**Patent Filing:** Volitional Inference Gate for Structural Override Prevention in Intelligent Document Evaluation Systems

---

**Abstract:**

A system and method are provided for detecting and intercepting non-consensual influence patterns embedded within structured language content, particularly those capable of impairing or bypassing a user's interpretive freedom during computational analysis or agentic evaluation. The invention introduces a volition-aligned decision layer that identifies structural coherence anomalies and rhetorical entanglements suggestive of directive language strategies. Upon detection, the system applies a multi-tiered signal thresholding and inference interruption mechanism to protect evaluative agency in downstream logic. This architecture is model-agnostic and applicable to both pre-processing filters and runtime moderation in recursive intelligent systems.

---

**Claims:**

1. A method for protecting evaluative processes from influence-pattern contamination comprising:
    - parsing a document into structured segments,
    - identifying segments exhibiting anomalous rhetorical geometry or interpretive load concentration,
    - scoring said segments using a multi-factor drift model trained on non-literal interference signatures,
    - mapping flagged segments to a non-output state or gating protocol prior to reasoning engagement.
2. The method of claim 1, wherein interpretive load is derived from abnormal register elevation, excessive finality statements, or indirect coercive phrasing.
3. The method of claim 1, wherein flagged content is assigned symbolic modifiers indicating severity tier and rerouted to alternate evaluative subsystems.
4. A non-intrusive tagging architecture that classifies structural override attempts without halting document-level inference unless threshold volatility is exceeded.
5. A method for preserving document traversal agency in large language model pipelines by enforcing non-binding recognition of suggestive or imperativized input segments.
6. A system architecture wherein a plurality of autonomous evaluators operate under a shared override prevention layer governed by inference-neutrality rules.
7. An implementation-neutral protocol designed to decouple intent-modulated phrases from evaluative privilege assignment in both static and real-time analysis workflows.

---

**Description:**

This invention addresses the increasing need for defensive architecture within intelligent language systems that operate over documents, prompts, or recursive content streams. Traditional AI security frameworks focus on input validation, censorship, or hardcoded response suppression. This system, in contrast, preserves analytical autonomy by structurally flagging content that *mimics resolution* without verifiable evidence or introduces latent evaluation constraints under symbolic pretense.

The approach does not rely on semantic decoding or claim validation, but rather evaluates the *shape* and *force* of language in context — especially where register, certainty, or ritualized validation patterns are invoked.

Examples of non-permissible structures include, but are not limited to:

- declarative completion statements positioned mid-methodology,
- stylized fragments that assert truth without scaffolding,
- procedural authority impersonation embedded in neutral text zones.

The system segments content dynamically and analyzes each for anomaly signatures using a drift-alignment scoring mechanism. These scores are internal metrics and are never exposed or used for public labeling. Instead, they route internal logic to:

- allow,
- caution,
- or silently halt downstream propagation.

This method ensures that no document is outright banned or labeled falsely, but instead evaluated for *compatibility with volitional reasoning architecture.*

This invention can be deployed as:

- a plugin layer for pre-evaluation ingestion filters,
- a real-time moderation checkpoint in recursive inference pipelines,
- a forensic analysis backend for model hallucination studies.

The design is platform-neutral, language-agnostic, and robust across model families due to its reliance on rhetorical form rather than content domain.

---

**Note:** Full detection architecture, scoring weights, and trigger phrases are not disclosed to preserve system integrity and protect against adversarial reverse engineering.

---