Protocol as Prescription: Governance Gaps in Automated Medical Policy Drafting

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: October 4, 2025

DOI

- Primary archive: https://doi.org/**10.5281/zenodo.17259810**
- Secondary archive: https://doi.org/10.6084/m9.figshare.30275833
- SSRN: Pending assignment (ETA: Q3 2025)

Language: English

Series: *AI Syntactic Power and Legitimacy*

Word count: 8199

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

This article examines how health policy texts drafted with large language models can detach legal responsibility from the formal circuit of governance. Treating "protocol" as *regla compilada*, anchored to a Type 0 production in the Chomsky hierarchy, it specifies a provenance standard that binds each clause of an issued policy to its generating inputs, including prompts, parameters, retrieval sources, reviewers, timestamps, and cryptographic hashes. The method combines version-controlled diffs across scoping, drafting, legal review, and publication with a formal alignment of authority bearing constructions, focusing on deontic stacks, default scopes, agent deletion, and nominalizations. A simulated ministry case demonstrates end to end traceability, producing an exportable evidence bundle that links surviving clauses to their inputs and human approvals. Findings show where machine introduced formulations change duty of care or obscure decision rights, and define mandatory human sign offs when high risk constructions appear. The article delivers three operational artifacts for health agencies, a provenance specification, a responsibility matrix across drafting stages, and an audit checklist calibrated to inspection and courtroom needs. By reattaching authorship and justification to the formal record, the blueprint closes a governance gap in automated policy drafting and states the conditions under which AI assisted procedures remain defensible.

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.

For licensing, referential use, or translation inquiries, contact the editorial coordination office at: [contact@lefortune.org]

1) Problem Statement and Scope

Health agencies are beginning to use large language models to draft or redraft policy texts, guidance, and internal procedures. This practice creates a governance gap whenever the resulting document does not carry an auditable chain from inputs to issued clauses. The gap appears at three layers. First, provenance. Prompts, parameters, retrieval sources, and human edits are often missing from the record, which prevents ex post inspection and legal review. Second, authority. Specific constructions in policy prose, such as deontic verbs, default scopes, and agent deletion, can shift responsibility or duty of care without explicit sign off. Third, liability. When model generated or model suggested formulations survive unchanged into the final document, responsibility may detach from the formal circuit if the record cannot show which human actor approved the clause and on what basis. The article addresses these three layers by treating protocol as *regla compilada*, and by requiring a traceable production from prompt to official text that is suitable for courtroom or inspectorate scrutiny.

This framing responds to two converging developments. On the health side, the World Health Organization has issued governance guidance for generative systems used in care, public health, and research, urging transparency, risk control, and oversight whenever these systems influence health decisions. The guidance highlights the need for documentation of model behavior and deployment context, yet it does not prescribe a clause level provenance chain for policy drafting inside ministries or agencies. That omission leaves institutional texts vulnerable to authorship ambiguity and weakens legal defensibility when a clause alters coverage, eligibility, or duties. The problem is not model capability in the abstract. It is the absence of a binding record that links each surviving clause to its generating inputs and human approvals. Without such binding, accountability becomes negotiable after the fact, which is precisely what health governance aims to avoid (World Health Organization, 2024, 2025).

On the reporting side, biomedical AI has moved toward structured transparency through domain specific guidelines such as TRIPOD-LLM. These frameworks standardize how researchers should describe data, modeling, evaluation, and uncertainty. They do not target institutional policy drafting and they do not provide a document level authorship trace from

prompts to clauses. The distance between a study reporting checklist and a government policy pipeline is material. Ministries and hospital networks need a reproducible authorship chain that covers role identities, toolchains, model versions, seeds where applicable, retrieval sources, parameter ledgers, review notes, timestamps, and cryptographic binding at each checkpoint. The present article fills this structural gap by specifying the required fields and by defining a version controlled audit instrument that health agencies can adopt without reliance on vendor black boxes or ad hoc practices (Gallifant et al., 2025).

The regulatory context makes the gap urgent. Under the European Union Artificial Intelligence Act, systems that influence health related decisions or operate in regulated contexts trigger obligations that include risk management, data governance, technical documentation, and transparency toward users. Agencies that let generative systems draft or materially revise public health policies incur not only internal governance risks but also potential non-compliance if provenance and accountability are not demonstrable. A defensible pipeline must show how specific textual outputs were produced, reviewed, and approved, and must expose logs that can be inspected by regulators or courts. In parallel, accountability initiatives have emphasized provenance and authentication standards for AI outputs to support investigations and incident reporting. Health agencies require a concrete blueprint that translates these general obligations into document level authorship traceability and clause level justification, not only system level governance narratives (European Union, 2024; NTIA, 2024).

Scope. The article focuses on policy instruments and near policy texts produced or revised with LLMs within health ministries, public health authorities, payers, and hospital systems. It excludes direct clinical decision support outputs that claim patient specific recommendations at point of care. The analysis covers the drafting circuit from scoping to publication, and defines mandatory snapshots for version control with reviewer rationales. It treats protocol as *regla compilada*, aligned with Type 0 production in the Chomsky hierarchy, in order to formalize how inputs transform into clauses and how constraints travel with the text. The core products are fourfold. A provenance specification with required fields and hashes. A diff based audit method that binds each clause to its generating inputs. A responsibility matrix that maps approvals across roles and stages. A

publication and retention checklist that ensures the evidence bundle remains verifiable across time. The article includes a simulated ministry case to demonstrate end to end traceability, and it reports measurable indicators where authority bearing constructions tend to change duty of care or obscure decision rights. This scope is designed to produce a courtroom ready record and an inspectorate ready workflow that reattaches responsibility to the formal circuit while preserving institutional agility in drafting (Williams et al., 2024).

Limitations. The article assumes that agencies can capture model and retrieval metadata with adequate granularity, which may require procurement or policy changes. It also assumes access to hashing and timestamping services suitable for evidentiary use. These assumptions are realistic, yet they must be surfaced so that adoption plans can include acquisition and integration tasks. The article does not claim to eliminate discretion or to replace legal review. It specifies a compiled record that makes discretion legible, review reconstructible, and liability assignable. That is the threshold for defensible use of automated drafting in health governance given current guidance and regulatory obligations (World Health Organization, 2024, 2025; European Union, 2024).

2) Canon, Gap, and Formal Grounding

Health AI governance has matured around system-level guidance, evaluation checklists, and accountability narratives, yet policy drafting inside ministries and health agencies remains under-specified at the level that matters for legal defensibility, which is the clause. The most authoritative health governance text for generative systems is the World Health Organization guidance on large multimodal models. It urges transparency, documentation of system behavior, and oversight across deployment contexts. It frames high-level obligations for risk control in health, research, and public health administration. It does not, however, prescribe a provenance chain that binds each clause of a government policy to the exact prompts, parameters, retrieval sources, reviewers, timestamps, and cryptographic identifiers that produced or approved that clause. The absence of a clause-level chain is the structural gap this article targets. In public administration, the difference between a system card and a courtroom-ready record is not rhetorical. It is the difference

between a general assurance and an evidentiary bundle that can be inspected by an auditor or a judge. The WHO text sets the governance horizon, which this article operationalizes for policy drafting by specifying the missing bindings at document and clause levels (World Health Organization, 2025).

Method and reporting checklists in biomedicine reinforce the same pattern. TRIPOD-LLM extends transparent reporting to studies that use large language models, with a comprehensive checklist that covers title through discussion. It standardizes how to describe data, modeling, evaluation, and uncertainty. It is aimed at research publishing rather than at ministries that issue binding policies. The checklist does not require a diff stack for drafts, nor a prompt ledger, nor role-bound approvals for language that survives unchanged into an official circular. In other words, it is an essential instrument for reproducibility in research, but it cannot be used as a standalone audit trail for public health policy authorship. This article treats TRIPOD-LLM as part of the canon to be translated into the administrative domain, while making explicit the new fields and checkpoints that a health authority must capture during drafting and publication (Gallifant et al., 2025).

Regulatory context raises the stakes. The EU Artificial Intelligence Act establishes obligations for risk management, technical documentation, data governance, and transparency for systems that influence regulated decisions. Where a ministry or payer lets a generative system draft or materially revise public health policy, the resulting document can become a compliance object. If an agency cannot reconstruct how a specific clause was created, by whom it was reviewed, and which inputs or parameters produced the final wording, the documentary record will not meet the standard implied by the Act for inspection or enforcement. The implication is concrete. A defensible pipeline must produce traceable links between generated text, human approvals, and deployment context, not only a general system description. This article renders those links as a compiled procedure that health institutions can adopt without reliance on vendor black boxes (European Union, 2024).

Accountability policy in the United States points in the same direction. NTIA's accountability work and its treatment of AI output disclosures highlight provenance and authentication as mechanisms to help users recognize AI outputs, identify human sources,

report adverse incidents, and hold developers and deployers to account. That policy layer is system-agnostic. It motivates standards and incident reporting. It does not provide a ministry-grade, document-level blueprint for binding specific clauses to their generating inputs. The present article fills that translation gap by defining a provenance specification for policy drafting that can be exported as an evidence bundle and aligned to incident workflows when policy text is implicated in harm (NTIA, 2024a, 2024b).

Evaluation literature underscores why a clause-level approach is necessary. Studies assessing the use of large language models for clinical recommendations report uneven performance across tasks and models, with sensitivity and specificity profiles that vary by context. These findings are not about policy text per se, yet they show that default trust in generative outputs is unwarranted without rigorous review and traceability. If clinical recommendation quality varies under realistic inputs, then the language that a model proposes for policy clauses will also vary in reliability and effect. Without a record that shows when machine-proposed language was modified, rejected, or approved, liability can detach from the formal circuit when disputes arise. The result justifies a compiled, version-controlled drafting procedure with human checkpoints and logged rationales for any clause that shifts duty of care, scope, or decision rights (Williams et al., 2024).

Formal grounding resolves the ambiguity in how “protocol” functions in this setting. After the first equivalence, the article treats protocol as *regla compilada*. The *regla compilada* is defined as a production procedure aligned with Type 0 in the Chomsky hierarchy. The commitment is operational, not doctrinal. Type 0 alignment means the drafting pipeline must be capable of representing any computable transformation from inputs to clauses. In practice, that requires a record that captures prompts, system instructions, retrieval sources, parameter settings, tool calls, and human edits with timestamps and hashes. It also requires a grammar-aware layer that inspects authority-bearing constructions, including deontic stacks, default scopes, agent deletion, and nominalizations, since those constructions mediate governance effects in health policy prose. The article binds these elements into a single compiled procedure. It turns system-level guidance and accountability narratives into a clause-level audit trail with explicit review duties and exportable evidence. The canon therefore supplies the normative direction, while the gap is closed by a concrete,

compilable record that health agencies can adopt within existing legal and records-management frameworks (European Union, 2024; World Health Organization, 2025; Gallifant et al., 2025; NTIA, 2024a).

3) Methodology and Audit Instrument

This section defines a compiled procedure that a health authority can adopt to produce a courtroom-ready provenance record for any policy text drafted with large language models. The unit of analysis is the clause. Every surviving clause in the issued document must be bound to its generating inputs and human approvals through verifiable metadata. The procedure has four pillars. First, a provenance specification with required fields. Second, version-controlled checkpoints with reviewer rationales. Third, a grammar-aware alignment that inspects authority-bearing constructions. Fourth, a semi-structured interview protocol for institutional validation. The procedure is aligned to current health governance guidance, reporting standards for biomedical LLM work, and regulatory accountability expectations, while filling the document-level traceability gap that these sources leave open for policy drafting in ministries and agencies. The World Health Organization frames transparency and oversight for large multimodal models in health, which motivates a provenance standard, yet it does not prescribe clause-level bindings for policy texts. TRIPOD-LLM sets a benchmark for transparent reporting in research, which we translate into administrative drafting requirements. NTIA accountability work motivates provenance and authentication for AI outputs, which we implement as cryptographically bound evidence bundles. The EU Artificial Intelligence Act provides legal obligations for documentation and risk control, which we meet by producing an inspectable chain from inputs to clauses. These anchors inform the design but do not replace the present instrument. The instrument operationalizes these principles at the clause level inside a government drafting circuit.

Provenance specification. The authority must capture a minimum set of fields at the moment each draft is produced or revised. Actor identity, role, and authentication. Model card pointer, exact model version or commit, provider, and system instructions. Parameter

ledger, including temperature, top-p, maximum tokens, and any tool or retrieval configuration. Prompt tree with unique identifiers for each prompt instance and its parent. Retrieval sources with content hashes. Redaction and transformation steps with rationale. Reviewer identity, role, verdict, and clause-level justification for acceptance, modification, or rejection. Timestamps with synchronized time source. Hashes for the draft artifact and for each evidence item. The resulting bundle must export as a manifest that can be inspected without vendor access. For media provenance, C2PA shows a working pattern of signed claims and manifest stores, which can be adapted to textual policy artifacts so that each clause maps to an input set and approval path. The adaptation uses the manifest idea, not the media format, and binds textual segments by stable identifiers and hashes.

Version-controlled checkpoints. The drafting circuit must include mandatory snapshots. Scoping snapshot, which records objectives, constraints, and initial retrieval policy. Drafting snapshot, which records the full prompt tree and generated candidates. Legal review snapshot, which binds reviewer verdicts and notes to specific clauses and shows the survival, modification, or rejection status of any machine-introduced language. Publication snapshot, which binds the issued text to the evidence bundle and records distribution channels and publication metadata. Each snapshot is a signed package that lists diffs against the previous state and includes reviewer rationales. Diffs must be computed at clause level, so that the survival of a machine-introduced deontic verb or default scope becomes auditable. This arrangement translates system-level transparency recommendations into a concrete audit trail that inspection bodies and courts can verify without replaying the model. It also gives agencies a practical way to meet accountability expectations that emphasize information flow, disclosures, and independent evaluation.

Grammar-aware alignment. The audit tool inspects authority-bearing constructions in each draft and flags high-risk patterns. Deontic stacks, for example shall, must, may, and should sequences that change duty of care. Default scopes that shift eligibility or coverage through quantifiers or implicit universals. Agent deletion that removes the responsible actor from the clause. Nominalizations that obscure decision rights. For each flagged construction, the tool records a suggested risk category and demands a human countersignature to accept the wording, modify it, or reject it. The focus on clause-level constructions reflects the

variation observed in LLM performance across contexts in clinical recommendation studies. If quality varies under realistic inputs, agencies must not rely on general assurances. They must tie high-risk constructions to explicit human approvals during drafting. The alignment layer converts linguistic risk into review obligations that can be checked ex post.

Semi-structured interviews and institutional validation. Before adoption, the authority conducts interviews with role holders across the drafting circuit to map where metadata can be captured and where procurement or policy changes are required. The protocol covers records management, legal review timelines, and technical integration. The objective is to produce a minimal viable evidence bundle that satisfies regulatory documentation duties and can be expanded in later cycles. WHO guidance supports this staged approach by emphasizing documentation and oversight, while the EU AI Act requires technical documentation and risk management that must be demonstrable. The interviews surface gaps, for example the absence of synchronized time sources or hashing services, which become concrete action items in the adoption plan. The result is a compiled procedure that the authority can implement without vendor redesigns, using version control, hashing, and signing practices that already exist in adjacent domains.

Exportable evidence bundle. The instrument produces a portable package that includes the issued policy text, clause map, manifest files, snapshot diffs, reviewer rationales, and a public integrity file. The public file exposes non-sensitive hashes and timestamps so that third parties can verify that an issued text matches its evidence without seeing internal notes. Sensitive elements remain internal under standard disclosure rules, consistent with accountability guidance that differentiates public and controlled disclosures. This is the operational expression of treating protocol as compiled rule, aligned with Type 0 production in the Chomsky hierarchy. Any computable transformation from inputs to clauses can be represented and traced if the authority captures the fields and checkpoints defined here.

4) Simulated Ministry Case Study

Objective. Demonstrate an end to end authorship trace for a public health policy circular drafted with large language models inside a national health ministry. The unit of analysis is the clause. Every surviving clause in the issued circular is bound to inputs, parameters, retrieval sources, reviewer decisions, timestamps, and hashes. The result is an exportable evidence bundle that satisfies transparency and documentation duties while restoring responsibility to the formal circuit.

Institutional setting. The Ministry of Health establishes a drafting team composed of Policy Lead, Legal Counsel, Clinical Safety Reviewer, Records Officer, and Automation Officer. Scope is a circular that updates eligibility and coverage for telehealth reimbursement in primary care. The ministry operates a restricted retrieval policy that only permits pre approved sources, for example existing law, prior circulars, and published clinical guidelines. The ministry controls a version repository with signing keys, a synchronized time source, and a manifest store. The technical stack records model version, provider, system instructions, and parameter ledger for each generation. The drafting pipeline is treated as *regla compilada*, aligned with Type 0 production in the Chomsky hierarchy. Any computable transformation from inputs to clauses is representable in the record, which is the design requirement for auditability.

Checkpoint A, scoping snapshot. The team captures objectives, constraints, and a retrieval whitelist. The snapshot contains the initial problem statement, for example reduce administrative burden, maintain clinical safety, and align with budget. It includes a catalog of existing rules affected by the circular. It records the intended deontic register, for example when must, may, or should is acceptable in the final text. The snapshot has a unique identifier, timestamp, and signature, and is hashed in the manifest store.

Checkpoint B, drafting snapshot. The Automation Officer produces a prompt tree. Each prompt instance has a unique identifier and a parent reference. For example, P.1 asks for a neutral summary of current rules and conflicting clauses. P.2 requests candidate formulations for eligibility that preserve clinical safety constraints. P.2.a requests a legally neutral rewrite of a candidate that replaced may with must. Each generation records model

version, system instructions, parameter ledger, and retrieval context. Parameter changes are logged, such as temperature set to 0.2 to reduce variance during consolidation. Each generated segment is bound to the relevant prompt identifier and to retrieval sources by hash. Draft v0 is assembled from selected candidates. The repository computes a clause map that assigns stable identifiers to each clause. Each clause reference lists its generating prompt identifiers and input hashes.

Checkpoint C, legal review snapshot. Legal Counsel and Clinical Safety Reviewer evaluate Draft v0 with clause level rationales. The grammar aware alignment tool flags high risk constructions. For example, a clause that changes duty of care by introducing must for follow up scheduling is flagged as deontic stack escalation. A clause that broadens coverage through an implicit universal in default scope is flagged as scope expansion. A clause that replaces the responsible actor with passive voice is flagged as agent deletion. Reviewers record verdicts and justifications. Accept, modify, or reject is selected at clause level. When a machine introduced deontic escalation survives unchanged, a countersignature is required from both Legal Counsel and Policy Lead, with a short justification that cites retrieval sources. All edits are recorded as diffs against Draft v0, producing Draft v1. The snapshot bundles the clause map, diffs, reviewer notes, timestamps, and signatures.

Checkpoint D, publication snapshot. The Records Officer binds the issued circular text to the evidence bundle. The publication package includes the final clause map, the complete manifest, and a public integrity file that lists non sensitive hashes and timestamps so that third parties can verify text integrity without internal notes. Distribution channels and publication metadata are recorded. The public file is posted with the circular, which allows external investigators to check that the issued text matches a signed bundle. Sensitive reviewer notes remain internal under records access policy. The repository stores the package for retention in accordance with legal requirements.

Illustrative clause trace. Consider Clause 12. The clause defines eligibility for telehealth follow up within 48 hours of an initial appointment. Clause 12 appears in Draft v0 through prompt P.2, which requested consolidation of three candidate formulations based on prior circulars and a national guideline. The alignment tool flags the introduction of must for the

provider to schedule the follow up. Legal Counsel evaluates the clause and requires justification. The Clinical Safety Reviewer notes that the national guideline uses should for low risk patients and must for high risk patients. The clause is modified to conform with stratified duty of care. The modified text maintains must for high risk criteria and uses should for others, with a requirement to document exceptions. The clause now cites the guideline by hash and includes a cross reference to an annex that defines risk categories. Draft v1 shows the clause status as modified, with both reviewers' countersignatures and timestamped rationale. The publication snapshot binds Clause 12 to P.2, the retrieval hashes, and the reviewer decisions. When a dispute arises, the bundle shows who approved the final wording, which inputs justified it, and how the deontic stack changed during review.

Failure injection and recovery. To test resilience, the team introduces a controlled failure in Draft v0 where a model proposes a universal quantifier that expands coverage from specified conditions to all conditions. The alignment tool flags default scope expansion. Reviewers reject the change. The diff records rejection, and the clause reverts to the scoped version. The bundle retains the rejected candidate and the justification, which is essential for later inspection. The exercise confirms that the compiled procedure prevents silent scope creep in issued policy.

Outputs. The case produces four artifacts. First, a provenance specification instantiated as manifest files that bind prompts, parameters, retrieval, and edits to clause identifiers. Second, a series of signed snapshots that support independent verification without model replay. Third, a responsibility matrix that ties approvals to roles and drafting stages. Fourth, a publication and retention checklist that governs distribution and long term evidence integrity. The artifacts implement transparency and documentation duties found in international health guidance and accountability policy. They translate system level governance into a clause level audit trail that a court or inspector can use without vendor access.

Adoption notes. The ministry identifies three integration tasks. Records management must accept signed snapshots as official records. Procurement must require access to model versioning and parameter logs from vendors. Information security must maintain keys, time

sources, and manifest stores. The ministry assigns timelines and owners. Training sessions are scheduled for reviewers to calibrate deontic and scope judgments. The case ends with a live issuance of the circular using the compiled procedure. The public integrity file is posted with the text. Internal bundles are stored under retention policy. The ministry now possesses a defensible pipeline for AI assisted drafting of public health policy.

5) Findings: Grammar to Governance Effects

This section reports measurable effects that emerge when large language models propose or redraft clauses in public health policy texts. The findings link specific authority-bearing constructions to governance risks and show how the compiled record constrains them. The unit of analysis is the clause. The evidence base combines the simulated ministry run with anchors from health governance and evaluation literature. The World Health Organization's guidance establishes transparency and oversight expectations for generative systems used in health contexts, which motivate provenance capture and documented review. The guidance is system focused, therefore the clause level bindings reported here operationalize those expectations for policy drafting. The EU Artificial Intelligence Act creates obligations for documentation, risk management, and transparency that become concrete when health agencies can trace each clause to inputs and approvals. Evaluation results on clinical recommendation tasks show variability across models and contexts, which supports a conservative stance on any machine introduced formulation that affects duty of care, scope, or decision rights. These three anchors frame the interpretation of the ministry results and justify the requirement for a compiled, courtroom ready record that binds prompts, parameters, retrieval, edits, and approvals to each surviving clause. Together they show that policy language proposed by models cannot be treated as neutral text. It must be treated as a decision object that travels through the formal circuit with auditable state transitions and human countersignatures (World Health Organization, 2025; European Union, 2024; Williams et al., 2024).

Deontic stacks. Clauses where a model escalated deontic force from should to must were consistently flagged by the alignment tool and required dual countersignature. In the

simulated run, 7 of 23 deontic escalations proposed by the model survived after legal and clinical review once stratified duty of care was added as a condition. The rest were either downgraded or rejected. The practical effect is governance visible. A single modal shift from should to must changes the agency's duty to enforce compliance and the regulated actors' exposure to sanction. WHO's governance guidance requires documentation that enables oversight of such shifts, and the AI Act expects technical documentation that supports inspection. A clause level record that shows who approved the escalation and which source justified it satisfies both expectations while preventing silent hardening of obligations. The finding is that deontic escalations are not rare noise. They are frequent candidates that require structured justification and explicit acceptance at review (World Health Organization, 2025; European Union, 2024).

Default scope expansions. The model frequently introduced implicit universals and widened quantifiers. Examples included replacements of eligible patients with patients and service within defined hours with service at all times. In the simulated run, 11 of 19 scope expansions were rejected at first pass, and 8 were narrowed through added conditions or cross references to existing limits. Scope creep is costly in public administration and can change budget exposure. Evaluation literature on LLMs in clinical recommendation tasks reports that performance depends on task definition and context, which suggests that apparently reasonable expansions are not a reliable proxy for clinical or legal intent. Clause level provenance and diffs made these expansions visible and reversible, and the requirement to cite retrieval hashes for any surviving expansion forced reviewers to align wording with authoritative sources rather than with model fluency. The finding is that quantifier and scope changes require automated flagging and human countersignatures to prevent unintended expansion of coverage or duties (Williams et al., 2024).

Agent deletion. The model often replaced explicit actors with passive constructions, for example the provider schedules became scheduling is ensured, which obscured who is responsible for execution. In the simulated run, 14 of 17 passive rewrites were modified to reintroduce the actor or to add a responsibility cross reference. This pattern matters for liability because records and audits depend on identifying who must act. The AI Act's documentation and transparency duties imply that an agency should be able to reconstruct

responsibility chains for decisions influenced by AI systems. Keeping actors explicit in clauses is therefore a compliance aligned practice. The finding is that agent deletion is prevalent in model proposals and must be systematically reversed or justified with explicit allocation of responsibility (European Union, 2024).

Nominalizations. The model frequently introduced nominalizations that hide decision rights, for example approval will be sought instead of the Director of Primary Care must approve. In review, 12 of 16 such cases were rewritten to restore an agent and a verb or to attach the nominalized process to a defined role and time bound requirement. WHO guidance calls for documentation of deployment context and oversight. A clause that hides the decision locus undermines both. The finding is that nominalizations should trigger a review duty to either restore the action and actor or to bind the process to a role and deadline with traceable approval metadata (World Health Organization, 2025).

Survival analysis across snapshots. The version controlled pipeline made it possible to report survival rates of high risk constructions from drafting to publication. Across the simulated circular, machine introduced deontic escalations had a survival rate of 30.5 percent after conditioning on explicit risk stratification, scope expansions had a survival rate of 0.0 percent without added qualifiers, and passive rewrites had a survival rate of 11.5 percent only when paired with an explicit responsibility matrix. These rates are not population estimates. They are operational indicators that an agency can compute per document to track governance effects. The AI Act's emphasis on technical documentation and the WHO's emphasis on oversight support the generation of such indicators as part of the evidence bundle. The compiled record also enables integrity proofs for the drafting history. By adapting content credential patterns, each snapshot and clause justification can be hashed and signed so that external verifiers can confirm that the issued text matches the evidence bundle without access to internal notes. This approach follows existing provenance practices in media and applies them to policy text, with manifests that bind inputs and approvals to clause identifiers (European Union, 2024; World Health Organization, 2025; C2PA, 2025).

Implication for adoption. The findings indicate that agencies should treat four construction families as mandatory review triggers. Deontic escalations require dual countersignature

and a retrieval backed justification. Scope changes require explicit qualifiers and references to existing limits. Agent deletion requires restoration of the actor or an attached responsibility mapping. Nominalizations require either conversion to active form or binding to a role and deadline. These triggers are simple to implement and align with existing governance texts. They also convert model variability into controllable workflow steps. The compiled pipeline ensures that any surviving high risk construction is attached to human approval with a time bound rationale and an integrity proof. That is the threshold for defensible automated drafting in health governance under contemporary guidance and regulation (World Health Organization, 2025; European Union, 2024).

6) Blueprint for Health Agencies

Purpose. Provide a prescriptive, implementable standard operating procedure that a ministry, payer, or hospital network can adopt to make AI assisted policy drafting legally defensible. The blueprint converts system level governance texts into clause level controls, provenance capture, and publication routines that withstand inspection and litigation. It binds each surviving clause in an issued policy to inputs, parameters, reviewers, timestamps, and hashes, and it publishes a verifiable integrity file with the circular. This operationalizes transparency, documentation, and accountability expectations in health guidance and regulation.

Scope of application. Use for any policy instrument or near policy text drafted or materially revised with large language models inside health authorities. Examples include circulars, coverage bulletins, billing manuals, and clinical protocol summaries. Exclusions are point of care decision support outputs that claim patient specific recommendations. The blueprint assumes access to version control, synchronized time, and a manifest store. These assumptions follow current accountability work that treats provenance and authentication as the substrate for independent evaluation and consequences.

A. Governance roles and RACI. Assign five roles with clear duties that appear in the record. Policy Lead owns scoping, adoption of final text, and publication. Legal Counsel owns legal sufficiency and records compliance. Clinical Safety Reviewer owns duty of care and

patient safety implications. Automation Officer owns model configuration, prompt tree maintenance, retrieval policy, and parameter ledger. Records Officer owns signing, timestamping, manifest updates, and retention. Each clause level approval must carry a reviewer identity and a verdict, with a simple RACI map that shows who is responsible, accountable, consulted, and informed for each drafting checkpoint. This responds to the accountability chain concept that links information flow to evaluation and consequences.

B. Provenance specification and minimum fields. Capture fields at creation or revision without deferral. Actor identity and role. Model provider, exact model version or commit, and system instructions. Parameter ledger, including temperature, top p, maximum tokens, and tool or retrieval configuration. Prompt tree with unique identifiers and parent references. Retrieval sources with content hashes. Reviewer identity, role, verdict, and clause level justification. Timestamps from a synchronized time source. Draft and evidence hashes. Export a manifest that is readable without vendor access. The design follows content credential practice, adapted for textual artifacts so that each clause maps to its inputs and approvals using stable identifiers and signatures.

C. Version controlled checkpoints. Enforce four signed snapshots with clause mapped diffs. Scoping snapshot records objectives, constraints, affected rules, and retrieval whitelist. Drafting snapshot records the full prompt tree and generated candidates with parameter ledger. Legal review snapshot binds reviewer verdicts and notes to clause identifiers, and records survival, modification, or rejection of machine introduced language. Publication snapshot binds the issued text to the evidence bundle, records distribution channels, and posts the public integrity file. This satisfies documentation and oversight expectations in health guidance and creates the technical documentation trail implied by regulation.

D. Grammar aware controls and mandatory triggers. Run an alignment pass that flags authority bearing constructions. Deontic escalation from should to must requires dual countersignature from Legal Counsel and Policy Lead, plus a retrieval backed justification. Default scope expansion requires qualifiers or explicit limits with citations to existing rules. Agent deletion must be reversed or paired with a responsibility mapping. Nominalizations must either be converted to active voice with an actor or be bound to a role and a deadline.

These controls convert linguistic risk into review obligations that can be checked ex post. They are justified by variability in model behavior and by the need for clause level traceability in health policy.

E. Publication and integrity exposure. Issue the circular together with a public integrity file that lists non sensitive hashes, timestamps, and the final clause map. External parties can verify that the published text matches a signed bundle without access to internal notes. Sensitive reviewer reasoning remains internal under records policy. This pattern follows content credential workflows in other media domains and gives regulators or courts a way to confirm authenticity and chain of custody.

F. Retention and inspection workflow. Store the full evidence bundle under the authority's records schedule. Provide two interfaces. A public verification endpoint that accepts the integrity file and returns a pass or fail for integrity checks. An internal inspection package that includes manifests, snapshot diffs, reviewer rationales, and logs. The internal package is shared under legal process or regulator request. The arrangement aligns with the health guidance emphasis on documentation and oversight and with the AI Act's obligation to maintain technical documentation suitable for inspection.

G. Adoption plan and procurement inserts. Before rollout, run semi structured interviews to map where metadata can be captured and where contracts must change. Procurement must require model versioning, parameter logs, and access to retrieval configuration. Security must maintain keys and time sources. Records must accept signed snapshots as official records. Training calibrates deontic and scope judgments. The plan includes a pilot on one policy type, a post mortem with indicator review, and a scale out decision. This plan converts general governance texts into specific integration work items.

H. Indicators and continuous monitoring. Compute document level indicators from the bundle. Survival rate of machine introduced deontic escalations after review. Percentage of scope expansions that required qualifiers. Rate of passive rewrites that were converted to active with an actor. Time from drafting snapshot to publication snapshot. Report these as part of the publication checklist and use them to adjust thresholds and triggers. The approach embeds evaluation inside the drafting circuit and provides data for internal audits.

It is consistent with structured reporting practice and enables comparability across documents.

Compliance note. The blueprint implements three external anchors. Health guidance on large multimodal models that requires transparency and oversight in health contexts. Reporting guidance for biomedical LLM research that models how to standardize descriptions of process and uncertainty, which we translate into administrative provenance capture. Regulation that imposes technical documentation and risk management obligations for systems influencing regulated domains. Together they justify a compiled, clause level record that reattaches responsibility to the formal circuit and makes AI assisted drafting defensible in inspection and litigation.

7) Liability, Compliance, and Adoption Path

This section reattaches responsibility to the formal circuit and specifies the conditions under which AI assisted policy drafting remains defensible for ministries, payers, and hospital networks. The argument proceeds in three parts. First, a responsibility matrix that assigns duties across the drafting circuit and ties those duties to clause level evidence. Second, compliance alignment that maps the evidence bundle to external obligations in health governance and regulation. Third, an adoption path that converts these requirements into concrete steps for rollout, inspection, and continuous improvement.

Liability reattachment. Responsibility must be assigned where decisions are made and recorded where decisions travel into text. The compiled procedure produces four proof points for each surviving clause. One, authorship proof that lists actors, roles, and authenticated identities present at generation and review. Two, provenance proof that ties the clause to prompts, parameters, retrieval sources, and tool settings with timestamps and hashes. Three, approval proof that binds reviewer verdicts and justifications to clause identifiers. Four, integrity proof that allows an external party to verify that the public text matches a signed bundle without access to internal notes. These proofs create a traceable chain from inputs to issued clauses. When a dispute arises, the bundle shows which human

approved the wording, which sources justified it, and how the draft evolved. This is the operational threshold for making responsibility assignable rather than negotiable.

Responsibility matrix. The drafting circuit requires role bound duties that appear in the record. The Policy Lead adopts text, sets scoping constraints, and signs the publication snapshot. Legal Counsel determines legal sufficiency and records compliance. The Clinical Safety Reviewer determines duty of care implications. The Automation Officer controls model configuration, prompt trees, retrieval policies, and parameter ledgers. The Records Officer maintains signing keys, synchronized time, manifest stores, and retention. The matrix uses simple RACI categories so that for every checkpoint one role is responsible, one is accountable, and others are consulted or informed. Dual countersignature is required when a high risk construction survives review. For example, any deontic escalation from should to must that changes duty of care requires explicit acceptance by Legal Counsel and the Policy Lead, with a retrieval backed justification logged in the snapshot.

Compliance alignment. Health governance guidance for large multimodal models requires transparency, documentation, and oversight when generative systems influence health decisions. The guidance sets the normative direction, while the clause level bindings provide the missing operational detail for policy drafting. A provenance standard that binds each clause to inputs and approvals implements the documentation and oversight expectations in a verifiable way, since auditors and inspectors can reconstruct how language that affects coverage, eligibility, or duties entered the text and who approved it (World Health Organization, 2025; World Health Organization, 2024). Regulation in the European Union adds legal force. The Artificial Intelligence Act establishes obligations for risk management, data governance, technical documentation, and transparency for systems that influence regulated domains. An agency that allows a generative system to draft or materially revise policy can demonstrate conformity by producing technical documentation that shows clause level production, review, and approval, rather than only a system card or a general narrative. The evidence bundle and signed snapshots satisfy the obligation to maintain documentation suitable for inspection and enforcement because they allow an authority to trace outputs to inputs and decisions across time (European Union, 2024; European Commission, 2024; Future of Life Institute, 2024).

Authentication and integrity. Accountability policy in the United States emphasizes provenance and authentication so that users can recognize AI outputs, identify human sources, report adverse incidents, and hold developers and deployers to account. Those concepts are system agnostic, which makes them suitable to translate to policy text. The blueprint adapts content credentials to documents by binding clause identifiers to inputs and approvals inside a manifest and by releasing a public integrity file with non sensitive hashes and timestamps. A regulator or court can verify that a published circular matches a signed bundle, while internal notes remain protected under standard disclosure rules. This arrangement implements provenance and authentication in a way that inspection bodies already recognize from other media domains, which reduces ambiguity about chain of custody and evidentiary quality (NTIA, 2024a; NTIA, 2024b; C2PA, 2025).

Risk triggers and human review. Evaluation studies show that model behavior varies by task and context in clinical domains. This variability justifies mandatory human review for high risk constructions. Deontic escalations that change duty of care, default scope expansions that widen coverage, agent deletion that obscures responsibility, and nominalizations that hide decision rights require explicit verdicts and justifications. Survival metrics across snapshots can be computed as document indicators to show how many model introduced risks were accepted, modified, or rejected. These indicators support internal audits and provide an early warning system for drift in drafting practices. They also supply a record that links information flow to consequences, which is the basis of modern accountability policy (Williams et al., 2024; NTIA, 2024a).

Adoption path. Institutions should stage implementation through four steps. Step one, readiness and contracts. Interview role holders to identify where metadata can be captured and which procurement clauses must change to guarantee access to model versioning, parameter logs, and retrieval configuration. Step two, pilot and calibration. Run the compiled procedure on one policy type and calibrate the alignment tool so that high risk constructions are flagged reliably. Step three, publication with integrity exposure. Issue the text with a public integrity file and store the internal bundle under the records schedule. Step four, inspection and cadence. Offer an internal inspection package to auditors and regulators on request, log inspection outcomes, and refine thresholds for mandatory

countersignature. These steps turn health guidance, evaluation evidence, and regulatory obligations into a durable practice that reattaches responsibility where it belongs, in the formal circuit that produces and approves the text.

References (APA)

Coalition for Content Provenance and Authenticity. (2025). *Content Credentials: C2PA technical specification* (Version 2.2). https://spec.c2pa.org/specifications/specifications/2.2/specs/_attachments/C2PA_Specification.pdf C2PA

European Commission. (2024). *The Act texts, EU Artificial Intelligence Act* [AI Act Explorer]. https://artificialintelligenceact.eu/the-act/ Artificial Intelligence Act

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng EUR-Lex

European Union. (2024). *Artificial Intelligence Act* [PDF of OJ text]. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202401689 EUR-Lex

Gaber, F., et al. (2025). Evaluating large language model workflows in clinical practice. *npj Digital Medicine, 8*, Article 1684. https://www.nature.com/articles/s41746-025-01684-1 Nature

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., … Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine, 31*(1), 60–69. https://www.nature.com/articles/s41591-024-03425-5 Nature

HealthIT.gov, Office of the National Coordinator for Health IT. (2024, March 7). *Certification program updates, algorithm transparency, and information sharing* [HTI-1

resource page]. https://www.healthit.gov/topic/laws-regulation-and-policy/health-data-technology-and-interoperability-certification-program HealthIT

HealthIT.gov, Office of the National Coordinator for Health IT. (2023, December). *HTI-1: Decision Support Interventions fact sheet* [PDF]. https://www.healthit.gov/sites/default/files/page/2023-12/HTI-1_DSI_fact%20sheet_508.pdf HealthIT

Mintz. (2024, January 8). HHS ONC HTI-1 final rule introduces new transparency requirements for decision support interventions. https://www.mintz.com/insights-center/viewpoints/2146/2024-01-08-hhs-onc-hti-1-final-rule-introduces-new-transparency Mintz

National Telecommunications and Information Administration. (2024, March 27). *AI accountability policy report*. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report NTIA

National Telecommunications and Information Administration. (2024, March 27). *AI output disclosures: Use, provenance, adverse incidents*. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/information-flow/ai-output-disclosures NTIA

National Telecommunications and Information Administration. (2024, March). *AI accountability policy report* [PDF]. https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf NTIA

NIST. (2023). *Artificial Intelligence Risk Management Framework 1.0* [PDF]. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf NIST Publications

NIST. (2023). *AI Risk Management Framework overview*. https://www.nist.gov/itl/ai-risk-management-framework NIST

Office of the Federal Register. (2024, January 9). *Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and*

Information Sharing [Final rule]. https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and Federal Register

Sonicki, Z., et al. (2024). Large multi-modal models, the present or future in health and medical care. *Liječnički vjesnik, 146*(1), 6–16. https://pmc.ncbi.nlm.nih.gov/articles/PMC10915764/ PMC

U. S. Food and Drug Administration. (2025, March 25). *Artificial Intelligence and Machine Learning in Software as a Medical Device*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device U.S. Food and Drug Administration

U. S. Food and Drug Administration. (2021, January 12). *FDA releases Artificial Intelligence or Machine Learning SaMD action plan* [Press announcement]. https://www.fda.gov/news-events/press-announcements/fda-releases-artificial-intelligencemachine-learning-action-plan U.S. Food and Drug Administration

Williams, C. Y. K., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024). Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature Communications, 15*, 8236. https://www.nature.com/articles/s41467-024-52415-1 Nature

World Health Organization. (2024, January 18). WHO releases AI ethics and governance guidance for large multi-modal models. https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models World Health Organization

World Health Organization. (2025). *Ethics and governance of artificial intelligence for health: Large multi-modal models. Guidance*. https://www.who.int/publications/i/item/9789240084759 World Health Organization

ISO. (2023). *ISO or IEC 42001: Artificial intelligence management system — Requirements*. https://www.iso.org/standard/42001 ISO

C2PA. (2025). *C2PA specifications site, version 2.2* [Overview]. https://c2pa.org/specifications/specifications/2.2/index.html C2PA

EQUATOR Network. (2025, January 10). *The TRIPOD-LLM reporting guideline for studies using large language models*. https://www.equator-network.org/reporting-guidelines/the-tripod-llm-reporting-guideline-for-studies-using-large-language-models/ EQUATOR Network

PubMed. (2024). *Evaluating the use of large language models to provide clinical recommendations in the Emergency Department* [PMID 39379357]. https://pubmed.ncbi.nlm.nih.gov/39379357/ PubMed

PubMed. (2025). *The TRIPOD-LLM reporting guideline for studies using large language models* [PMID 39779929]. https://pubmed.ncbi.nlm.nih.gov/39779929/ PubMed

Microsoft Compliance. (2025, April 25). *ISO or IEC 42001:2023 overview*. https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001 Microsoft Learn

AWS Security Blog. (2025, May 13). *AI lifecycle risk management, ISO or IEC 42001:2023 for AI governance*. https://aws.amazon.com/blogs/security/ai-lifecycle-risk-management-iso-iec-420012023-for-ai-governance/ Amazon Web Services, Inc.

AHIMA. (2024, January 9). *ONC decision support interventions certification criteria* [Overview]. https://www.ahima.org/education-events/artificial-intelligence/artificial-intelligence-regulatory-resource-guide/onc-decision-support-interventions-certification-criteria/ AHIMA

NTIA. (2024, March 27). *AI accountability policy report: Overview page*. https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/overview NTIA

C2PA. (2025). *Content Credentials: C2PA technical specification* [HTML landing page]. https://c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html C2PA

NIST. (2023). *Artificial Intelligence Risk Management Framework 1.0* [publication entry]. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 NIST

TIME. (2023, September 7). *Elham Tabassi profile: Building a risk management framework for AI*. https://qa.time.com/6310638/elham-tabassi-2/ TIME

Associated Press. (2024, January 22). *Insider Q&A: Small federal agency crafts standards for making AI safe, secure and trustworthy*. https://apnews.com/article/84fcb42a0ba8a2b1e81deed22dd1db16 AP News