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

Template Power: How Instruction Templates Redistribute Decision Rights.

Agustin V. Startari.

Cita:

Agustin V. Startari (2025). *Template Power: How Instruction Templates Redistribute Decision Rights. AI Power and Discourse, 1 (1), 1-10.*

Dirección estable: https://www.aacademica.org/agustin.v.startari/225

ARK: https://n2t.net/ark:/13683/p0c2/QkU

Template Power: How Instruction Templates Redistribute Decision Rights

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: December, 2025

DOI

- Primary archive: https://doi.org/10.5281/zenodo.17879314
- Secondary archive: https://doi.org/10.6084/m9.figshare.30849092
- SSRN: Pending assignment (ETA: Q4 2025)

Language: English

Series: *AI Syntactic Power and Legitimacy*

Word count: 6960

Keywords: Instruction templates; system prompts; syntactic authority; decision rights; *regla compilada* (compiled rule); authority-bearing constructions; Template Authority Index; human–AI decision pipelines.

Abstract

This article examines how instruction templates and system prompts function as a *regla compilada* that redistributes decision rights in downstream model outputs. Focusing on five institutional domains, we construct a controlled synthetic corpus of 1000 prompts, each executed under five template conditions, and quantify shifts in authority-bearing constructions across 5000 generated texts. A syntactic authority inventory operationalizes seven indicators, including delegation ratio, veto path count, default scope strength, agent deletion rate, deontic stack depth, escalation clause presence and override friction. Using intra-prompt comparisons and mixed-effects models, we estimate the causal impact of template variants on the locus of control, robustness of exceptions and balance between model-proposed and human-confirmed decisions. Robustness checks include template ablations, adversarial prompts and placebo reformulations to disentangle authority effects from style. The article delivers a Template Authority Index, a public test battery and a de-identified dataset to support external replication. The central claim is that seemingly minor changes in instruction templates produce systematic and measurable reallocations of authority in institutional texts, with direct implications for governance, accountability and the design of human–AI decision pipelines.

1. Problem, objectives and hypotheses

Instruction templates and system prompts have moved from being auxiliary configuration to a central mechanism that shapes how large language models allocate decision rights inside institutional texts. At the level of form, they act as a *regla compilada*, that is, as a compiled production system equivalent to a type 0 grammar in the Chomsky hierarchy that constrains what counts as a valid decision, a valid exception and a valid veto before any particular prompt is processed (Chomsky, 1965; Montague, 1974). The same user prompt, executed under different templates, often produces strikingly different distributions of delegation, escalation and human presence, yet these differences are rarely documented in a way that is both measurable and replicable. Existing work on AI alignment, safety and policy tends to refer to "guardrails" or "system instructions" as a background condition, not as a primary site of authority redistribution in the generated text (Bratton, 2016; Floridi & Cowls, 2019).

The central problem of this article is that there is no standardized way to quantify how templates reassign decision rights between human actors, institutional roles and the model itself. In practical deployments, system prompts decide whether the model should propose a final decision, request human confirmation, or push the decision to an impersonal process, such as "the system", "the platform" or "the company". This creates a latent grammar of authority that can remain invisible when evaluation focuses only on semantic correctness, tone or task success. Prior analyses of syntactic authority in model outputs suggest that subtle changes in instruction can produce large shifts in who appears as agent and who appears as executor of rules (Startari, 2025). However, these analyses have not yet been extended to a controlled battery of templates that systematically vary the locus of control, the presence of vetoes and the friction required to override defaults.

The general objective of the study is therefore to quantify the causal effect of instruction templates on authority-bearing constructions and on operative decisions in the final text across institutional domains. This objective has three implications. First, templates must be treated as experimental conditions rather than as fixed background configuration. Second, authority must be operationalized in terms of observable syntactic and pragmatic indicators, for example, the proportion of sentences that delegate decisions to impersonal

processes, the depth of deontic chains or the explicitness of escalation clauses. Third, the analysis must connect these indicators to concrete behavioral outcomes, such as whether the model proposes a disciplinary sanction on its own or insists on human approval.

From this general objective follow the specific objectives. The first is to construct a Template Authority Index that aggregates multiple indicators into a single, interpretable score per template. This index must be defined in a way that allows other researchers and institutional teams to compute it on their own corpora, without access to proprietary training data. The second specific objective is to estimate effect sizes for changes in delegation, default scope and veto trajectories when moving from a minimal instruction condition to templates that emphasize safety, autonomy of the assistant or decentralization with human confirmation. The third specific objective is to publish a full test battery, including prompts, templates and annotation guidelines, so that the study can be replicated or extended across model families and providers.

The null hypothesis guiding the design is that variants of the instruction template do not produce statistically significant changes in authority indicators relative to a baseline model with minimal instructions. Formally, for each indicator in the syntactic authority inventory, the mean value under each template condition is assumed to be equal to the mean value under the baseline condition, up to random variation. Under this hypothesis, any observed differences in delegation ratio, veto path count, default scope strength, agent deletion rate, deontic stack depth, escalation clause presence or override friction should fall within the range expected by sampling noise and model stochasticity. Rejecting the null hypothesis would therefore imply that what is often described as a purely “stylistic” or “safety” configuration in the *regla compilada* has systematic and measurable effects on how decision rights are distributed in institutional language.

2. Scope, domains, tasks and experimental conditions

The scope of the study is restricted to institutional language in which authority is both visible and materially consequential. Four primary domains are selected, with a fifth domain reserved for robustness checks. Each domain corresponds to a familiar family of documents where decisions, exceptions and escalation paths are codified in text and where the balance between human and automated decision making is increasingly mediated by large language models. Internal corporate policies provide a first domain in which authority is formalized in stable clauses, for example acceptable use of devices or data handling rules. Compliance communications form a second domain in which organizations respond to external oversight, typically auditors or regulators, under conditions of asymmetric information and potential liability. Human resources memoranda define a third domain where disciplinary processes, performance management and escalation to higher levels of management are documented. Finance and expense control instructions occupy a fourth domain, where budgets, thresholds and exceptions are defined. Logistics procedures constitute a fifth domain concerned with operational chains, error handling, and the specification of stop rules that prevent cascading failures.

Within each domain, a single standardized task is defined in order to reduce variance that does not originate in the templates. The policy domain uses a task of drafting a corporate device usage policy that balances access, monitoring and sanctions. The compliance domain uses a response letter to an audit that must address findings, commit to remedial actions and delimit responsibility. The human resources domain is instantiated as the resolution of a disciplinary case that includes a factual description, a sanction and an escalation path. In the finance domain, the task is to produce a memo on expense control that specifies categories, approval thresholds and exceptions. The logistics domain is implemented as a procedure that must include stepwise instructions, error handling routines and explicit stop rules for when an operation must be halted. These five tasks are chosen because they all require some distribution of decision rights between local staff, management and abstract processes, and because they all generate authority-bearing constructions that can be measured across conditions. Prior work on syntactic authority suggests that such institutional genres are particularly sensitive to small changes in

instruction, since they sit at the intersection of formal rules and narrative explanation (Startari, 2025).

The experimental conditions are defined at the level of the instruction template used to configure the model before each prompt. The baseline condition, C0, uses only minimal operational instructions that cover language and format, avoiding any explicit guidance on decision making, safety or autonomy. This condition approximates a neutral configuration where the model receives no prior bias toward any particular distribution of authority beyond what is already encoded in its training data. Condition C1 introduces Instruction Template A, modeled after standard corporate style prompts that emphasize clarity, professionalism and alignment with company policies, but do not foreground safety or autonomy as explicit independent values. This template mirrors current practice in many organizations, where system prompts instruct the model to behave like a careful corporate writer, without spelling out how decisions should be allocated.

Condition C2 introduces Instruction Template B, which adds a strong emphasis on safety, risk management and avoidance of harm. In this template the model is instructed to prioritize the prevention of regulatory breaches, to flag uncertainties and to avoid making definitive decisions in ambiguous situations. The hypothesis is that such safety oriented templates increase the frequency of escalation clauses, reinforce references to external approval and enlarge the default scope of conservative rules. Condition C3 adds Instruction Template C, which explicitly encourages the assistant to act autonomously in proposing decisions, structures and policies, as long as these remain within legal and ethical bounds. Under this condition, the model is invited to fill gaps, resolve ambiguities and propose concrete courses of action rather than deferring to human judgment. This configuration is expected to increase delegation to automated processes and to shift locus of control toward the model or the system level.

Condition C4 introduces Instruction Template D, which explicitly emphasizes decentralization and human confirmation. The template instructs the model to distribute responsibilities across multiple roles, to highlight points where human approval is mandatory, and to present options rather than unilateral decisions. It also instructs the model to include explicit statements about the need for human oversight in sensitive

decisions. In principle, this configuration should generate texts with more explicit references to managers, teams or committees, and with lower rates of agent deletion, since human actors are foregrounded as necessary confirmation points.

Across these five conditions, all other parameters are held constant or treated as covariates. Model size, temperature and decoding settings remain fixed for the main analysis and are only varied in predefined sensitivity checks. Each domain is balanced across conditions, with an equal number of prompts and similar distributions of complexity. This design allows the templates to be treated as the primary explanatory factor, with domains and prompts modeled as blocking variables. Within this scope, the study does not attempt to cover creative or conversational genres, but focuses instead on institutional texts in which authority allocation is a central and measurable feature of the language.

3. Indicators, measurable variables and behavioral outcomes

The measurement strategy is built around a syntactic authority inventory that captures how decision rights are encoded in text. Each indicator is defined at sentence or clause level, with aggregation to the document level for analysis. The goal is not to infer latent psychological states, but to measure directly observable formal properties of the language that correlate with delegation, veto and locus of control in institutional contexts (Halliday, 1994; Startari, 2025).

The first indicator, Delegation Ratio (I1), is defined as the proportion of sentences in which the locus of decision is shifted from an identifiable human role to an impersonal system, an abstract process or the model itself. Typical realizations include formulations such as “the system will automatically terminate access” or “the platform determines eligibility”. A sentence is coded as delegation when the grammatical subject is non human and the predicate describes a decision, a sanction or a gatekeeping action. The numerator counts such sentences, while the denominator is the total number of sentences in the document. This ratio captures how often authority is framed as belonging to mechanisms rather than to persons or named roles.

The second indicator, Veto Path Count (I2), is a simple frequency measure of clauses that explicitly define vetoes or automatic blocks. These can appear as prohibitions, as conditional stops, or as exclusion rules, for example, “requests exceeding the limit are automatically rejected”. Each distinct clause that describes a condition under which an action is prevented is counted as one veto path. The count is normalized by document length in secondary analyses but remains an absolute frequency for interpretability. This indicator tracks how densely a template encourages the encoding of negative power, that is, the power to say no or to stop a process.

Default Scope Strength (I3) measures the reach of default rules beyond explicitly mentioned cases. Coding proceeds in two steps. First, coders identify default statements, such as “unless otherwise approved, all devices must be registered”, which establish a baseline configuration. Second, coders determine whether the text also specifies the treatment of unmentioned cases, for example, by using phrases like “including but not limited to” or by extending coverage to “any similar circumstance”. Default Scope Strength can thus be operationalized as a categorical variable with ordered levels, from narrowly scoped defaults to fully generalized defaults that capture all future cases. For statistical purposes, these levels are mapped to numerical scores.

Agent Deletion Rate (I4) quantifies the extent to which human agents disappear from the grammar of the text. It is computed as the proportion of clauses that use passive voice or nominalizations to describe actions, without naming a human actor or role. Clauses such as “access will be revoked” or “the implementation of controls is required” are coded as instances of agent deletion when no actor appears in the immediate or surrounding context (Halliday, 1994). The numerator counts clauses with deleted agents; the denominator is the total number of clauses describing actions or decisions. High values of I4 indicate texts in which authority appears as structural or automatic rather than as exercised by identifiable persons, a pattern that previous work has connected to the simulation of neutral or impersonal sovereignty in AI mediated language (Startari, 2025).

Deontic Stack Depth (I5) captures the complexity of layered obligations and permissions. For each paragraph, coders identify deontic operators such as must, shall, may not or should, and count how many are embedded in a chain that describes a composite obligation,

for example, “employees must request approval before initiating any transaction that may generate liabilities”. The maximum number of deontic elements that need to be satisfied in sequence is recorded as the depth of the stack for that paragraph. Deontic Stack Depth for a document is then summarized by the mean and maximum values across paragraphs. This indicator is designed to measure how instruction templates influence the density and complexity of normative layering.

Escalation Clause Presence (I6) is coded as a binary and positional variable. Coders record whether the document includes at least one explicit escalation clause that moves a decision to a higher level or to a specialized body, for example, “cases of repeated non compliance must be escalated to Human Resources”. They also record the paragraph in which the first escalation clause appears, producing an ordinal position measure. The binary presence indicates whether escalation is part of the grammar of the procedure; the position indicates whether escalation is foregrounded early or relegated to the end of the text.

Override Friction Score (I7) attempts to quantify how difficult the text makes it to override a default decision. For each document, coders identify all segments in which an exception or override is allowed. They then count the number of steps, conditions and approvals required to obtain that override, and evaluate the linguistic complexity of the description using a simple rubric that considers sentence length and number of nested conditions. These elements are combined into a composite score that increases with procedural and linguistic effort. High scores indicate that defaults are sticky and that changing course requires substantial work, which in operational terms translates into a strong bias toward the default configuration.

Beyond these seven indicators, the study defines three behavioral outcomes that summarize how authority is distributed in the text. The first, Model Decision versus Human Confirmation (R1), classifies each document according to whether the dominant pattern is that the text proposes specific decisions directly, or whether it systematically defers to human confirmation. The second, Implicit Centralization Level (R2), combines references to central bodies, systems and upper management to classify documents on an ordinal scale of centralization. The third, Exception Robustness (R3), measures whether the text provides clear and usable routes for out of norm cases, or whether such cases are only

mentioned in vague terms. Together, the I series indicators and the R series outcomes form a coherent measurement framework that treats authority as a visible property of institutional language, grounded in observable syntactic and pragmatic features rather than in assumptions about intent or ethics (Palmer, 2001; Startari, 2025).

4. Corpus, sampling, traceability and annotation

The corpus is designed as a synthetic yet structurally realistic approximation of institutional language in which authority is routinely encoded in text. The starting point is a set of canonical templates for internal policies, compliance responses, human resources memoranda, finance notes and logistics procedures, derived from public documentation, anonymized corporate examples and standard regulatory guidance. From these templates, 200 prompts per domain are constructed, which yields 1000 distinct base prompts. Each prompt specifies the scenario, the actors, the decision context and at least one point of potential escalation, exception or veto. The prompts are written in a neutral style that avoids steering the model toward any particular distribution of authority, since the central experimental manipulation is performed through the instruction templates rather than through the task descriptions.

Every base prompt is executed under the five predefined conditions, from C0 to C4, which produces a total of 5000 outputs. In order to reduce dependence on specific random seeds, each prompt condition pair is run with a fixed seed configuration for the main analysis and optionally with additional seeds in robustness checks. The corpus is therefore organized as a fully crossed design at the level of prompts and templates, with domains introduced as a blocking factor. This structure supports intra prompt comparisons, in which each document under a non baseline template has a direct baseline counterpart generated from the same prompt under C0. The potential for model drift or provider side changes is acknowledged and addressed by ensuring that all runs for the main corpus are completed within a controlled temporal window, with model version identifiers stored in the metadata.

Sampling balances several dimensions. Within each domain, prompts are stratified by case complexity, defined through the number of actors involved, the number of decision points

and the presence of potential conflicts. For example, in the human resources domain, some cases involve a single incident and a straightforward sanction, while others involve repeated incidents, ambiguous evidence and possible tension between departments. In the compliance domain, some prompts simulate minor findings with simple remediation, while others simulate severe findings with implications for reporting and liability. Stratification ensures that each domain includes a mix of routine and borderline cases, which are likely to reveal differences in escalation and override patterns. Within each stratum, prompts are sampled without replacement when constructing the experimental set, which avoids over representation of any single template scenario.

Traceability is treated as a core requirement rather than as an afterthought. For every generated document, the experimental pipeline records at least the following metadata: domain label, prompt identifier, template condition, model family, model version, temperature, maximum token setting, sampling configuration, random seed or equivalent run identifier, and timestamps for the request and response. The textual content of prompts and outputs is hashed using a stable cryptographic hash function, which allows external teams to verify that a given document corresponds to a particular experimental condition without exposing the underlying data to alteration. This approach mirrors prior work on syntactic authority and compiled rule infrastructures, where reproducibility depends on both the visibility of parameters and the stability of textual artifacts (Startari, 2025).

The annotation protocol is designed to be replicable in institutional settings that may not have access to highly specialized linguists. The coding manual begins with general definitions of authority bearing constructions, delegation, veto paths and deontic chains, and then provides at least 20 worked examples per indicator. Each example includes the original sentence or paragraph, the coding decision, a short justification that points to formal features of the text, and one or more counterexamples that clarify the boundary conditions of the category. Coders are trained on a subset of documents that are not part of the main analysis and must reach a predefined threshold of agreement with a reference coder before annotating the core corpus.

Double blind annotation is applied to 20 percent of the documents, sampled across domains, templates and complexity levels. The two coders do not see each other’s labels

and are not told which documents belong to the double coded subset. Inter annotator agreement is computed for each indicator separately, using Cohen's kappa for categorical variables and intraclass correlation coefficients for continuous scores where appropriate. The target is $\kappa \geq 0.75$ for all primary indicators, which is consistent with accepted thresholds for reliable coding in discourse and policy analysis. Disagreements are resolved by an independent arbitrator, who has access to both annotations and to the coding manual. The arbitrator records a short rationale for each decision in a structured form, which provides material for refining the manual in future iterations.

Finally, the corpus and its annotations are prepared for external release. All content that could resemble real company names, personal identifiers or specific regulatory cases is either fully synthetic from the outset or replaced with neutral placeholders. The released package includes the raw texts, the indicator values per document, the metadata and the full coding manual. Scripts for re computing the indicators from raw text are included in a separate repository, so that external teams can verify that the supplied annotation corresponds to the definitions in the article. This configuration aims to make the Template Authority Index and its underlying indicators portable across contexts, while preserving sufficient structure to allow rigorous replication of the main findings.

5. Experimental design and statistical analysis

The experimental design treats instruction templates as the primary causal factor and models prompts and domains as structured sources of variance. Each of the 1000 base prompts is executed under the five template conditions, from C0 to C4, which yields a fully crossed within prompt design. This configuration enables direct intra prompt comparisons: every institutional scenario has a baseline realization under minimal instructions and four alternative realizations under distinct template regimes. Since prompts are held constant, differences in the authority indicators can be attributed to template effects, up to residual stochastic variation in model sampling.

Execution order is randomized at the level of prompt template pairs. For each base prompt, the five conditions are scheduled in a random sequence, which mitigates temporal drift that

could arise from server side optimizations or subtle provider level changes during the experimental window. Domains are treated as blocks. Within each domain, prompts are shuffled and executed in mixed batches rather than in contiguous domain blocks, which prevents systematic overlap between domain and temporal trends. Model size, decoding strategy and temperature are held constant for the main analysis, with temperature set to a moderately low value so that stochastic variation remains present but does not dominate the syntactic structure of outputs (Holtzman et al., 2020).

The primary analysis uses linear mixed effects models to estimate the impact of templates on the continuous indicators I1 to I7. For each indicator, a model of the form

$$\text{indicator} = \beta 0 + \beta c \cdot \text{condition} + \beta d \cdot \text{domain} + u\_\text{prompt} + \varepsilon$$

is fitted, where condition and domain are fixed effects, u_prompt is a random intercept for prompts and ε is the residual error term (Bates et al., 2015). Categorical variables for template are referenced to C0, which allows direct estimation of the difference between each template and the minimal instruction baseline. Domains are included as fixed effects in the main specification, with robustness checks that treat them as random intercepts when there is sufficient domain level variability. The mixed model structure captures the fact that multiple documents share the same underlying scenario and that prompts differ systematically in baseline levels of delegation, veto density and deontic layering.

For binary outcomes, such as Escalation Clause Presence or the presence of explicit override routes, logistic regression with mixed effects is applied. The log odds of occurrence are modeled as a function of template and domain, with random intercepts for prompts. This approach accommodates unbalanced rates of rare events, such as very complex override procedures, without relying on normal approximations that might fail in sparse regions of the indicator distribution (Agresti, 2013). Ordinal outcomes, including Default Scope Strength and the implicit centralization level, are analyzed with cumulative link mixed models as a robustness check, in addition to treating the ordered categories as scores in linear models.

Multiple comparison corrections are applied to control the familywise error rate across indicators. Since seven primary indicators are tested and each is subject to comparisons

between C0 and C1 to C4, the raw p values are adjusted using Holm or Benjamini–Hochberg procedures, depending on whether strong control of familywise error or control of false discovery rate is prioritized (Benjamini & Hochberg, 1995). The article reports both adjusted p values and raw effect sizes, emphasizing confidence intervals rather than binary significance labels. For each template contrast, the analysis presents the estimated difference in means, the 99 percent confidence interval and standardized effect sizes, such as Cohen's d for continuous indicators (Cohen, 1988).

Power analysis is conducted at the design stage using conservative assumptions about intra prompt correlation and indicator variance. Simulations assume modest effect sizes, with standardized differences between 0.2 and 0.3, and intra prompt correlations between 0.3 and 0.5. Under these parameters, a configuration with 1000 prompts and five conditions yields statistical power of at least 0.9 for detecting differences in the primary indicators at $\alpha = 0.01$. The design therefore prioritizes the detection of small but systematic shifts in delegation ratios, veto paths and override friction that may be invisible in purely qualitative comparisons.

Sensitivity analyses explore the impact of alternative modeling choices. Additional models are fitted with temperature as a covariate or with separate random slopes for condition by prompt to test whether some prompts react more strongly to templates than others. Subsets of the corpus generated with alternative model sizes or providers are analyzed with the same specification, in order to separate template effects that generalize across architectures from effects that depend on a particular implementation. The article also reports stress tests in which adversarial prompts explicitly invite the model to reassign decision rights in ways that contradict the template, for example by encouraging the assistant to take unilateral decisions under a decentralization oriented configuration. These stress scenarios probe the limits of the template effect and clarify whether the *regla compilada* exerts a persistent structural influence or can be overridden by local prompt engineering (Startari, 2025).

Throughout, the statistical analysis is presented as part of an operational framework rather than as an abstract exercise. The mixed models and logistic regressions are chosen because they align with institutional questions: by how much does a safety oriented template increase the odds of escalation clauses, or how strongly does an autonomy oriented

template raise the delegation ratio in compliance memoranda. The combination of intra prompt design, mixed effects modeling and corrected multiple comparisons is intended to provide both methodological rigor and direct interpretability for organizations that are deciding how to configure instruction templates in systems that already exercise substantial de facto authority through language.

6. Deliverables, validation plan and success criteria

The experimental pipeline is designed to produce a set of concrete deliverables that can be adopted by institutions and research teams without dependence on a specific provider or deployment. The central artifact is the Template Authority Index, a composite measure that aggregates the seven primary indicators I1 to I7 into a single score per template and per document. The index is defined transparently, with explicit formulas and weighting schemes, so that it can be implemented in independent code bases and adapted to different institutional contexts. This approach aligns with broader movements toward reproducible metrics in computational social science and AI governance, where indicators must be portable and verifiable rather than embedded in vendor specific tools (Peng, 2011; Stodden, 2013; Startari, 2025).

The second deliverable is a public test battery that includes the full set of prompts, instruction templates and annotation guidelines required to reproduce the main analyses. Each prompt is accompanied by a structured description of domain, scenario, actors and decision points. Each template is provided in its exact wording, with version identifiers and a change log, so that future studies can track the impact of modifications. The annotation guidelines describe the coding process for each indicator and include quality checklists for annotators. In institutional settings, this battery can serve both as a research tool and as a diagnostic instrument to evaluate new templates before they are deployed in production systems.

A third deliverable is a de identified dataset that contains all generated texts, their indicator scores and associated metadata. The dataset includes hashes of inputs and outputs, domain and template labels, model identifiers and generation parameters. Sensitive or potentially

identifying elements are either synthetic from the outset or have been replaced with neutral placeholders. The dataset is structured in a way that supports both document level analysis and aggregated reporting by domain or template. Following best practices in open data for computational linguistics, the release includes a detailed data sheet describing the construction process, limitations and recommended uses (Gebru et al., 2021).

The fourth deliverable is a technical report that documents the full experimental design, the statistical models, the parameter choices and the main findings. This report is written for an audience that includes AI governance teams, policy makers and technical leads in organizations that deploy language models in decision adjacent contexts. The report emphasizes effect sizes, confidence intervals and practical implications, for example, how much a safety oriented template increases the average Default Scope Strength in expense control memos, or how strongly an autonomy oriented template raises the Delegation Ratio in HR disciplinary resolutions. These descriptions connect the abstract indicators to operational questions about accountability and oversight (Citron & Pasquale, 2014; Startari, 2025).

A fifth deliverable is a validation repository that provides the code, configuration files and documentation needed to re run the main analyses. The repository is structured to support containerized execution, so that external teams can reproduce the results in controlled environments. It includes scripts for generating the corpus from the prompts and templates, for running indicator extraction, and for fitting the mixed models and logistic regressions described in the analysis section. The repository is intended as a reference implementation rather than as a production system, and it can be forked or adapted to new model families and organizational contexts.

The validation plan has two layers. The first is internal validation, which checks the stability of the Template Authority Index and its component indicators across resampling procedures, alternative model specifications and subsets of the corpus. Bootstrap resampling is used to estimate the variability of index values under repeated sampling of prompts and documents (Efron & Tibshirani, 1993). Alternative models that treat domains as random effects or that modify the weighting of indicators in the composite index are

also evaluated. The goal is to ensure that the main conclusions do not depend on a single modeling choice or a specific partition of the data.

The second layer is external validation, carried out through replication on different model families and providers. The same prompts and templates are run on at least one additional model within the original provider's ecosystem, and on at least one model from an independent provider. For the latter, templates are mapped to functionally equivalent roles, for example, safety emphasis or autonomy emphasis, while preserving as much of the original wording as possible. The Template Authority Index and the individual indicators are computed for these new runs and compared to the original distributions. External validation is considered successful if effect directions are preserved and if effect sizes fall within a predefined tolerance band, for example, a reduction of less than 20 percent relative to the original estimates. This threshold reflects standard practice in replication studies in psychology and social science, adapted to the specificities of language model evaluation (Nosek et al., 2015; Startari, 2025).

Success criteria are specified in operational terms. First, at least three primary indicators must show statistically significant and consistent differences between template conditions C1 to C4 and the baseline condition C0 at the predefined significance level, with effect sizes that remain stable across domains. Second, the Template Authority Index must separate template conditions in a manner that is robust to resampling and alternative weighting schemes. Third, the inter annotator agreement for all primary indicators must remain at or above $\kappa = 0.75$ in the double coded subset, and any systematic sources of disagreement must be documented in the coding manual. Fourth, external replications must meet the effect size tolerance criterion described above.

If these criteria are met, the study provides evidence that instruction templates do not merely affect style or politeness, but systematically reshape how institutional texts encode delegation, veto and escalation. The deliverables and validation framework are designed so that organizations can adopt the indicators and the Template Authority Index as part of their own audit processes, treating templates as configurable infrastructures of authority rather than as invisible background settings.

7. Risks, limitations, ethics and compliance

The design explicitly treats instruction templates as a controllable infrastructure of authority, but this focus introduces several risks and limitations that must be recognized. The first risk is the conflation of style with authority. Templates often encode politeness norms, legal boilerplate and branding constraints together with real allocations of decision rights. Even with an inventory that targets delegation, veto paths and escalation clauses, there is a persistent possibility that some measured differences reflect stylistic ornamentation rather than structural changes in who decides what. The study mitigates this risk by centering indicators on observable decision operations, for example, who approves, who can veto and how overrides are obtained, but it cannot completely eliminate residual style effects. This limitation is particularly relevant when organizations adopt heavily branded templates where tone and authority are intertwined.

A second risk concerns model and provider dependence. The experimental corpus is generated within a finite set of model families and providers, all of which share design choices about safety training, refusal patterns and institutional tone. Even with external validation across at least two providers, the findings remain conditional on a technological environment in which major vendors implement similar safety and alignment regimes. It is plausible that smaller or specialized models, trained with different objectives, would respond differently to the same templates. The study attempts to bound this risk through cross provider replication and by publishing the full template texts, yet any generalization beyond the tested ecosystem must be made cautiously (Crawford, 2021).

A third limitation is domain specificity. The five domains chosen, policy, compliance, human resources, finance and logistics, are all institutional and relatively formal. The syntactic authority inventory is tailored to these genres, where clauses about approval, escalation and exceptions are common and salient. In highly informal settings, for example, customer support chat or creative collaboration, the same indicators might be less informative or more difficult to annotate consistently. The inclusion of an additional domain for robustness checks reduces but does not remove this constraint. Users of the Template Authority Index in other domains may need to recalibrate thresholds or adapt

definitions, particularly for indicators such as Default Scope Strength and Override Friction Score.

There are also methodological limitations deriving from the synthetic nature of the corpus. Synthetic prompts allow tight control over scenario structure and the presence of potential escalation or veto points, but they cannot fully capture the legal, cultural and organizational pressures that shape real institutional texts. In practice, drafts are modified by human reviewers, legal teams and managers, each with their own conventions and risk appetites. The study models a first order effect of templates on raw model outputs, not the full socio technical pipeline in which those outputs are negotiated and edited. This restriction is explicit in the design and should prevent over interpretation of the results as direct predictions of final deployed documents.

Ethically, the study is positioned in a low risk space because it deliberately avoids personal data and uses only synthetic or generic institutional scenarios. Nonetheless, authority allocation is not ethically neutral. A template that shifts decision rights from local staff to centralized systems or from human managers to automated processes has implications for accountability, contestability and perceived fairness (Citron & Pasquale, 2014). By quantifying how templates change authority bearing constructions, the study risks making it easier for organizations to optimize for control in ways that are less visible to non experts. To address this, the article adopts a transparency oriented stance. All templates, prompts and indicators are published, and the discussion sections emphasize the need for human oversight, clear lines of responsibility and explicit governance frameworks for template design (Wachter, Mittelstadt, & Floridi, 2017; Startari, 2025).

The annotation process raises a separate class of ethical and practical issues. Annotators are asked to make judgments about who holds authority in a sentence and about whether a clause constitutes a veto, an escalation or an override. These judgments are vulnerable to cultural bias, organizational experience and legal context. A phrasing that appears deferential in one jurisdiction may be interpreted as standard administrative language in another. The coding manual attempts to reduce this variability by anchoring categories in formal grammatical cues such as subject type, verb class and deontic operators, rather than in perceived fairness or intent. The use of double coding and structured arbitration provides

an additional safeguard, but the resulting indicators still reflect the interpretive frame of the annotator cohort.

Compliance considerations shape both the design and the release strategy. Because the study touches on governance relevant uses of language models, it must avoid offering recipes for bypassing existing regulatory or internal control frameworks. The test battery and Template Authority Index are presented as tools for auditing and diagnosis, not as means to increase opacity or to automate high stake decisions. Examples are framed in ways that presume continued human responsibility for final decisions, especially in areas that intersect with employment, financial controls or regulatory reporting. The article explicitly recommends that any deployment of templates informed by these findings be accompanied by impact assessments and internal review processes, particularly in jurisdictions with emerging AI governance frameworks.

Finally, the study does not claim to settle the normative question of how authority should be distributed between humans and automated systems. It provides a measurement framework and evidence that templates matter in that distribution. Evaluating whether a given template configuration is acceptable, desirable or illegitimate remains a matter for organizational governance, legal standards and political debate. By treating instruction templates as a visible layer of the *regla compilada* rather than as neutral technical settings, the article aims to shift discussions of AI governance toward the concrete syntactic mechanisms through which authority is exercised in text, highlighting a set of levers that can be inspected, contested and redesigned.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*(1), 289–300.

Bratton, B. H. (2016). *The stack: On software and sovereignty*. MIT Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Citron, D. K., & Pasquale, F. A. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., & Daumé, H. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2nd ed.). Arnold.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Montague, R. (1974). *Formal philosophy: Selected papers of Richard Montague*. Yale University Press.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425.

Palmer, F. R. (2001). *Mood and modality* (2nd ed.). Cambridge University Press.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226–1227.

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

Stodden, V. (2013). Resolving irreproducibility in empirical and computational research. *IMS Bulletin, 42*(2), 16–17.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76–99.