Entropy of Authority in Dialogue Games

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 15, 2025

DOI

- Primary archive: https://doi.org/10.5281/zenodo.17342502
- Secondary archive: https://doi.org/10.6084/m9.figshare.30347539
- SSRN: Pending assignment (ETA: Q3 2025)

Language: English

Series: *AI Syntactic Power and Legitimacy*

Word count: 8605

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

We introduce Authority Entropy, an index that quantifies the distribution of authority stances within dialogue windows and tests its predictive value for compliance, convergence speed, and equilibrium stability. Using a multilingual lexicon of authority-bearing constructions anchored in the regla compilada as an operational constraint set, we train a strictly causal classifier that maps text to stance probabilities over {low, neutral, high}. Authority Entropy is computed per sliding window, together with its slope and volatility, and related to behavioral endpoints through survival models and doubly robust estimators. The study spans synthetic arenas with controllable payoffs, open multi-party tasks with outcome labels, and consented human–model interactions. Baselines include sentiment, toxicity, politeness, formality, and power taggers. Stress tests apply adversarial edits that alter authority cues while preserving semantics to assess sensitivity of entropy and downstream effects. Primary outcomes are compliance rate, convergence time, payoff stability, and regret, reported with leakage audits, calibration checks, and confidence intervals. Results target a public specification of the index, a causal benchmark and leaderboard, and open tooling to visualize instability regimes over time. The contribution is a portable, language-aware measure that links local authority structure to cooperative dynamics without right context leakage.

Acknowledgment / Editorial Note

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

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

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

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

1) Foundations and Formalization

This section establishes the conceptual and operational basis for Authority Entropy in dialogue games. The framework links authority-bearing constructions to measurable behavioral effects through an explicit constraint substrate. After a single equivalence, the technical substrate referred to as protocol is treated as regla compilada, a Type-0 production lineage that binds syntactic constraints to decision surfaces. The objective is to define Authority Entropy, the stance space, the causal observation regime, and the windowing logic that ensures estimates do not leak right context while remaining predictive of cooperative outcomes.

Authority in this work is not a property of speakers, intentions, or roles in isolation. It is an operational property of language that travels through recurrent constructions. These constructions include, among others, agent deletion, nominalizations that compress agency, enumerative stacks that produce default scopes, deontic clusters, and formatting signatures that anchor obligation and control. Prior research shows that such forms function as portable carriers of executable power across domains and languages when specified as a compiled constraint set with auditable selection rules and thresholds (Startari, 2025a, 2025b, 2025d, 2025g). The regla compilada provides the formal conduit between local linguistic shape and downstream behavior. It defines what is eligible for activation, how activations aggregate, and which transitions follow when a construction profile crosses a verifiable bound.

Authority Entropy is a windowed measure of stance uncertainty. The stance space r is trinary by design, with values low, neutral, and high. The classifier maps a dialogue window x, defined as a contiguous slice of turns, to stance probabilities p(r | x). Entropy H is the negative sum of p times log p over r, using the natural log base. Low H indicates a concentrated stance distribution, therefore a strong and coherent authority signal. High H indicates dispersion, therefore uncertainty or competing authority pulls. The intuition is straightforward. When the local authority structure is settled, agents have fewer degrees of freedom, transitions narrow, and compliance or convergence can occur with lower negotiation cost. When the local authority structure is unsettled, the game samples more branches, delay rises, and coordination may degrade. This interpretation aligns with the

broader claim that legitimacy in modern systems travels through form and its compiled constraints, not through speaker essence or intention alone (Startari, 2025b, 2025e, 2025g).

The measurement regime is strictly causal. Windows are constructed with left context only. No tokens to the right of the decision point enter the input to the stance classifier. This constraint is essential for using Authority Entropy as an early-warning or steering signal. If the estimator were to see right context, it would convert from a predictive diagnostic into a descriptive summary. Causal masking is verified with unit tests that assert the absence of future indices in attention patterns and that fail if window boundaries are breached. This ensures that any association observed between entropy dynamics and outcomes is not confounded by lookahead leakage. The regime is compatible with human–human, human–model, and model–model arenas and supports mono and cross-lingual settings. Cross-lingually is accommodated by building the lexicon as a multilingual inventory with dialectal variants and by evaluating calibration per language family, since stance probabilities must be reliable across heterogeneous morphosyntactic carriers of authority (Startari, 2025a, 2025c).

Windowing follows a fixed turn count k with stride s. The default k is five turns with stride one. Each window is tagged with speaker identifiers, language, task state, and a snapshot of payoffs when available. The classifier consumes the text with optional speaker embeddings but without outcome tokens or task identifiers that would leak labels. The result is a time series H1 to HT that captures local authority uncertainty over the episode. Two derived quantities are critical for downstream use. Slope S is the rate of change of entropy across adjacent windows, estimated by a robust linear fit over a short horizon. Volatility V is either the variance of H over a horizon or the realized absolute change between successive windows. Low absolute slope and low volatility indicate steady authority regimes that are likely to sustain stable cooperative equilibria. High absolute slope or high volatility indicate instability regimes that may precede failure modes such as coordination breakdown, regret spikes, or error cascades.

The stance classifier is trained under weak supervision guided by the lexicon and adjudicated samples. Weak supervision supplies silver labels using construction matches and compositional rules. Human annotators then adjudicate stratified subsets to calibrate

per class precision and to refine edge cases, for example polite imperatives that express soft authority or enumerations that only signal structure without obligation. Targets for quality are set at agreement κ of at least 0.70 and per class F1 of at least 0.80 on the adjudicated test. Calibration uses temperature scaling on a validation split. These settings aim to constrain epistemic uncertainty so that entropy reflects genuine stance dispersion rather than model miscalibration. The lexicon is versioned, local scoped, and traceable to ensure that expansions can be audited and that ablations removing the lexicon meaningfully degrade performance. Such degradations are expected if authority indeed travels through form, since the lexicon encodes the form layer that the *regla compilada* elevates into decision relevance (Startari, 2025b, 2025d, 2025f).

This foundation yields a portable diagnostic. Authority Entropy is computed without right context, grounded in a multilingual construction inventory, and interpreted through slope and volatility to flag stability or instability regimes. The diagnostic is designed to be sector agnostic and to integrate with survival and causal outcome models. The central claim is testable. If lower entropy windows are associated with higher compliance, faster convergence, and more stable payoffs after controlling for sentiment, toxicity, politeness, and formality, then authority in dialogue is not merely sentiment or tone. It is a formal structure with measurable dynamics, compiled into constraints that shape behavior at the point of use. This completes the formal ground required to proceed to the lexicon and labeling protocol.

2) Lexicon and Labeling Protocol

This section specifies the linguistic inventory of authority-bearing constructions and the protocol used to label authority stance within dialogue windows. The objective is to obtain calibrated probabilities for $r \in \{\text{low}, \text{neutral}, \text{high}\}$ without right-context leakage, with full traceability of rules and dialectal variants. The lexicon is treated as an instance of *regla compilada*, namely an operational constraint set that links formal patterns to expected behavioral transitions. The working hypothesis is that authority does not primarily reside

in personal intentions. It resides in recurrent forms that activate, inhibit, or redirect courses of action during interaction.

2.1. Lexicon design

The lexicon is multilingual and dialect sensitive. Each entry follows a minimal schema: unique identifier, construction family, formal pattern, language and dialect variants, activation conditions, counter-cues, positive and negative examples, and ambiguity notes. Construction families cover at least the following operational categories. First, agent deletion and passive formats that foreground obligation or outcome. Second, nominalizations that compress agency and open space for implicit rules. Third, deontic and modal stacks that densify prescription. Fourth, enumerations with default scope and delimiters that function as a control structure. Fifth, normative formats and metadiscursive markers that announce decision or closure. Sixth, mitigated directives and politeness formulas that still carry effective illocutionary force. Seventh, it appeals to authority and source citations that anchor a decision. Each family includes parameterized patterns for languages with distinct morphology or canonical order and documents the pragmatic saturation in formal and colloquial registers.

Inclusion criteria require operational evidence on tasks, for example reduction of dissent, faster agreement, or reframing of perceived costs. Traceability is mandatory. Each new entry records provenance, the context in which it triggered a transition, and its performance in sensitivity tests. The lexicon is versioned with semantic change control. Every expansion undergoes interference audits to avoid collinearity with trivial signals such as affective polarity. Prior work shows that these forms transport effects beyond propositional content when modeled as activable, measurable constraints across heterogeneous corpora, which justifies their codification as a formal layer distinct from surface semantics or sentiment (Startari, 2025a, 2025b, 2025g).

2.2. Positive cues, counter-cues, and local context

Each formal pattern is documented with positive cues that reinforce its authority reading and counter-cues that weaken or cancel it. A deontic stack may degrade if it co-occurs with broad-scope conditionals or with deliberation metatags. Coding includes local context

windows and maximum distances for pattern by counter-cue interaction, so that the causal classifier never requires future tokens. Dialectal effects are recorded. In varieties where direct politeness maintains perlocutionary force, mitigation formulas are labeled as neutral or even high if dialogue history exhibits repeated compliance. This reduces cultural transfer bias and supports cross-lingual robustness of the entropy index.

2.3. Weak supervision and human adjudication

Labeling of r proceeds in two stages. First, weak supervision generates silver labels through lexicon matches and compositional rules. These rules consider the density of constructions, co-occurrence with closure or execution signals, and turn position. The preliminary window label is produced by robust aggregation, with family weights and penalties for ambiguous patterns. Second, human adjudication is performed on stratified samples by language, dialect, construction family, and model confidence level. The annotation guide defines operational criteria for low, neutral, and high. High stance is expected when executable directives or unequivocal normative closures are present. Low stance is expected when deliberative openings, broad conditionals, or attenuators establish symmetric footing. Neutral stance covers informative or coordination structures without clear prescription.

Quality thresholds on the adjudicated set are $\kappa \geq 0.70$ and per-class $F1 \geq 0.80$. Recurrent disagreements trigger revisions to the lexicon and the guide, never blind majority votes. Edge examples are archived with rationale and included as hard cases in training. Outcome tokens and task labels are excluded from adjudication inputs to prevent annotators from conflating authority with knowledge of the result.

2.4. Calibration, leakage audits, and class balance

After training the causal classifier with silver labels and adjudicated batches, $p(r \mid x)$ is calibrated per language on a validation split. Temperature scaling is applied, and reliability is verified through confidence versus accuracy curves, with low expected calibration error both per class and macro average. Leakage audits enforce strict window boundaries and left-context masking across the entire pipeline. Class balance is handled by stratified oversampling for low and high windows, without inflating trivial lexical shortcuts.

Coverage statistics by construction family and dialect are reported to prevent the model from confusing authority with dominant idiolects.

2.5. Decision criteria and version traceability

For time-series use, each window yields p(r | x) and entropy H. Decision rules raise alerts when the density of high-authority entries exceeds family-specific thresholds or when the family mixture matches configurations historically associated with convergence or conflict. Traceability includes lexicon version hash, seed set, and split boundaries. Any public release of the index lists the lexicon version, language coverage, weak supervision rules, and calibration estimators, enabling external labs to replicate results and compare models under equivalent causal regimes.

This design supplies an operational basis for reliable and comparable authority stance estimates. A versioned lexicon controlled human adjudication, and language-aware calibration jointly supports the validity of p(r | x). Authority entropy then becomes an informative and stable measure, ready to be linked to behavioral endpoints and to survival or causal effect models in the following sections.

3) Causal Classifier and Calibration

This section defines the model that maps a dialogue window x to stance probabilities p(r | x) over r ∈ {low, neutral, high} under a strict left-context regime, together with the procedures that verify masking, quantify uncertainty, and calibrate outputs. The design goal is to obtain probabilities that are both causally valid and decision useful, so that Authority Entropy H and its derivatives reflect genuine stance concentration rather than artifacts of leakage or miscalibration.

3.1 Input representation and causal masking

A window x is a contiguous slice of k turns with stride s. Each turn contains text and a speaker token. Optional features include language id and a bounded history of task state that excludes outcome tokens. Text is serialized as a flat sequence with segment delimiters

that mark turn boundaries. The model uses a unidirectional attention mask that prevents each token from attending to future tokens within the window. Windows are constructed only from left context relative to the decision point. No tokens to the right of the window boundary are admitted during training or inference. This constraint is verified by static and dynamic tests. Static tests check that the attention mask is strictly lower triangular for all batches. Dynamic tests delay the final t tokens of multiple windows, run forward passes with and without the delayed suffix, and assert that logits for positions preceding the delay remain bitwise identical within numerical tolerance. Any deviation flags a pipeline violation.

3.2 Model class and training objective

Two families are admissible. First, a compact causal transformer with rotary position embeddings, shared across languages. Second, a masked-to-causal adapter that converts a bidirectional encoder into a causal surface by zeroing right-context attention and pruning residual connections that bypass the mask. The output head is a three-class classifier trained with cross-entropy. To improve robustness, the loss includes class-balanced weights estimated from adjudicated frequencies, a small label-smoothing term ε in the range 0.02 to 0.05, and a focal factor only when minority classes fall below ten percent of samples. Regularization uses dropout on attention and feedforward layers, stochastic depth at low rates for deep variants, and weight decay tuned on a language-stratified validation grid. Early stopping monitors negative log likelihood. All splits enforce zero overlap of speakers and tasks between train, validation, and test.

3.3 Lexicon-aware features and ablations

The lexicon informs training in two ways. First, token spans that match authority-bearing constructions are marked by a binary feature stream. The model is not allowed to see lexicon confidence scores to avoid shortcut learning. Second, curriculum scheduling increases the proportion of windows that contain rare construction families during the first third of training, then reverts to the empirical distribution. Ablations remove span markers, randomize their positions, and drop the lexicon entirely. A meaningful drop in calibrated

F1 and a rise in entropy error when the lexicon is ablated support the claim that authority travels through form rather than sentiment proxies (Startari, 2025b, 2025g).

3.4 Uncertainty, reliability, and calibration

Raw softmax outputs are not assumed to be calibrated. The pipeline estimates Expected Calibration Error at class level and macro level, Brier score, and negative log likelihood. Temperature scaling is the default post-hoc method applied per language on the validation split, with a single temperature parameter T that minimizes NLL and is frozen before test time. When class imbalance or nonlinearity of miscalibration warrants it, isotonic regression is used as a sensitivity check, but temperature scaling remains the primary estimator to preserve monotonicity and prevent overfitting on small bins (Guo, Pleiss, Sun, & Weinberger, 2017). Reliability diagrams are reported for each class with binning chosen by the Freedman–Diaconis rule, and results are accompanied by bootstrap confidence intervals. The target is a macro ECE under two percentage points and classwise ECE under three percentage points on held-out data. To measure stability across distribution shift, calibration is re-estimated on stress partitions that vary register, stakes, and dialogue length. Large drift in T or ECE triggers a review of span features and sampling.

3.5 Leakage and proxy audits

Leakage audits include three tests. First, a right-suffix permutation test that appends neutral tokens to windows and asserts invariance of preboundary logits. Second, a delayed-outcome test that removes any token n-grams known to encode success or failure. If removal changes stance predictions materially, the windowing or weak-supervision rules are revised. Third, an influence-function probe that estimates token-level contributions to the loss. Tokens outside lexicon spans should rarely dominate attributions once content words are controlled. If outcome tokens or explicit reward numerals show high influence, the dataset filters are corrected. These audits ensure that causal validity is not compromised by shortcuts to results.

3.6 Cross-lingual control and dialectal parity

Since the lexicon is multilingual, the classifier exposes a language embedding. Calibration and performance are reported per language and macro averaged. Dialectal parity is monitored by grouping samples with shared dialect tags and computing gaps in ECE, Brier, and calibrated F1. Acceptable parity requires that gaps remain within two standard errors across dialect groups after controlling for construction family. If parity fails, additional dialectal variants are added to the lexicon, and the curriculum is adjusted to avoid overrepresentation of a dominant idiolect. This procedure reduces the risk that Authority Entropy reflects language imbalance rather than stance uncertainty (Startari, 2025a, 2025c).

3.7 Decision surfaces and entropy computation

At inference, each window yields calibrated p(r | x). Entropy H is computed with natural log. To stabilize H over short windows, probabilities are optionally smoothed by a convex combination of current and previous window outputs with a small coefficient α below 0.2. Slope and volatility are computed on the smoothed series only if calibration passes predefined thresholds. Thresholds for low-entropy regime are set by maximizing the Youden index on validation against a compliance outcome, then held constant on test. All decisions are logged with seed, model checksum, lexicon version, language id, and window boundary indices for replication.

3.8 Reproducibility controls

Training uses fixed random seeds at framework and CUDA levels, deterministic convolution and attention kernels when available, and exact recording of tokenization versions. Checkpoints are saved at the epoch with minimum validation NLL. The release includes scripts that reconstruct attention masks and replay leakage audits. These controls enable third parties to validate that Authority Entropy is causally derived, calibrated, and portable across languages and registers, which is necessary for the outcome models in the next section.

4) Metrics and Outcome Models

This section defines the measurement layer that links authority structure in dialogue to behavioral endpoints. The measurement layer outputs windowed indicators derived from stance probabilities p(r | x), then estimates their association with compliance, convergence, payoff stability, and regret under identification assumptions that exclude right context and control for lexical confounds. The aim is to make Authority Entropy H and its temporal dynamics decision useful for early warning and steering.

4.1 Windowed indicators

Each window x yields calibrated p(r | x) over r ∈ {low, neutral, high}. Authority Entropy is $H(x) = -\Sigma_r p(r \mid x) \log p(r \mid x)$, with natural log. Lower H indicates concentrated stance. Higher H indicates dispersion. Two temporal derivatives summarize dynamics. Slope S is the rate of change of H over contiguous windows estimated by a robust linear fit in a short horizon. Volatility V is either var(H) over the horizon or realized volatility defined as the mean absolute difference between adjacent windows. When calibration passes predefined thresholds, a smoothed series $\hat{H}t$ may be computed as $\alpha Ht + (1 - \alpha)Ht{-}1$ with α below 0.2. Low H with low |S| and low V characterizes stable authority regimes. High H or large |S| or large V characterizes instability regimes.

4.2 Behavioral endpoints

Compliance rate is the probability that an agent executes the relevant directive within a bounded turn budget after a window. Convergence time Tconv is the number of turns until the first stable joint policy is reached and maintained for a fixed dwell period. Payoff stability index measures post convergence dispersion, defined as one minus the coefficient of variation of payoffs over a fixed post convergence window. Regret is the gap between an oracle or best observed payoff and the realized payoff, averaged per episode. These endpoints are recorded at the episode level with precise turn indices, language tags, and stakes level.

4.3 Identification and controls

All analyses enforce left context inputs for the authority layer and exclude outcome tokens from features. Baselines include sentiment, toxicity, politeness, and formality. The objective is to establish that H and its derivatives contribute predictive lift beyond these baselines. Covariates include language, dialect, dialogue length, and domain. To mitigate confounding by topic or register, models include fixed effects for task template and stakes. All splits prevent speaker and task leakage across train, validation, and test.

4.4 Survival models for timing outcomes

Timing outcomes use discrete time survival analysis. The hazard of compliance at turn t is modeled as a function of Ht, St, Vt, and covariates. Two estimators are reported. First, a Cox proportional hazards variant discretized to turns, with baseline hazards stratified by language to absorb cross linguistic speed differences (Cox, 1972). Second, an additive hazards model to improve interpretability when proportionality is questionable (Aalen, 1989). Goodness of fit includes Schoenfeld style checks adapted to discrete time, calibration of predicted cumulative incidence, and time dependent AUC. Results are reported as hazard ratios or additive effects with 95 percent confidence intervals. Standard errors for proportions are also reported where applicable using $SE = \sqrt{[p(1 - p) \div n]}$, which makes explicit the uncertainty of window level compliance estimates when aggregated.

4.5 Causal effect estimation for payoffs and regret

To estimate the effect of entropy regimes on payoffs and regret, the analysis defines a low entropy indicator $L\tau$ that equals one if $H \leq \tau$ for a window. The threshold τ is selected on validation by maximizing the Youden index against a compliance label and then held constant. Causal effects are estimated with doubly robust learners that combine propensity models and outcome models with cross fitting to control bias (Chernozhukov et al., 2018). Causal forests are used as a nonparametric alternative that supports heterogeneous treatment effect exploration across languages and stakes levels (Wager & Athey, 2018). Identification assumes no hidden confounders after conditioning on observed covariates and the baseline tagger outputs. Sensitivity to unmeasured confounding is reported through Rosenbaum style bounds.

4.6 Thresholds, alarms, and decision rules

Three decision rules are defined for operational use. First, a low entropy regime alarm triggers when $H \leq \tau$ for m out of n consecutive windows. Second, a rising risk alarm triggers when S exceeds a positive threshold or when V over the last n windows crosses a historical percentile calibrated per domain. Third, a family mix alarm triggers when the distribution of construction families in the window matches a profile previously associated with conflict or error cascades. Decision rules are evaluated for precision, recall, and time to alarm relative to outcome onset. The expected cost of false alarms is estimated by measuring downstream intervention overhead in synthetic arenas where interventions can be scripted.

4.7 Model evaluation and ablations

Predictive lift is quantified by comparing models that use only baselines against models that add H, S, and V. Metrics include log loss for compliance prediction, time dependent Brier score, and concordance for survival tasks. For payoffs and regret, models are compared using mean absolute error and R^2 on held out episodes. Ablations remove the lexicon features, randomize span positions, or scramble right context positions while keeping token counts constant. If lexicon ablation wipes out most of the lift while sentiment controls remain unchanged, this supports the claim that authority travels through form rather than affect.

4.8 Uncertainty, calibration, and robustness

All predictive models report calibration curves and expected calibration error at class level and macro average for compliance, as well as reliability of time to event predictions through calibration belts. Bootstrap with at least one thousand resamples produces confidence intervals for effect sizes, regret gaps, and hazard differences. Multiple comparisons are controlled using Holm correction when families of hypotheses are tested. Robustness includes language specific re estimation, stakes stratification, and random seed perturbations to verify stability of findings across initializations.

4.9 Reporting and replication

Public artifacts include the exact lexicon version, windowing parameters, entropy base, split boundaries, seed sets, and evaluation scripts. Survival and causal notebooks specify preprocessing steps and hyperparameters. Leaderboard entries are accepted only when leakage audits pass, calibration errors are within predefined bounds, and replication scripts reconstruct the reported scores on a clean environment.

This measurement and modeling layer transforms local authority structure into actionable signals for coordination. By design, the outputs are portable across languages, causally valid by construction, and benchmarked against strong baselines. The next section specifies datasets and experimental design that realize these models in controlled and open settings.

5) Datasets and Experimental Design

This section describes datasets, collection protocols, and experimental factors that support identification without right context leakage and enable cross linguistic generalization. The design separates synthetic arenas, open multi party corpora with outcome labels, and consented human model dialogues that target realistic tasks. All artifacts are versioned and released with seeds, splits, and leakage audits. The goal is to measure the relation between Authority Entropy and behavioral endpoints while controlling for sentiment, toxicity, politeness, and formality baselines.

5.1 Dataset families and inclusion criteria

The program uses three dataset families. First, synthetic dialogue arenas with controllable goals, stakes, cost structures, and payoff matrices. These arenas instantiate cooperative, mixed motive, and adversarial games with explicit success conditions that are observable from turn structure alone. The synthetic layer supports counterfactual interventions, for example scripted insertion or deletion of authority cues at specified turns, and fine grained perturbation budgets for adversarial tests. Second, open multi party dialogue sets with public outcome labels. Representative sources include task oriented team chats, collaborative instruction following, and moderated debates where success, failure, or

stalemate is annotated at the episode level. These corpora provide ecological variation and language diversity that the synthetic layer cannot fully emulate. Third, consented human model dialogues collected under bounded tasks. Tasks include coordination to a shared plan, conflict resolution with time limits, and information triage under resource constraints. All human participation follows consent procedures, removal of personally identifying information, and documented redaction of administrative content.

Inclusion criteria are operational. Each candidate corpus must expose turn boundaries, allow construction of windows from left context only, provide episode level outcomes or proxies for convergence, and be licensable for research redistribution or reproducible extraction. Corpora that encode outcomes directly in the text of the final turns are admissible only if filters can remove those tokens before modeling. Each dataset receives a version number, a short textual rationale, and a summary of language and dialect coverage.

5.2 Data schema and preprocessing

All datasets are normalized to a common schema. At the turn level, records include a speaker identifier, timestamp if available, raw text, language tag, and optional task state without outcome tokens. At the window level, records include contiguous k turn slices with left context only, stance labels when available, and construction family counts derived from the lexicon. Preprocessing applies language aware normalization and tokenization, removes quoted administrative boilerplate, and redacts email addresses, phone numbers, and names. A deterministic pipeline produces train, validation, and test splits with zero overlap of speakers and tasks across splits. Split boundaries are recorded as index ranges to enable exact reconstruction.

5.3 Arms, factors, and blocking

The experimental matrix crosses arms and factors. Arms include human human, human model, and model model interactions. Factors include attention regime, stakes, and language condition. The attention regime toggles between causal only and bidirectional reading for comparison, although only causal inputs feed the Authority Entropy estimator. Stakes vary between low and high cost of error, with high stakes defined by longer dwell

times to convergence or larger payoff penalties for failure. Language condition covers monolingual experiments and cross lingual transfer where training and testing languages differ. Blocking is applied by domain and by dataset family. This reduces variance due to topic or platform idiosyncrasies, and it permits stratified reports that isolate language and stakes effects.

5.4 Windowing, sampling, and balance

Window size k and stride s are pre specified by validation studies and held constant for primary analyses. The default is k equal to five turns and s equal to one turn. Sampling procedures equalize class exposure for stance labels where feasible, and enrich rare construction families during model warm up. To avoid trivial lexical shortcuts, sampling does not oversample specific word types. Instead, it uses construction family tags and dialect tags to target underrepresented combinations. Each batch generator receives a random seed that is stored alongside the dataset hash to guarantee replicability.

5.5 Outcomes and annotation

Primary endpoints include compliance within a bounded turn budget after a window, convergence time to first stable joint policy, payoff stability after convergence, and regret versus an oracle or best observed policy. Secondary endpoints include error cascades, rework, and intervention counts for synthetic arenas where interventions are scripted. For open and human model corpora, adjudicators validate episode outcomes on stratified samples. The adjudication guide requires two independent labels and a tie breaking protocol. Agreement targets are Cohen’s κ of at least 0.70 with per endpoint reliability reported per language. When open corpora already contain outcome labels, the program validates a sample to document alignment with the present definitions.

5.6 Preregistration and evaluation protocols

Analyses are preregistered before final model training. The registration states windowing parameters, baselines, covariates, primary and secondary endpoints, and exclusion rules. The registration also specifies survival model types, calibration targets, and thresholds for low entropy regimes. Evaluation scripts are frozen and hashed. Leaderboard submissions

must pass leakage audits and replicate scores with the released scripts. The preregistration and the evaluation protocol prevent outcome drift, reduce analytic flexibility, and anchor external comparison.

5.7 Leakage, privacy, and risk controls

Leakage audits verify that the authority layer never ingests right context. Static checks validate lower triangular masks in all forward passes. Dynamic checks delay suffix tokens and assert invariance of preboundary logits. Privacy controls remove personally identifying data and redact administrative content. The pipeline stores only pseudonymous speaker identifiers. All releases document collection prompts, sampling settings, and model checkpoints to support downstream audits. Risk controls cover dialectal parity checks and language specific calibration. If gaps in expected calibration error exceed predefined margins, the release includes a corrective note and a plan for lexicon extension.

5.8 Stress tests and perturbation suites

Stress tests operate on synthetic and open corpora. Perturbation suites modify authority cues while preserving propositional content. Edits include deletion of deontic stacks, substitution of agentive for passive frames, injection of hedges, and modality flips that attenuate or intensify directive force. Each edit has a budget that limits character or token changes and a locality bound that restricts edits to the window. Tests report changes in Authority Entropy, compliance probability, convergence time, and regret. Sensitivity profiles are plotted per construction family and per language to reveal where the index is most informative.

5.9 Releasing artifacts and replication package

Public artifacts include raw to normalized transformation scripts, lexicon version and coverage, windowing parameters, split files, seeds, and checksums for all datasets. Synthetic arenas ship with generators that reproduce episodes from seeds. Open corpora are referenced by stable identifiers and accompanied by extraction and filtering scripts when redistribution is restricted. Human model dialogues are released in anonymized form where consent enables redistribution. The replication package includes notebooks for

survival and causal analyses, calibration reports, and ablation runners. All scripts are runnable in a clean environment and produce the reported tables and figures.

5.10 Rationale and relation to prior work

The dataset and design choices follow the view that authority travels through recurrent form and that the technical substrate for decision relevance is a compiled constraint set with auditable rules and thresholds. Prior work argues that formal mechanisms of authority can be measured independently from sentiment and from propositional meaning when linguistic structures are treated as operational carriers of effect (Startari, 2025a, 2025b, 2025c, 2025g). The present design makes that claim testable across languages and interaction regimes, with explicit controls for leakage, calibration, and fairness. By separating synthetic control from open variability and human model realism, the program provides both internal validity and external relevance.

6) Results, Stress Tests, and Ablations

This section specifies how results are produced, audited, and interpreted. It reports the predictive contribution of Authority Entropy H and its temporal derivatives to compliance, convergence, payoff stability, and regret. It then presents sensitivity analyses under adversarial edits that manipulate authority cues while preserving propositional content. Finally, it documents ablations that remove or perturb the formal layer provided by the lexicon in order to test whether measured effects depend on recurrent constructions rather than sentiment, toxicity, politeness, or formality baselines.

6.1 Main effects on behavioral endpoints

Analyses use the calibrated stance probabilities p(r | x), the corresponding H per window, the local slope S, and the realized volatility V. Compliance timing is evaluated with discrete time survival models that include baselines as controls. The primary specification is a Cox formulation with language stratification and fixed effects for task template and stakes level. A secondary additive hazards model provides effect sizes on an interpretable scale when proportionality is uncertain (Aalen, 1989, pp. 907 to 915; Cox, 1972, pp. 187 to 220).

Convergence time uses identical covariates and goodness of fit checks adapted to discrete time. For payoffs and regret, doubly robust learners estimate the marginal effect of low entropy regimes after conditioning on covariates and baseline taggers. Causal forests provide heterogeneous treatment effect profiles by language and stakes group (Chernozhukov et al., 2018, pp. C1 to C68; Wager and Athey, 2018, pp. 1228 to 1242).

The reporting standard is as follows. For compliance hazards, include hazard ratios per standardized unit decrease in H and per standardized unit increase in absolute slope. For convergence, provide median time differences between low entropy and non low entropy regimes with bootstrap confidence intervals. For payoff stability, report the change in the coefficient of variation within post convergence windows after exposure to low entropy segments, together with the effect on average regret per episode. All proportions include standard errors SE with SE equal to the square root of p times (1 minus p) divided by n. Calibration of probability outputs is reported with expected calibration error at class level and macro average, with temperature parameters per language and bootstrap belts around reliability curves (Guo, Pleiss, Sun, and Weinberger, 2017, pp. 1321 to 1330).

6.2 Early warning and lead time analysis

To assess steering value, the program quantifies lead time. An alarm is raised when H is at or below a fixed threshold τ in m out of n consecutive windows. Lead time is the difference in turns between the alarm and the first compliance event or the first transition to a stable joint policy. Precision, recall, and alarm time are reported across domains and languages. A useful early warning system must trade precision against lead time in a way that reduces regret after an allowed intervention cost. The analysis reports the net regret reduction that can be achieved with a fixed intervention budget on synthetic arenas where interventions are scripted.

6.3 Stress tests under adversarial edits

Stress tests operate on synthetic and open corpora. They target the authority layer by manipulating form while preserving propositional content. The following families are mandatory. Deontic deletion replaces layered modal stacks with neutral paraphrases. Agent restoration converts passives into active clauses with explicit agents. Hedge injection

inserts scope widening conditionals and softeners in directive contexts. Modality flips convert strong necessity into weak advisability or the reverse. Each edit adheres to a token budget and a locality bound within the evaluation window. The protocol measures changes in H, compliance probability, convergence time, payoff stability, and regret. The sensitivity profile of each construction family is summarized by the average change in H and the corresponding change in endpoint metrics. If authority travels through form, edits that neutralize formal carriers should elevate H and diminish compliance hazards, while intensifying edits should reduce H and accelerate convergence. Reports are stratified by language and stakes level to reveal cross linguistic and context dependent sensitivity.

6.4 Ablations and evidence against shortcuts

Ablations remove or perturb the formal layer in order to test dependence on recurrent constructions rather than affect or topic. Three ablations are mandatory. First, remove all lexicon span markers from the feature stream. Second, randomize the positions of span markers while preserving counts, which destroys local alignment between form and effect. Third, drop the lexicon entirely and retrain the classifier. For each ablation, compare calibrated F1 on stance classes, expected calibration error, the distribution of H, and the predictive lift on endpoints relative to a baseline that includes sentiment, toxicity, politeness, and formality. If the lift collapses when the formal layer is removed while affective baselines remain stable, the analysis supports the claim that the measured signal depends on constructional form rather than sentiment proxies. Leakage audits are re run after each ablation to ensure that changes in performance are not artifacts of right context contamination.

6.5 Cross linguistic generalization and dialectal parity

Results are reported per language and macro averaged. For stance prediction, provide classwise F1, macro F1, and calibration errors by language. For endpoint models, report hazard ratios and effect sizes by language with confidence intervals. Dialectal parity is assessed by grouping windows with shared dialect tags and computing gaps in calibration error, Brier score, and calibrated F1. Acceptable parity requires gaps within two standard errors after conditioning on construction family frequencies. When parity fails, document

which construction families drive the discrepancy and whether dialectal variants in the lexicon are under specified. The release includes an action plan to extend variants and re run calibration.

6.6 Error analysis and residual diagnostics

Residual diagnostics identify where the index is least informative. The analysis inspects failure clusters that share high H with positive outcomes or low H with negative outcomes. Two categories tend to arise. The first is strategic resistance, where a counterpart complies only after external incentives change. The second is formal mimicry, where superficial markers of authority appear without commitment to execution. For each category, annotate characteristic surface patterns and propose lexicon updates. Model diagnostics include influence function estimates to verify that tokens outside lexicon spans do not dominate losses once content words are controlled. If outcome numerals, explicit success markers, or administrative boilerplate show high influence, dataset filters are revised and the affected runs are flagged.

6.7 Robustness and multiple comparisons

Robustness checks include random seed perturbations, alternative window sizes and strides, and re estimation of calibration per language on stress partitions that vary register and dialogue length. When families of hypotheses are tested in parallel, p values are adjusted with Holm correction. This practice controls the family wise error rate without undue loss of power for a small set of primary endpoints. All bootstrap confidence intervals use at least one thousand resamples. Scripts that reproduce every figure and table are part of the public release.

6.8 Summary of evidentiary standards

A result is considered decision useful when four conditions hold. First, the authority layer passes leakage audits and exhibits classwise calibration with expected calibration error under the predefined bounds. Second, H and its derivatives add predictive lift beyond sentiment, toxicity, politeness, and formality on held out data with language stratification. Third, adversarial edits that neutralize authority cues increase H and degrade endpoints in

the predicted direction, with effects that exceed bootstrap uncertainty. Fourth, ablations that remove or randomize lexicon spans erase most of the lift while baseline taggers remain stable. If these conditions hold, the evidence supports the claim that local authority structure, measured as stance concentration and its dynamics, is a portable, language aware predictor of cooperative outcomes.

7) Validation, Public Specification, and Reproducibility

This section defines how the artifact is validated, specified for public release, and made reproducible by independent teams. The objective is to guarantee that Authority Entropy and its derivatives are causally valid, calibrated, fair across languages and dialects, and replicable on clean environments with the same scores and figures.

7.1 Validation goals and acceptance criteria

Validation targets four properties. First, causal validity. All stance estimates must come from left context only. Second, probabilistic reliability. Classwise and macro calibration errors must remain within predefined bounds. Third, fairness. Language and dialect gaps must be limited under conditioning on construction families. Fourth, replicability. A third party must obtain the reported scores with the provided seeds, splits, and scripts. Release acceptance requires passing leakage audits, meeting calibration targets, documenting fairness gaps and corrective actions, and reproducing all tables and figures on a clean runner.

7.2 Leakage audits

Two families of tests are mandatory. Static audits verify that attention masks are strictly lower triangular for every batch and that no residual connection bypasses the mask in masked to causal adapters. Dynamic audits delay suffix tokens and assert invariance of logits for positions before the delay within numerical tolerances. A right suffix permutation test appends neutral tokens and verifies that predictions for preboundary positions remain unchanged. Any violation blocks release until the pipeline is corrected and retested.

7.3 Calibration reports

Probability outputs p(r | x) are evaluated with reliability diagrams, expected calibration error by class and macro average, Brier score, and negative log likelihood. Temperature scaling is applied per language on the validation split and frozen for test. Reports include the fitted temperature, confidence versus accuracy plots with bootstrap belts, and sensitivity checks with isotonic regression as a secondary estimator when warranted. The target is macro ECE under two percentage points and classwise ECE under three percentage points on held out data (Guo, Pleiss, Sun, & Weinberger, 2017).

7.4 Fairness and parity audits

Parity is tested at two levels. Language level results include classwise F1, macro F1, and calibration errors. Dialect level results group windows by dialect tags and compute gaps in ECE, Brier, and calibrated F1 after conditioning on construction family frequencies. Acceptable parity requires that gaps lie within two standard errors. Failures are cataloged with the construction families that drive the discrepancy, the hypothesized cause, and the remediation, usually lexicon variant extensions and curriculum adjustments. All parity analyses are re run after remediation and attached to the release.

7.5 Specification files and artifact registry

Every public release ships a specification bundle with machine readable and human readable components. The machine readable component is a JSON document with fields for lexicon version and hash, language and dialect coverage, windowing parameters, entropy base, stance model checksum, training and inference seeds, split boundaries as index ranges, calibration temperatures, and alarm thresholds. The human readable component explains the assumptions, defines the indicators, and lists inclusion and exclusion rules. Both files are signed and versioned. Checksums cover raw to normalized transformations, trained weights, and evaluation scripts. A registry maintains version lineage, deprecations, and backward compatibility notes.

7.6 Replication package

The replication package includes five elements. First, data preparation scripts that recreate normalized datasets from raw sources or from documented extraction procedures when redistribution is restricted. Second, training scripts for stance models with exact random seeds, tokenizer versions, and environment files. Third, evaluation notebooks for survival models, causal learners, calibration reports, and fairness audits. Fourth, stress test and ablation runners that apply perturbation suites and feature removals. Fifth, a figure and table builder that regenerates all plots and tables in the paper. A single make target rebuilds the entire artifact on a clean machine image. Successful rebuild is a release condition.

7.7 Benchmark and leaderboard governance

Public benchmarking requires a submission template that enforces equivalence of causal access. Submissions must declare whether stance models are causal or bidirectional. Only causal submissions are eligible for the Authority Entropy leaderboard. Each entry must pass leakage audits and report per language calibration with temperatures learned on validation only. Submissions must include a short system card with training compute, energy estimate, and licensed dependencies. Scores are frozen with a hash of the evaluation scripts. The benchmark repository stores all accepted entries, the scripts used to verify them, and a changelog describing any later corrections. This governance follows the recent push in machine learning toward structured reproducibility commitments and checklists (Pineau et al., 2021).

7.8 Statistical reporting and uncertainty

The release reports standard errors for proportions with SE equal to the square root of p times one minus p divided by n, confidence intervals at the ninety five percent level, and calibration belts for probability reliability. Survival analyses report hazard ratios or additive effects with confidence intervals and diagnostics for proportionality when applicable. Causal effect estimates report doubly robust scores with cross fitting and bootstrap intervals. Multiple comparisons are adjusted with Holm correction when families of hypotheses are tested. All resampling uses at least one thousand draws.

7.9 External validation

External labs are encouraged to validate on additional corpora that satisfy inclusion criteria. The package includes a template for adding new datasets with language tags, dialect tags, and outcome definitions. Validation reports must replicate leakage audits, calibration checks, and fairness tests. Deviations are documented with suspected sources such as register, stakes, or construction family distribution shifts. When deviations are systematic and traceable to the lexicon, the next minor version incorporates new variants and updates calibration parameters. External validations are logged in the registry with data references and reproducible scripts.

7.10 Ethical, privacy, and licensing

All releases remove personally identifying information and redact administrative content. Human model dialogues include consent records and a description of task boundaries. Licenses for code and data are explicit. Where redistribution is restricted, extraction and normalization scripts are provided to enable third party recreation. Conflict of interest statements and funding disclosures are included in the human readable specification. Energy and compute budgets are reported as part of the system card when available.

7.11 Change management

Major versions correspond to structural changes such as stance space redefinition or windowing changes. Minor versions capture lexicon expansions, calibration adjustments, or bug fixes. Each change entry lists motivation, expected impact on scores, migration notes, and deprecation schedules. When a change affects comparability, the leaderboard is annotated with a boundary and legacy scores are archived under their original specification.

7.12 Summary

Validation, specification, and reproducibility controls are designed to make Authority Entropy a dependable measurement layer. Causal masking prevents lookahead contamination, calibration turns scores into reliable probabilities, fairness audits constrain language and dialect gaps, and a complete replication workflow enables independent teams to rebuild the artifact. Public governance aligns data, code, and specification so that evidence accumulates across versions rather than fragmenting into incomparable variants.

References

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine, 8*(8), 907–925.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double or debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

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

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B, 34*(2), 187–220.

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

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L^2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57*(4), 453–476. https://doi.org/10.1007/BF01025868

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

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

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

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Lin, H.-T., & Hutter, F. (2021). Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research, 22*(32), 1–20.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* (pp. 61–74). MIT Press.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-3692-2

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

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

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

Startari, A. V. (2025d). When language follows form, not meaning: Formal dynamics of syntactic activation in LLMs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5285265

Startari, A. V. (2025e). TLOC – The irreducibility of structural obedience in generative models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5303089

Startari, A. V. (2025f). Ethos without source: Algorithmic identity and the simulation of credibility. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5313317

Startari, A. V. (2025g). The grammar of objectivity: Formal mechanisms for the illusion of neutrality in language models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5319520

Startari, A. V. (2025h). Executable power: Syntax as infrastructure in predictive societies. *Zenodo*. https://doi.org/10.5281/zenodo.15754714

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839