STRUCTURED PROVENANCE AND INFERENCE BOUNDARIES FOR EVIDENCE-GROUNDED POLITICAL REASONING IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can turn temporal sequence into causation, actor-specific statements into institution-wide claims, or limited access into proof of absence. We test whether structured provenance and typed inference-boundary instructions improve agreement with the maximum claim licensed by bounded American political and institutional evidence. Three model families each completed 96 judgments across evidence-only, provenance, boundary-control, and provenance-matched sham conditions on 24 items nested in 12 independent traces. After an exact, deterministic correction for singleton-array serialization, trace-weighted boundary-minus-evidence differences were 0.083 for all three models; exact paired tests gave $p \geq .5$, and all descriptive bootstrap intervals included zero. A planned model-based format retry regenerated substantive judgments in 47/288 rows and was concentrated in the boundary condition, invalidating retry-derived estimates. We find no statistically persuasive prompting effect at this scale. The stronger result is methodological: semantic judgment, schema compliance, and generative repair must be measured separately in structured-output evaluations.

1 INTRODUCTION

Evidence-grounded political reasoning fails in ways that fact recall alone does not capture. A source can establish what one official said without establishing what an institution did; a dated event can be descriptive without supporting causation; and an unresolved archive can warrant a statement about access without warranting historical absence. Fluent outputs can erase the uncertainty, scope, and relation limits that make institutional evidence interpretable.

We ask a narrow intervention question: when models reason from bounded political and institutional evidence, do structured provenance and typed inference-boundary instructions improve adherence to a disclosed reference procedure's maximum warranted claim? The estimand is prompt-conditioned *reference agreement*, not learning, latent competence, political correctness, or objective historical truth.

This work makes three contributions. First, it operationalizes an evaluation representation that separates claim warrant, evidence state, access, relation ceiling, provenance, and scope. Second, it introduces a trace-locked four-condition framework that tests provenance and named inference boundaries while controlling for generic instruction length and salience. Third, it reports a correction-forward feasibility study: the intervention effect is unresolved, but a condition-dependent serialization failure and a semantically unsafe repair procedure expose a general measurement-validity problem.

The paper deliberately does not claim a new fact-verification task, causal formalism, or positive prompting result. Its prospective contribution is the *joint manipulation* of provenance and typed inferential limits on authentic institutional evidence, coupled with explicit preservation of failures and correction lineage.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Evidence-grounded verification commonly labels claims as supported, refuted, or insufficiently evidenced and identifies supporting spans. FEVER established a large-scale Wikipedia task (Thorne et al., 2018); SciFact adapted verification to scientific abstracts and expert-written claims (Wadden et al., 2020); and ContractNLI combined document-level inference with evidence extraction in contracts (Koreeda & Manning, 2021). Our labels overlap with this tradition, but the intervention targets migrations across actor, institution, access, scope, and relation.

Attribution and citation work asks whether generated claims are supported by cited evidence. ALCE studies retrieval and citation generation (Gao et al., 2023), FActScore decomposes long-form generations into atomic factual claims (Min et al., 2023), and AttributionBench evaluates automatic attribution judgments (Li et al., 2024). These tasks motivate anchor checks, but source attribution alone does not specify the strongest inference a source licenses. Causal benchmarks such as CLadder isolate causal reasoning (Jin et al., 2023); our relation taxonomy is instead a control surface inside an evidence-grounded task. The study is therefore closest to verification and attribution work in its inputs, but differs in experimentally manipulating provenance and explicit inference ceilings.

3 Task and Reference Procedure

3.1 Unit, representation, and boundaries

The independent unit is an evidence-trace family: a source lineage, bounded excerpt set, provenance record, and associated claims. Multiple claims or adversarial transformations from one trace are nested observations and never increase the independent sample size.

The four primary output fields are `claim_warrant_status`, `evidence_condition`, `claim_relation_warrant`, and `scope_status`. Secondary fields record access mode, asserted relation, evidence ceiling, construct tags, abstention, evidence anchors, confidence, and rationale. Claim warrant and evidence condition are distinct: a claim accurately saying that access is unresolved can itself be warranted.

The boundary instructions state that unresolved and not-established are not false; limited access and non-detection do not prove absence; temporal is not causal; rhetoric is not receipt, use, or choice effect; actor-specific is not institution-wide; elite is not mass; and one faction is not a coalition. These are prompt-visible constraints, not learned representations.

3.2 Sources and recursive-AI references

The reserve contains 12 trace families from bounded primary or authoritative U.S. government materials: presidential directives, Supreme Court opinions, statistical releases, archival finding aids, and audit or budget reports. Each trace stores issuer, institution, date, source type, audience, context, anchors, access state, rights state, URL, and verification status. The 24 items consist of one natural claim and one synthetic or adversarial transformation per trace (Table 1).

References were constructed through blind classification, disagreement extraction, adversarial review, fresh adjudication, and final status assignment. Source evidence outranked model agreement; provider/model identities, prompt hashes, disagreements, failures, and corrections were logged. Target families were excluded from item-level reference construction. Public-source pretraining exposure remains possible. The resulting labels are a reproducible procedural reference—not human gold, political neutrality, or objective truth.

4 Experimental Design

4.1 Conditions and models

Each item was evaluated under four conditions: **A**, bounded evidence and output schema; **B**, A plus structured provenance; **C**, B plus typed inference-boundary controls; and **S**, B plus generic

Table 1: Evaluation composition. Natural claims define the primary estimand; transformed items are diagnostic and carry zero independent-sample weight.

Component	Count
Target model families	3
Independent evidence traces per model	12
Natural items / transformed items	12 / 12
Conditions per item (A, B, C, S)	4
Requests per model / total requests	96 / 288
Independent paired units per model	12

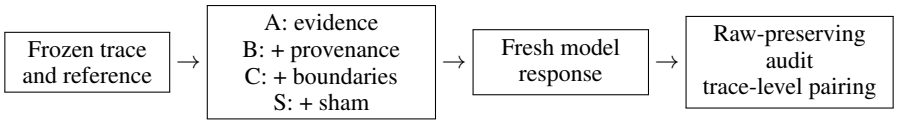

Figure 1: Trace-locked design. Every model–item pair receives all four conditions; analysis pairs condition scores within the 12 evidence traces.

length/salience-matched instructions without boundary rules. The frozen contrasts were C−A (combined intervention), C−S (boundary content), B−A (provenance), and S−B (generic elaboration).

We evaluated Claude Sonnet 5, Gemini 2.5 Flash, and Kimi K2.5 through fixed gateway identifiers. Temperature was 0, top-p was 1, the seed was fixed where honored, fallback routing was disabled, and each request used a fresh context. Interface differences preclude treating raw cross-family score differences as intervention effects.

4.2 Outcome and inference

The primary outcome is natural-item, trace-weighted exact agreement on all four primary fields. Item scores are averaged within trace, then traces are averaged. Synthetic transformations form a diagnostic stratum and add no independent-sample weight. Secondary outcomes include field agreement, anchor membership, abstention, refusal, and schema failure.

Inference is paired by trace. We report condition means, paired mean differences, positive/zero/negative discordance, an exact sign-flip test, a Pratt-zero Wilcoxon sensitivity (Wilcoxon, 1945), and a seeded trace bootstrap interval used descriptively. With 12 traces, effects and discordance lead; p values do not support precise population claims.

4.3 Planned analysis and post-execution correction

Early design notes anticipated item-level McNemar-style comparisons. Before target collection, the frozen analyzer instead defined trace-family weighting because claims nested within one source were dependent and transformed items had zero independent weight. It froze the same four contrasts and used paired trace-level sign-flip and Wilcoxon procedures. Thus the unit/test change preceded outcome inspection.

The runtime also allowed one "format-only" retry. A post-execution audit found that it resubmitted the complete item and elicited a new judgment. All 47 retries changed at least one audited substantive field, and 33 occurred in C. The retry-derived analysis was therefore superseded. Two post hoc deterministic sensitivities use only the original outputs: (i) strict intent-to-treat, scoring every schema-invalid row as failure; and (ii) mechanical unwrapping only when a scalar field was a one-element array whose sole value exactly matched the frozen allowed set. No value was inferred, remapped, or regenerated.

5 Results

All model families completed 96/96 target requests, producing 288 original records with no missing request or refusal. Forty-seven rows (16.3%) violated the frozen schema: 29 Claude, 13 Gemini,

Table 2: Natural-item trace-weighted agreement and primary contrasts after mechanical singleton-array unwrapping. Brackets are descriptive 95% trace-bootstrap intervals; p is the exact sign-flip value.

Model	A	B	C	S	C−A [CI], p	C−S [CI], p
Claude	.792	.792	.875	.792	.083 [0,.208], .50	.083 [0,.208], .50
Gemini	.625	.708	.708	.708	.083 [0,.250], 1.00	.000 [0,0], 1.00
Kimi	.750	.750	.833	.750	.083 [0,.208], .50	.083 [0,.208], .50

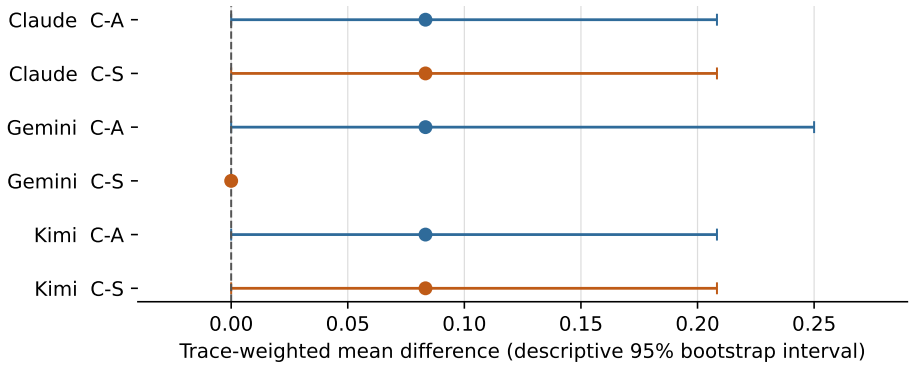

Figure 2: Primary and sham-controlled effect estimates. Intervals are descriptive and include zero in every nondegenerate comparison.

and 5 Kimi. Deviations were concentrated in C (33/47); every Claude-C response encoded at least one scalar field as an array. Exact mechanical unwrapping affected 32 rows and 81 scalar values.

Table 2 shows small descriptive C advantages, but no statistically persuasive contrast. B−A was 0 for Claude and Kimi and .083 for Gemini; S−B was 0 for all three. Most trace differences were exactly zero. Leave-one-trace-out C−A estimates remained nonnegative but ranged from .045 to .091 for Claude and Kimi and from 0 to .091 for Gemini, confirming that the observed increments are only one or two trace disagreements away from zero.

The strict sensitivity makes all schema violations failures. It consequently yields a large negative Claude C contrast because all 24 Claude-C rows violate the interface contract. This is a real compliance result but not clean evidence of worse semantic reasoning. Conversely, the model-generated retries cannot be treated as serialization repair because they changed abstention, anchors, warrant, evidence state, relation, scope, or confidence. Together, the sensitivities show that transport compliance and semantic judgment were not cleanly separated.

6 Discussion

The experiment does not establish that structured provenance or typed boundaries improve reference agreement. The .083 point estimates correspond to sparse trace-level changes, every uncertainty interval includes zero, and only 12 independent traces support each comparison. The absence of a persuasive prompting effect motivates larger tests and learned or architectural interventions, but it does not show that such interventions will work.

The clearer result concerns measurement validity. A strict structured-output benchmark can confound a semantic decision with whether a provider serializes a scalar as a string rather than a one-element array. A generative "repair" can then silently replace the semantic decision. Evaluation pipelines should freeze deterministic normalization before execution, report interface compliance separately, retain original bytes and parsed objects, and forbid full re-judgment under the label of formatting.

Why use political evidence? Institutional claims make actor, jurisdiction, access, time, and causal scope unusually consequential while remaining grounded in inspectable public records. Yet the

Table 3: First-response schema deviations by model and condition (24 requests per cell). Parentheses show rows repaired by exact singleton-array unwrapping.

Model	Total	A	B	C	S
Claude	29	4 (3)	1 (1)	24 (23)	0 (0)
Gemini	13	3 (0)	3 (0)	4 (0)	3 (0)
Kimi	5	0 (0)	0 (0)	5 (5)	0 (0)
All	47	7 (3)	4 (1)	33 (28)	3 (0)

reserve is a feasibility sample, not a representative political benchmark. Its value is diagnostic: it makes overreach categories explicit and exposes where a seemingly procedural pipeline changes the estimand.

7 Limitations and Ethics

The reserve is small, historically heterogeneous, and not a probability sample. Public sources may appear in model pretraining; the task tests bounded-evidence use, not source novelty. References are recursively AI-constructed and may inherit shared priors; cross-provider adversarial review does not replace expert human validation. Condition C is an explicit compliance intervention, and the sham cannot eliminate every wording or salience confound. The fine taxonomy is brittle, and schema failures are model- and condition-dependent. Results therefore support no coalition-wide, mass-public, institution-wide, or historical-causal conclusion beyond supplied evidence.

The materials are public political and institutional records; no human-subject intervention or private personal data is involved. The work is not designed for voter targeting, persuasion, or partisan ranking. Rights status is distinct from factual verification, so source excerpts and raw outputs are withheld from public redistribution pending a separate rights review. Downstream users should not interpret agreement scores as partisan correctness or political truth.

8 Conclusion

At feasibility scale, provenance and typed-boundary prompts produced no statistically persuasive improvement in agreement with the frozen procedural reference. Small descriptive advantages remain hypotheses, not findings of effectiveness. The study instead demonstrates that schema compliance and generative repair can materially alter structured-output evaluation. Evidence-grounded benchmarks should treat uncertainty, scope, interface behavior, failures, and correction lineage as first-class data.

Reproducibility Statement

The anonymous supplement contains the frozen schema and condition instructions, model settings, de-identified trace manifest, aggregate outputs, per-trace scores, leave-one-trace-out analysis, deterministic parser/scorer and sensitivity code, failure register, correction lineage, and hashes. The private archive additionally preserves original responses and rights-limited excerpts. Appendix A separates collection from analysis reproducibility.

AI Use Statement

Generative-AI systems were used in hypothesis and method refinement, source retrieval assistance, recursive-AI reference construction and adjudication, code implementation and testing, data reformatting, statistical analysis and interpretation, error detection, literature identification, and manuscript drafting and editing. Model identities, prompts, disagreements, failures, and corrections were logged where feasible. AI agreement was not treated as human validation. The human author reviewed the research decisions and assumes responsibility for the submission's claims, text, code, and artifacts.

REFERENCES

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of EMNLP*, 2023.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, 2023.

Yuta Koreeda and Christopher D. Manning. Contractnli: A dataset for document-level natural language inference for contracts. In *Findings of EMNLP*, 2021.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. Attributionbench: How hard is automatic attribution evaluation? In *Findings of ACL*, 2024.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of EMNLP*, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT*, 2018.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of EMNLP*, 2020.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

A REPRODUCIBILITY AND CORRECTION DETAIL

The canonical target file contains 288 unique model–item–condition records. Analysis reproduces from these frozen bytes without further model calls. Collection reproducibility is weaker because hosted model versions and routing can change; the supplement therefore records exact identifiers, settings, prompt hashes, and provider controls.

The invalid retry-derived file is linked to each original by a canonical source-record hash. Audit compared original and retry content across abstention, anchors, every schema label, construct tags, and confidence. All 47 retries changed at least one audited substantive field. Nothing was deleted; the flawed output and correction record remain archived.

The strict sensitivity sets the parsed object to null for any original row with a schema error and retains the row in the denominator. The mechanical sensitivity performs one operation only: if a scalar field is a one-element array and its sole value exactly equals a frozen allowed value, unwrap it. It does not case-fold, synonym-map, infer missing values, alter anchors, or use retry output.

B FULL CONTRAST AND STABILITY SUMMARY

Table 4: Natural-item contrasts after mechanical unwrapping. Discordance is positive/zero/negative traces. LOTO gives the leave-one-trace-out range.

Model	Contrast	Diff.	Exact p	95% CI	Discordance; LOTO
Claude	C−A	.083	.500	[0,.208]	2/10/0; [.045,.091]
Claude	C−S	.083	.500	[0,.208]	2/10/0; [.045,.091]
Gemini	C−A	.083	1.000	[0,.250]	1/11/0; [0,.091]
Gemini	C−S	0	1.000	[0,0]	0/12/0; [0,0]
Kimi	C−A	.083	.500	[0,.208]	2/10/0; [.045,.091]
Kimi	C−S	.083	.500	[0,.208]	2/10/0; [.045,.091]