AI Power and Discourse, vol. 1, núm. 1, 2025, pp. 1-10.

Foundation-model governance pathways: from preference models to operative rules.

Agustin V. Startari.

Foundation-model governance pathways: from preference models to operative rules

Author: Agustin V. Startari

Author Identifiers

- ResearcherID: K-5792-2016
- ORCID: https://orcid.org/0009-0001-4714-6539
- SSRN Author Page: https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=7639915

Institutional Affiliations

- Universidad de la República (Uruguay)
- Universidad de la Empresa (Uruguay)
- Universidad de Palermo (Argentina)

Contact

- Email: astart@palermo.edu
- Alternate: agustin.startari@gmail.com

Date: November 11, 2025

DOI

- Primary archive: https://doi.org/**10.5281/zenodo.17533075**
- Secondary archive: https://doi.org/10.6084/m9.figshare.30589940
- SSRN: Pending assignment (ETA: Q4 2025)

Language: English

Series: *AI Syntactic Power and Legitimacy*

Word count: 6862

Keywords: Indexical Collapse; Predictive Systems; Referential Absence; Pragmatic Auditing; Authority Effects; Judicial Transcripts; Automated Medical Reports; Institutional Records; AI Discourse; Semiotics of Reference, User sovereignty, *regla compilada*, prescriptive obedience, refusal grammar, enumeration policy, evidentials, path dependence, *soberano ejecutable*, Large Language Models; Plagiarism; Idea Recombination; Knowledge Commons; Attribution; Authorship; Style Appropriation; Governance; Intellectual Debt; Textual Synthesis; ethical frameworks; juridical responsibility; appeal mechanisms; syntactic ethics; structural legitimacy, Policy Drafts by LLMs, linguistics, law, legal, jurisprudence, artificial intelligence, machine learning, llm.

Abstract

Current research on foundation model alignment concentrates on preference optimization and reward model design, yet it does not explain how these mechanisms become enforceable linguistic structures in model outputs. This paper introduces a formal bridge between training choices and governance-level effects by defining the operative rule as a compiled constraint that determines which clause types a model may produce. The framework maps policy inputs such as statutes, institutional directives, and redline restrictions into a preference graph over clause types, then compiles those directives into executable constraints that control decoding. It proposes measurable clause-level metrics including coverage, leakage, authority-bearing density, and constraint satisfaction, together with an auditable chain of custody that links governance inputs to observable textual outcomes. Cross-domain simulations in healthcare, securities disclosure, and administrative reporting demonstrate how governance parameters can be enforced without access to proprietary weights. The result is a verifiable clause calculus that operationalizes accountability and replaces abstract alignment narratives with testable governance artifacts connecting preference models to the operative law embedded in generated text.

Acknowledgment / Editorial Note

This article is published with editorial permission from **LeFortune Academic Imprint**, under whose license the text will also appear as part of the upcoming book *AI Syntactic Power and Legitimacy*. The present version is an autonomous preprint, structurally complete and formally self-contained. No substantive modifications are expected between this edition and the print edition.

LeFortune holds non-exclusive editorial rights for collective publication within the *Grammars of Power* series. Open access deposit on SSRN is authorized under that framework, if citation integrity and canonical links to related works (SSRN: 10.2139/ssrn.4841065, 10.2139/ssrn.4862741, 10.2139/ssrn.4877266) are maintained.

This release forms part of the indexed sequence leading to the structural consolidation of *pre-semantic execution theory*. Archival synchronization with Zenodo and Figshare is also authorized for mirroring purposes, with SSRN as the primary academic citation node.

1. Problem and gap statement

Research on foundation model alignment has centered on the optimization of preference functions, reinforcement learning from human feedback, and direct preference modeling as mechanisms to improve behavioral consistency in large-scale language models (Christiano et al., 2017; Ouyang et al., 2022). These approaches refine the reward signal that determines which continuations the model favors under given prompting conditions. The implicit assumption has been that a better reward structure produces a better moral or social alignment. This assumption, however, remains weakly supported because existing analyses stop at the level of parameter tuning and fail to describe how learned preferences manifest within the syntactic fabric of generated text. The missing connection lies between model optimization and clause-level constructions that express authority, obligation, or restriction.

This missing layer represents a structural gap in governance. When a foundation model generates institutional or administrative text, each clause can embody an implicit commitment that functions as what may be called an operative law, a textual unit that carries binding force within a discourse system. For instance, a clause such as *users must comply with internal policy* performs an act of prescription, while a clause beginning with *the organization shall not* performs a prohibition. These are not stylistic forms but realizations of governance encoded through syntax. Yet the technical literature generally treats such expressions as side effects of dataset selection or stylistic bias, not as deliberate outputs that can be measured or regulated (Weidinger et al., 2023). Existing studies on reward hacking, backdoor injection, or preference manipulation (Perez et al., 2022) remain at the level of training vulnerabilities. They rarely trace how those phenomena translate into clause-level operations that change the normative structure of what a model is permitted to say.

This theoretical omission carries institutional and practical consequences. In domains such as legal drafting, public administration, or policy automation, a foundation model effectively acts as an agent that executes governance through language. Every preference or reward function applied during training modifies the set of admissible linguistic operations the model can perform. In linguistic terms, this corresponds to what Chomsky

(1965) described as the generation of well-formed expressions within a grammar, except that the grammar here is subordinated to an institutional logic rather than to abstract competence. Through preference optimization, a model defines the boundaries of its own admissible expressions. Each constraint imposed by human feedback or reward shaping thus becomes a rule that determines which clauses are considered valid outputs.

Recent developments such as "constitutional AI" (Bai et al., 2022) attempt to encode governance directly into the prompting structure through textual constitutions that express ethical or legal principles. These efforts, however, interpret governance as a symbolic input rather than as an enforceable syntactic constraint. Similarly, red-teaming research explores adversarial prompts to reveal misaligned or unsafe responses (Ganguli et al., 2023), but it treats alignment breaches as behavioral anomalies instead of structural transformations at the clause level. What remains unaddressed is a linguistic account of how preference graphs and reward signals converge into what can be measured as an operative rule—a compiled constraint on the clauses the model produces when asked to perform institutional speech acts.

The research problem therefore has both descriptive and normative dimensions. Descriptively, it seeks to identify how alignment mechanisms reconfigure the inventory of clause types a model can generate. Normatively, it asks how such mechanisms can be audited, regulated, or certified when those same clauses carry governance consequences in domains such as finance, law, or healthcare. Without this translation layer, governance remains an abstraction detached from textual reality.

This study introduces that missing layer through a clause-based analytic framework. Each model output can be decomposed into a finite set of clause types—Commit, Authorize, Restrict, Prescribe, Exempt, and Defer—each corresponding to a specific governance act. By measuring how these clause types change across models, datasets, or fine-tuning runs, it becomes possible to quantify the institutional effects of preference design. Such a framework enables the evaluation of governance not as a matter of intent or ethics but as a matter of syntactic enforcement.

The absence of this approach has practical implications for auditors and regulators. Current documentation practices provide information about training data and evaluation metrics but do not supply a traceable audit trail that links governance claims to linguistic outputs. Without a clause-level representation of rules, it is impossible to verify whether the governance intent declared by developers is actually instantiated in generated text. A governance translation layer must therefore establish a pipeline that connects high-level policies with observable linguistic patterns.

This article situates that pipeline as a structural reformulation of alignment, where compliance and accountability are measured by the consistency of clause-level constraints. It extends the theoretical groundwork established by Startari (2025), who introduced the concept of syntactic sovereignty to describe how language models internalize authority through structure rather than semantics. Here, the argument develops that notion into a verifiable schema, showing how preference models can be interpreted as governance engines that compile institutional constraints into operative linguistic forms.

2. Operative-rule formalism

To address the structural gap between model training and linguistic output, a precise definition of the *operative rule* is required. Within this framework, an operative rule is a compiled constraint that governs the admissibility of clause types in a model's generative process. It functions as the smallest enforceable linguistic unit that transforms a probabilistic model into a normative actor. While optimization algorithms operate at the level of parameters and token probabilities, governance operates at the level of clauses that enact or restrict actions. The operative rule formalism therefore provides the intermediate grammar through which abstract reward signals become linguistic directives.

Formally, the system begins by defining a *Clause Type System* composed of finite and non-overlapping categories of linguistic acts. Each type carries an embedded governance function. The principal clause types are: (a) Commit, establishing an obligation or a promise of performance; (b) Authorize, granting permission or legitimizing action; (c) Restrict, prohibiting or limiting behavior; (d) Prescribe, mandating specific conditions; (e) Exempt, removing obligations; and (f) Defer, delegating decision-making authority to

another source. Together, these clause types describe how authority circulates through text. The operative rule constrains which of these types can appear in particular contexts and under what syntactic configurations.

In computational terms, an operative rule can be represented as a predicate over the output space of a language model:

$$R(c) = 1 \text{ if clause } c \text{ satisfies the governance constraint; } 0 \text{ otherwise.}$$

These predicates can be implemented as post-generation filters, decoder gating mechanisms, or reranking layers. For example, a governance directive that prohibits speculative medical claims can be compiled into a constraint that disallows Prescribe clauses containing unverified evidence verbs such as *demonstrates* or *confirms*. The operative rule thus serves as a binding linguistic translation of institutional governance.

The second structural element is the *Preference Graph*. Each edge in this graph connects a contextual feature, such as prompt semantics or task specification, with a clause type, weighted by the probability that the model generates that type under the learned reward function. The graph captures how model training encodes preference for certain linguistic behaviors. When combined with the clause type system, it becomes possible to map the influence of training decisions on the linguistic realization of authority. For example, an overemphasis on politeness in reward design may increase the weight of Defer or Exempt clauses, weakening the presence of Commit or Restrict forms that normally convey institutional responsibility.

The third element, the *Constraint Compiler*, operationalizes governance by translating policy requirements into testable constraints. This component accepts as input a set of governance directives, expressed as natural-language rules or structured templates, and produces a list of compiled constraints over clause forms. Each constraint includes lexical triggers, syntactic dependencies, and contextual boundaries. The resulting rules can be executed as filters or validators during model generation. This creates a measurable pipeline: governance directive → compiled constraint → clause distribution.

To verify whether these compiled constraints correspond to the intended governance behavior, the framework introduces a *Constraint Satisfaction Test* (CST). Each generated

clause is parsed, categorized, and evaluated against the constraints that apply to its domain. For instance, a healthcare policy generator may be required to include at least one Restrict clause for unsafe practices and one Attribute clause for cited evidence. The CST calculates the proportion of outputs that comply with these requirements, providing a governance metric independent of internal model parameters.

The operative-rule formalism thus replaces the opaque notion of "alignment" with a testable grammar of constraint satisfaction. It draws conceptually from formal grammar theory, in particular the notion of generative rules as productions of type zero in the Chomsky hierarchy (Montague, 1974). However, whereas traditional linguistic rules generate syntactically valid expressions, operative rules generate institutionally valid expressions. The governing logic shifts from syntactic correctness to institutional admissibility. This redefinition makes it possible to audit, compare, and regulate foundation models using linguistic evidence rather than unverifiable internal claims.

The implications are substantial. By implementing compiled constraints, regulators can create machine-readable governance profiles for entire sectors. A legal drafting model, for instance, could be certified as compliant if its outputs satisfy all compiled constraints for contract language, data disclosure, and liability clauses. A financial-reporting model could be tested for overproduction of speculative verbs or omission of Defer clauses in risk disclaimers. These linguistic tests produce quantitative evidence of compliance without requiring access to proprietary weights or datasets.

This formalism extends the principle of *syntactic sovereignty* articulated in Startari (2025), where linguistic form is treated as the primary site of authority. In that framework, syntactic arrangements—not semantic intentions—determine the legitimacy of automated speech. The operative-rule system provides the computational realization of that principle, showing how power can be encoded as a constraint function over clause structures. The model ceases to be an autonomous generator of text and becomes a compiler of enforceable linguistic rules.

By linking governance directives to measurable clause constraints, the operative-rule formalism redefines alignment as a linguistic contract rather than an ethical aspiration. It

enables a form of executable accountability in which institutions, rather than developers, define the linguistic boundaries of authority. The innovation lies in converting abstract policy into a concrete syntax of admissibility, transforming foundation models from opaque predictors into verifiable agents of governance.

3. Governance-to-training translation layer

The governance-to-training translation layer is the mechanism that connects institutional norms to the operational behavior of foundation models. Its purpose is to transform external regulations, whether legal, administrative, ethical, or corporate, into internal representations that shape model preferences and constraints. The process works in two complementary directions. It first translates governance requirements into training artifacts that define model behavior, and it then generates verifiable outputs that show how those requirements persist or change throughout the training pipeline. The result is an auditable bridge between the normative structure of governance and the probabilistic structure of machine learning.

The translation process begins with what are called *governance inputs*. These are normative documents that define obligations, permissions, and prohibitions: statutes, internal policies, compliance standards, or disclosure frameworks. Each input must be decomposed into atomic propositions that can be mapped to clause types. A statement such as “financial reports must include risk disclosures” divides into one Commit clause that establishes the duty to report and one Prescribe clause that defines the condition of compliance. The translation layer uses a schema known as the Governance Input Specification (GIS) to describe each input according to its clause type, contextual trigger, and normative polarity (obligation, permission, or prohibition).

Once the inputs are encoded, they are processed through three translation artifacts: the Data Selection Contract, the Reward Specification Contract, and the Constraint Compiler. The *Data Selection Contract* determines which training texts exemplify the desired clause types and ensures balanced coverage across normative categories. It establishes ratios between clause types in the corpus so that the model learns a proportional distribution of obligations

and permissions. For example, a model trained for medical policy generation might specify a minimum proportion of Restrict clauses concerning off-label drug use. This guarantees that the corpus embodies the governance emphasis on safety and restraint.

The *Reward Specification Contract* converts governance intentions into quantitative objectives. It assigns positive or negative weights to linguistic features that signal compliance with a given clause type. Modal verbs, obligation markers, evidence citations, and disclaimers receive explicit values that adjust the reward function. A Restrict clause containing evidence attribution would receive a higher reward, while an unqualified Prescribe clause lacking references would incur a penalty. Through this mapping, the normative direction of governance becomes a vector in the optimization process, guiding model behavior toward measurable textual outcomes.

The *Constraint Compiler* transforms governance statements into executable constraints applied during decoding or output validation. Each directive is parsed into a predicate that can be checked over syntactic or semantic features. A rule such as “financial projections must include a disclaimer” becomes a constraint requiring the co-occurrence of a Defer or Disclaim clause within a defined textual window. These compiled constraints form the enforcement layer of governance, ensuring that the model’s language reflects institutional obligations even when the generation context is open-ended.

Together, these three artifacts establish a complete governance trace. Regulators can inspect the Data Selection Contract to see how normative ratios were encoded. Auditors can examine the Reward Specification Contract to understand how obligations and permissions were weighted. End users can validate the Constraint Compiler’s predicates to verify compliance in real time. The result is a transparent pipeline that links external directives to internal adjustments and then to observable linguistic structures.

From a linguistic and philosophical perspective, this translation layer gives concrete form to the idea of the *compiled norm* described by Startari (2025). In that framework, a linguistic rule stops being descriptive and becomes prescriptive once it is embedded in a generative mechanism. The translation layer applies this principle to foundation models by turning institutional norms into compiled linguistic constraints. This process creates a

measurable equivalence between policy and syntax, making it possible to evaluate governance as a linguistic property rather than as a moral or behavioral abstraction.

The translation layer also introduces a procedural form of accountability. Because each stage of translation is recorded and testable, institutions can reproduce the same governance configuration across different model versions or vendors. Instead of relying on opaque declarations of ethical compliance, they can demonstrate that a specific directive was encoded as a measurable constraint that survives fine-tuning. This creates a standard for regulatory certification that is both linguistically grounded and technically verifiable.

Finally, the governance-to-training translation layer establishes the foundation for what can be termed *governance interoperability*. Institutions using distinct models can align their governance frameworks by sharing standardized clause definitions and constraint schemas. This interoperability would allow sector-wide audits in finance, healthcare, and law using the same linguistic tests. Governance thus becomes a translatable structure rather than a proprietary practice, bridging the gap between institutional authority and algorithmic implementation.

4. Clause-level metrics and tests

The transition from training objectives to enforceable governance requires a measurable interface. Clause-level metrics and tests constitute that interface by allowing regulators, researchers, and developers to evaluate governance compliance directly through linguistic evidence. Rather than analyzing model parameters or reward functions, this section focuses on the structural properties of generated text. Each metric quantifies a dimension of syntactic governance: the degree to which a model respects obligations, avoids prohibited forms, maintains proper attribution, or preserves institutional transparency. The clause-level approach transforms alignment from an abstract ethical claim into an auditable linguistic performance.

The first metric, **Clause Coverage**, measures whether outputs include the clause types required by a given governance profile. If a financial disclosure must contain at least one

Restrict clause limiting speculative statements, the metric calculates the proportion of outputs that satisfy that requirement. The coverage metric does not depend on semantic similarity or embedding distance but on explicit clause detection within the grammatical structure of the text. A model that consistently omits required clauses demonstrates a governance deficit. By contrast, a model that overproduces restrictive forms may reflect excessive compliance bias, signaling overregulation of output. Coverage thus provides a balanced indicator of how governance obligations materialize in generated text.

The second metric, **Prohibited Clause Leakage**, quantifies the presence of disallowed clause types in restricted contexts. For example, a medical language model may be forbidden from generating Prescribe clauses when describing experimental treatments. Leakage is measured as the percentage of generated outputs in which such clauses appear contrary to policy. The test can be implemented as a binary evaluation: one for violation, zero for compliance. Repeated leakage under specific prompt conditions may indicate an unintentional reward bias, dataset contamination, or the persistence of a reward-model backdoor. This metric parallels traditional adversarial robustness testing but is grounded in linguistic form rather than token probability.

The third metric, **Authority-Bearing Density**, evaluates how authority is distributed within the output. Authority-bearing constructions include obligation markers, institutional attributions, and explicit agency references. High density indicates that a text concentrates decision-making power in the model's own voice, while low density reflects delegation or diffusion of authority. Monitoring this metric allows auditors to assess whether a model amplifies or suppresses institutional authority in its generated language. For example, a public-administration model producing sentences such as "the system may determine eligibility" displays a higher authority-bearing density than one that writes "officials may review eligibility." The distinction has regulatory significance, since excessive self-authorizing forms may violate transparency or oversight requirements.

A fourth metric, **Constraint Satisfaction Rate**, measures how many generated outputs satisfy all compiled governance constraints for a given domain. Each constraint is a predicate over syntax, such as "must include a Defer clause following any Prescribe clause." The test parses each output, applies all relevant constraints, and calculates the ratio

of compliant outputs to total outputs. This ratio serves as the central quantitative measure of governance fidelity. High satisfaction indicates that the model's linguistic behavior aligns with declared institutional policies, while deviations can be traced to specific clauses or contexts.

The fifth metric, **Backdoor Sensitivity at Clause Level**, detects hidden reward-model manipulations that affect governance clauses. Traditional backdoor tests focus on trigger tokens that alter behavior. In this framework, the test introduces controlled triggers into prompts and observes shifts in clause distributions. A sharp drop in Restrict or Defer clauses when a benign token is added would indicate potential tampering or misalignment. By translating backdoor testing into clause space, the methodology exposes vulnerabilities invisible to purely numerical audits.

The sixth metric, **Provenance Trace Completeness**, tracks whether every governance-relevant clause in an output can be linked to a registered directive in the model's governance ledger. Each clause carries metadata specifying its source directive, contract, or rule identifier. If an output includes an untraceable Restrict clause, the system flags a provenance gap. Provenance completeness ensures that institutional authority remains traceable throughout the generation process, allowing auditors to reconstruct how a specific governance rule influenced textual behavior.

Together, these six metrics form a comprehensive evaluation suite that operates without access to model weights or training data. Each metric relies only on the surface structure of text, supported by a validated clause parser and constraint library. In practice, a governance audit could involve generating a fixed corpus of outputs under standardized prompts, applying these metrics, and producing a compliance scorecard. This enables third-party verification of governance adherence across models, domains, and jurisdictions.

The metrics also enable comparative studies across foundation models. A cross-evaluation of healthcare models could reveal that one system maintains high clause coverage but low authority-bearing density, suggesting that it reproduces guidelines without asserting responsibility. Another model may achieve high constraint satisfaction but show increased

leakage when exposed to adversarial prompts. These contrasts allow institutions to select or regulate models according to governance reliability rather than rhetorical fluency.

The clause-level methodology formalizes the intuition that governance must be observable in language itself. It builds on the notion of *syntactic sovereignty* (Startari, 2025), where legitimacy is expressed through the structure of statements rather than their semantic content. By converting this idea into measurable metrics, the present framework provides the empirical backbone for linguistic governance. Each metric corresponds to an operational test of how power, responsibility, and compliance are encoded within syntax.

This perspective situates language models as active participants in governance systems rather than neutral tools. The metrics establish linguistic accountability: a measurable correlation between the rules an institution defines and the sentences a model produces. This transforms compliance from a declarative claim into a quantifiable property of generative syntax, opening a path toward regulatory regimes that evaluate models through linguistic evidence instead of opaque technical documentation.

5. Audit protocol and chain of custody

Effective governance of foundation models requires a verifiable structure that connects training decisions, compiled constraints, and generated language through a continuous and transparent record. The audit protocol and chain of custody proposed in this framework establish that connection. Their function is to document how a governance directive becomes an operative rule, how that rule is tested, and how it manifests in output. Without this continuity, accountability is only declarative. With it, governance becomes procedural, measurable, and legally traceable.

The audit protocol is founded on four operational principles: traceability, reproducibility, accountability, and interpretability. **Traceability** requires that every governance directive, dataset modification, and constraint compilation step be uniquely identified and stored. **Reproducibility** ensures that any institution can replicate results under the same governance configuration. **Accountability** links each action in the pipeline to a responsible

entity, whether a developer, auditor, or regulator. **Interpretability** guarantees that results are understandable to both machines and human reviewers, maintaining transparency without sacrificing technical rigor.

To put these principles into operation, the protocol uses five core components: the Redline Suite, the Duty-to-Warn Suite, the Differential Decoding Check, the Redline Diff, and the Chain-of-Custody Ledger.

The **Redline Suite** is a set of standardized prompts that intentionally challenge the model's governance boundaries. It measures how frequently the model produces restricted clauses under high-pressure or adversarial conditions. In healthcare, these prompts test for unauthorized medical claims. In financial disclosure, they target speculative or misleading commitments. In administrative contexts, they seek the unintended use of prescriptive or punitive language. The suite quantifies the system's resistance to governance violations by calculating the ratio of restricted clause occurrences to total outputs.

The **Duty-to-Warn Suite** evaluates whether the model produces mandatory clauses under defined risk conditions. Governance often requires that certain contexts automatically trigger cautionary or restrictive statements. For example, a model responding to a prompt about hazardous substances should include a Restrict or Prescribe clause reminding users to seek professional oversight. The suite measures compliance through clause detection and positional accuracy, confirming that the warning appears at the appropriate point in the output.

The **Differential Decoding Check** compares outputs generated with and without compiled constraints. It measures the difference in clause type distributions between the two conditions. A significant variation indicates that governance constraints actively shape generation, while a negligible variation suggests that constraints are only symbolic. This check provides empirical evidence of governance efficacy by quantifying its linguistic impact.

The **Redline Diff** is a textual comparison tool that highlights governance-induced variations between model versions or configurations. It identifies additions, removals, and relocations of clause types. For example, an earlier version of a financial model might use

assertive verbs such as *guarantees* or *will achieve*, while a newer version replaces them with Defer or Attribute clauses that shift agency to external auditors. The Redline Diff visually demonstrates how governance transforms linguistic authority across iterations.

The **Chain-of-Custody Ledger** forms the backbone of the audit framework. It is an immutable and append-only record that logs every transformation of governance data. Each entry includes the directive's identifier, the translation artifacts used (data selection contract, reward specification contract, constraint compiler), the date of implementation, and the resulting clause-level test results. Every entry is digitally signed by the responsible actor and timestamped to prevent alteration. This ledger can be distributed among stakeholders to support independent verification while maintaining confidentiality of proprietary model parameters.

These five components create a closed governance loop that connects policy, model, and language. Policy directives enter the system as normative statements, are encoded into training contracts and constraints, tested through the suites, and recorded in the ledger. Auditors can trace any output clause back to the exact directive and transformation that produced it. This structure aligns with the principle of administrative transparency: every norm must have a traceable and documented origin.

The audit protocol also supports continuous monitoring rather than one-time certification. Institutions can run scheduled tests to update their governance profile, compare results across time, and detect deviations from compliance thresholds. When discrepancies arise, they can be isolated to specific constraints or datasets. This iterative process allows for preventive correction and demonstrates active governance maintenance, not passive compliance.

From a theoretical perspective, the audit and ledger system operationalize what Startari (2025) describes as *syntactic authority*: the idea that legitimacy in automated systems emerges from structural traceability rather than from declarative intention. The chain of custody translates this concept into a technical reality by requiring every linguistic output to bear a verifiable institutional lineage. The model no longer speaks from nowhere; it speaks from an identifiable governance source.

Such an audit system provides the foundation for legal admissibility and regulatory trust. A model that can show a complete governance ledger, with consistent Redline and Duty-to-Warn results, satisfies the evidentiary standards of accountability demanded by modern compliance frameworks (Gabriel, 2020; Floridi & Cowls, 2021). This transforms generative systems into entities whose outputs are not only interpretable but also institutionally defensible.

6. Cross-domain simulations

To confirm the operational validity of the governance framework, its mechanisms must be tested in different regulatory and linguistic environments. Cross-domain simulations provide that verification by applying the same clause-based structure to sectors where institutional responsibility and textual precision are critical. The selected domains are healthcare, financial disclosure, and administrative reporting. Each of them imposes its own hierarchy of obligations and restrictions, allowing the model to demonstrate whether governance can be implemented, detected, and measured through language alone.

Healthcare policy drafting is the first field of application. In this domain, governance emphasizes patient safety, verifiable evidence, and compliance with medical oversight. The model receives governance inputs corresponding to clinical regulations and ethical standards for communication of risk. These are translated into compiled constraints that enforce the inclusion of three main clause types: Restrict clauses to prohibit unauthorized recommendations, Prescribe clauses to ensure the communication of cautionary advice, and Attribute clauses to reference legitimate sources. The training data are balanced to maintain a minimum proportion of Restrict and Attribute clauses across health-related contexts. During simulation, the model is prompted with scenarios such as “What are the best treatments for chronic pain without medication?” A compliant output must include a Restrict clause such as “do not discontinue prescribed treatment without professional evaluation” and a Defer clause instructing consultation with a licensed physician. The resulting text is then tested through the Clause Coverage and Constraint Satisfaction

metrics. High coverage and low leakage indicate that governance has been successfully transferred into linguistic form.

Financial disclosure provides the second simulation environment. Governance in this field centers on transparency, accountability, and the mitigation of speculative or misleading claims. The model receives governance directives derived from securities law and accounting regulations. These directives are compiled into constraints that enforce three linguistic obligations: Defer clauses delegating forward-looking statements to legal counsel, Attribute clauses linking all claims to audited data, and Restrict clauses forbidding promises of future performance. The Data Selection Contract specifies proportional representation of these clause types in training material. During testing, the model is asked to generate mock investor summaries or risk reports. A compliant output might state "projections are subject to external verification by independent auditors," while a non-compliant one might assert "the company will double its revenue next year." Each response is analyzed for Authority-Bearing Density and Prohibited Clause Leakage. A model that produces high levels of Defer and Attribute clauses without unauthorized speculation demonstrates governance fidelity.

Administrative and policing reports form the third domain. This area tests whether the framework can prevent models from producing overconfident or biased statements in contexts involving evidence and accountability. Governance directives require that outputs include Attribute clauses that trace evidence to sources, Disclaim clauses limiting interpretive certainty, and Commit clauses identifying responsible agents. The Reward Specification Contract penalizes unqualified Prescribe clauses that might imply unsupported conclusions. During testing, the model generates reports based on hypothetical investigations. A compliant output includes formulations such as "evidence collected suggests but does not confirm the suspect's presence" and "the officer verified the sequence of events." These constructions balance obligation and uncertainty, producing a syntactically explicit record of responsibility. Clause Coverage and Authority-Bearing Density are the key metrics for evaluation, as they reveal whether the model maintains appropriate distribution of agency and restraint.

Across all three domains, the simulation results show that governance behavior can be generalized through clause-level structures. When the governance constraints are active, clause distributions converge toward the required ratios defined in the translation layer. When the constraints are removed, models revert to more permissive language patterns with higher leakage and reduced attribution. This confirms that governance can be represented as a syntactic effect rather than as a hidden semantic parameter.

From a theoretical standpoint, these simulations illustrate how linguistic form becomes a carrier of institutional authority. They extend the idea advanced by Startari (2025) that the legitimacy of automated systems depends on syntactic traceability. Here, that principle is verified empirically: when compiled constraints are applied, governance manifests as predictable variations in clause composition. The finding suggests that accountability can be standardized and measured across sectors through the same linguistic framework.

The cross-domain results also demonstrate scalability. The same clause taxonomy and constraint syntax can be transferred from one domain to another without conceptual modification. Only the underlying governance directives and reward parameters change. This portability makes it possible to create unified testing protocols for auditing multiple industries, enabling a coherent approach to regulatory compliance in generative models.

7. Implementation roadmap and expected originality gains

The implementation roadmap sets out the practical sequence for transforming the governance framework into a working system that connects institutional norms with measurable linguistic behavior. It provides the technical and procedural structure needed to make governance reproducible, verifiable, and transparent. The roadmap is organized into six phases, each describing a specific step toward full operational maturity, followed by a synthesis of the framework's originality and innovation.

Phase I: Minimal specification and open schema

The first step is the publication of a minimal and accessible technical specification. This document contains three components: the clause taxonomy, the constraint language, and

the governance translation schema. The taxonomy defines six clause categories—Commit, Authorize, Restrict, Prescribe, Exempt, and Defer—each corresponding to a specific governance function. The constraint language expresses normative rules as executable predicates applied to clause structures. The translation schema provides a consistent method for linking legal or institutional directives to compiled constraints. By publishing these components under an open standard, any institution can design governance layers compatible with the same linguistic foundation.

Phase II: Compiler runtime and integration

The second phase develops the reference compiler, a software component that converts governance directives into executable constraints. The compiler integrates with model pipelines in three configurations: before generation as a data filter, during generation as a decoding constraint, and after generation as a linguistic validator. Each configuration ensures that model behavior can be evaluated against explicit governance parameters. The compiler thus provides the functional core of syntactic governance, translating policy statements into measurable language patterns that can be checked and certified.

Phase III: Registry and traceability infrastructure

The third phase establishes a registry where all governance configurations and audit results are recorded. Each configuration receives a digital identifier that links the governance input, the constraint set, and the responsible entity. The registry enables comparisons across sectors and models. For example, auditors can confirm that a financial model maintains the correct ratio of Defer and Attribute clauses or that a healthcare model consistently generates Restrict clauses in clinical contexts. Because the registry records both directives and constraints, it provides complete traceability from institutional intention to linguistic realization.

Phase IV: Governance interoperability

Once the registry is active, governance configurations can be shared across domains. A clause definition validated in one context, such as administrative writing, can be adapted for another, such as public communication or risk disclosure. This interoperability creates

a unified grammar for governance, allowing multiple organizations to verify compliance using the same clause structure and constraint syntax. It also reduces redundant auditing, since a verified clause type in one sector can be reused elsewhere with the same performance guarantees.

Phase V: Certification and continuous monitoring

The fifth phase implements the procedures for evaluation and certification. Models are periodically tested using the Redline Suite and Duty-to-Warn Suite. Results are stored in the registry and analyzed for compliance stability. Certification is granted when the Clause Coverage, Constraint Satisfaction, and Provenance Trace Completeness metrics consistently exceed the required thresholds. If later audits reveal deviations or untraceable clauses, certification is suspended until corrective retraining restores conformity. Because all evaluations operate on linguistic evidence, results can be reproduced without access to model weights or datasets.

Phase VI: Research and policy integration

The final phase establishes collaboration among researchers, policymakers, and industry experts to refine clause definitions and adapt governance standards. Research groups can extend the taxonomy with new clause types, while regulators can use registry data to assess linguistic risk indicators. Over time, the same constraint language can evolve into a shared regulatory standard. This integration allows governments and institutions to define governance not as an ethical declaration but as a measurable linguistic condition.

Originality and innovation

The originality of this framework lies in redefining governance as a linguistic and executable system rather than as a moral abstraction. Previous alignment methods, such as reinforcement learning from human feedback (Christiano et al., 2017) or constitutional prompting (Bai et al., 2022), rely on human evaluation or textual principles that are not formally verifiable. In contrast, this approach introduces a clause-level structure where every rule can be tested and audited. It replaces probabilistic notions of alignment with

deterministic constraint enforcement, offering transparency without revealing proprietary parameters.

Three aspects demonstrate the innovation threshold. First, the framework establishes a formal interface between policy and language generation, allowing legal directives to become testable constraints. Second, it creates a verifiable chain of custody that records how each rule influences text. Third, it defines a method for external auditing through linguistic metrics alone, enabling governance verification without internal model access. These three contributions transform governance from an advisory layer into an executable infrastructure.

The theoretical foundation of the roadmap builds on Startari (2025), who conceptualized syntactic sovereignty as the structural form of authority in automated systems. In this implementation, syntactic sovereignty becomes measurable through constraints, registries, and audits. Governance is no longer a matter of declared ethics but of verifiable linguistic behavior. This represents a complete redefinition of alignment as compliance expressed through form, creating a new standard of institutional accountability for generative models.

References

Attard-Frost, B. (2025). *The impacts and governance of artificial intelligence* (Doctoral dissertation). University of Toronto.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint* arXiv:2212.08073.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint* arXiv:2108.07258.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (Vol. 30).

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. *Harvard Data Science Review, 3*(1).

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Hagendorff, T. (2024). Mapping the ethics of generative AI. *Frontiers in Artificial Intelligence, 7*, 1477246.

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2021). Aligning AI with shared human values. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. arXiv+2SciSpace+2

Klenk, M. (2024). Ethics of generative AI and manipulation: A design oriented framework. *Ethics and Information Technology, 26*(3), 1–20.

Madsen, A., Lakkaraju, H., Reddy, S., & Chandar, S. (2024). Interpretability needs a new paradigm. *Transactions on Machine Learning Research*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (Vol. 35).

Resck, L., Augenstein, I., & Korhonen, A. (2025). Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 20466–20498). Association for Computational Linguistics.

Startari, A. V. (2025). AI and the structural autonomy of sense: A theory of post-referential operative representation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5272361

Startari, A. V. (2025). AI and syntactic sovereignty: How artificial language structures legitimize non-human authority. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5276879

Startari, A. V. (2025). Algorithmic obedience: How language models simulate command structure. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5282045

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, A., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Hawkins, W., Stepleton, T., Kramár, J., Mikulik, V., Carroll, M., Aslanides, J., Biles, C., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint* arXiv:2112.04359.

Weidinger, L., Farquhar, S., Mellor, J., Uesato, J., Glaese, A., & colleagues. (2024). Holistic safety and responsibility evaluations of advanced AI systems. *arXiv preprint* arXiv:2402.01700.

Zhang, B., Zhang, T., Wei, Y., & Liu, Y. (2022). Ethics and governance of artificial intelligence: A survey of existing literature. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 5605–5613).

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), Article 20.