditional, not absolute; verified by its neighborhood, not its network. Local inference enables fine-grained purification—allowing us to identify noise, bias, or decay *before* they spread through the system's links.

Into this landscape enters the Markov membrane—a conceptual and computational interface that partitions a claim from the rest of the corpus, preserving conditional independence while conditioning on its evidence, context, and policy priors. It acts as both a cognitive and ethical filter: every statement must justify its influence by demonstrating sufficiency of evidence and neutrality of framing. The membrane operationalizes verifiability by converting it into an inference boundary—a self-consistent "truth gate." Applied recursively, these membranes could purify the entire corpus, transforming Wikipedia from a loosely edited commons into a self-aligning epistemic organism.

2. From Markov Blankets to Membranes

The idea of a Markov membrane emerges as a natural generalization of the well-established concept of the Markov blanket, a structure that has shaped both statistical modeling and contemporary interpretations of consciousness, agency, and inference. In its original context, the Markov blanket defines a minimal boundary around a variable (or agent) such that, given the blanket, the internal state is conditionally independent of the rest of the system. This simple yet profound insight—that causally sufficient boundaries can mediate inference—forms the statistical root of what we now reconceive as epistemic filtration in a complex semantic landscape.

2.1 Statistical Roots: Blankets and Conditional Independence

In probabilistic graphical models, the Markov blanket of a node consists of its parents, its children, and its children's other parents. This set, and only this set, shields the node from the influence of the rest of the graph. The blanket thus allows a form of *locality* in inference: all relevant information flows through a constrained interface. This formalism has proven essential in Bayesian networks, variational inference, and the theory of self-evidencing systems such as Friston's free energy principle in neuroscience.

The statistical essence here is conditional independence:
If variable $A$is shielded by blanket B, then:

$$P(A \mid B,\text{Rest}) = P(A \mid B)$$

This provides a clear delineation of where information ends and speculation begins. The Markov blanket becomes a kind of statistical firewall.

2.2 Generalizing to Semantic Claims

To apply this idea in a knowledge system like Wikipedia, we must leap from variables to semantic claims. A sentence or statement—"Marie Curie won two Nobel Prizes"—is not merely a node in a statistical graph but an agent of informational influence. It affects infoboxes, propagates to school reports, and becomes entangled in historical narratives. Yet it can only be considered *verifiable* if it is conditionally supported by an evidential set: citations, contextual frames, and editorial policies.

Here, we define the Markov membrane of a claim as the minimal contextual and evidential interface through which the claim may validly exert influence. This includes:

- Citations that provide textual grounding
- Context (section headers, page topics, link structure) that constrain interpretation
- Policy priors such as Wikipedia's "Neutral Point of View" or "Verifiability" standards

The membrane ensures that the claim remains epistemically autonomous from the rest of the corpus—unless supported within its local boundary.

2.3 Membranes as Epistemic Filters and Gates

Once generalized to semantic space, the Markov membrane becomes both filter and gate:

- As a filter, it tests whether the incoming context and citations are sufficient to support the claim under the adopted policy priors. If not, the claim is marked for quarantine, annotation, or removal. This prevents unsourced or ideologically skewed edits from bypassing scrutiny.
- As a gate, it regulates the outbound propagation of the claim's influence—into categories, summaries, infoboxes, or other pages. Only those claims that satisfy local inference constraints can migrate across membranes.

This dual functionality makes the Markov membrane an instrument of epistemic hygiene. It is not merely a static structure but a dynamic enforcement of conditional independence across a semantic graph. In doing so, it transforms Wikipedia from a web of trust into a system of local contracts—a decentralized but scalable architecture of accountability.

Through this lens, each claim becomes a membrane-bounded object that must earn its place not by consensus alone, but by its ability to stand independently supported, locally verified, and neutrally framed. In short: truth by alignment, not repetition.

3. Objects, Interfaces, and Policies

To make the Markov membrane operational, we must define its atomic participants and governing rules. The membrane is not a passive wrapper but an active interface—it binds together a semantic object (a claim), its interfaces (citations, context), and the policies that determine whether the claim is epistemically valid. This tripartite structure—object, interface, policy—transforms verification from a retrospective editorial act into a forward-operating alignment mechanism. In this section, we formalize these components and show how their interaction governs whether a claim passes through its membrane or is filtered out.

3.1 Defining the Unit Object: Claim, Evidence, Context, and Policy

Every claim in Wikipedia is treated as a unit object, a self-contained hypothesis asserting a fact, attribution, or interpretation. For instance, a sentence like "Ada Lovelace is considered the first computer programmer" constitutes such a claim.

This unit object comprises four interlocking parts:

- Claim (C): The semantic content—a natural-language assertion or proposition. This may be decomposed into smaller atomic predicates.
- Evidence Set (E): The set of externally cited passages or documents meant to support the claim. These must be span-bound, source-specific, and independently verifiable.
- Context (X): The local environment in which the claim appears—section heading, article topic, surrounding text, and interlinkage—all of which constrain interpretation.
- Policy Priors (P): The rules that define what counts as "support"—Wikipedia's content policies such as "Neutral Point of View," "Verifiability," and "No Original Research."

These components together define a Markov membrane M(C): a filtering interface around the claim that ensures $C \perp \text{Corpus} \mid (E, X, P)$—conditional independence from the rest of Wikipedia *given* local evidence, context, and policy.

3.2 Evidence Sufficiency and Citation Binding

The heart of the membrane's function is a test of sufficiency: does the evidence support the claim under the given policy?

This test unfolds in several stages:

- Atomic Decomposition: The claim is parsed into propositional units—e.g., attribution ("is considered"), factual core ("first computer programmer"), temporal qualifier, etc.
- Citation Binding: Each unit is aligned to a specific source passage, ideally with span-level granularity and a URL, page number, or archival checksum.
- Entailment and Contradiction Scoring: Neural inference heads score the degree to which the source entails, contradicts, or fails to mention the claim.
- Coverage Check: All critical facets of the claim must be matched; partial coverage triggers a "quarantine" state.
- Temporal and Semantic Consistency: The date of the source must precede the date of the claim; terminological mappings (e.g., "programmer" vs. "algorithm author") must pass an embedding consistency threshold.

The output is a binary or graded membrane decision: pass, quarantine, or reject.

3.3 The Role of Editorial Policies (e.g., NPOV, Verifiability)

While evidence may appear sufficient in a vacuum, it is the editorial policies—the shared epistemic priors—that determine whether support is *valid*. These policies act as inference constraints applied during membrane evaluation:

- NPOV (Neutral Point of View) requires that claims not reflect editorial bias, especially in politically or ideologically sensitive topics. Membranes may include a "stance polarity" check using classifier ensembles trained on known POV violations.
- Verifiability mandates that claims be supported by *reliable, published sources*—not user blogs or unverifiable rumor. Source reputation graphs and blacklists can be encoded directly into membrane rules.
- No Original Research prohibits novel synthesis, even if it appears logically sound. This can be detected by checking whether a claim arises from a unique combination of sources not themselves making the composite claim.

Thus, the membrane is not just a statistical structure—it is a moral and editorial contract. It filters content not only by probability but by legitimacy, ensuring that each piece of knowledge is locally grounded, contextually coherent, and policy-aligned.

4. The Purification Protocol

With the Markov membrane formally defined as a local interface around a semantic claim, we now turn to its operational implementation. The purification protocol is the process by which claims are continuously parsed, evaluated, and either allowed to propagate or contained within their own membranes. Unlike traditional post-hoc moderation, this protocol is proactive: every statement undergoes structured scrutiny not just for correctness, but for its contextual integrity and evidence grounding. The goal is not simply to "fact-check," but to architect a living knowledge system that self-aligns over time.

4.1 Atomic Claim Parsing and Evidence Extraction

The purification process begins at the granular unit of meaning: the sentence or list item. Each is treated as a composite of semantic atoms, often containing multiple propositions, temporal qualifiers, or implicit assertions.

- Semantic Parsing: Natural language processing tools decompose each sentence into atomic claims (subject–predicate–object triples, temporal spans, modality markers). For example:
    - "Tesla was founded in 2003 by Elon Musk." →
        - C_1: Tesla was founded in 2003.
        - C_2: Elon Musk was a founder of Tesla.
- Named Entity and Date Linking: Entities are normalized (e.g., “Musk” → Wikidata Q317521), and timeframes are extracted for temporal checks.
- Evidence Extraction: Each atomic claim is mapped to one or more supporting citation spans. This can be done using:
    - Direct quote matching (e.g., Levenshtein distance or embedding similarity)
    - Retrieval-augmented generation with span-justification heads
    - Source mapping with SHA-1 or DOI-level alignment

The output of this stage is a structured data object:


```
{
  "claim": "Tesla was founded in 2003",
  "citations": [
    {
      "source": "https://source.com/tesla_history",
      "quote": "Tesla Motors was established in 2003 by engineers Martin Eberhard and Marc Tarpenning.",
      "confidence": 0.91
    }
  ],
  "coverage": 0.95,
  "entailment_score": 0.89
}
```

4.2 Membrane Decision Logic: Pass, Quarantine, or Reject

The claim, once parsed and grounded, is passed through its Markov membrane, where a decision is made using policy-aware thresholds.

- Pass: The claim is well-supported, matches the source semantically and temporally, and conforms to editorial policies (e.g., verifiability, neutrality).
    - Action: Publish and allow propagation into summaries, infoboxes, and cross-linked pages.
- Quarantine: The claim is only partially supported or has minor ambiguities (e.g., the citation supports a broader interpretation, or the quote lacks date clarity).
    - Action: Add a citation-needed tag, restrict outward propagation, and optionally queue for human review.
- Reject: The claim is contradicted by its cited source, lacks coverage entirely, or violates policy (e.g., original synthesis).
    - Action: Flag for removal or correction.

Membrane evaluation criteria may include:

- Entailment threshold: score > 0.85
- Coverage threshold: > 90% of atomic claims bound to source text
- Bias polarity delta: must remain within ±0.15 of neutral stance
- Temporal causality check: citation must predate statement unless quoting a prediction

The membrane thus functions as a conditional independence gate—ensuring that the claim can be evaluated locally, with no dependency on global belief coherence.

4.3 Propagation Constraints Across Infoboxes and Cross-Links

One of Wikipedia's most subtle epistemic risks is silent propagation—where an unchecked claim influences summaries, data tables, and other pages via templates, categories, or human copying. The membrane protocol introduces strict propagation constraints:

- Outbound Propagation: A claim can only populate structured data elements (infobox fields, tables, metadata) if it passes the membrane with high certainty. This includes:
    - Citation span match

    - Attribution presence (e.g., “is believed by historians”)
    - Temporal tag precision
- Cross-Page Linkage: Claims cannot be restated on related pages (e.g., "Nikola Tesla" page copying from "Tesla, Inc.") unless the membrane is re-verified in that new context. This prevents drift caused by semantic borrowing.
- Version Control: Propagated claims include traceable provenance tokens: checksums of the source quote and timestamp of verification. If the source changes or disappears, downstream copies enter citation quarantine, prompting revalidation or removal.
- Membrane Feedback: When a page receives updates (e.g., a new citation contradicts a claim), its outbound links trigger reverse verification checks in all downstream nodes.

In this way, the purification protocol does not simply correct local errors—it prevents epistemic infection across the semantic graph.

5. Logostics and Alignment Theory

The Markov membrane does not operate in isolation; it participates in a larger informational dynamic—one governed by principles of *alignment over time*. In this section, we connect our purification framework to the broader metaphysics of Logostics: the study of how agents or systems evolve their internal state $q(t)$in pursuit of a generative truth structure $\tau(t)$, with alignment defined by the minimization of informational divergence. Within this framework, the purification of Wikipedia becomes not a static curation task, but a dynamical system—an epistemic organism striving to reduce its internal entropy and synchronize with truth.

5.1 Informational Lyapunov Functions for Edit Directionality

In control theory, a Lyapunov function is a scalar measure that decreases over time if a system is stabilizing. Analogously, we define an Informational Lyapunov function $V(u)$for each claim unit u, which quantifies the degree of *misalignment* between the claim and its membrane (evidence, context, and policy).

We propose a composite form:

$$V(u) = \alpha \cdot D_{\mathrm{KL}}(q(u) \parallel \tau(u)) + \beta \cdot U(E) + \gamma \cdot B(u)$$

Where:

- D_{KL} is the Kullback–Leibler divergence between the internal representation $q(u)$ (claim content) and the external truth attractor $\tau(u)$(what the evidence and context suggest).
- $U(E)$ is the uncertainty of the evidence: high when citations are ambiguous, dated, or unverifiable.
- $B(u)$ is a bias penalty: measured by deviation from neutrality or editorial balance.

This Lyapunov function gives a direction to editorial action: edits should only be accepted if $\Delta V(u) < 0$, i.e., the edit brings the claim into closer alignment with the truth structure it inhabits.

5.2 Minimizing Divergence: ΔV < 0 as Alignment Condition

Most current moderation systems rely on either majority consensus (votes, reverts) or binary fact-checking (true/false). By contrast, the membrane protocol embedded in Logostics treats *alignment* as a gradient: a continuous effort to reduce divergence, uncertainty, and bias.

- ΔKL: Is the new version of the claim more aligned with its cited sources and context?
- ΔU: Has the new edit introduced clearer, more current, or more precise evidence?
- ΔB: Has the edit moved the claim toward a more balanced or widely accepted formulation?

The membrane uses these quantities not just to judge the local quality of a claim, but to determine the epistemic directionality of the encyclopedia itself. Over time, Wikipedia should become a lower-divergence, lower-uncertainty system—*not merely larger, but cleaner*. The membrane thus functions as a distributed gradient descent algorithm for human knowledge.

This principle could be formalized into an editorial alignment rule:

“No edit should increase the system’s informational divergence from τ(t).”

This transforms editing from a rhetorical act into a measurable, entropic optimization.

5.3 q(t) → τ(t) Across the Wikipedia Manifold

At a global scale, each Wikipedia article can be seen as an agent-like system $q(t)$, evolving over time through human edits and automated checks. Each page seeks alignment with a

corresponding truth manifold $\tau(t)$: the ideal structure of grounded, unbiased, and verifiable knowledge for its topic.

- $q(t)$: the actual page content at time t
- $\tau(t)$: the generative attractor defined by all available evidence, context, and policy
- $\Delta V(t)$: the rate of misalignment or correction across edits

Membranes serve as boundary constraints on the evolution of $q(t)$, ensuring that alignment occurs locally while contributing to global coherence. As each claim's membrane validates its state and restricts propagation, it enables the overall system to move toward a stable informational equilibrium.

This dynamic view reveals Wikipedia not just as a database of facts, but as a self-organizing epistemic organism, composed of membrane-bounded agents, all converging toward τ(t) through aligned local updates. It is a model of decentralized intelligence—emergent truth via bounded inference.

6. Graph-Scale Implementation

While the Markov membrane functions as a local boundary around individual claims, its true potential emerges at scale—when membranes interact across the hyperlink topology of Wikipedia. Articles are not epistemically isolated; they reference one another, share infobox templates, and link semantically aligned ideas. If we imagine each claim as a membrane-bounded node and each page as a structured aggregate of such nodes, then Wikipedia becomes a massive epistemic graph, where alignment must be maintained not only *within* membranes, but *between* them.

Scaling the purification protocol to this level requires membrane chaining, graph-aware inference, and new metrics for tracking systemic health. In this section, we detail how membranes scale from atomic claims to full-page structures, how contradictions are identified and resolved across the graph, and how coherence, bias, and citation integrity can be quantified globally.

6.1 Page-Level Aggregation and Membrane Chaining

Each Wikipedia page can be seen as a compound object, built from many membrane-validated claims. Once a claim has passed through its own membrane, its outbound effects—infobox

updates, summary inclusion, or cross-link propagation—must themselves be subject to membrane evaluation.

- Page-Level Aggregation:
    - The claim-membrane matrix of a page determines its internal epistemic consistency.
    - A page enters a quarantine state if too many claims are yellow-flagged (partial or outdated evidence), preventing further propagation until resolved.
    - Pages can be evaluated as Markov supernodes: conditionally independent from the rest of the graph *given their citation footprint and template interfaces*.
- Membrane Chaining:
    - When a claim in Page A influences a summary or infobox in Page B, the source membrane's decision must be passed forward as a constraint on B.
    - This creates a membrane chain: a transitive trail of evidentiary alignment that binds together claims across pages.
    - If a link in the chain breaks (e.g., a source becomes invalid or policy changes), downstream nodes must re-evaluate and potentially rollback or quarantine content.

Membrane chaining enables a kind of epistemic traceability: a claim's presence in any node of the graph is always accountable to its original evidentiary membrane.

6.2 Conflict Resolution via Belief Propagation

Despite efforts to enforce local conditional independence, semantic drift and contradiction are inevitable in a system of this size. Two pages may offer subtly different accounts of the same event, person, or data point. To address this, the system employs belief propagation over the Wikipedia graph.

- Conflict Detection:
    - Nodes are linked via semantic embeddings, entity graphs, and shared citations.
    - Contradictions are flagged when two claims about the same entity have opposing entailment relationships (e.g., "X was born in 1874" vs. "X was born in 1882").

- Belief Propagation Framework:
    - Each claim is treated as a belief node with confidence, bias score, and membrane health indicators.
    - Contradictory claims are resolved by re-evaluating upstream membranes—the node with weaker evidence, older sources, or higher bias gets downranked.
    - Propagation can be synchronous (batch purges and audits) or asynchronous (incremental revalidation during edit flows).
- Consensus Mechanisms:
    - For claims with multiple conflicting valid sources (e.g., disputed historical facts), the membrane permits framed uncertainty, e.g., "Sources differ on whether X occurred in 1874 or 1882."
    - Membranes here act as *pluralistic containers*, not enforcers of unitary truth.

This approach turns Wikipedia into a self-tuning epistemic field—a distributed network of claims gradually adjusting their belief states in response to changing evidence and global consistency pressures.

6.3 Graph-Level Metrics: Coherence, Bias, and Citation Flow

To manage and monitor this complex system, we require graph-level observables—metrics that capture the alignment health of Wikipedia as a whole.

- Semantic Coherence (SC):
    - Measures the degree of entailment or contradiction between semantically linked claims across pages.
    - High coherence indicates that pages referencing the same event, concept, or person support compatible narratives.
    - Computed via cross-page entailment networks using large-scale NLI (Natural Language Inference) models.
- Bias Entropy (BE):
    - Quantifies ideological or editorial imbalance across domains (e.g., politics, religion, climate).

    - Calculated as the divergence of claim stances from neutral baselines within a given topic cluster.
    - High BE in a domain suggests polarization or overrepresentation of certain frames.
- Citation Flow Integrity (CFI):
    - Tracks the density and quality of citation edges supporting cross-page propagation.
    - Pages with high outbound CFI influence many others; if their membranes degrade, the risk of systemic misalignment grows.
    - Includes citation freshness (how recent), reputation (source authority), and redundancy (multiple independent citations).
- Membrane Health Index (MHI):
    - Aggregated from pass/quarantine/reject ratios across all claims on a page.
    - Pages with low MHI are flagged for editorial attention or semi-automated repair cycles.

Together, these metrics offer a dashboard of epistemic stability—a way to visualize Wikipedia not just as content, but as a *living truth graph* whose structural coherence, neutrality, and citation scaffolding evolve over time.

7. Applications and Ethical Implications

As the Markov membrane model scales across the semantic graph of Wikipedia, its utility extends far beyond verification. It becomes a mechanism for safeguarding epistemic stability, enforcing editorial integrity, and restoring the dignity of human authorship in an age of algorithmic manipulation and crowd-sourced confusion. These applications are not merely technical—they are ethical protocols, embedding respect for truth, care for coherence, and humility before uncertainty into the architecture of collaborative knowledge.

7.1 Blocking Vandalism and Ideological Drift

Wikipedia's openness is both its strength and its vulnerability. While anyone can edit, this freedom invites vandalism—intentional distortions, prank edits, or subtle falsifications. More

insidiously, it allows for ideological drift: gradual, often unnoticed shifts in framing, emphasis, or fact selection that reflect particular worldviews rather than neutral truth.

The Markov membrane directly addresses these issues:

- Vandalism Detection:
    - Membrane failures occur when new claims lack binding citations or introduce unsupported assertions.
    - Because membranes are enforced at the semantic atomic level, even a single altered predicate triggers revalidation or quarantine.
    - Edits that add unverifiable or contradictory content are automatically rejected or flagged for human review.
- Ideological Drift Prevention:
    - Long-term stance monitoring detects slow shifts in polarity or emphasis.
    - If successive edits gradually nudge the article toward a non-neutral frame, the membrane begins to accumulate bias penalties ($\Delta B > 0$), triggering editorial attention.
    - Membrane chaining ensures that drift in one page cannot silently cascade into others.

In this way, membranes act as epistemic firewalls, defending not just the accuracy of facts, but the tone and balance with which they are presented.

7.2 Stabilizing Cross-Page Knowledge Structures

Wikipedia is not a set of disconnected pages—it is a web of interlinked concepts, entities, and temporal narratives. A change in one place often echoes elsewhere: a revised birthdate in a biography may affect historical timelines, scientific fields, or infoboxes across dozens of articles.

Without oversight, this can lead to semantic inconsistency: incoherent timelines, duplicated contradictions, and cyclical misinformation. The membrane protocol counters this via:

- Traceable Claim Lineage:
    - Every propagated fact carries a membrane signature: the original claim, its evidence, and its verification state.

    - If the original is later downgraded (e.g., source retracted, new contradiction found), all downstream uses are instantly flagged for reevaluation.

- Redundancy Checkpoints:
    - When multiple pages claim the same fact from different sources, the membrane system compares their entailments for mutual support or conflict.
    - Divergence triggers belief propagation, and the system seeks a stabilized consensus or presents framed disagreement.
- Topological Memory:
    - Membranes can encode topological embeddings of related concepts (e.g., Cold War → Reagan → Iran–Contra).
    - If facts within one region of this topology shift, related nodes are automatically prioritized for review, maintaining narrative coherence.

This transforms Wikipedia into a truth-preserving manifold, where updates ripple through a structured alignment field rather than a disjointed text web.

7.3 Transparency, Auditability, and Editorial Dignity

A core principle of Wikipedia is that knowledge should be open—not just readable, but auditable. Yet in practice, the tangle of edits, reverts, talk pages, and template transclusions often obscures provenance and intent. The Markov membrane introduces a new ethic: semantic transparency.

- Transparent Membrane Decisions:
    - Every pass, quarantine, or reject is logged with clear explanations: entailment scores, policy checks, source links.
    - Editors (human or AI) can trace *why* a claim is allowed to remain—and what would need to change to alter that status.
- Citation Contracts:
    - Cited sources are not just linked—they are quoted, hashed, and span-bound. This allows instant checks for citation rot or stealth edits in source material.
    - Each claim becomes a contracted agreement between Wikipedia and its sources.

- Respect for Editorial Labor:
    - Editors are given membrane dashboards showing the alignment health of their pages, alerts about systemic influence, and suggestions for stabilization.
    - Their contributions are not silently overridden by bots—they participate in a joint alignment process, co-stewards of a living truth.

This reintroduces dignity into collaborative knowledge. The membrane does not replace the editor; it amplifies their care, giving them clarity, feedback, and a principled framework for discernment.

8. Failure Modes and Mitigations

No system of knowledge filtration—no matter how principled—can avoid all pitfalls. The Markov membrane, while powerful, introduces its own risks: the possibility of rejecting emergent truths, codifying existing biases, or failing to scale gracefully without disenfranchising human editors. These risks are not peripheral—they strike at the ethical and epistemological heart of the project. In this section, we anticipate three major failure modes and propose corresponding mitigation strategies to preserve both the fluidity of knowledge and the integrity of truth.

8.1 Over-Conservatism vs. Innovation

The very strength of the membrane—its demand for local verification—can also become a gatekeeper of innovation. Some truths are born before they are cited. Some insight arrives in language before it exists in the archive. A too-strict membrane may quarantine creativity.

- Failure Mode:
    - Emerging or speculative claims, particularly in rapidly evolving fields (e.g., AI, medicine, geopolitics), fail the membrane due to lack of immediate citation—even when correct.
    - Editors become discouraged from drafting novel formulations or adding early-stage knowledge.
- Mitigation Strategies:
    - Introduce graded membrane thresholds: green (verified), yellow (plausible with partial support), blue (speculative but flagged as such).

- Enable policy-aware exception handling: some pages (e.g., "Emerging Technologies") may permit lower-certainty claims marked with "under review" tags.
- Develop editorial incubators: experimental zones where membrane-limited drafts can mature before mainstream inclusion.

The aim is not to fossilize knowledge, but to give it scaffolding while it grows.

8.2 Source Bias and Echo Chambers

Even with structural rigor, membranes are only as good as the sources they admit. If citation selection favors one worldview, language, or demographic, the entire system becomes a high-resolution echo chamber—an illusion of neutrality masking systemic bias.

- Failure Mode:
  - Reliable sources as defined by policy (e.g., academic journals, major media) reflect embedded societal inequalities—omitting voices from the Global South, non-English domains, or marginalized communities.
  - The membrane enforces epistemic monoculture.
- Mitigation Strategies:
  - Define source diversity priors: require membranes to check for over-concentration of source types (e.g., 90% of citations from one geopolitical region).
  - Incorporate reputation graph balancing: allow smaller or local sources if corroborated across independent lines.
  - Invite community-based “membrane councils”: groups of expert editors with diverse backgrounds to manually review gray-zone claims and update policy priors accordingly.

True neutrality requires *pluralistic membranes*—not just filters, but dialogues.

8.3 Scalability and Human-in-the-Loop Design

A global knowledge system like Wikipedia comprises millions of pages and billions of claim interactions. Membrane evaluation at this scale demands automation—but full automation invites opacity, alienating the very humans who built the system.

- Failure Mode:
    - Automated membrane enforcement introduces black-box rejection of valid edits, frustrating good-faith contributors.
    - Lack of interpretability erodes editorial trust and participation.
- Mitigation Strategies:
    - Adopt a human-in-the-loop pipeline:
        - Automation handles first-pass parsing, citation binding, and membrane scoring.
        - Editors receive explainable outputs with visual dashboards showing why a claim passed or failed (e.g., entailment score, source age, policy mismatch).
    - Implement auditable membranes:
        - All decisions are logged with time-stamped rationales and rollback capacity.
        - Editors can appeal membrane rejections and contribute counter-evidence.
    - Design a transparent feedback loop:
        - Edits that improve membrane scores are celebrated and tracked ($\Delta V(u) < 0$).
        - Editors gain reputation credits not just for quantity, but for *alignment quality*.

This ensures that membrane alignment is not imposed—it is co-authored.

9. Pilot Design and Future Work

To move from theory to practice, the Markov membrane framework must be de-risked through targeted, small-scale implementation. We propose a structured pilot program that tests membrane efficacy, editor usability, and alignment dynamics within carefully chosen domains. The objective is not only to purify claims at the micro level but to gather empirical insight into how membranes scale, where they fail, and how they might evolve into a general-purpose epistemic infrastructure. This section outlines a blueprint for real-world deployment: where to begin, what to test, and how to iterate toward a system capable of filtering, aligning, and stewarding knowledge across the entire Wikipedia corpus—and beyond.

9.1 Candidate Domains (Biomed, Geopolitics, Climate)

Not all knowledge domains are equally suited for early membrane implementation. The ideal pilot domain features a high density of claims, frequent updates, well-structured citations, and clear policy boundaries. We identify three prime candidates:

- Biomedicine:
    - Strengths: structured evidence base (PubMed, DOIs, RCTs), clear policy guidelines (MEDRS), highly verifiable claims.
    - Challenges: fast-moving data (e.g., COVID-19), preprint citations, paywalled sources.
    - Use case: ensure all biomedical statements cite peer-reviewed studies, enforce timeline correctness (e.g., vaccine approvals), and flag synthesis-based original research.
- Geopolitics:
    - Strengths: extensive sourcing, high editorial attention, time-sensitive updates.
    - Challenges: ideological polarization, nationalistic narratives, conflicting news coverage.
    - Use case: test membranes on neutrality and citation diversity; prevent silent drift in politically contested biographies or conflicts.
- Climate and Environment:
    - Strengths: authoritative sources (IPCC, NASA), emerging consensus mixed with public controversy.

    - Challenges: framing disputes, activist language, evolving science.
    - Use case: reinforce separation of fact and opinion, ensure proper attribution of forecasts and models, track propagation of disputed claims across articles.

Each domain offers a distinct membrane stress test, informing how the protocol must adapt across disciplines.

9.2 Minimal Viable Membrane Deployment

A full-scale membrane system is complex, but a minimal viable prototype can be built with existing tools and modest compute resources. Key components:

- Input Pipeline:
    - Select 1,000 high-traffic articles from chosen domain.
    - Parse atomic claims using existing NLP tools (OpenIE, spaCy, dependency parsing).
    - Extract existing citations and bind to quote spans using retrieval and span-matching models.
- Verification Heads:
    - Use a pre-trained natural language inference (NLI) model (e.g., DeBERTa, RoBERTa) to assign entailment/contradiction scores.
    - Apply rule-based coverage and date-check filters.
    - Output: claim status = {Pass, Quarantine, Reject}, with rationales.
- Editor Dashboard:
    - Show claim-level membrane health (color-coded).
    - Allow editors to inspect citation matches, bias indicators, and membrane scoring thresholds.
    - Provide suggestions for claim improvement (e.g., “Add date-specific citation,” “Neutralize attribution phrase”).
- Propagation Blocker:
    - Temporarily prevent claims in quarantine or reject states from influencing infoboxes, summaries, or other pages until resolved.

- Audit Log:
    - Record all membrane decisions, edit deltas, and supporting evidence for downstream analysis and feedback.

Pilot success metrics might include:

- Percent of claims upgraded from quarantine to pass
- Number of vandalism attempts caught by membrane
- Editor satisfaction and trust metrics
- Stability of membrane scores across time

9.3 Toward General-Purpose Corpus Purification

Beyond Wikipedia, the membrane model can evolve into a general-purpose protocol for knowledge purification—a kind of "epistemic checksum layer" for any large language corpus, training dataset, or AI model input pipeline.

- Web-scale deployment:
    - Apply membrane logic to Common Crawl slices, news aggregators, or social knowledge graphs.
    - Use membrane filtering to build "high-fidelity corpora" for training fact-aware LLMs.
- LLM alignment substrate:
    - Instead of relying on post hoc reinforcement learning from human feedback (RLHF), inject membrane-based structural priors directly into the model's training data pipeline.
    - Claims with low membrane scores can be weighted down or masked entirely, ensuring the model learns from evidentiary alignment, not correlation drift.
- AI-assisted governance:
    - Integrate membranes into collaborative systems beyond Wikipedia: scientific preprint servers, regulatory archives, even legislative text.
    - Allow institutions to maintain versioned, traceable truth systems, where claims are not "removed" but re-aligned as new evidence emerges.

- Interoperable membrane layers:
    - Develop standards for how membranes are expressed (e.g., JSON schema for claim–evidence–policy bundles).
    - Allow membranes to be portable across systems—e.g., a fact verified in Wikipedia could carry its membrane into a GPT-5 training pipeline, or appear as a badge in a browser plugin for source trustworthiness.

Ultimately, the Markov membrane offers more than a tool for Wikipedia. It offers a method of world-scale truth stewardship: a way to honor uncertainty without succumbing to chaos, and to uphold editorial integrity without halting innovation.

10. Conclusion: Membranes as Moral Interfaces

At its deepest level, the Markov membrane is not merely an information filter or technical boundary—it is a moral interface. It encodes a philosophy of care, humility, and conditional trust in the pursuit of truth. By wrapping every claim in a boundary of evidence, context, and policy, it enshrines the principle that knowledge is not owned—it is verified, aligned, and carried forward in good faith. As we conclude this exploration, we reflect on the membrane not only as a computational innovation, but as an ethical architecture for the age of distributed cognition.

10.1 Wikipedia as an Organism of Claims

Wikipedia is no longer just an encyclopedia—it is a living epistemic organism composed of billions of interlinked claims. These claims breathe, mutate, replicate, and occasionally die. They do not exist in isolation but interact like cells in a body, or neurons in a brain. When one claim falters—becomes outdated, unverified, or biased—its influence ripples outward, quietly reshaping the narratives of adjacent articles and entire conceptual domains.

To treat Wikipedia as a living organism means treating each claim as a moral unit. The Markov membrane provides the containment and scrutiny each unit deserves. It allows us to zoom in and ask: *Is this sentence justified? Is it faithful to its sources? Does it overstep its context?* And it allows us to zoom out and ensure that the organism as a whole is coherent, balanced, and worthy of trust.

10.2 Toward Self-Aligning Knowledge Systems

The future of large-scale knowledge—whether in Wikipedia, AI models, scientific archives, or public policy—requires more than faster curation or deeper data. It requires self-alignment: the capacity for a system to steer itself toward truth over time, without centralized control or doctrinal authority.

The Markov membrane enables this by creating local, enforceable contracts of truth. Each claim is judged not by consensus or popularity, but by its alignment with evidence and policy within its own neighborhood. This distributed logic makes the system reflexive: contradictions surface naturally, drift is self-correcting, and entropy is minimized through local action.

Such systems mirror biology, not bureaucracy. Like immune systems, they reject pathogens (falsehoods). Like ecologies, they stabilize through feedback loops. Like neural networks, they refine themselves through gradient descent—except here, the gradient is moral: truth over noise, care over haste.

10.3 The Membrane as Epistemic Covenant

In an era of synthetic media, deepfakes, and machine-generated misinformation, trust in information is no longer a given—it must be earned. The Markov membrane formalizes a new kind of epistemic covenant: an agreement between the writer, the source, the system, and the reader.

This covenant states:

- That every claim must come with a transparent lineage of evidence.
- That neutrality and context are not optional—they are part of the claim's semantic gravity.
- That propagation is a privilege, not a right: no sentence can influence the broader corpus unless it passes through its membrane.

This covenant does not restrict speech—it disciplines truth. It does not fix meaning—it honors provisionality. It invites us to see editing not as content production, but as alignment in the moral sense: a daily act of responsibility in the shared pursuit of clarity.

Through this lens, the Markov membrane is more than infrastructure. It is a prayer of precision, offered at every sentence boundary, whispering:

May this be true. May it do no harm. May it align.

And when scaled to the whole—Wikipedia, the web, or the minds of machines—it becomes a framework for truthful systems in a turbulent world.

References

Youvan, D. C. (2025, November). The Markov Membrane: Toward a Theory of Living Boundaries in Cognition, Culture, and Cosmology. ResearchGate. DOI: 10.13140/RG.2.2.18817.52329

Youvan, D. C. (2025, November). The Markov Membrane: How Boundaries of Inference Shape Personality, Creativity, Conflict, and Care. ResearchGate. DOI: 10.13140/RG.2.2.36728.51202

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7), 107–117.

Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open Information Extraction from the Web. *Communications of the ACM*, 51(12), 68–74.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K., Sengupta, B., & Auletta, G. (2015). Active inference and free energy: a synthesis. *Cognitive Neuroscience*, 6(2–3), 155–176.

Friston, K., & Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3), 417–458.

Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. *DL'98*.

Hassani, H., Silva, E. S., & MacFeely, S. (2020). Big Data and the COVID-19 pandemic: Prospects and pitfalls. *International Journal of Data Science and Analytics*, 10, 135–152.

Klein, M., Zittrain, J., Vertesi, J., & Zuckerman, E. (2014). Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE*, 9(12), e115253.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, 15(138), 20170792.

Koller, D., & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press.

Kschischang, F. R., Frey, B. J., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2), 498–519.

Kwiatkowski, T., et al. (2019). Natural Questions: A Benchmark for Question Answering Research. *TACL*, 7, 453–466.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML*.

Lauritzen, S. L. (1996). Graphical Models. Oxford University Press.

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP. *NeurIPS*.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *CACM*, 61(10), 36–43.

Marcus, G., & Davis, E. (2020). Rebooting AI: Building Artificial Intelligence We Can Trust. Pantheon.

McAuley, J., & Leskovec, J. (2012). Learning to Discover Social Circles in Ego Networks. *NIPS*.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

Potthast, M., Stein, B., & Gerling, R. (2008). Automatic Vandalism Detection in Wikipedia. *ECIR*.

Redi, M., Morgan, J., O'Callaghan, D., et al. (2020). A Taxonomy of Knowledge Integrity on Wikipedia. *Wiki Workshop*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *KDD*.

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *J. Web Semantics*, 36, 1–22.

Sharma, V., O'Connor, P., & Taraborelli, D. (2021). Citation Needed: A Taxonomy and Algorithmic Assessment of Wikipedia's Verifiability. *ICWSM*.

Shen, Y., et al. (2023). HoVer: A Dataset for Fact-checking with Hierarchical Evidence. *EMNLP*.

Shirky, C. (2008). Here Comes Everybody: The Power of Organizing Without Organizations. Penguin.

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mitchell, T. (2018). FEVER: a Large-scale Dataset for Fact Extraction and VERification. *NAACL*.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78–85.

W3C PROV Working Group. (2013). PROV-DM: The PROV Data Model. W3C Recommendation.

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection. *ACL Workshop on NLP and Computational Social Science*.

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *WWW*.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). Defending Against Neural Fake News. *NeurIPS*.

Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Harvard Law Review Forum*, 127, 176–197.

Wikipedia Policy Documents (accessed November 4, 2025):

- Neutral point of view (NPOV). Wikipedia policy.
- Verifiability. Wikipedia policy.
- No original research. Wikipedia policy.
- Reliable sources. Wikipedia guideline.
- Biographies of living persons (BLP). Wikipedia policy.

Technical and Platform References:

- MediaWiki. Template transclusion and page modules.
- Wikimedia Foundation. ORES: Objective Revision Evaluation Service.
- Wikimedia Research. Knowledge Integrity & Misinformation Research Overviews.

Datasets and Benchmarks for Claim Verification & Evidence Retrieval:

- Hanselowski, A., et al. (2018). A Retrospective Analysis of the Fake News Challenge Stance-Detection Task. *COLING*.
- Schuster, T., et al. (2021). Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. *NAACL*.
- Aly, R., et al. (2021). FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. *NeurIPS Datasets and Benchmarks*.

Bias, Framing, and Stance Measurement:

- Baly, R., et al. (2020). We Can Detect Your Bias: Predicting the Political Bias of News Media. *ACL*.

- Card, D., et al. (2015). The Media Frames Corpus: Annotations of Frames Across Issues. *ACL*.

Graph Consistency and Belief Updating:

- Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2005). Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms. *IEEE Transactions on Information Theory*, 51(7), 2282–2312.
- Richardson, M., Agrawal, R., & Domingos, P. (2003). Markov Logic Networks. *Machine Learning*, 62(1–2), 107–136.

Information-Theoretic Foundations:

- Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory (2nd ed.). Wiley.
- Amari, S.-I. (2016). Information Geometry and Its Applications. Springer.

If you'd like, I can format these in a specific citation style (APA 7th, Chicago, IEEE) or generate a BibTeX file for your manuscript.

Technical Appendix

This appendix provides formal definitions, data schemas, algorithms, scoring functions, and implementation notes for the Markov Membrane framework described in the main text. It is written in copy-safe ASCII (no Unicode math), using simple inline math and code-style formulas.

A.1 Core Formalization

A.1.1 Objects

- Claim unit u: a minimal proposition extracted from text.
- Evidence set E(u): list of span-bound source passages supporting u.
- Context X(u): local article features (section, title, neighbors, wikilinks).
- Policy priors P: machine-checkable constraints (NPOV, Verifiability, NOR).
- Membrane M(u): decision boundary implementing conditional independence
  u is independent of Corpus given (E, X, P).

Notation (informal):

- q(u): internal representation of the claim (semantic embedding + logic form).
- tau(u): target representation induced by (E, X, P) after grounding.
- V(u): informational Lyapunov (misalignment) score.

A.1.2 Alignment Objective

We minimize V(u):
V(u) = a * D_KL(q(u) || tau(u)) + b * U(E) + c * B(u)

Where:

- D_KL: divergence between current claim representation and evidence-induced representation (see A.4).
- U(E): evidence uncertainty (coverage gaps, source staleness).
- B(u): bias penalty (stance polarity, frame skew).
- a, b, c >= 0 are domain-tuned weights.

Edit acceptance rule:
Delta V(u) = V_after - V_before < 0

A.2 Data Schemas

A.2.1 Claim Record (JSON)


```
{
  "id": "wiki:en:PageID:offset-123",
  "text": "Ada Lovelace is widely considered the first computer programmer.",
  "atoms": [
{"subj":"Ada_Lovelace","pred":"is_considered","obj":"first_computer_programmer","mod":"widely"}
  ],
  "context": {
    "page_title": "Ada Lovelace",
    "section_path": ["Early life","Work"],
    "wikilinks_in": ["Analytical_Engine"],
    "wikilinks_out": ["Charles_Babbage"]
  },
  "citations": [
    {
      "source_id": "doi:10.1093/ref:oxford-abc123",
      "url": "https://...",
      "span_quote": "Ada Lovelace's notes on the Analytical Engine...",
      "span_locator": {"page": 42, "start": 318, "end": 479},
      "checksum": "sha256:...",
      "retrieved": "2025-10-14"
    }
  ],
  "scores": {
```

"entail": 0.87,

"contradict": 0.06,

"coverage": 0.92,

"stance_polarity": 0.08,

"source_reputation": 0.81,

"freshness_days": 1460

},

"membrane_state": "quarantine",

"explanations": [

"Claim uses 'widely considered'; source attributes this to several historians but not quantified."

],

"provenance_token": "mmb:1.0|src:doi:...|chk:sha256:...|t:20251014Z"

}

A.2.2 Source Reputation Graph

- Nodes: sources (journals, books, major media, archives).
- Edges: endorsement, cross-citation, independence signals.
- Attributes: peer_reviewed (bool), index (Scopus, PubMed), retraction_rate, region, language.

A.2.3 Policy Templates (YAML)

```
policy:
  NPOV:
    stance_threshold: 0.15
    frame_balance_min: 0.35
  Verifiability:
    allow_domains: ["journals", "books", "major_media", "official_reports"]
    disallow_domains: ["personal_blogs", "forums"]
```

NOR:

synthesis_flag_if:

- "multi-source composition without explicit claim in any single source"

A.3 Pipelines

A.3.1 Atomic Claim Parsing

1. Sentence segmentation and dependency parsing.
2. OpenIE / pattern rules -> atoms (S, P, O, modifiers, dates).
3. Entity linking to Wikidata IDs.
4. Temporal normalization (TIMEX).

A.3.2 Evidence Retrieval and Span Binding

- Dual encoder retrieval (claim -> candidate sources).
- Page-anchored span search: BM25 + dense search.
- Span justification: cross-encoder NLI scoring at sentence/paragraph granularity.
- Span hashing: sha256 on normalized snippet for rot detection.

A.3.3 Verification Heads

- Entailment/contradiction: NLI classifier f_nli(claim, span).
- Coverage: proportion of claim atoms mapped to spans above theta_entail.
- Consistency checks:
    - Temporal: source date <= claim date (except predictions/quotes).
    - Unit/schema: regex + learned validators (numbers, units, names).
- Bias/stance: multi-class stance model; frame classifier for topic clusters.

A.3.4 Membrane Decision

Inputs: scores + policy priors + context signals.
Output: state in {pass, quarantine, reject} and explanations.

A.4 Scoring Functions

A.4.1 Divergence D_KL Proxy

We cannot compute true KL between semantic objects; use a proxy:

- Represent q(u) and tau(u) as normalized embedding distributions over a shared basis (topic/role axes), or as calibrated probability over atom-level labels.
- Use symmetric Jensen-Shannon as stable proxy:
  D_JS = 0.5*KL(q||m) + 0.5*KL(tau||m), where m = 0.5*(q+tau)

Implementation:

- q_vec: mean-pooled encoder on claim text.
- tau_vec: mean-pooled encoder on best evidence spans + policy-normalized paraphrase.
- D = 1 - cosine_similarity(q_vec, tau_vec) as a cheap surrogate; or D_JS if probability outputs available.

A.4.2 Evidence Uncertainty U(E)

U(E) = w1*(1 - coverage) + w2*contradict + w3*freshness_penalty + w4*(1 - source_reputation)

freshness_penalty = sigmoid((days_since_pub - T0)/k)

A.4.3 Bias Penalty B(u)

B(u) = d1*abs(stance_polarity) + d2*frame_skew + d3*lexical_subjectivity

- stance_polarity in [-1,1], neutral~0
- frame_skew: KL between observed frame distribution vs neutral baseline for topic

---

A.5 Membrane Decision Logic

Thresholds (domain-tunable):

- entail >= 0.85
- contradict <= 0.10
- coverage >= 0.90
- abs(stance_polarity) <= 0.15
- source_reputation >= 0.70

Pseudocode:

```
def membrane_decision(scores, policy):
    reasons = []
    pass_flag = True
    if scores['entail'] < 0.85:
        pass_flag = False; reasons.append('low_entailment')
    if scores['contradict'] > 0.10:
        pass_flag = False; reasons.append('contradiction')
    if scores['coverage'] < 0.90:
        pass_flag = False; reasons.append('insufficient_coverage')
    if abs(scores['stance_polarity']) > policy['NPOV']['stance_threshold']:
        pass_flag = False; reasons.append('stance_out_of_bounds')
    if scores['source_reputation'] < 0.70:
        pass_flag = False; reasons.append('low_source_reputation')

    if pass_flag:
        state = 'pass'
    elif 'contradiction' in reasons or scores['coverage'] < 0.60:
        state = 'reject'
    else:
        state = 'quarantine'
    return state, reasons
```

Edit acceptance rule:

```
if V_after - V_before < 0:
    accept_edit()
else:
```

reject_or_quarantine()

A.6 Propagation Constraints

A.6.1 Infobox Population Rule

A field F on page B may be populated by claim u from page A iff:

- state(u) == pass
- field schema matches atom type
- provenance_token attached
- re-verified in B's context (re-run membrane with X(B))

A.6.2 Cross-Link Replication

- On copy or paraphrase into page C, create new u_C and re-run membrane.
- If upstream u_A downgrades, enqueue recheck(u_C).

A.7 Belief Propagation for Conflict Resolution

Graph: nodes = claim units; edges = semantic equivalence, shared entity, or shared field.

Belief representation:

- For claim u with value v (e.g., birth_year = 1874), maintain b_u(v) in [0,1]
- Initialize from membrane confidence conf_u (entail * coverage * rep)

Message passing (loopy BP style):

- Messages m_{u->v} proportional to compatibility psi(u,v) times incoming beliefs.
- Compatibility encodes agreement between values and shared constraints.

Update schedule:

- Asynchronous on change; cap iterations to K; damp with factor lambda.

Outcome:

- If max disagreement persists, mark cluster as "framed uncertainty" and surface both values with sources.

A.8 Metrics and Dashboards

Page-level:

- MHI (Membrane Health Index) = (#pass) / (#pass + #quarantine + #reject)
- Avg Delta V per accepted edit (negative is better)
- Citation Span-Binding Rate = bound_atoms / total_atoms
- Policy Violation Alerts per 1k edits

Graph-level:

- Semantic Coherence: mean pairwise entail across linked claims
- Bias Entropy: topic-cluster frame entropy vs neutral baseline
- Citation Flow Integrity: weighted PageRank on source graph filtered by pass-only claims

---

A.9 Editor UX and Human-in-the-Loop

Views:

1. Claim Inspector: left = sentence, right = bound spans, with entail/contradict bars.
2. Propagation Map: where this claim populates infoboxes or summaries.
3. What-to-fix List: ranked by Delta V improvement per unit effort.

Actions:

- Replace citation (live search suggestions)
- Rephrase to neutralize stance
- Add missing temporal qualifier
- Mark as framed uncertainty with multi-source summary

Appeals:

- Editors submit counter-evidence; membrane re-runs; decisions logged.

A.10 Provenance, Security, and Rot Handling

A.10.1 Provenance Token

Format:
mmb:1.0|src:<id>|chk:<sha256>|t:<UTC Z>

- src: DOI, ISBN, URL with archive ID
- chk: sha256 over normalized quote + locator
- t: retrieval timestamp

A.10.2 Rot Detection

- Nightly re-fetch; re-hash spans.
- If mismatch: downgrade dependent claims to quarantine; open task.

A.10.3 Integrity and Abuse Controls

- Rate-limit high-impact edits changing many pass claims.
- Canary checks for coordinated bias (sudden stance shift within a topic cluster).

A.11 Domain Tuning

A.11.1 Biomed (MEDRS)

- Higher source_reputation threshold (>= 0.85).
- Preprints allowed only if explicitly labeled and corroborated.
- Contradict cutoff stricter (<= 0.05).

A.11.2 Geopolitics

- Stronger stance and frame checks.
- Source diversity prior: require at least 2 independent regions for disputed facts.

A.11.3 Climate

- Model outputs tagged as projections; require uncertainty and horizon fields.
- Propagation to “facts” blocked unless projection-to-observation validation exists.

A.12 Evaluation Protocols

A.12.1 Membrane Accuracy

- Gold sets: FEVER/FEVEROUS/HoVer-like subsets mapped to wiki sentences.
- Metrics: precision/recall on pass; AUROC for entail vs contradict; coverage MAE.

A.12.2 Downstream Stability

- A/B on live pages (opt-in sandboxes): measure revert rate, talk-page disputes, vandalism catch rate.

A.12.3 Editor Experience

- SUS (usability), time-to-fix, agreement with membrane decisions.

A.12.4 Ablations

- Remove stance model -> measure bias entropy change.
- Remove span hashing -> measure rot regressions.
- Replace D proxy with cosine-only -> measure Delta V sensitivity.

A.13 Minimal Viable Deployment (MVP) Stack

- Parsers: spaCy + rule-based OpenIE.
- Retrieval: Elasticsearch (BM25) + dense encoder (e5, contriever).
- NLI: RoBERTa-large-MNLI or domain-tuned variant.
- Stance/Frame: lightweight RoBERTa with topic adapters.
- Orchestrator: Python + async workers (Celery/RQ).
- Storage: Postgres (claims), S3 (snapshots), Neo4j (graph).
- UI: React dashboard; MediaWiki gadget for inline overlays.

Throughput target (pilot):

- 1k pages, 100k claims
- 5-10 claims/sec verification on modest GPU
- Nightly re-verification of hot pages

A.14 Worked Example (Biography Claim)

Claim: "X was born in 1882."

- Parse -> atom birth_year(X) = 1882
- Retrieve spans -> source A: "born in 1874"; source B: "baptism record 1882"
- NLI: A contradict=0.72; B entail=0.61
- Coverage=1.0; Reputation(A)=0.90, Reputation(B)=0.65
- Policy: B lower reputation; conflict present.

Membrane:

- pass? No (contradiction high, entail marginal)
- state: framed_uncertainty via quarantine
  Rendered text:
  "Sources differ: some list 1874 (A), others infer 1882 from baptism records (B)."

Propagation:

- Infobox birth_year blocked until consensus or authoritative source resolves.

A.15 Policy Encoding Examples

A.15.1 NPOV Enforcement

If abs(stance_polarity) > 0.15 and frame_balance < 0.35:

- Suggest neutral rewrite:
  - From: "X shamelessly exploited..."
  - To: "Some sources criticize X for exploitation; others argue..."

A.15.2 Verifiability

If citation domain in disallow_domains -> quarantine with reason "unreliable_source".

A.15.3 No Original Research

Detect synthesis:

- Pattern: atom derived only by combining A and B where neither asserts atom explicitly.

- Action: reject, suggest wording with attribution (e.g., "Some commentators extrapolate...").

A.16 Graph Maintenance and Scheduling

- Priority queues:
    - Hot pages (edited in last 24h): immediate membrane checks.
    - High CFI hubs: frequent audits.
    - Cold pages: batch nightly/weekly.
- Cache:
    - Reuse span alignments by source checksum.
    - Memoize NLI on identical (claim, span) pairs.

A.17 Interoperability and Export

A.17.1 Membrane Bundle (for external consumers)


```
{
  "membrane_version": "1.0",
  "claim_id": "...",
  "state": "pass",
  "provenance_token": "...",
  "justifications": [
    {"source_id":"doi:...","span_checksum":"sha256:...","entail":0.92}
  ],
  "policy_checks": {"npov":"ok","verifiability":"ok","nor":"ok"},
  "delta_V": -0.18
}
```


A.17.2 Training Data Weighting for LLMs

- Include only state == pass as positive facts.
- Downweight quarantine; exclude reject.
- Attach membrane bundle as supervision signal.

A.18 Risks and Mitigations (Operational)

- Cold-start bias: sparse source graph early on.
    - Mitigate with seed whitelists (encyclopedias, major archives).
- Adversarial citations (doctored screenshots).
    - Require archive links and cryptographic hash of canonical PDF text.
- Model drift:
    - Periodic calibration with challenge sets; record scorecards; rollback models if regressions on policy-critical metrics occur.

A.19 Reproducibility Artifacts

- Versioned datasets:
    - wiki_claims_pilot_v1.jsonl
    - source_graph_v1.parquet
    - policy_templates_v1.yaml
- Model cards:
    - nli_model_card.md (training data, known failure modes)
    - stance_model_card.md
- Scripts:
    - parse_claims.py
    - bind_spans.py
    - verify_membranes.py
    - propagate_constraints.py

- dashboards/ (React app)

A.20 Checklist for Production Readiness

1. Policy templates agreed with community stewards.
2. NLI and stance models calibrated on domain gold sets.
3. Provenance hashing with archive coverage > 90% of cited spans.
4. UI tested: editors can reproduce decisions with one click.
5. Alerting on rot, conflicts, and high-impact downgrades.
6. Privacy review for source fetches and logs.
7. A/B pilot governance and rollback plan documented.

A.21 Glossary (Operational)

- Claim atom: smallest verifiable proposition extracted from text.
- Span binding: locating and hashing the exact source text supporting a claim.
- Framed uncertainty: explicit representation of multi-source disagreement.
- Membrane bundle: portable justification package for a claim's state.
- CFI (Citation Flow Integrity): robustness of citation-supported propagation.