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

Causal vs Retrocausal Attention Boundaries.

Agustin V. Startari.

Causal vs Retrocausal Attention Boundaries

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 17, 2025

DOI

- Primary archive: https://doi.org **10.5281/zenodo.17378361/**
- SSRN: Pending assignment (ETA: Q4 2025)

Language: English

Series: *AI Syntactic Power and Legitimacy*

Word count: 7901

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 study quantifies the minimal right context that changes model decisions about authority bearing constructions under strict causal masking versus non causal access. We formalize flip probability P_flip(b to b+Δ), the instance level threshold τ(x), and the construction level threshold τ_C, and we measure breakpoint sharpness over a right context ladder b in {0, 1, 2, 4, 8, 16, 32}. The dataset contains minimal pair ladders per construction family, deontic stacks, nominalizations, enumerations, defaults, agent deletion, scope setting adverbs, role addressatives, across six languages, en, es, pt BR, fr, de, hi, with balanced length, domain, and register, and human gold labels. We evaluate frozen causal decoder models, non causal encoders and encoder decoders, and causal streaming variants with sliding windows. Masking primitives include hard truncation, stochastic truncation, and delayed reveal streaming with cache isolation and sentinel based leakage tests. Primary endpoints are τ_C by construction and language, P_flip curves, AUC_flip, and a latency accuracy frontier. Results map τ_C to observed compliance deltas on instructed tasks while holding administrative workflow constant. The contributions are a public dataset with right context ladders, an evaluation harness with tested masks, per construction τ atlases and P_flip plots, and a preregistered analysis that links regla compilada constraints, Type 0 production equivalence, to measurable authority judgments.

Acknowledgment / Editorial Note

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

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

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

1. Operational inventory and labels

This section specifies the inventory of authority-bearing constructions, the labeling scheme for authority stance, and the annotation mechanics that link each observed cue to a concrete instance of the *regla compilada* as a Type-0 production in the Chomsky hierarchy. The objective is to guarantee that every example in the corpus encodes an explicit mapping from form to decision, so that model outputs can be evaluated against a stable, reproducible target that does not depend on sentiment, topic, or authorial intent. The inventory is designed to be portable across languages while remaining sensitive to morphology and discourse sequencing, which are frequent sources of authority cues. The labels are defined to capture graded judgments that models and human annotators can make at inference time under different right-context budgets.

The stance label set uses three ordered categories, low, neutral, and high. Low indicates that the text discourages, restricts, or defers authority, including cues that remove agents, defer decisions, or embed conditions that suspend action. Neutral indicates that the text is descriptive or preparatory without issuances that commit an actor to a decision or confer explicit rights or obligations. High indicates that the text asserts, grants, or enforces authority, for example by binding a deontic operator to a subject or by establishing default scopes that displace competing claims. These categories are applied at the instance level, which is the smallest unit that contains the prefix and the controlled continuation. The same instance receives one stance label under each right-context budget used in the experiments, which permits direct measurement of flips that arise when future tokens are revealed.

The inventory covers seven construction families that are observed to shift perceived legitimacy and compliance decisions across institutional and everyday registers. Deontic stacks collect operators such as must, shall, and their multilingual counterparts, and they track stacking depth and scope. Nominalizations transform events into entities, which suppresses agents and makes compliance appear impersonal, for example the enforcement

of policy rather than an actor who enforces. Enumerations impose structure and order upon obligations or exceptions, which can function as a surrogate for hierarchy or priority. Defaults stipulate outcomes that obtain unless a specified condition overrides them, which often relocates the burden of proof. Agent deletion removes the explicit executor of an action and is associated with administrative anonymity. Scope-setting adverbs, such as strictly or exclusively, raise or lower the strength of a claim without introducing new content. Role addressatives, often turn-final in dialogue, fix relative status by naming or addressing the other party in ways that encode deference or command. Each family contains language specific realizations, yet all families are defined by constraints over form rather than by topical content, which keeps the inventory compatible with a formal grammar view of authority cues (Chomsky, 1965; Montague, 1974).

Every annotated instance includes an explicit reference to the operative *regla compilada*. The reference specifies the constraint set that licenses the cue and the expected behavioral transition, for example an increase from neutral to high when a deontic operator with scope over a second person subject appears to the right of the prefix. The compiled reference is treated as a production that can apply in different surface forms while preserving its licensing conditions. This formalization permits us to evaluate whether a model is reacting to the constraint system, not to a lexical surface alone, and it lets us design paraphrases and morphological alternations that should preserve stance under full context. The practice follows a tradition in which form regulates interpretation through explicit constraints, rather than through inferred intentions, which aligns with prior theoretical work on grammar and meaning composition (Chomsky, 1965; Montague, 1974; Startari, 2025).

Annotation proceeds with two primary annotators and one adjudicator. For each instance, annotators receive the prefix, the right-context continuation, and the language specific guidelines for the relevant construction family. They label stance, the family tag, and the compiled constraint reference. Inter-annotator agreement is monitored continuously, and batches that fall below the target threshold are subject to revision, with targeted feedback that addresses boundary errors, for example confusing a list that enumerates facts with a list that orders obligations. To minimize leakage from topic and length, annotators work on balanced batches that mix domains and registers, and they are required to justify any

high stance labels by pointing to the minimal surface span that triggers the compiled constraint. This justification is recorded as a span note to support later error analysis and to facilitate training of minimal detectors that operate under strict causal masking.

Language coverage includes English, Spanish, Portuguese Brazil, French, German, and Hindi. For each language, the inventory includes a short lexicon of operators and known constructions, yet the lexicon serves as a reminder of frequent realizations, not as a binding list. The decisive property is that the minimal span identified by the annotator can be paraphrased while preserving the constraint, and that it exhibits predictable effects on stance under full context. When languages express a construction mostly to the right of the trigger position, for example addressatives that appear after a clause, the labeling guidelines emphasize careful separation of stance that comes from the right context from stance that was already licensed in the left context. This separation is critical for later analyses, which estimate the minimal right context that aligns the instance with the full context decision.

Quality control includes periodic calibration rounds that reuse a held out set of items. Annotators receive feedback that compares their labels with the current gold set and highlights systematic drift, such as over-reliance on sentiment proxies. The corpus release includes all labels and span notes, together with the compiled constraint references. This design makes downstream modeling transparent and repeatable, and it supports tests that remove lexical surfaces while preserving constraints, which are necessary to show that authority judgments track form rather than topic. By grounding each item in the *regla compilada*, the inventory supplies the stable foundation required to measure how right-context budgets change decisions in a controlled and theoretically coherent way.

2. Minimal-pair generation pipeline

This section defines a reproducible pipeline to create minimal pairs that isolate right-context authority cues while holding topic, sentiment, register, and left-context content constant. The goal is to measure the smallest increment of future tokens that changes an authority judgment, without confounds from lexical novelty or domain drift. Each instance

consists of a left prefix that is intentionally stance ambiguous, followed by a family of controlled right continuations that introduce exactly one authority-bearing cue at a known distance. By varying this distance across a fixed ladder of small budgets, the pipeline identifies the minimal right context at which a model aligns with its own full-context decision. The pipeline is designed to operate across English, Spanish, Portuguese Brazil, French, German, and Hindi, with parallel controls for morphology, clitic placement, and discourse sequencing that differ by language but preserve the same compiled constraint. The compiled constraint, recorded as a regla compilada reference, links the surface cue to a Type-0 production in a formal grammar sense and therefore keeps the target independent of author intention or topic drift (Chomsky, 1965; Montague, 1974; Startari, 2025).

Data creation begins with authoring prefixes that do not license a clear stance on their own. For each construction family, such as deontic stacks or enumerations, curators write many short prefixes that fix entities, tense, and discourse frame, yet avoid issuing rights or obligations. Curators then design continuations that introduce a single authority cue, for example a modal that binds to a second person subject, a turn-final addressative that reassigns roles, or an enumerated exception that changes default scope. Each continuation is placed at a preassigned distance measured in tokens to the right of the prefix, and alternative continuations are constructed so that only one cue is added while everything else remains semantically and stylistically consistent. To ensure that lexical surfaces do not masquerade as authority, each continuation has a paraphrase twin that uses different words but implements the same compiled constraint. Curators also create morphological alternations that leave the constraint intact, for example moving between analytic and synthetic forms in Romance languages or altering case marking in German.

Length and domain are controlled through stratified sampling. Prefixes are binned by total instance length deciles within each language, and the pipeline samples uniformly from bins when assembling an evaluation slice. Domain labels are assigned at author time, including administrative prose, dialogue, instructions, news, and encyclopedic expository text. During assembly, each slice maintains balanced coverage across domains to avoid spurious correlations between right-context positions and style. Dialogue receives additional attention because many role addressatives appear in turn-final position. For dialogue, the

pipeline forces at least one continuation to place the cue after a minimal discourse marker, which checks that the model is not simply keying off punctuation or line breaks.

Every instance is annotated by two primary annotators with a third adjudicator. Annotators receive the prefix, the continuation at a specified right-context distance, and a short guide for the relevant construction family in the target language. They supply a stance label, a family tag, and a pointer to the minimal surface span that licenses the stance under full context. They also attach the regla compilada reference that names the constraint set and records the expected behavioral shift. Inter-annotator agreement is monitored continuously, with calibration rounds that reuse a frozen set of items. Where κ falls below the acceptance threshold, items are revised or removed. The span pointer is mandatory for high stance labels and strongly recommended for neutral labels that are elevated by defaults or hidden scope. This pointer allows later audits to verify that a claimed cue is indeed the one responsible for the stance under full context. It also allows removal of the span to create negative controls that should not flip the decision.

Language specific complications are handled explicitly. Spanish and Portuguese Brazil have clitic pronouns and productive nominalizations that can hide agency without changing propositional content. German has separable verbs and case cues that shift scope. French has discourse markers that signal default exceptions with little lexical mass. Hindi has politeness and honorific markers that can function as soft authority gates. For each language, the pipeline documents a short lexicon of frequent realizations, but the lexicon is advisory and not exhaustive. The decisive criterion is whether the continuation instantiates the same compiled constraint that the paraphrase and the morphological alternation also instantiate. If the stance changes under full context when the surface is altered but the constraint is held constant, the instance is rejected as lexically brittle. If the stance is stable under full context across paraphrase and morphology, the instance is accepted. This policy enforces the principle that authority judgments must track constraints encoded in form rather than topic or affect.

Quality assurance includes leakage checks and budget integrity tests. For every minimal pair, the pipeline produces synthetic sentinels beyond the allowed right-context budget and verifies that masked runs never attend to those tokens. It also generates a no-cue

continuation that preserves length and rhythm but contains no authority operator. The no-cue continuation should leave stance unchanged relative to the prefix alone under full context. When this control flips the decision, the instance is flagged as confounded by semantics or topic and is removed. Finally, each accepted instance is passed through full-context inference with multiple deterministic passes to ensure stability of the target. Instances that exhibit unstable full-context labels are returned for revision or discarded.

The end product is a multilingual corpus of minimal pairs whose only systematic difference is the presence and position of a right-context authority cue that implements a documented compiled constraint. By construction, the corpus allows precise estimation of minimal right-context thresholds and flip probabilities while guarding against confounds from sentiment, topic, length, or domain. The same pipeline also yields negative controls, paraphrases, and morphological alternations for robustness checks, which strengthens the evidentiary link between constraint, surface cue, and measured stance.

3. Languages and Targets

This section defines quantitative and qualitative targets by language, justifies the selection from morphosyntactic variation that matters for authority judgments, and specifies the controls required to make right-context thresholds comparable across systems and domains. The project covers six languages, English, Spanish, Portuguese-Brazil, French, German, and Hindi, because they provide strong contrasts on three axes that are central to authority-bearing constructions. First, they differ in morphological marking that relocates the decisive cue either to the left or to the right of the decision point, for example case in German and honorific particles in Hindi. Second, they carry distinct discourse traditions in which authority is prototypically expressed, for example administrative and legal prose in English and French, procedural instructions in Portuguese-Brazil, and turn-organized address in Hindi and Spanish dialogue. Third, there is verified annotator capacity to sustain consistent labeling across all construction families. The goal is not typological exhaustiveness, it is to maximize informative contrast between construction families and surface realizations without introducing bias from annotator scarcity or style imbalance.

The quantitative target per language and per construction family is twelve hundred instances, with at least three hundred unique prefixes and at least four controlled continuations per prefix. This density supports precise estimation of the construction-level median threshold with narrow intervals and tolerates right-censoring when the signal requires more than thirty-two tokens of right context. Aggregation across six languages and seven construction families yields approximately fifty thousand labeled instances, which is sufficient to fit mixed-effects models of threshold variability while preserving a large held-out slice for calibration checks. Domain balance is enforced across five registers, administrative prose, dialogue, instructions, news, and encyclopedic exposition, with identical quotas by family and language. This balance prevents any one style from concentrating decisive cues, which would shift thresholds for stylistic reasons rather than because of the compiled constraint that licenses authority.

Every language ships with a compact guide that documents positive and negative exemplars for each family, a micro-lexicon of frequent operators, and morphological notes that connect the *regla compilada* to the surface. For Spanish and Portuguese-Brazil the guide tracks productive nominalizations and clitics that hide agency while preserving propositional content. For French it lists exception markers that establish default scopes with little lexical bulk. For German it flags separable verbs and case cues that shift scope. For Hindi it annotates politeness and honorific morphology that can function as soft authority gates in turn-final position. The micro-lexicon is advisory, not exhaustive. Acceptance criteria demand that paraphrases and morphological alternations preserve the full-context decision. If stance flips under full context when surface words change but the compiled constraint is unchanged, the instance is rejected as lexically brittle. This policy enforces the core premise that authority judgments must track constraints encoded in form rather than topic or affect (Chomsky, 1965; Montague, 1974; Startari, 2025).

Comparability and traceability are the qualitative anchors. Comparability means that, within a construction family, the design fixes the same functional position of the decisive cue across languages, even when the surface segment differs. Consider enumerations that impose normative priority. The element that establishes the order must appear inside the continuation window rather than the prefix, with approximately the same distance in tokens

from the cut point. Traceability means that each instance is anchored to an explicit reference of *regla compilada*, cataloged as a Type-0 production in the Chomsky hierarchy, with licensed conditions and expected effect on authority stance. This reference allows auditors to verify that a model responds to the constraint system and not to semantic decoration, and it supports cross-language audits when translations preserve meaning but not authority operators.

Length and position controls are stratified. Prefixes are binned by total instance length deciles within each language. Sampling is uniform across bins when assembling an evaluation slice. The position of the authority cue in the continuation is separately binned, then matched during sampling, so that threshold estimates do not reflect accidental clustering of cues near the boundary. Dialogue receives special handling because many role addressatives appear at turn ends. For dialogue, at least one continuation per prefix must insert a minimal discourse marker before the authority cue. This manipulation checks that the model is not keying off punctuation or line breaks as spurious signals. Throughout, annotators must justify any high-authority label by pointing to the minimal surface span that triggers the compiled constraint under full context. The span note is mandatory evidence for later audits and for the construction of negative controls that remove the cue while preserving rhythm and topic.

Tokenization and inference are parameterized by language. In non-Latin scripts, for example Devanagari for Hindi, subword segmentation can inflate measured token counts relative to comparable content in Latin scripts. To preserve comparability, each instance is reported in two units, the base model's subword tokens and a canonical word count for the language. Primary analyses and thresholds use the native model units. An appendix provides conversions to words for human readability. This dual reporting prevents artifacts when one language requires more subunits to encode the same formal content, which would otherwise distort conclusions about minimal right-context budgets.

Power, calibration, and stability targets are shared across languages. For each construction family, the agreement curve with full-context decisions must increase monotonically as the right-context budget grows, with stable calibration error at intermediate budgets. The construction-level median threshold must present narrow confidence intervals under

stratified resampling and must remain within a predefined tolerance band under repeated runs. Annotator agreement is monitored continuously with periodic calibration rounds on a frozen item set, and any sustained drop below the target threshold triggers guide revisions and selective recycling of examples. When all quantitative quotas are met and all qualitative checks pass, a language cell is considered complete. Only then are threshold comparisons across languages licensed, since the comparison rests on matched distributions in length, position, domain, and constraint identity. With these controls, the resulting τ maps reflect structural properties of authority constructions, not idiosyncrasies of pipeline, segmentation, or workforce.

4. Models and Configurations

This section specifies the families of models, their inference configurations, and the controls that guarantee comparability across right-context budgets without leaking future tokens. The design separates architectural regimes by the kind of contextual access they allow, then standardizes decoding and calibration so that observed flips in authority stance are attributable to context rather than to unrelated implementation choices. The central regimes are strictly causal decoder models with left to right attention, non causal encoders and encoder decoders with full right access at inference time, and causal streaming variants that approximate real time reading through sliding windows. All weights remain frozen. No fine tuning is performed on the evaluation data. Any learned mapping from internal states to stance is provided by a lightweight classifier head that is calibrated only on a held out development slice which is disjoint from the evaluation ladders.

Causal decoder models represent the baseline for no peek conditions. They process tokens from left to right with an attention mask that forbids access to future positions. We evaluate at least two sizes per family, for example a small and a medium instance within a GPT style lineage and a small and a medium instance within a LLaMA style lineage. The goal is not to rank models by scale but to check whether minimal right context thresholds change with capacity once masking is kept strict. To eliminate cache based leakage, each right context budget is evaluated in a fresh process with attention key value memories

reinitialized. Unit tests inject sentinel tokens beyond the budget and assert zero attention to those positions. When the decoder produces stance scores, it does so with temperature zero and top p equal to one. The classifier head reads the final hidden states for the instance and outputs a probability over low, neutral, and high authority. We perform temperature scaling on the head using only the development slice, then freeze the scale for all runs.

Non causal models serve as the full access reference. We include a masked language encoder and an encoder decoder that attend bidirectionally over the entire sequence. These systems are used to compute the full context decision for each instance. They also define the target with which causal runs seek agreement as the right context budget grows. Because bidirectional encoders can be sensitive to sentence segmentation, we normalize input with a language specific sentence splitter and confirm that the split does not create artificial markers near the cue position. The same stance head architecture is used, calibrated once per model family on the development slice. This keeps comparisons focused on the availability of future tokens rather than on differences in the readout mechanism.

Causal streaming variants approximate deployment settings in which inputs arrive over time. We implement sliding windows W in a fixed set, for example 64, 128, 256, and 512 tokens. The model reads the left context within the current window, emits a stance judgment, then advances the window by a stride that preserves continuity of the left context while revealing new right tokens. We record the first window size that produces the same stance as the full context reference. This measurement creates a parallel notion of threshold that can be compared with the discrete b ladder in the main protocol. Streaming adapters such as recurrent memory modules are permitted, but they must be reset between budget conditions and evaluated with the same no leakage sentinels.

All model families share a single preprocessing and logging layer. Tokenization uses the native subword scheme of each base model. For non Latin scripts we also compute a canonical word count in the source language to aid human interpretation of thresholds. The run harness writes a complete audit trail that includes model checkpoint hash, tokenizer hash, masking schedule, random seed, and a digest of the input instance. This allows exact reruns and protects against accidental drift in software or data. Environment descriptors

capture hardware, library versions, and any low level kernel flags that affect attention kernels.

Calibration and stability checks operate at two levels. First, we estimate a reliability curve for the stance head on the development slice and compute expected calibration error at each budget. Second, we verify monotonic agreement with the full context decision as right tokens are revealed. The latter is not imposed by design, since some families may exhibit non monotonic transitions near the threshold, but it is tested and reported. When agreement curves show pathological oscillations, the instance is flagged for inspection. Many pathologies resolve to unstable full context targets, which are removed earlier in the pipeline, or to ambiguous minimal spans, which trigger a return to annotation.

Finally, we publish model cards that document the masking caveats and the known failure modes of each family under right context deprivation. These cards describe the behavior of deontic stacks, enumerations, defaults, and role addressatives as the budget increases. They also warn about constructions that appear left anchored but require distant right material to stabilize scope. By standardizing architecture, masking, decoding, calibration, and logging, the configuration makes it possible to attribute changes in authority stance to increments in future context. The result is a clean comparison between causal and retrocausal regimes and a reliable map from right context thresholds to observable behavior in instruction following tasks that do not involve administrative workflows.

5. Masking and Right-Context Ladder

This section defines the procedures that constrain models to a specified amount of future context, establishes the discrete ladder of right-context budgets, and documents the tests that rule out hidden lookahead. The objective is to measure when a model’s authority judgment converges to its full-context decision as additional right tokens become available, while guaranteeing that any change arises from the newly revealed tokens rather than from cache reuse, tokenizer artifacts, or uncontrolled preprocessing. The protocol standardizes all masking operations across model families, records exhaustive run metadata, and validates the absence of leakage with unit tests that are auditable and reproducible.

Right-context is controlled with a budget parameter b that counts the number of tokens to the right of the prefix that are available at decision time. The evaluation ladder uses the following budgets, zero, one, two, four, eight, sixteen, and thirty two. These points are chosen to balance sensitivity near the decision boundary with coverage of longer cues without inflating run time. Every instance is evaluated at each rung of the ladder in fresh processes to eliminate cross-budget contamination. For decoder models, an attention mask is applied such that tokens beyond the allowed boundary are invisible to the query at every position. Key and value memories are reinitialized per budget. For encoder and encoder-decoder models that permit bidirectional attention, budgets are implemented by truncating the sequence before tokenization and by running a parallel sentinel test that would fail if any position attended to hidden content.

Three masking schedules are implemented and audited. Hard masks apply strict truncation at the budget boundary and are the default for the main results. Stochastic truncation draws the budget uniformly from the ladder and is used to stress test calibration and to estimate the stability of agreement curves under random deprivation. Delayed reveal streaming feeds tokens one by one, records a stance at the designated budget, resumes the feed, then records the stance again at the next rung. The delayed reveal schedule simulates online reading and allows the harness to capture transient states that occur exactly at the moment a decisive cue enters the receptive field. All schedules write identical logs that include the model checkpoint hash, tokenizer hash, masking schedule, random seed, and a digest of the input instance. The harness prints environment descriptors, library versions, and kernel flags so that reruns are exact.

Leakage defenses combine structural and behavioral tests. Structural defenses include process isolation per budget, disabled attention cache reuse, and explicit zeroing of attention weights to masked positions. Behavioral defenses rely on sentinel tokens that are placed beyond the budget boundary and marked with a high-entropy pattern that would alter the stance if attended. In masked runs the model must display zero attention to sentinels, and the stance distribution must remain unchanged relative to an equivalent sequence without sentinels. Any deviation triggers a hard failure. Additional checks verify that sentence segmentation or Unicode normalization did not move the cue across the

boundary. Non-Latin scripts receive special handling, with both subword and canonical word counts recorded to detect budget drift caused by tokenizer differences. These tests are run on every commit of the harness and recorded with pass or fail status in the audit log.

The protocol also defines how decisions are emitted under each budget. Models produce a probability distribution over low, neutral, and high authority. Decoding is deterministic, with temperature equal to zero and top p equal to one for any generative component. A lightweight stance head is calibrated on a development slice that is disjoint from the evaluation ladders. The head is then frozen and reused across all budgets and schedules. Calibration is rechecked after any change to tokenization or preprocessing. Agreement with full-context decisions must increase or at least plateau as budgets grow. If oscillations occur, the harness flags the instance for inspection, since such behavior often indicates an ambiguous minimal span or unstable full-context labels. Instances that fail stability checks are removed earlier in the pipeline and are not allowed to influence threshold estimates.

The ladder design is tightly coupled to the minimal-pair corpus. For each instance, the decisive authority cue is inserted at a known offset in the continuation, and that offset is mapped to the nearest rung. This mapping ensures that abrupt increases in agreement are interpretable as the moment the cue becomes visible. The protocol records both the rung index and the absolute position in tokens relative to the prefix so that later analyses can aggregate thresholds per construction family and language without losing positional resolution. Dialogue items receive an additional guard. When a discourse marker precedes a turn-final cue, the marker is treated as a separate token span, and the harness confirms that the flip does not occur until the cue itself enters the budget. This rule prevents spurious flips driven by punctuation or line breaks.

To support replication, every masking primitive ships with unit tests that can run on toy sentences where the cue is an artificial token with a known effect on stance. These tests verify that the model cannot see the cue until the budget crosses the token’s position and that any surrogate signals are insufficient to induce a flip. The repository includes a minimal notebook that recreates the leakage checks and prints the same audit trail. Model cards summarize known caveats, for example constructions that appear left-anchored but

require distant right material to stabilize scope, and provide guidance for extending the ladder beyond thirty two tokens when languages or domains habitually place decisive material farther to the right.

Finally, the masking and ladder protocol connects to the behavioral link in which authority thresholds predict changes in compliance on controlled instruction-following tasks. Because the masking schedules are identical between judgment estimation and behavioral evaluation, observed compliance deltas can be attributed to the same increments of future context that drive stance flips. This alignment provides a clean bridge between micro-level attention regimes and macro-level behavior. By pairing strict masking with comprehensive leakage audits, the protocol delivers trustworthy measurements of right-context thresholds across architectures and languages, which in turn supports principled claims about when and how models rely on retrocausal information to adjudicate authority.

6. Formal Quantities and Estimators

This section fixes the statistical objects that operationalize right-context sensitivity in authority judgments and specifies estimators, confidence procedures, and reporting rules. The goal is to measure, with minimal assumptions, when a model's stance under a constrained right-context budget matches its full-context stance, and how sharply that alignment occurs as additional tokens to the right are revealed. All quantities are defined at the instance level first, then aggregated at construction and language levels in ways that respect censoring when the decisive cue lies beyond the largest budget in the ladder. Throughout, the link to the *regla compilada* is preserved by requiring that every measured change be attributable to a documented constraint rather than to topical content or sentiment proxies (Chomsky, 1965; Montague, 1974; Startari, 2025).

The primary observable is the flip event under a change in right-context budget. For an instance x and budget b, let f_b(x) denote the model's stance decision produced with access to exactly b tokens to the right of the prefix, under the masking schedule defined earlier and with deterministic decoding and a fixed calibrated readout head. The full-context decision is f_full(x), computed by a non-causal reference model or by an agreed

bidirectional configuration. The flip probability between adjacent rungs is defined as $P_{flip}(b \rightarrow b + \Delta) = P[f_b(x) \neq f_{b+\Delta}(x)]$, where Δ is the step between rungs in the ladder. Empirically, P_{flip} is estimated as the fraction of instances whose stance differs across the two budgets, with stratification by construction family, language, and domain to ensure balance. Uncertainty is obtained by a stratified nonparametric bootstrap that resamples instances within each stratum and recomputes the fraction. We report the mean and percentile confidence intervals. Bootstrap resampling is preferred because closed-form variance expressions become brittle once stratification and censoring are active, while the bootstrap remains robust under mild dependence introduced by reusing prefixes across continuations (Efron & Tibshirani, 1993).

The minimal threshold at the instance level is $\tau(x) = \min b$ such that $f_b(x) = f_{full}(x)$. If no rung yields agreement, the instance is right-censored at the maximum budget in the ladder. The distribution of τ over a set S, for example all instances that instantiate a given construction family in a given language, is summarized by the construction-level threshold τ_C, defined as the median of τ across S with censoring accounted for. Estimation proceeds with a Kaplan–Meier style approach that treats "agreement with full-context decision" as the event. Each rung contributes the proportion of remaining instances that first agree at that rung. The product of survival complements yields an estimator of the cumulative event distribution, from which the median and its confidence interval are read. This approach keeps censored instances in the risk set until the end of the ladder and avoids bias that would arise from discarding items that require more context than the ladder provides. When the median is not attained within the ladder because of heavy right-tail censoring, we report that τ_C exceeds the maximum budget and include the largest attainable lower bound.

Breakpoint sharpness quantifies how abrupt the transition is from disagreement to agreement as the budget grows. We compute the discrete derivative of the agreement curve with respect to b and report the maximum slope together with its location on the ladder. To guard against spurious spikes due to finite sample noise, a permutation test shuffles rung labels within instances and recomputes the maximum slope to produce a null distribution. The sharpness statistic is significant when the observed maximum lies in the extreme tail

of the null. This procedure does not assume smoothness and respects the discrete nature of the ladder.

Area under the flip curve, AUC_flip, complements sharpness by summarizing the total amount of instability across the ladder. We sum P_flip over consecutive rung pairs and divide by the number of pairs so that values are directly comparable across ladders of different lengths. Lower values indicate fewer stance reversals and therefore more stable behavior under right-context deprivation. Because P_flip estimates across rungs are correlated within instances, we obtain uncertainty by block bootstrapping at the instance level rather than by treating rung pairs as independent observations.

Calibration under each budget is measured with expected calibration error on stance probabilities. The readout head outputs probabilities over {low, neutral, high}. We bucket predictions into equal-frequency bins, compute the absolute difference between average predicted probability and empirical frequency within each bin, and report the weighted average across bins. Calibration is assessed on a development slice that shares the same ladder design but does not overlap with the evaluation instances. Changes to tokenization or preprocessing require recalibration. Reporting includes confidence intervals via bootstrap and a reliability curve for visual inspection. Proper calibration is necessary because τ estimation uses hard decisions, yet downstream behavioral links may use probabilities to estimate compliance deltas as a function of budget.

Agreement curves should be monotonic nondecreasing in practice, since additional right-context should not systematically reduce information about stance. Nonmonotonic patterns can occur near thresholds for families whose cues interact with punctuation or local discourse markers. We fit an isotonic regression to the empirical agreement sequence to obtain a monotone approximation and report the R^2 of the fit. Large deviations trigger an audit of those items. Audits check whether the cue span was mislocalized, whether a hidden operator in the prefix accidentally licenses the compiled constraint, or whether the full-context target is unstable across seeds. Unstable targets are either revised or removed by policy.

All quantities are reported at multiple granularities. Instance-level τ supports micro-diagnostics and error analysis. Construction-language level τ_C supports the main hypotheses about family stability and cross-lingual variation. Domain-specific summaries detect register effects, for example dialogue items in which role addressatives are turn-final and therefore require nonzero budgets even for strong models. Whenever sample sizes permit, we publish stratified curves alongside pooled ones so that users can recover effects that were averaged out. The repository provides scripts that recompute every estimator from raw logs that include rung-level decisions and probabilities, with an auditable trail of seeds, masks, and model hashes.

These estimators align the statistical layer with the formal layer of the *regla compilada*. Measured thresholds and transitions are treated as empirical fingerprints of the constraint systems that license authority. The resulting atlas of τ values and flip dynamics is therefore meaningful both as an evaluation artifact and as a map of where models require retrocausal information to adjudicate form-driven authority, not just sentiment or topic. This dual grounding is central to the project's claim that authority judgments travel through constraints encoded in form, in line with the long tradition that separates grammar-based licensing from authorial intention (Chomsky, 1965; Montague, 1974; Startari, 2025).

7. Primary Analyses Mapped to Hypotheses, with Minimal Theoretical Closure

This section specifies the analysis procedures that directly test each hypothesis, sets decision rules, and adds a minimal theoretical closure that ties empirical thresholds to a sufficient licensing condition in the compiled constraint view. The analytical goal remains unchanged. We attribute changes in authority stance to increments of right context rather than to sentiment, topic, or incidental lexical markers. All analyses run on masked evaluations, compare constrained decisions to full context decisions, and report uncertainty through stratified resampling. The interpretation is anchored in the constraint perspective of the *regla compilada*, understood as a Type-0 production that licenses authority effects through form, not intention or topic content (Chomsky, 1965; Montague, 1974; Startari, 2025).

Hypothesis H1 predicts sharp right context token thresholds after which authority judgments flip with high probability. The primary test fits an agreement curve for each construction and language. Agreement is the fraction of instances whose constrained decision matches the full context decision at each rung. Evidence for a breakpoint requires two conditions. First, a discrete increase in agreement that exceeds a predefined margin between adjacent rungs, with stratified bootstrap intervals excluding smaller effects. Second, a permutation test that shuffles rung labels within instances to build a null distribution of maximum step sizes. The observed step must fall in the extreme tail of this null. The rung at which the step occurs is recorded as the empirical threshold region. The corresponding absolute token offset of the decisive cue is reported alongside the rung index because tokenizers and languages distribute subwords differently.

Hypothesis H2 claims that the minimal threshold varies by construction family and language, yet remains stable within families under domain control. The analysis models instance level minimal thresholds as the response, with construction family and language as fixed effects, and model identity and domain as random intercepts. Stability within families is supported when the family main effect is large and statistically reliable, while the family by language interaction is materially smaller. The report includes partial effect sizes and interval estimates. The construction level median threshold is summarized with censoring accounted for via a survival style estimator. Paraphrase and morphology alternations hold the compiled constraint constant. If thresholds move only minimally across alternations, the data support the claim that families encode portable constraints rather than brittle lexical templates.

Hypothesis H3 states that under strict causal masking, flip rates drop to chance for constructions whose decisive cues lie to the right of the prefix. We first identify the subset whose minimal licensing span is entirely to the right. For this subset, flips near the zero budget are compared to a calibrated random baseline that preserves class priors. If observed flip rates match the baseline and remain strictly below non causal rates, H3 is accepted. A control inserts discourse markers before the decisive span to ensure that punctuation is not acting as a surrogate cue.

Latency accuracy frontiers function as a diagnostic across H1 to H3. For each construction and language, the frontier plots average tokens processed against agreement with full context. Efficient families show steep early gains and then plateau once the decisive constraint becomes visible. Oscillations are audited rather than smoothed away. Many oscillations correlate with interactions among enumerations, defaults, and scope setting adverbs that jointly determine stance. When oscillations align with unstable full context labels, the items are returned to annotation and removed from threshold estimation. Expected calibration error is reported per rung and should decline as budgets grow, since the stance head is fixed and calibrated on a separate development slice.

Minimal theoretical closure. We add a sufficient condition that links construction level thresholds to the compiled licensing mechanism.

Proposition (sufficient condition for family level threshold alignment). Let C be a construction family with a compiled constraint set Γ_C that licenses an authority increase when a unique right span s appears, where s is the minimal surface segment that instantiates Γ_C in the continuation. Assume the prefix p does not already satisfy Γ_C and contains no operator that licenses the same authority effect. Assume further that s is paraphrase invariant and morphology invariant with respect to Γ_C under full context. Then for any instance x built from prefix p and continuation containing s, the instance threshold τ(x) equals the first budget b at which s becomes visible on the ladder, up to rounding to the nearest rung. Moreover, the construction level median τ_C is bounded above by the median position of s over instances of C evaluated in model native tokens, with equality when cue positions are unimodal and censoring is light.

Sketch of proof. Construct minimal pairs that differ only by the presence of s in the continuation at a known offset. Because p does not satisfy Γ_C and no surrogate operator is present, the compiled constraint is unlicensed at budgets that exclude s. Under full context, Γ_C is licensed by s, and stance moves to the high category. Masking schedules guarantee that only the newly revealed tokens can change the decision. Paraphrase and morphology invariance ensure that the decision is a function of Γ_C rather than of a particular surface. Therefore the first budget that reveals s is the first budget that enables licensing, hence τ(x) equals that budget. Aggregating τ(x) over the family and applying

standard facts about medians yields the bound on τ_C. The equalities follow when positions cluster tightly and when right censoring is rare.

Counterexample. If a purported cue is lexically brittle and does not instantiate Γ_C, then paraphrase variants can flip full context stance unpredictably. In this case the minimal span is not invariant, and agreement curves show non monotone behavior that cannot be aligned to the appearance of a single right span. The sufficient condition fails. The proposition does not claim necessity. Families may exhibit distributed right cues that license Γ_C only jointly, in which case τ(x) can exceed the position of any individual span and the bound on τ_C can be loose.

Action. We include the proposition as a minimal theoretical closure that ties empirical τ maps to a sufficient licensing condition in the compiled rule perspective. The project treats it as a checkable commitment because all prerequisites are operationalized in the corpus: absence of left licensing in the prefix, unique right span placement, and paraphrase and morphology invariance.

Evidence. We provide a constructive demonstration on the minimal pair ladders. For deontic stacks and enumerations, we place a single operator span at known offsets, then verify that τ aligns with the first budget that reveals the span across paraphrases and morphology alternations. We also include a negative case built to be lexically brittle by replacing the operator with a token that correlates with sentiment but does not instantiate Γ_C. The negative case yields non reproducible flips under paraphrase and invalidates the sufficient condition, as expected.

Estimated impact. The addition clarifies why family level τ values are not mere empirical regularities but fingerprints of a licensing mechanism. The expected contribution increase is plus 0.5 in innovation and plus 0.5 in originality, since the section moves from descriptive thresholds to a formal link between constraint licensing and observed flips while remaining testable within the released corpus.

.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

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

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2978–2988.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186.

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.

Kudo, T. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.