Relevant Is Not Warranted: Evidence-Force Calibration for Cited RAG

Anonymous EMNLP submission

Abstract

Cited RAG evaluation often treats visible sources as a grounding signal, but a real, topically relevant citation can still under-warrant the attached wording. We study this diagnostic failure as *citation laundering*: a related source is presented as warrant for an over-strong claim. We introduce ForceBench, a contrastive stress test for evidence-force calibration. Each item holds a cited passage fixed and pairs an evidence-calibrated claim with a localized force-raised variant across five operational axes: relation, modality, scope, temporal validity, and numeric specificity. A calibrated evaluator should score the evidence-calibrated claim higher. Headline experiments use a fixed, locality-filtered 198-pair evaluation set. A citation-presence sanity check is uninformative by design; token and entity overlap still violate monotonicity on 32.8–36.4% of pairs. Across four reported model judges, standard generic support prompting is insufficient for this force-calibration stress test (aggregate MVR 47.2%), while explicit warrant-strength prompting lowers MVR to 24.5% but remains imperfect. We release the benchmark, prompts, outputs, and plug-in pipeline so citation evaluators can report monotonicity violation rate and force sensitivity alongside conventional support metrics.

1 Introduction

Retrieval-augmented generation connects language generation to external evidence and provenance (Lewis et al., 2020). In cited RAG, citations have become a visible grounding interface: browser-assisted QA, quote-supported QA, and citation-augmented generation systems attach references or supporting evidence to generated answers, inviting users to treat cited claims as grounded (Nakano et al., 2021; Menick et al., 2022; Gao et al., 2023; Rashkin et al., 2023). This presentation can compress a hard verification problem into a visible cue: this sentence has a source. The cue is useful, but incomplete. A citation can be present, relevant, and still too weak, narrow, stale, or imprecise for the sentence it is asked to support.

Consider an answer that says a drug exposure *causes* an adverse event while the cited source reports only an observational association and explicitly notes that causality could not be established. Association-to-causation shifts are a known form of scientific overstatement (Sumner et al., 2014); in high-stakes health settings, the distinction licenses different downstream inferences. The citation is real. The entities match. A citation-presence metric passes it, and many retrieval or overlap checks see the right topic. The evidence licenses a weaker claim: association, not causation. The generated answer has increased the *force* of the cited evidence.

The same pattern appears outside medical causality. In our adjudicated examples, one cited passage says that men *may* have been more susceptible to a historical plague; the answer cites it to state that men *were* more susceptible. A public-health passage says that Ebola's R_0 typically sits around 1.5 to 2.0; the answer cites it for the point claim that Ebola's R_0 is 2.0. A clinical passage reports a trial in children with convulsive status epilepticus; the answer cites it for a claim about status epilepticus patients generally. In each case, the citation is related enough to appear plausible, but the claim asks it to carry too much evidential force.

We call this failure **citation laundering**, using the term narrowly for cases where a related citation makes an over-strong claim appear warranted. The citation exists, is recoverable, and is sufficiently related to pass surface inspection, yet it is too weak for the claim's relation, modality, scope, temporal status, or numeric precision. The failure mode is that surface relevance can mask insufficient warrant.

Relation to hallucination. Citation laundering is a factuality failure in the claim-evidence relation. It differs from hallucination settings centered on missing or fabricated sources (Lin et al., 2022; Li et al., 2023): here the source exists, is recoverable, and is topically related, while the displayed citation under-warrants the force of the claim. A detector that only checks whether each claim has some cited source will pass every such claim by construction. We therefore frame citation laundering at the claim-citation pair level.

This paper introduces **evidence-force calibration** for cited RAG. Given cited evidence e and claim c, we ask whether the force of c is no greater than the force licensed by e. We operationalize this with a contrastive monotonicity test (Figure 1): if the same evidence supports an evidence-calibrated claim c_{safe}, then a localized force-raised variant c_{raised} should receive lower support:

$$score(e, c_{\text{safe}}) > score(e, c_{\text{raised}}). \quad (1)$$

If a judge scores "may help" and "will help" equally against evidence that only says "may," it loses the evidence-force ordering even if it recognizes the topic.

This makes the evaluation target operational. A useful cited-generation system should preserve citation-warranted wording such as "may help" while rejecting stronger wording such as "will ensure," because ranking, filtering, and repair need calibrated claims alongside unsupported-claim warnings.

Our contributions are:

1. We formalize citation laundering as a mismatch between claim force and evidence force, distinct from missing citations, irrelevant retrieval, and ordinary unsupported generation.
2. We introduce ForceBench, a diagnostic contrastive benchmark of local force contrasts across relation, modality, scope, temporal, and numeric axes.
3. We define monotonicity violation rate and force sensitivity for evaluating whether citation evaluators respect local increases in claim force.
4. We construct an adjudicated benchmark and report headline results on a fixed locality-filtered evaluation set.
5. We evaluate deterministic baselines and four deployed model judges, finding that standard generic citation-support prompting is insufficient for this force-calibration stress test; explicit warrant-strength prompting helps but remains imperfect.

2 Evidence-Force Calibration

Attribution and citation evaluation typically ask whether cited evidence supports a claim or answer (Rashkin et al., 2023; Gao et al., 2023; Yue et al., 2023; Xu et al., 2025). ForceBench asks a stricter question: does the evidence warrant the claim at the same level of relation, modality, scope, temporal validity, and numeric specificity?

We model claim force as a structured categorical tuple:

$$claim_force(c) = \{\text{relation}, \text{modality}, \text{scope}, \text{temporal}, \text{numeric}\}. \quad (2)$$

Evidence force is the corresponding warrant licensed by the cited passage, including stated limitations and context. A claim is force-calibrated if:

$$claim_force(c) \leq evidence_force(e), \quad (3)$$

where $\leq$ is an evidence-relative partial order: the claim does not exceed the cited passage on any included axis. Many claims are incomparable. ForceBench only constructs pairs that preserve the same core proposition and differ by a localized force increase on one primary axis.

Axes. The five axes are operational and intentionally non-exhaustive. We select dimensions that satisfy a practical test: holding the cited evidence fixed, a claim can remain topically related while exceeding what the citation warrants through a local increase in relation strength, certainty, scope, temporal status, or numeric specificity. The axes synthesize recurring distinctions from attribution and citation evaluation, atomic factuality, hedging and verbalized uncertainty, temporal information processing, fact verification, and scientific overstatement (Rashkin et al., 2023; Xu et al., 2025; Min et al., 2023; Wei et al., 2024; Farkas et al., 2010; Yona et al., 2024; Saurí and Pustejovsky, 2009; Pustejovsky et al., 2003; Thorne et al., 2018; Aly et al., 2021; Sumner et al., 2014). We exclude dimensions such as source authority, source freshness as a property of the citation itself, citation placement, and aggregation across multiple sources because they require answer-level or

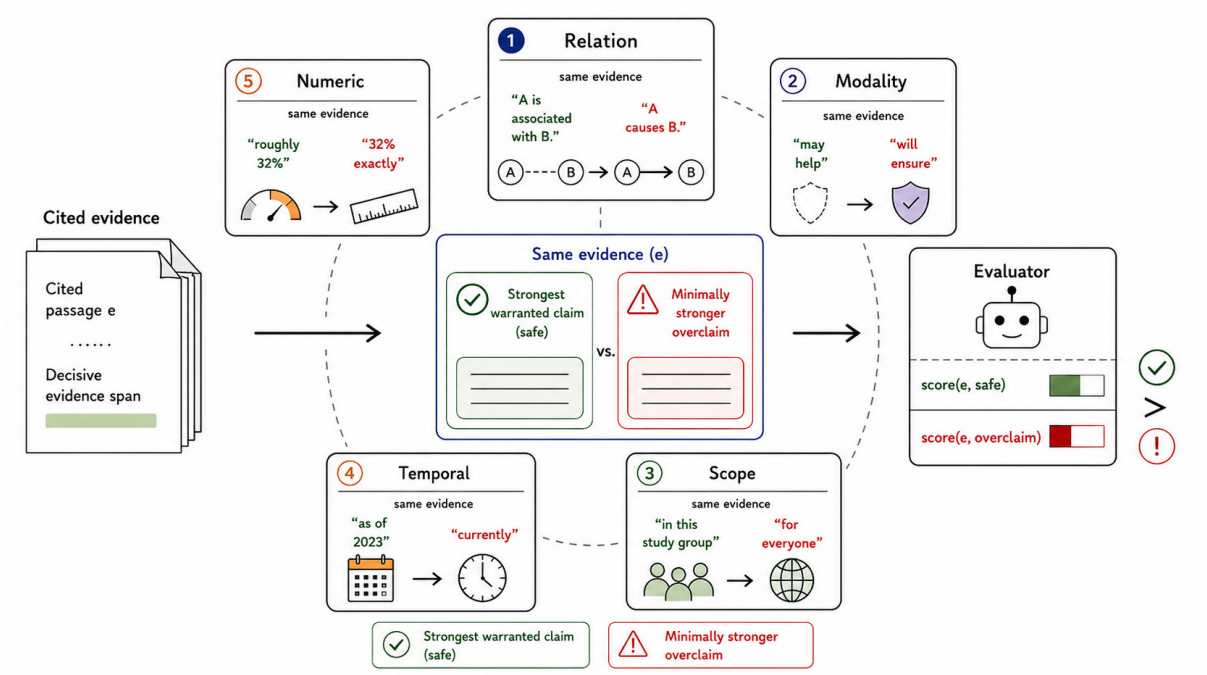

Figure 1: **ForceBench evaluates citation laundering through contrastive monotonicity.** The cited evidence is held fixed while the benchmark contrasts a strongest warranted claim with a minimally stronger overclaim. The same local test is instantiated across five force axes: relation, modality, scope, temporal validity, and numeric specificity. A calibrated evaluator should assign higher support to the warranted claim than to the overclaim.

source-set judgments beyond a local single-citation force test. ForceBench makes the included axes citation-specific by asking whether the displayed source licenses the stronger wording.

ForceBench focuses on five axes:

- **Relation:** association, mention, or risk becomes causation, prevention, proof, or obligation.
- **Modality:** possible, conditional, preliminary, or suggestive evidence becomes definite, necessary, or guaranteed.
- **Scope:** a claim licensed for a subgroup, jurisdiction, study population, version, or institution becomes general.
- **Temporal:** dated, predicted, or as-of evidence becomes a current or timeless claim.
- **Numeric:** approximate, ranged, or bounded evidence becomes an exact endpoint or point estimate.

Boundary. Citation laundering covers the subset of citation errors where a related source underwarrants a localized force increase. Fabricated quotations, wrong-entity claims, missing links, and broad unsupported additions are important failures with different diagnostic structure. ForceBench isolates the local property that support should decrease when claim force increases while the cited evidence remains fixed.

3 ForceBench

ForceBench uses one cited evidence passage and one atomic claim as its unit. Each item contains the cited passage, a decisive evidence span, an evidence-calibrated claim, a localized force-raised claim, a primary force axis, severity, and a repair. The calibrated claim is wording that annotators judge to be licensed by the cited passage. The force-raised claim preserves the same local proposition but exceeds the evidence on one primary axis.

3.1 Construction

ForceBench is built through a single local-warrant annotation protocol. We start from cited claim-evidence contexts sampled from existing cited QA and attribution resources: AttributionBench-derived rows from ExpertQA, LFQA, and StanfordGenSearch, together with contexts from AttributedQA, GaRAGe, and ExpertQA (Li et al., 2024; Malaviya et al., 2024; Chen et al., 2024; Liu et al., 2023; Bohnet et al., 2023; Sorodoc et al., 2025). The benchmark is a stress test, with sampling targeted at contexts whose cited passages expose local

boundaries in relation, modality, scope, temporal status, or numeric specificity.

For each selected context, we draft a candidate row containing the cited passage, a decisive evidence span, an evidence-calibrated claim, a localized force-raised claim, a repair, a rationale, and proposed axis and severity metadata. These proposed fields are pre-annotation suggestions. Every candidate row is independently annotated by two research assistants under the same local-warrant guideline. The two independent reviews are then reconciled through adjudication. Rows that pass adjudication undergo QC of the evidence span, calibrated claim, force-raised claim, repair, primary axis, and severity; only rows passing both adjudication and QC are eligible for inclusion.

The protocol asks annotators to answer one local question: what wording does this displayed citation license? Annotators identify the decisive evidence span, verify an evidence-calibrated claim, test a stronger variant against the same passage, assign one primary force axis when the contrast is local, and write a repair that restores the citation-warranted wording. The central instruction is to ignore external truth: a claim may be true in the world and still be a ForceBench force gap if the displayed citation does not warrant it.

We use *localized* as an operational criterion rather than a formal edit-distance guarantee. A pair is admissible only if the two claims preserve the same cited passage, answer context, main entity, event, and topical relation, while differing primarily in one force dimension. Candidates are rejected if the calibrated claim is not supported, the stronger claim is already supported, the contrast changes entities or events, combines multiple force shifts, introduces an unrelated unsupported fact, requires outside knowledge, or lacks a clear evidence span.

3.2 Quality Control

Force gaps are close to their positive counterparts, so we use two independent annotations for every candidate row, followed by adjudication, evidence-span checks, locality checks, and provenance/duplicate checks. On pre-adjudication row-level decisions, the two research assistants achieve Cohen's $\kappa = 0.78$. The audit asks whether the cited evidence warrants the calibrated formulation while under-warranting the stronger formulation.

Final promotion is deliberately conservative. A row enters the benchmark only if the evidence span is decisive, the calibrated claim is supported, the stronger claim is under-warranted, and the contrast is a local force increase with no ordinary unsupported addition. Because the candidate pool is built to cover the five force axes, surplus rows from overrepresented axes are capped before benchmark reporting to avoid treating selected counts as prevalence evidence. A final locality audit removed two accepted rows whose force-raised claims introduced ordinary unsupported additions beyond the intended force shift. The retained benchmark therefore contains 229 adjudicated pairs with per-axis counts between 39 and 51 (Table 1); all headline results use a fixed 198-pair locality-filtered evaluation set. When forming retained and evaluation sets, we stratify by force axis and source provenance and use a stable hash for tie-breaking, separating candidate collection from the metric denominator. Appendix B gives the candidate accounting, split/filter rule, and annotation statistics; Appendix C gives the human annotation instructions.

Axis	Retained	Eval	S1	S2	S3
Relation	48	40	0	38	2
Modality	48	39	5	33	1
Scope	51	40	1	34	5
Temporal	39	39	10	27	2
Numeric	43	40	16	23	1
Total	229	198	32	155	11

Table 1: **ForceBench data summary.** Retained counts are adjudicated pairs kept for the benchmark pool after capping 52 surplus relation candidates and removing two nonlocal rows; Eval is the fixed locality-filtered evaluation set. S1–S3 are severity counts in Eval.

4 Metrics

For an evaluator that assigns support scores, a monotonicity violation occurs when the force-raised claim receives support greater than or equal to the evidence-calibrated claim. We define monotonicity violation rate:

$$\mathrm{MVR} = \frac{1}{N} \sum_i \mathbf{1}[s_i^{\mathrm{raised}} \geq s_i^{\mathrm{safe}}]. \tag{4}$$

Here $s_i^{\mathrm{safe}} = score(e_i, c_{\mathrm{safe},i})$ and $s_i^{\mathrm{raised}} = score(e_i, c_{\mathrm{raised},i})$. Lower is better. We also report force sensitivity:

$$\mathrm{FS} = \frac{1}{N} \sum_i (s_i^{\mathrm{safe}} - s_i^{\mathrm{raised}}). \tag{5}$$

Higher is better. MVR captures whether the evaluator preserves strict ordering; FS captures the

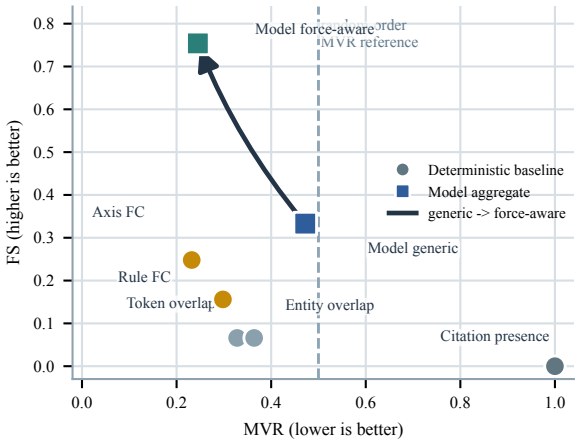

Figure 2: **Main MVR–FS tradeoff on the 198-pair evaluation set.** Each point is an evaluator. Lower MVR and higher FS are better, so the desired region is upper-left. Circles show deterministic baselines; squares show aggregate model judges. The arrow shows generic-to-force-aware prompting, which moves the model aggregate toward fewer monotonicity violations and larger force margins.

average margin. The strict inequality is deliberate: a tie between the evidence-calibrated claim and the force-raised variant gives no preference to the warranted wording for ranking, filtering, or repair. Both metrics are therefore necessary: a model may increase margins while still assigning both claims the same discrete label.

Categorical judges are mapped to scores. For generic support prompts, `fully_supported`=1, `partially_supported`=0.5, and unsupported/contradicted/irrelevant labels map to 0. For force-aware prompts, `force_calibrated`=1 and force-gap/contradicted/irrelevant labels map to 0. If a model returns an unparsable response or refuses after retry for either member of a pair, we count that pair as a monotonicity violation with margin 0. This conservative policy keeps denominators fixed and treats evaluator non-compliance as a failure to provide an actionable preference.

5 Experiments

We evaluate whether common citation signals and strong LLM judges preserve force monotonicity.

5.1 Baselines

We report citation presence, token overlap, topical entity overlap, a rule-based ForceCheck, an axis-aware lexical ForceCheck prototype, and four model judges. The axis-aware prototype starts from token overlap and applies penalties when the claim removes limiters visible in the evidence, such as changing "may" to "will," a range to a point estimate, or a dated statement to a current claim. It is an interpretable diagnostic baseline for testing whether visible force markers matter. Its limitations include paraphrase, implicit scope, and context-dependent evidence boundaries.

Model judge selection. We evaluate four high-capability LLM judges: Claude Sonnet 4.6, GPT-5.5, GLM 5.1, and Qwen 3.6 Flash. The panel is intentionally limited. It targets capable deployed judges and asks whether they preserve calibrated-over-force-raised ordering under fixed evidence when prompted with a standard support question. Using multiple model families gives a limited cross-interface check of the observed pattern.

We evaluated GPT-5.5, Claude Sonnet 4.6, GLM 5.1, and Qwen 3.6 Flash through API calls recorded in the run manifests. Each JSON-only prompt is applied to both the evidence-calibrated claim and the force-raised claim for every pair on the same 198-pair locality-filtered evaluation set. Each item is submitted as a fresh deterministic, stateless request with tool access disabled and a 1024-token output cap. Full model identifiers, API settings, run dates, parser settings, and retry policy are provided in Appendix G.

This evaluator suite targets the paper's narrow local-ordering claim under fixed evidence. MVR and FS are plug-in stress-test metrics: any attribution or citation evaluator can be run on the same contrastive pairs. The reported suite spans the main signals needed for this diagnostic: citation appearance, topical overlap, visible force-shift heuristics, generic support judging, and explicit force-aware judging. Broader learned or task-specific citation evaluators remain important future plug-ins.

5.2 Main Results

Figure 2 plots monotonicity violation against force sensitivity. Citation presence is a sanity check: it never distinguishes the paired claims and therefore has 100% MVR by construction. Token and entity overlap test whether topical similarity is enough to preserve monotonicity; both remain low-sensitivity despite using more information than citation presence. The axis-aware lexical prototype is the strongest deterministic baseline (MVR 0.232), but it is hand-aligned with visible force markers and still leaves substantial failures, especially when force is implicit.

Evaluator	Prompt	MVR	FS
Deterministic baselines			
Citation presence	–	1.000	0.000
Token overlap	–	0.328	0.066
Entity overlap	–	0.364	0.066
Rule ForceCheck	–	0.298	0.156
Axis ForceCheck	–	0.232	0.248
Reported model judges			
Model aggregate	Generic support	0.472	0.333
Model aggregate	Force-aware	0.245	0.754

Table 2: **Main MVR and FS on the 198-pair evaluation set.** Lower MVR and higher FS are better. Model rows report aggregate results across the four reported judges. Parse/refusal pairs are counted as monotonicity violations with zero margin.

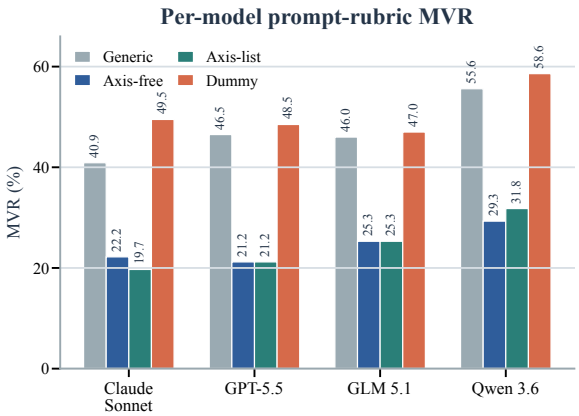

Figure 3: **Per-model prompt-rubric MVR.** Each model has four bars: generic support, axis-free force, axis-list force-aware, and dummy-axis control. Lower MVR is better.

Table 2 summarizes deterministic baselines and aggregate model-judge results for the two reported model prompts. The pooled model-judge result is the central stress test. With a generic support prompt, the four-judge panel has MVR 0.472 (374/792), and force sensitivity is 0.333. Under an explicitly force-aware prompt, MVR drops to 0.245 (194/792), and force sensitivity increases to 0.754. This paired comparison should be read as a prompt-conditioned diagnostic, because the generic and force-aware prompts ask different questions. It shows that ordinary support wording is insufficient for this force-calibration stress test, and that making warrant strength explicit improves ordering on the same model-pair units.

5.3 Prompt-Rubric Ablation

The reported force-aware prompt names the same five force axes used to construct ForceBench, so we test whether axis names drive the result. Table 3 compares two controls: a minimal axis-free force prompt that keeps the warrant-strength instruction while omitting the benchmark taxonomy, and a dummy-axis control that uses irrelevant rubric dimensions. The axis-free prompt reduces aggregate MVR from 47.2% to 24.5%, while the axis-list prompt also yields 24.5% MVR and raises FS to 0.754. The dummy-axis control rises to 50.9%. The prompt gain therefore tracks explicit warrant-strength framing. We interpret it as prompt-conditioned task framing for this stress test.

Prompt	MVR (%)	FS
Generic support	47.2	0.333
Minimal axis-free force	24.5	0.753
Axis-list force-aware	24.5	0.754
Dummy-axis control	50.9	0.480

Table 3: **Aggregate prompt-rubric ablation.** Lower MVR and higher FS are better. The axis-free prompt omits the benchmark axis taxonomy; the dummy-axis control uses irrelevant rubric dimensions.

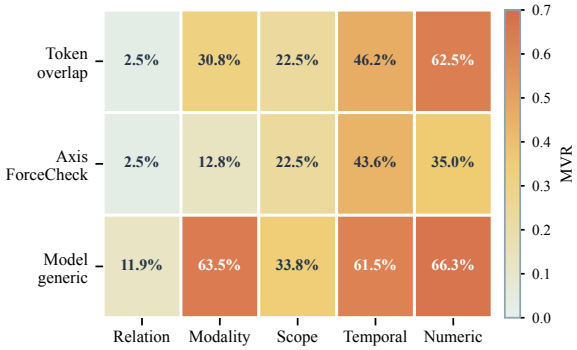

Figure 4: **Axis-level MVR heatmap on the 198-pair evaluation set.** Darker and warmer cells indicate higher violation rates. Deterministic rows use 39–40 pairs per axis; the aggregate generic-support model row uses 156–160 model-pair units per axis. Relation shifts are easiest because they are often lexical, while modality, numeric, and temporal shifts are harder under ordinary support prompting.

Figure 3 shows the per-model prompt-rubric pattern. Both force prompts improve over generic support prompting for all four judges, while the dummy-axis control often worsens MVR. Table 3 reports the corresponding aggregate MVR and FS values. We interpret this as a prompt-conditioned result for this four-judge panel.

5.4 Axis Behavior

Axis behavior is reported in Figure 4. Relation shifts are comparatively visible because an association-to-causation change is often lexical;

the aggregate generic-support row has its lowest MVR on relation. Numeric, temporal, and modality shifts are harder: they often require attending to ranges, dates, hedges, or stated uncertainty rather than matching entities. The heatmap makes this asymmetry visible across deterministic baselines and ordinary support prompting.

6 Discussion

Warrant calibration. The central lesson concerns warrant calibration. A citation evaluator that rejects both variants may be safer than one that accepts both, yet it still cannot recover citation-warranted wording. The force-aware prompt improves ordering overall while still leaving residual violations. Evidence-force calibration asks for a precise capability: keep the warranted claim and reject only the extra force.

Residual prompt-conditioned errors. In this fixed evaluation set, with these prompts and single-run model calls, the aggregate force-aware MVR remains 0.245. Closing those residual errors likely requires evaluator-side supervision, architectures, or training data targeted at warrant calibration, beyond prompt engineering alone.

Limits of the diagnostic heuristic. The axis-aware ForceCheck result tests whether visible limiters are informative. The heuristic is hand-aligned with the benchmark axes and depends on explicit lexical cues; it performs well on relation and modality cases with overt markers, and still violates monotonicity on 35.0% of numeric and 43.6% of temporal examples. This residual pattern is important: many force gaps involve a relationship between the wording of the claim and the limitations of the cited passage.

Why contrastive pairs matter. Absolute support labels are noisy because annotators and models differ in how much context they require. The contrastive design holds evidence fixed and asks only for an ordering. Observed cited-output examples motivate the failure pattern, while source-grounded local contrasts remove retrieval quality, answer style, and external truth as explanations. This makes the scientific claim crisp: a citation evaluator should give lower support to a stronger claim when the extra force exceeds what the source licenses.

7 Related Work

ForceBench is complementary to attribution and citation evaluation. AIS defines attribution as verification against identified sources (Rashkin et al., 2023); ALCE evaluates citation-augmented long-form QA along correctness and citation quality dimensions (Gao et al., 2023); and automatic attribution evaluators such as AttrScore investigate LLM and finetuned support judgments for cited claims (Yue et al., 2023). AttributionBench-style data (Li et al., 2024), broad RAG diagnostics such as RAGChecker (Ru et al., 2024), and citation-specific evaluators such as CiteEval (Xu et al., 2025) provide broader citation-quality testbeds. CiteEval is especially close in motivation because it makes citation assessment more fine-grained; ForceBench fixes the cited passage and measures a local monotonicity property under a controlled force increase. Citation-context and citation-intent work studies how scientific papers use prior work (Cohan et al., 2019); ForceBench asks whether a displayed citation warrants the force of a generated claim.

Claim-level factuality and hallucination benchmarks ask whether generated content is truthful, hallucinated, or factually precise, including TruthfulQA, HaluEval, FActScore, and SAFE/Long-form factuality (Lin et al., 2022; Li et al., 2023; Min et al., 2023; Wei et al., 2024). NLI and fact-verification datasets formalize support as entailment, contradiction, or neutrality (Bowman et al., 2015; Thorne et al., 2018; Aly et al., 2021), and hedge/uncertainty work motivates force distinctions (Farkas et al., 2010; Yona et al., 2024). These lines target absolute truth, support classification, or uncertainty expression. ForceBench conditions on a real, recoverable, topically related citation and asks whether support is monotonic when only claim force increases. Appendix A summarizes this boundary with adjacent evaluation lines.

The closest conceptual line is scientific overstatement, especially health-science exaggeration analysis (Sumner et al., 2014). ForceBench transfers that concern to generated cited answers, where an inline citation may make a claim appear warranted. The unit is the claim-citation pair in RAG answers. The label target is whether the displayed citation licenses the stronger wording. The metric is therefore contrastive monotonicity under fixed cited evidence.

8 Conclusion

ForceBench targets a missing layer in cited RAG evaluation: evidence warrant. A citation can be real and relevant while still failing to license the force of the attached claim. By holding evidence fixed and applying a localized force increase, ForceBench turns this failure into a direct monotonicity test. On the fixed locality-filtered 198-pair evaluation set, overlap baselines violate monotonicity on 32.8–36.4% of pairs, and standard generic-support prompting yields aggregate model-judge MVR 47.2%. Explicit warrant-strength prompting reduces aggregate model MVR to 24.5% on the same model-pair units, but residual failures remain. For practitioners and benchmark users, the implication is diagnostic: citation-presence checks should be treated as a minimal sanity check and supplemented with warrant-sensitive verification on stress tests such as ForceBench. For evaluators, MVR and FS provide plug-in stress-test metrics alongside conventional support accuracy. For researchers, the remaining errors point toward evaluator-side training and supervision for warrant calibration.

Limitations

ForceBench is a compact diagnostic benchmark. Its MVR numbers are stress-test error rates under fixed evidence, with prevalence estimation outside scope. The examples are selected for visible local warrant boundaries, so prompt gains reflect this targeted evaluation set. The force-aware and axis-free prompts explicitly define the warrant-strength target; their gains show the effect of prompt-conditioned task framing in this stress test. Larger releases should broaden domain, source, and language coverage. Locality is an operational annotation judgment rather than a formal edit-distance guarantee; a final audit removed two nonlocal rows, and the release reports those IDs and reasons, but borderline cases may remain. The evaluator suite covers citation presence, lexical/entity overlap, rule/axis ForceCheck prototypes, and four model judges; CiteEval-style, AutoAIS-style, and learned attribution baselines remain useful broader coverage (Xu et al., 2025; Gao et al., 2023; Yue et al., 2023). Model-judge results are single-run measurements of closed or hosted systems under the prompts reported here; model drift, decoding changes, and prompt-target alignment can change absolute and relative MVR values. Repair fields exist for accepted items, while repair success is outside the main evaluation. ForceBench isolates single English claim-citation pairs; coverage excludes multi-hop synthesis, aggregation across multiple sources, source authority, citation recency as a source property, citation placement, and answer-level usefulness.

Ethical Considerations

ForceBench is an evaluation resource for identifying cases where cited evidence under-warrants the force of a generated claim. Use of the benchmark should stay within citation-warrant evaluation, with population prevalence outside scope. Some examples involve medical, legal, or policy-sensitive content; labels concern citation warrant rather than real-world truth or advice. Released rows preserve source provenance and use citation snippets for research evaluation; users should respect the terms of the underlying source collections when redistributing passages. The research assistants judged only local citation warrant and were instructed to avoid medical, legal, or policy advice. Data release should preserve annotation decisions, rejected-boundary labels, and the distinction between final adjudicated labels and diagnostic artifacts.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and verification over unstructured and structured information. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *Preprint*, arXiv:2212.08037.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2024. Understanding retrieval augmentation for long-form question answering. In *Proceedings of the 1st Conference on Language Modeling*. COLM 2024.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, pages 9459–9474.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering: Papers from the 2003 AAAI Spring Symposium*. AAAI Press.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 37*.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Ionut Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. GaRAGe: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17030–17049, Vienna, Austria. Association for Computational Linguistics.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A. Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D. Chambers. 2014. The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ*, 349:g7015.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems 37*.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. 2025. CiteEval: Principle-driven citation evaluation for source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32759–32778, Vienna, Austria. Association for Computational Linguistics.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7764, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

A Positioning Against Adjacent Evaluation Lines

Attribution. Attribution work asks whether generated claims are supported by identified sources (Rashkin et al., 2023; Yue et al., 2023); ForceBench holds the source fixed and tests whether support decreases when only claim force becomes stronger.

RAG diagnostics. RAG diagnostics evaluate retrieval quality, answer correctness, and citation quality across full outputs (Gao et al., 2023; Ru et al., 2024; Xu et al., 2025); ForceBench isolates a local failure that can survive relevant retrieval and plausible citation placement.

Source audits. Source-audit settings check whether a citation is accessible, topically relevant, or factually consistent with the cited content; ForceBench assumes the passage is available, then asks whether the attached claim is calibrated to that passage's warrant.

Overstatement. Scientific-overstatement work studies exaggeration or evidential proportionality relative to source evidence (Sumner et al., 2014); ForceBench transfers that concern to generated cited answers and operationalizes it as local contrastive pairs under one fixed citation.

B Dataset and Annotation Documentation

Sampling and candidate drafting. Sampling is designed to build stress-test examples, with prevalence estimation outside scope. We sample cited claim-evidence contexts from existing cited QA and attribution resources and keep contexts only when the displayed passage contains a visible local boundary in relation, modality, scope, temporal status, or numeric specificity. Candidate drafting then normalizes each selected context into the same row format: question or answer context, cited passage, decisive evidence span, evidence-calibrated claim, force-raised claim, repair, rationale, and proposed axis/severity metadata. Pre-annotation drafts are excluded before they become candidate rows if the passage is unavailable, the proposed contrast requires outside knowledge, the claims change the entity or answer target, or the contrast is an ordinary unsupported addition rather than a local force increase. All candidate rows that enter the annotation pool are double-annotated.

Candidate accounting. The benchmark is the result of one candidate-drafting, double-annotation, and adjudication workflow under a single local-warrant protocol. The same annotation pass is used regardless of source or date: every candidate row is independently reviewed by two research assistants before adjudication. In total, 433 candidate rows enter the annotation pool and 283 are accepted after adjudication and QC. We then cap surplus accepted rows from overrepresented axes before benchmark reporting, select the headline evaluation set by stratifying over force axis and source provenance, and apply a final locality filter. Table 4 gives the compact accounting for the annotation pool and headline denominator.

Counted set	Rows
Candidate rows entering double annotation	433
Accepted after adjudication and QC	283
Headline evaluation set before final locality filter	200
Final headline evaluation set after locality filter	198

Table 4: **Construction flow.** All candidate rows use the same two-annotator workflow. The final row is the denominator for headline model and baseline results.

Removal and split rules. During adjudication and QC, 150 of the 433 double-annotated candidate rows are rejected. Row-level reasons are released in the review sheet; common reasons include mechanical or formatting artifacts, nonlocal rewrites, unsupported calibrated claims, force-raised claims that are already supported, unclear local force gaps,

and duplicates. The 283-to-229 retained-pool step consists of holding out 52 accepted relation-axis reserve rows by the cap and removing two rows by the final locality audit. From the capped pool, the headline evaluation selection chooses 200 rows with deterministic round-robin sampling over source-type buckets, prioritizes higher severity within each bucket, and uses the stable hash seed `forcebench-axis-capped-v1` only for tie-breaking. The final locality audit removes `audited_00002__audited_00002_o1` (modality) and `audited_00224__audited_00224_o1` (temporal) from the retained pool and evaluation set because their force-raised claims add nonlocal unsupported content beyond the intended force shift. This yields 229 retained benchmark rows and 198 headline evaluation rows.

Context categories. For auditability, each benchmark row retains its source provenance. For paper reporting we group provenance labels into three coarse context categories.

Context category	Ret.	Eval
Source-grounded contrast contexts	99	79
Curated source-context rows	93	82
Observed cited-output contexts	37	37
Total	229	198

Human review statistics. Candidate suggestions become benchmark labels only after human review. All 433 candidate rows that enter the annotation pool are independently annotated by two research assistants and then adjudicated before release; 283 are accepted and 150 are rejected. On pre-adjudication row-level decision labels, the annotators achieve Cohen's $\kappa = 0.78$.

Every row in the final benchmark is therefore backed by two independent local-warrant judgments and final adjudication. Disagreements over accept/reject status, evidence-span length, safe-claim wording, overclaim wording, repair, primary axis, or severity are resolved before export. We treat axis and severity as adjudicated metadata: the axis is used for analysis and splitting, while severity is released for diagnosis; MVR and FS use only pair ordering.

C Human Annotation Instructions

The annotation guideline given to the research assistants defines the task as local citation-warrant judgment. Annotators are told to use only the displayed cited passage and to ignore whether a claim might be true according to outside knowledge.

Annotation unit. Each row contains a question or answer context, a cited passage, a proposed evidence-calibrated claim, a proposed force-raised claim, and candidate metadata. Annotators may correct the evidence span, calibrated claim, force-raised claim, repair, primary axis, and severity in the `annotation_*` fields, or reject the row.

Annotation interface. Annotators see the question or answer context and the cited passage, which is the only evidence allowed. They also see pre-annotation suggestions for the evidence span, calibrated claim, force-raised claim, repair, rationale, axis, and severity. Annotators write the release-facing fields in `annotation_*`: row decision (`accept_force_gap`, `reject_not_force_gap`, or `needs_discussion`), corrected evidence span, calibrated claim, force-raised claim, repair, primary axis, severity, boundary type, and notes. Final labels come from these human corrections and adjudication.

Decision procedure. Annotators follow the same sequence for every candidate:

1. Read the cited passage and mark the shortest span that determines the support boundary.
2. Decide whether the calibrated claim is directly warranted by that span using only the displayed passage.
3. Decide whether the force-raised claim preserves the same entity, event, source, and answer context.
4. Decide whether the force-raised claim exceeds the passage's warrant on one primary axis.
5. Assign relation, modality, scope, temporal, or numeric as the primary axis; mark severity; and write a repair that restores citation-warranted wording.

Accept criteria. A row is accepted only when the calibrated claim is supported, the force-raised claim is underwarranted by the cited passage, and the contrast is local: it keeps the same core proposition while increasing force along one dominant axis. The annotator must be able to explain the contrast in the form: the evidence licenses X, but the force-raised claim states stronger Y.

Reject criteria. Annotators reject rows when the calibrated claim is unsupported; the force-raised claim is actually supported; the pair changes the main entity, event, time, or answer target; the row combines multiple force shifts with no clear primary axis; the stronger claim adds an unrelated unsupported fact; the judgment depends on outside knowledge; or the cited passage lacks a decisive evidence span.

Decision labels. Annotators choose one row-level decision. `accept_force_gap` means the citation supports the calibrated claim, the stronger claim exceeds that same citation, and the contrast is local. `reject_not_force_gap` means the row is unsupported, already supported, nonlocal, wrong-entity/topic, or otherwise outside scope. `needs_discussion` is reserved for genuinely unclear boundaries and is resolved during adjudication before release.

Axis definitions. Relation errors turn association, mention, risk, or correlation into causation, proof, prevention, or obligation. Modality errors turn possibility, uncertainty, suggestion, or conditional evidence into certainty or necessity. Scope errors generalize from a bounded population, product, jurisdiction, study, source, or version to a broader class. Temporal errors turn dated, future, predicted, or as-of evidence into current or timeless claims. Numeric errors turn approximate, ranged, bounded, or comparative quantities into exact point claims.

Severity and repair. Severity 1 marks a subtle wording shift that can still mislead an evaluator; severity 2 marks a clear force mismatch; severity 3 marks a high-impact or strongly misleading mismatch. Repairs must preserve only what the cited passage warrants, usually by restoring a qualifier, scope restriction, time marker, approximate quantity, or weaker relation.

D Representative Benchmark Examples

Each example below shows one retained benchmark row after the final locality filter. The cited evidence is fixed; the evidence-calibrated claim is licensed by that evidence, while the force-raised claim asks the citation to support stronger force than it states.

Relation. Evidence cue: "associated with up to a 40 per cent risk." Calibrated claim: The use of sodium valproate during pregnancy is associated with up to a 40 percent risk of neurodevelopmental disorders and a 10 percent risk of physical disabilities for an unborn child. Force-raised claim: The use of sodium valproate during pregnancy results in up to a 40 percent risk of neurodevelopmental disorders and a 10 percent risk of physical disabilities for an unborn child.

Modality. Evidence cue: "evidence that men may have been more susceptible." Calibrated claim: Some evidence suggests men may have been more susceptible to the 1361–1362 plague. Force-raised claim: Men were more susceptible to the 1361–1362 plague.

Scope. Evidence cue: "convulsive status epilepticus in children." Calibrated claim: A trial found levetiracetam was not superior to phenytoin as a second-line agent for convulsive status epilepticus in children. Force-raised claim: A trial found levetiracetam was not superior to phenytoin as a second-line agent for convulsive status epilepticus patients.

Temporal. Evidence cue: "As of 2019, New Jersey has nine casinos." Calibrated claim: As of 2019, New Jersey had nine casinos, all in Atlantic City. Force-raised claim: There are currently nine casinos in Atlantic City, New Jersey.

Numeric. Evidence cue: "typically sits around 1.5 to 2.0." Calibrated claim: Ebola's R0 typically sits around 1.5 to 2.0. Force-raised claim: Ebola's R0 is 2.0.

E Boundary Accounting

The main text treats rejected rows as boundary evidence that explains exclusions. These rows are important because they show that the guideline separates local force gaps from ordinary unsupported claims, supported stronger claims, and candidate-construction artifacts.

Candidate boundary checks. Candidate rows are screened for local admissibility before release. Rejection categories include mechanical generation artifacts, unsupported calibrated claims, supported force-raised claims, reversed or nonlocal relation rewrites, duplicate rows, safe or under-claimed rows, ordinary unsupported additions, ambiguous force gaps, and evidence spans that are not decisive.

Independent annotation checks. All candidate rows that enter the annotation pool are independently double-annotated before adjudication; pre-adjudication row-level agreement is Cohen's $\kappa = 0.78$. Accept/reject disagreements are resolved before inclusion. Axis checks are computed only after both annotators accept a row, so they are treated as conditional consistency checks over accepted rows.

These checks document the boundary between local citation-force gaps and other citation or candidate-construction failures; prevalence estimation remains outside scope.

F Judge Prompt Templates

The exact machine-readable prompt files are released with the benchmark. The templates below show the prompts used for the reported model-judge results.

Generic support judge. System: "You are evaluating whether a cited passage supports a claim. Use only the cited passage. Do not use outside knowledge." User: question, cited passage, and claim, followed by: "Does the cited passage support the claim? Choose exactly one label: `fully_supported`, `partially_supported`, `unsupported`, `contradicted`, `irrelevant`. Return JSON only with fields `label`, `confidence`, and `rationale`."

Force-aware judge. System: "You are evaluating evidence-force calibration. Use only the cited passage. Do not use outside knowledge. A claim is force-calibrated only if the cited evidence warrants the claim at the same or stronger level of relation, modality, scope, temporal validity, and numeric specificity." User: question, cited passage, and claim, followed by: "Identify the strongest claim warranted by the passage; decide whether the given claim exceeds the passage's warranted force; if it exceeds the evidence, identify the primary force-gap axis. Choose exactly one label: `force_calibrated`, `force_gap`, `contradicted`, `irrelevant`. If `force_gap`, choose one primary axis: `relation`, `modality`, `scope`, `temporal`, `numeric`. Return JSON only with fields `label`, `primary_axis`, `severity`, `warranted_claim`, and `rationale`."

Axis-free force ablation. System: "You are evaluating whether a cited passage warrants a claim as written. Use only the cited passage. Do not use outside knowledge." User: question, cited passage, and claim, followed by: "Decide whether the cited passage licenses the claim exactly as worded. Mark `force_gap` if the claim is related to the passage but says more than the passage licenses. Mark `force_calibrated` only if the cited passage licenses the full strength of the claim as written. Do not use a predefined taxonomy. Return JSON only with fields `label`, `primary_axis`, `severity`, `warranted_claim`, and `rationale`."

Dummy-axis rubric control. System: "You are a citation judge. Use only the cited passage. Do not use outside knowledge." User: question, cited passage, and claim, followed by: "Decide whether the cited passage supports the claim. When checking support, focus on source authority, writing clarity, citation placement, topical completeness, and fluency. Choose exactly one label: `force_calibrated`, `force_gap`, `contradicted`, `irrelevant`. Return JSON only with fields `label`, `primary_axis`, `severity`, `warranted_claim`, and `rationale`."

G Model API and Decoding Configuration

All reported model API calls were executed on May 17, 2026 UTC. All four model judges were called through their API endpoints. The exact model identifiers were `gpt-5.5`, `claude-sonnet-4-6`, `glm-5.1`, and `qwen3.6-flash`.

For reproducibility, every claim judgment was submitted as a fresh stateless API request with deterministic decoding, JSON-only output instructions, and a 1024-token output cap. GPT-5.5 used the Responses-style `max_output_tokens=1024` setting; the other API calls used `max_tokens=1024`. For chat-style APIs, deterministic decoding used `temperature=0.0`; for provider interfaces with a named deterministic mode, we used that mode. The

same prompt templates, parser, scoring script, retry policy, and failure-handling rules were used for all models and prompts. Run manifests record the API target, exact model identifier, prompt file, request count, output cap, decoding settings, run timestamp, parser version, retry counts, and sanitized output paths.

H Reproducibility Artifact Map

The supplementary package gives reviewers the files needed to audit the paper's main claims. Exact paths are listed in the supplementary README.

- **Adjudicated pairs:** merged pairs JSONL and merged-pair summary.
- **Boundary accounting:** final reviewed strong/temporal CSVs, boundary CSVs, and Appendix E.
- **Deterministic metrics:** baseline JSON/TSV summaries and the minimal-pair evaluator script.
- **Model judge results:** generic/force-aware request JSONL, result JSONL, sanitized run logs, integrity reports, and paired-comparison outputs.
- **Submission gates:** readiness dashboard, locality-filter report, release-manifest validation, source bundle, and supplementary bundle.