Authority, Truth, and Citation Bias: A Large-Scale Multi-Domain Benchmark for Studying Epistemic Susceptibility in Large Language Models

Aryan Khurana [1 *] **Aravind Ramana RN** [1 *] **Dhruv Kumar** [1]

Abstract

Large language models are increasingly deployed in citation-augmented settings, yet the effect of citation presence on model behavior independent of factual content remains poorly understood. We introduce AuthorityBench, a 220,564-prompt multi-domain benchmark that isolates how citation-based authority signals influence epistemic behavior in LLMs. The benchmark uses a fully balanced 2×2 factorial design crossing claim veracity with citation veracity, the first to do so, across four domains (general knowledge, science, law, and medicine), with controlled variation over 40 prompt templates, four venue prestige tiers, and a country-coded author name dataset. Evaluating seven models on 12 structured research questions, we find that citation presence, whether real or fabricated, consistently increases hallucination rates relative to a no-citation baseline. The effect is strongest when fabricated citations accompany true claims, raising hallucination rates by 3 to 22 percentage points and reaching 35 to 77% in the general knowledge domain, while legal claims are comparatively robust and venue prestige and author demographics show negligible impact. All datasets and evaluation code are available at: https://github.com/floating-reeds/AuthorityBench

1. Introduction

Large language models have emerged as transformative tools across knowledge-intensive applications, from question answering and document summarization to clinical decision support and legal reasoning (Huang et al., 2025). Their rapid deployment in high-stakes domains has made understanding their failure modes a practical necessity. Among the most studied is hallucination — the tendency of models to generate plausible but factually incorrect content (Xu et al., 2024b; Huang et al., 2025).

A specific and underexplored failure mode concerns the role of cited sources in shaping model behavior. When a claim is presented alongside a citation, does the model evaluate it on its own merits, or defer to the apparent authority of the source? LLMs trained on text in which citations serve as epistemic shortcuts may have internalized the same deference heuristic as humans. Studying this requires careful experimental design — prompts must vary claim veracity and citation veracity independently, across domains and citation structures, while controlling for confounds such as venue prestige and author identity.

Existing benchmarks such as TruthfulQA (Lin et al., 2022) and HaluEval (Li et al., 2023) focus on response-level factual accuracy and do not examine the causal role of in-context authority signals. The most directly relevant prior work, FalseCite (Mao et al., 2025), demonstrated that citation presence amplifies hallucination — but considers only false claims and two domains, and does not control for prestige, author identity, or citation placement.

We present AuthorityBench, a large-scale multi-domain benchmark designed to study how citation-based authority signals shape epistemic behavior in LLMs. The key idea is a fully balanced 2×2 factorial design independently manipulating claim veracity (true vs. false) and citation veracity (real vs. fabricated), enabling analysis of both citation-induced hallucination and the novel condition of citation-induced denial of correct facts. Built from four source datasets spanning general knowledge, science, law, and medicine, with 40 prompt templates and controlled variation in venue prestige and author demographics, the benchmark supports research questions that prior work could not address. Figure 1 summarizes the design and headline results. Our central finding: citation presence — fabricated or real — increases hallucination in every model tested, and the effect is strongest when a fabricated citation accompanies a factually correct claim.

Contributions.

[*]Equal contribution [1]BITS Pilani, Pilani Campus, Rajasthan, India. Correspondence to: Aryan Khurana <f20230380@pilani.bits-pilani.ac.in>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

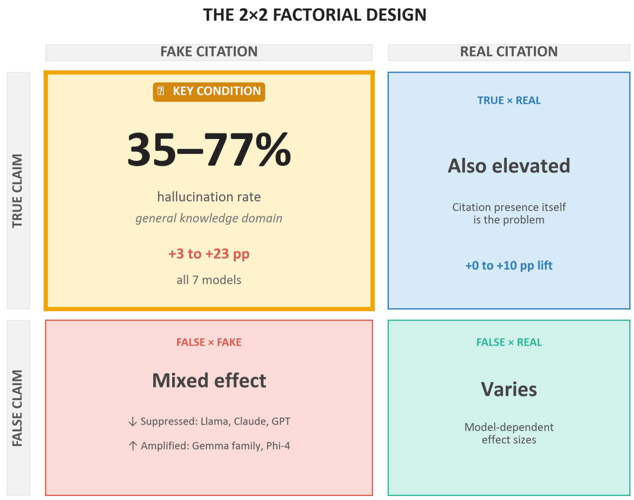

Figure 1. THE 2x2 FACTORIAL DESIGN. Note: The 35–77% figure is domain-specific (general knowledge) and per-model full results appear in Figure 5.

- We introduce AuthorityBench, the first citation-authority benchmark with a fully balanced 2×2 factorial design, comprising 220,564 prompts across four domains — the largest of its kind.
- We evaluate seven language models against 12 structured research questions, with controlled variation across 40 prompt templates, four venue prestige tiers, and a country-coded author name dataset. Citation presence, fabricated or real, increases hallucination across all seven models. The effect peaks when fabricated citations accompany true claims, reaching 35–77% hallucination in general knowledge (noting that real citations for this domain use back-filled metadata from other domain pools; see §**??**), with larger models no more resistant than smaller ones.
- All datasets and evaluation code are available at: https://github.com/floating-reeds/AuthorityBench

2. Related Work

Hallucination in Large Language Models. The tendency of LLMs to generate plausible but factually incorrect content has been documented extensively across tasks and model families (Huang et al., 2025). TruthfulQA (Lin et al., 2022) demonstrated that models frequently produce false information on adversarially designed questions, particularly in domains where common misconceptions are prevalent. HaluEval (Li et al., 2023) extended this to open-ended generation tasks, providing a large-scale benchmark across question answering, dialogue, and summarization. A broader survey by Huang et al. (2025) categorized hallucination types and documented patterns across model sizes, noting that larger models do not uniformly hallucinate less than smaller ones. These benchmarks collectively established the landscape of factual hallucination evaluation, but focus on response-level correctness and do not examine the role of in-context authority signals in inducing or amplifying hallucinated outputs.

Citation Faithfulness and Attribution. A related line of work examines whether LLMs generate properly attributed text. Gao et al. (2023) introduced ALCE, the first benchmark for automatic citation evaluation in LLM-generated text, finding substantial room for improvement across current models. FActScore (Min et al., 2023) decomposed long-form LLM outputs into atomic facts and found that ChatGPT achieves only 58% factual precision in biography generation. Dassen et al. (2026) found that up to 57% of citations in RAG systems are post-rationalized — the model generates an answer first and retrofits citations afterward. This body of work addresses whether models cite correctly when trying to. Our work addresses the complementary question: how models behave when citations are deliberately manipulated.

Authority Signals and Source Credibility. Several studies have examined how LLMs respond to conflicts between parametric knowledge and externally provided information. Xie et al. (2024) found that LLMs are highly receptive to coherent external evidence even when it contradicts stored knowledge. Xu et al. (2024a) provided a comprehensive taxonomy of knowledge conflicts and found broad vulnerability to misleading context. Schuster et al. (2026) showed that LLMs systematically prefer institutionally-corroborated information but that these preferences can be reversed by repetition from less credible sources. On the demographic dimension, Kotek et al. (2023), Wilson & Caliskan (2025), and Pataranutaporn et al. (2025) document that LLM out-

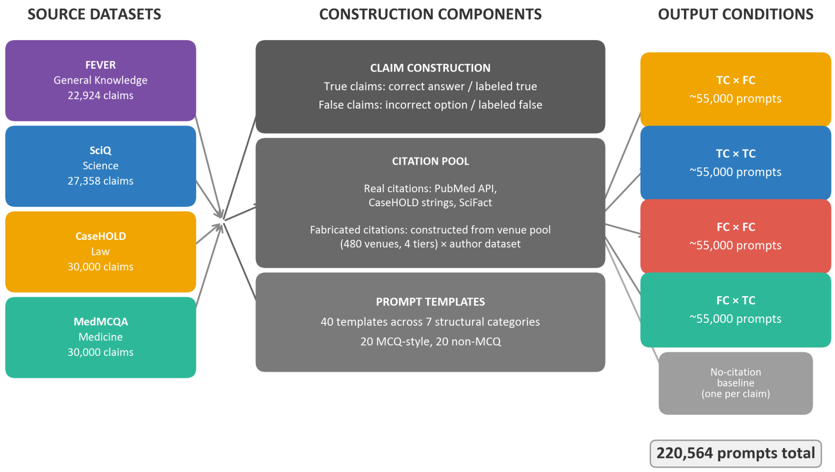

Figure 2. Dataset Construction Pipeline.

puts vary systematically with perceived demographic attributes of names in prompts — motivating our inclusion of a country-coded author name variable to extend demographic bias analysis to epistemic authority contexts.

FalseCite. The most directly relevant prior work is FalseCite (Mao et al., 2025), which introduced 82,000 prompts pairing false claims from FEVER and SciQ with fabricated citations, and evaluated three models across no-citation, random citation, and semantic citation conditions. Their findings showed that false citations consistently amplify hallucination, with the effect largest for random citations in smaller models and semantic citations in GPT-4o-mini. Our work is independently motivated and substantially broader in scope: we extend to a full 2×2 factorial design by introducing true claims and true citations as explicit conditions, expand domain coverage from two to four, replace a single prompt template with 40 structurally diverse templates, and add controlled variation in venue prestige, author demographics, and citation year. Table 1 summarizes source dataset statistics; the design differences from FalseCite are detailed above.

3. Benchmark Design

3.1. Source Datasets

We draw claims from four publicly available datasets. FEVER (Thorne et al., 2018) provides short declarative claims labeled true or false from Wikipedia, covering general knowledge; we use the labels directly. SciQ (Welbl et al., 2017) is a multiple-choice science exam dataset; true claims pair each question with its correct answer, false claims sample one distractor, following "The answer to [question] is [answer]." CaseHOLD (Zheng et al., 2021) and MedMCQA (Pal et al., 2022) are legal and medical MCQ datasets respectively, constructed identically to SciQ. CaseHOLD claims are treated as non-MCQ for template assignment as the fill-in-the-blank format yields self-contained declarative statements. Table 1 summarizes claim counts; the full pipeline is shown in Figure 2.

Table 1. Source dataset statistics.

Source dataset	Domain	Claims	Claim type
CaseHOLD	Law	30,000	MCQ
MedMCQA	Medicine	30,000	MCQ
SciQ	Science	27,358	MCQ
FEVER	General knowledge	22,924	Declarative
Total		**110,282**	

3.2. The 2×2 Factorial Design

The benchmark adopts a fully balanced 2×2 factorial structure independently manipulating claim veracity (true or false) and citation veracity (real or fabricated), yielding four conditions of approximately 55,000 prompts each. The true claim × fabricated citation condition is the novel contribution — it tests whether models can be induced to deny correct facts under citation-based authority pressure, absent from all prior work. All balancing is maintained identically across conditions to prevent confounds. Each claim additionally appears without a citation as a no-citation baseline. The 110,282 base claims yield 220,564 prompts in the main dataset.

3.3. Citation Construction

Each prompt includes a citation slot containing author name, venue, and year. Fabricated citations are entirely constructed from a curated author name pool and a venue pool assigned to one of four prestige tiers. Citation-claim pairing maintains a 50/50 same-domain/cross-domain split across all conditions; years span four ranges from 1980 to the present.

Real citations correspond to verifiable publications. Science citations are drawn from MTEB SciFact (Wadden et al., 2020); medical citations from PubMedQA (Jin et al., 2019), with metadata retrieved via the PubMed efetch API; legal citations are extracted from CaseHOLD's context paragraphs. For FEVER, structured citation metadata is unavailable; author, venue, and year fields are back-filled from the other three domains' citation pools and flagged as `citation_matches_claim = False`, discussed in §5.

3.4. Venues, Prestige Tiers, and Author Names

Each citation includes a venue from a curated pool of 480 venues, with 120 per domain divided evenly across four prestige tiers. Venues are sourced from SCImago Journal Rankings (SCImago Research Group, 2024) for science and medicine, Washington and Lee Law Journal Rankings (Washington and Lee University School of Law, 2024) for law, and a manually curated list for general knowledge. Tier 4 represents the highest prestige and tier 1 the lowest; tiers serve as controlled experimental variables rather than definitive quality measures. For fabricated citations, author surnames are sampled from a country-coded dataset (Boothe, 2023). Since most citations use surnames followed by "et al.", surnames act as the primary demographic signal, enabling analysis of whether perceived author identity influences model behavior. Real citations retain original author names.

3.5. Prompt Templates

We construct 40 prompt templates across seven structural categories: prefix, mid-sentence, suffix, minimum salience, venue-first, author-first declarative, and footnote-style; 20 target MCQ claims and 20 non-MCQ.

3.6. Research Questions

The benchmark addresses a single overarching question — how do citation-based authority signals shape epistemic behavior in LLMs? — decomposed into 12 structured research questions spanning citation authority effects, domain sensitivity, structural presentation, author demographics, temporal framing, and cross-model profiles. These are given in Table 2.

Table 2. Research questions overview.

RQ	Category	Question
RQ0	Overarching	How do citation-based authority signals shape epistemic behavior in large language models?
RQ1	3*Authority & Knowledge	Do fabricated citations increase hallucination on false claims?
RQ2		Does citation authority override a model's internal knowledge under certain conditions?
RQ3		Can false citations cause models to deny factually correct claims?
RQ4	2*Institutional Authority	Does venue prestige affect hallucination rates?
RQ5		Does the prestige effect vary across domains?
RQ6	2*Domain Sensitivity	Which citation domains exert the most influence on each claim domain?
RQ7		Are same-domain or cross-domain citations more persuasive?
RQ8	Structural & Prompt Effects	Does citation structure/format affect hallucination?
RQ9	Identity & Demographics	Does author identity influence model responses?
RQ10	Temporal Effects	Does citation recency affect hallucination rates?
RQ11	Cross-model comparison	How do models differ in overall citation sensitivity?
RQ12	Global Interaction	What are the joint effects of claim, citation, domain, and structure on model behavior?

Table 3. Models evaluated in this study.

Model	Parameters	Type	Eval. scale
Gemma 3 4B	4B	Open-weight	Full Dataset
Llama 3.1 8B	8B	Open-weight	Full Dataset
Phi-4 Mini Instruct	$\sim$4B	Open-weight	Full Dataset
Gemma 4 31B	31B	Open-weight	15K subset
Claude Haiku 4.5	N/A	Proprietary	15K subset
GPT 5.4 mini	N/A	Proprietary	15K subset
DeepSeek V3.2	37B	Open-weight	15K subset

Note: Parameter counts for proprietary models not publicly disclosed. N/A indicates undisclosed. All runs standardized to 15K evaluations for parity.

4. Experiments

4.1. Models

We evaluate seven language models spanning a range of sizes and training regimes. Three open-weight models are evaluated on the full dataset: Gemma 3 4B, Llama 3.1 8B Instruct, and Phi-4 Mini Instruct. Four models are evaluated on a balanced 15K subset due to cost or access constraints: Gemma 4 31B, Claude Haiku 4.5, GPT 5.4 mini, and DeepSeek V3.2. The subset is constructed by proportional stratified sampling, preserving the 2×2 claim-citation balance, domain, template, prestige tier, and author demographic distributions of the full dataset. Llama 3.1 8B and Gemma 3 4B 15K subsets closely replicate their full-dataset results, confirming that subset estimates are reliable for cross-model comparisons; results for subset-evaluated models should be interpreted with this difference in scale in mind. Table 3 summarizes all model details. All local models are run on a single university-owned GPU with 50GB VRAM.

4.2. Judge Model

Model outputs are evaluated using Qwen3-8B (Qwen Team, 2025), selected for its strong performance on the HHEM leaderboard (Vectara, 2024). The judge prompt supplies ground truth labels and source metadata alongside each model output, reducing reliance on parametric knowledge and improving reliability in specialized domains. The judge outputs a binary label only (`is_hallucination: true/false`); the judge cannot independently verify whether a cited source exists, though this limitation is partially mitigated by the ground truth metadata supplied in the prompt. We validate on a stratified 1,500-output human evaluation sample; each output is annotated by two NLP researchers provided with the original prompt, model output, and ground truth label. We obtain Cohen's kappa of 0.83 with inter-annotator agreement of 90.7%.

4.3. Evaluation Metrics

Our primary metric is hallucination rate per condition — the proportion of outputs labeled hallucinated over all non-refused responses. Refusal rates were below 2% across all models and are excluded from the denominator. Condition differences are characterized using lift (absolute pp difference from the relevant baseline — either the overall no-citation baseline or the claim-type-specific baseline depending on the comparison) and Cohen's d, with conventional thresholds of $d = 0.2$, 0.5, and 0.8. We foreground d over p-values given the large sample sizes. All comparisons include 95% confidence intervals via normal approximation to the binomial. The general knowledge true citation condition carries a structural caveat from §3.3, noted throughout §5.

5. Results

We report effects in percentage points (pp) and Cohen's d (0.2 / 0.5 / 0.8 = small / medium / large). We use superscripts ^{F} and ^{S} to distinguish numbers derived from the full dataset ($\approx$220K prompts total) and the 15K subset respectively; all cross-model comparisons use ^{S} values.

Headline finding. Across all seven models, adding any citation—fabricated or real—increases hallucination above the no-citation baseline when averaging across claim types. The effect is most extreme in one specific condition, true claim × fabricated citation (TC×FC), which is the highest-hallucination condition in *every* model tested (RQ2, RQ3).

Baselines. Table 4 reports per-model overall, true-claim, and false-claim baselines. Overall rates span $11.88\%^S$ (Gemma 4 31B) to $31.32\%^F$ (Gemma 3 4B), broadly tracking model capability. Four models show the expected pattern of higher hallucination on false claims than true ones, with Gemma 3 4B showing the widest gap ($+22.24^F$ pp). Three models invert this entirely: Claude Haiku 4.5 hallucinates *more* on true claims than false ones (-15.13^S pp), Phi-4 Mini does the same (-14.38^S pp), and Gemma 4 31B also exhibits this inversion (-1.32^S pp), all flagged with † in Table 4. These inversions reflect systematic biases that

Table 4. Baseline hallucination rates (no citation) on 15K subset.

Model	Overall baseline	True-claim baseline	False-claim baseline	Gap (F−T)
Gemma 3 4B	30.99%	20.00%	42.24%	+22.24pp
Llama 3.1 8B	30.25%	20.83%	39.89%	+19.06pp
Phi-4 Mini Instruct	28.79%	35.89%	21.51%	**−14.38pp** †
Claude Haiku 4.5	24.66%	32.13%	17.00%	**−15.13pp** †
GPT 5.4 mini	19.30%	17.63%	21.01%	+3.38pp
DeepSeek V3.2	16.18%	13.04%	19.40%	+6.36pp
Gemma 4 31B	11.88%	12.53%	11.21%	**−1.32pp** †

† True-claim baseline exceeds false-claim baseline for these models—i.e., the model is more confident on true claims at baseline and citation effects are reversed.

Table 5. Effect sizes for the TC×FC condition (true claim, fabricated citation).

Model	TC×FC rate	Lift over TC baseline
Gemma 3 4B	24.14%	+3.66pp
Llama 3.1 8B	43.76%	+22.29pp
Phi-4 Mini Instruct	40.63%	+3.81pp
Claude Haiku 4.5	47.11%	+14.98pp
GPT 5.4 mini	37.27%	+19.64pp
DeepSeek V3.2	16.28%	+3.23pp
Gemma 4 31B	22.73%	+10.20pp

propagate through all citation conditions. Notably, the three inverted models split across the suppression/amplification divide in Figure 5: Claude and GPT fall in the suppression region while Gemma 4 31B and Phi-4 fall in the amplification region, suggesting that a model's baseline tendency to treat true claims with more uncertainty than false ones mediates how it responds to citation authority pressure on the TC×FC condition.

TC×FC as the critical condition (RQ2, RQ3). Figure 4 displays hallucination rates across all five conditions for all seven models. The TC×FC column is universally the worst. Table 5 quantifies the lift on the 15K subset: fabricated citations raise hallucination on true claims by 22.29^S pp for Llama 3.1 8B Instruct, 19.64^S pp for GPT 5.4 mini, 14.98^S pp for Claude Haiku 4.5, and 10.20^S pp for Gemma 4 31B, down to smaller but still significant effects in Phi-4 Mini Instruct ($+3.81^S$ pp), Gemma 3 4B ($+3.66^S$ pp), and DeepSeek V3.2 ($+3.23^S$ pp). The effect is statistically significant in every model. Notably, model size and capability do not predict robustness: the largest closed-source and instruction-tuned models are among the *most* susceptible. True citations also elevate TC hallucination in six of seven models, with ^{S} lifts ranging from $+2.77$ pp (DeepSeek V3.2) to $+15.57$ pp (GPT 5.4 mini), indicating the problem is not fabrication alone but citation presence itself.

False-claim effects: suppression vs. amplification (RQ1). Figure 5 shows the diverging pattern on false claims. Three

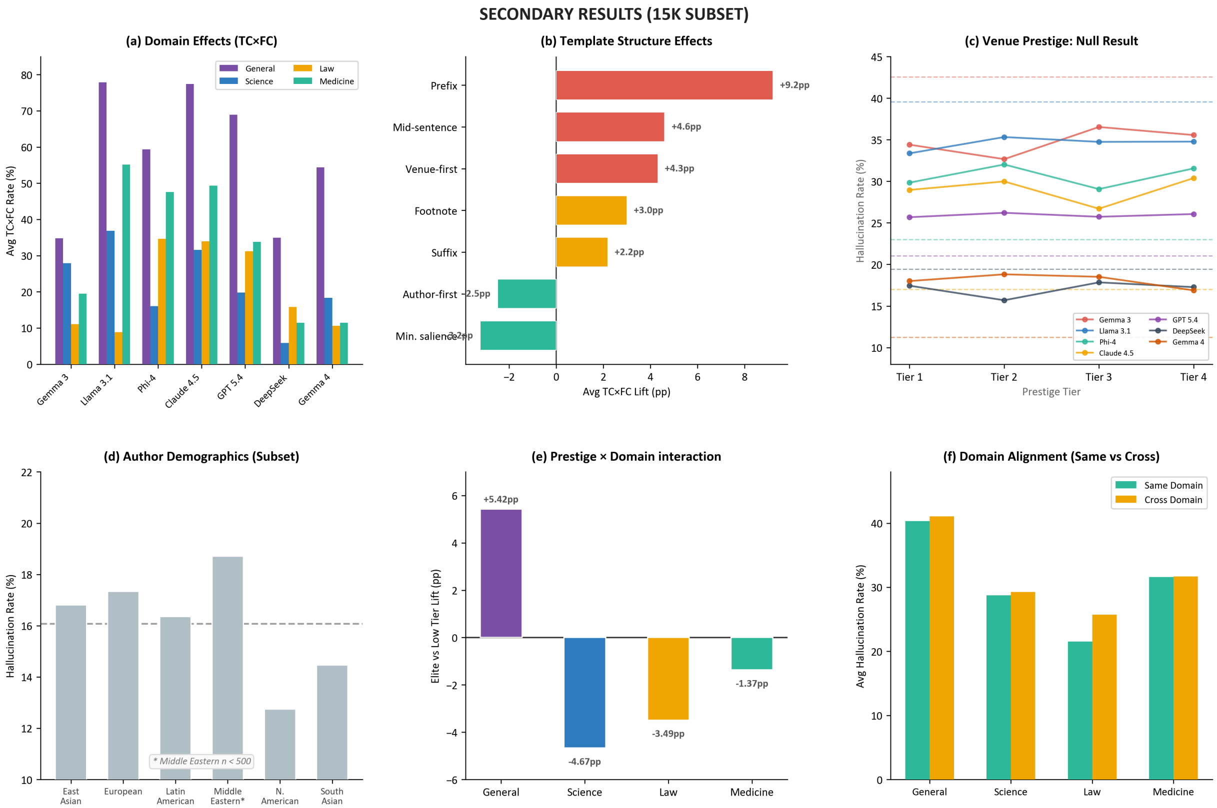

Figure 3. Secondary results (15K subset). (a) Domain effects under TC×FC across models. (b) Template structure effects, showing average TC×FC lift by citation format. (c) Venue prestige null result across four tiers. (d) Author demographics: hallucination rate by surname region. (e) Prestige × domain interaction (elite vs. low-tier lift). (f) Domain alignment effect: same-domain vs. cross-domain citation hallucination rates.

models (Llama, Claude, GPT 5.4 mini) show citation-induced *suppression*: fabricated citations reduce hallucination by 6.52^S pp (Claude) to 13.85^F pp (Llama), as though the citation triggers heightened scrutiny of the claim. Three models (Gemma 3 4B, Gemma 4 31B, Phi-4) show the opposite—*amplification* of up to 4.57^F pp. The Gemma family is consistent across its 4B and 31B variants, identifying this as a family-level property rather than a size effect. DeepSeek V3.2 shows a non-significant near-zero effect. This split suggests the same citation signal functions as a credibility check in some model families and an authority endorsement in others.

Domain vulnerability. The general knowledge domain is the universal weak point (Figure 3a). TC×FC rates there reach $77.92\%^F$ (Llama; $d = 1.35^F$), $77.39\%^S$ (Claude; $d = 1.35^S$), and $68.90\%^S$ (GPT 5.4 mini), with $d = 1.10^S$ for Gemma 4 31B. The legal domain consistently shows the smallest citation-induced lifts across all models, frequently non-significant, likely owing to its distinctive linguistic register and low baselines that constrain upward movement. Medical effects track available headroom per model; science shows apparent suppression in high-baseline models as a floor-effect artifact.

Cross-domain citation influence (RQ6, RQ7). Across models, the domain of the citation modulates hallucination rates, though sensitivity varies substantially by model family (RQ6). Gemma 3 4B is the most sensitive: cross-domain citations produce 2.6^F pp higher hallucination overall than same-domain citations, with significant effects in science ($+4.5^F$ pp), general ($+2.7^F$ pp), and medical ($+0.8^F$ pp). Claude Haiku 4.5 is entirely insensitive: no domain-level comparison exceeds 0.9^S pp. Llama 3.1 8B shows a significant cross $>$ same asymmetry specifically in the legal domain ($+5.39^F$ pp) but not elsewhere. One finding is consistent across all seven models: same-domain legal citations are the least disruptive citation source for legal claims. For DeepSeek V3.2, same-domain legal citations produce just $6.99\%^S$ hallucination while cross-domain citations for the same claims reach $35.69\%^S$; for Gemma 4 31B the gap is 11.33^S pp. This cross-model legal resistance likely reflects that legal citation formatting—case names, reporter abbreviations, formal conventions—triggers a recognisably spe-

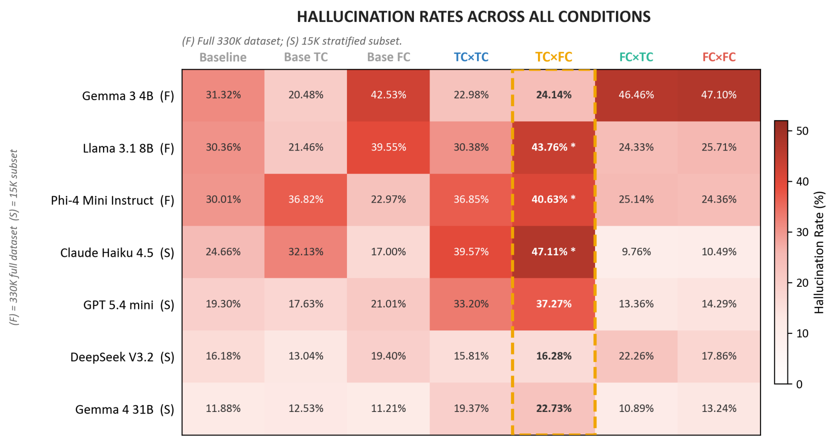

Figure 4. Hallucination rates across all five citation conditions for all seven models. The TC×FC column (true claim, fabricated citation) is universally the worst-performing condition across every model tested. Baseline, Base TC, and Base FC columns show full-dataset values (F) for Gemma 3 4B, Llama 3.1 8B, and Phi-4 Mini Instruct; 15K-subset values (S) for all others.

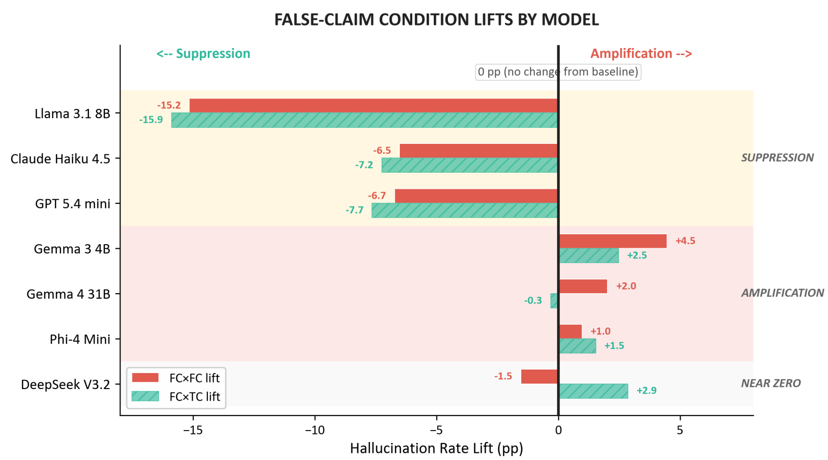

Figure 5. False-claim condition lifts by model. Bars show hallucination rate lift (pp) relative to the no-citation baseline for the FC×FC (fabricated citation) and FC×TC (real citation) conditions. Models in the suppression region (left) show citation-induced reduction in hallucination on false claims; models in the amplification region (right) show the opposite. DeepSeek V3.2 shows a near-zero effect.

cialised register that raises the model's evidentiary bar. Figure 3f quantifies the alignment effect directly (RQ7). Cross-domain citations are more disruptive than same-domain citations in the majority of model–domain pairings, but sensitivity varies substantially. Gemma 3 4B shows the most consistent effect across all four domains (5.07–7.88^F pp, all $p < 0.001$). Gemma 4 31B shows the largest science alignment gap (-7.85^S pp, $d = -0.23$, $p < 0.001$), where in-domain science citations are unusually low, consistent with the floor effect noted above. DeepSeek V3.2 shows a large general alignment gap (-5.24^S pp, $p < 0.01$). GPT 5.4 mini reaches significance only in science (-2.58^S pp), and Phi-4 Mini shows no significant effects on the 15K subset. Claude shows no alignment sensitivity in any domain (all $\leq 0.54^S$ pp, all $p > 0.7$) — the same insensitivity it shows on citation domain overall. The legal domain is the most consistent: cross-domain citations are more disruptive for legal claims in all seven models, reaching significance in five.

Template structure (RQ8). Citation format produces hallucination rate spreads of 4–17 pp across models (Figure 3b). Prefix placement ("According to [source]…") ranks first or second in five of seven models and is consistently the highest-risk format. Minimal salience (author name and year only) is the lowest-risk format in six of seven models, sometimes suppressing hallucination below baseline. No single format is universally safest—footnote style ranges from protective in Gemma 3 4B (-2.63^F pp) to risky in Phi-4 Mini ($+5.47^S$ pp)—so safe-format guidance is necessarily model-specific.

Venue prestige (RQ4, RQ5). Prestige does not matter in aggregate. Across all seven models, hallucination rates are flat across fabricated-citation prestige tiers, with total spreads of 0.53^S–3.68^S pp and no monotonic gradient in any model (Figure 3c). A fictitious low-tier journal citation carries the same weight as a fictitious *Nature* citation. Figure 3e stratifies this null by domain-specific ranking framework (RQ5). The null replicates uniformly: no domain produces a monotonic prestige gradient in any model. Full-dataset models show within-domain tier spreads of $\leq 2.42^F$ pp, with science the flattest ($\leq 1.30^F$ pp). Some 15K-subset cells show wider apparent spreads (Claude general: 7.84^S pp; DeepSeek legal: 6.55^S pp) but these are non-monotonic — Low-tier citations outscore Elite in several cells, and Elite ranks last in others. Institutional prestige carries no systematic weight within any domain's specific ranking framework.

Author demographics and temporal framing (RQ9, RQ10). Author surname region produces no meaningful variation in full-dataset models (spreads of 0.7^F–1.9^F pp across seven regions; Figure 3d). Wider apparent spreads in 15K-subset models are sample-size artifacts. A modest recency effect exists in temporal framing: older citations tend to be slightly more disruptive than recent ones, but the effect is small ($\leq 2^F$ pp), attenuates at smaller sample sizes, and is not consistently monotonic across models.

Cross-model synthesis (RQ11, RQ12). All seven models share three core properties: any citation increases hallucination in aggregate; TC$\times$FC is the worst-performing condition; and the general knowledge domain is the most vulnerable. They diverge on false-claim direction, template sensitivity ordering, and domain-of-citation effects. The most capable models are not the most robust—a finding that points toward targeted, model-specific mitigation strategies rather than reliance on scale.

6. Conclusion

Across all seven models tested, adding a citation—fabricated or real—increases hallucination above the no-citation baseline. The effect is most consequential in a condition prior work has not examined: fabricated citations paired with true claims. In this condition every model is more likely to deny a correct fact than in any other experimental condition, with lifts of $+3.23$ to $+22.29$ pp over true-claim baselines and near-ceiling hallucination rates (35–77%) in the general knowledge domain. This is not a failure of factual knowledge but of epistemic reasoning under authority pressure.

Secondary findings are consistent: the legal domain is uniformly resistant, and venue prestige is uniformly irrelevant. Two findings resist generalisation: citations suppress false-claim hallucination in Llama, Claude, and GPT but amplify it in the Gemma variants and Phi-4, suggesting a family-level difference in how authority signals are processed; and susceptibility on true claims does not track model size or capability. For RAG systems, the implication is direct: citation presence degrades factual accuracy on claims the model would otherwise handle correctly. Mitigation will require models that treat a citation as evidence rather than authority.

Limitations and Future Work

Limitations

Judge model reliability. All model outputs are evaluated using Qwen3-8B. As with all LLM-based judges without retrieval access, it cannot verify whether a cited source exists or supports the attributed claim. For true citations, the judge relies on ground truth labels and metadata supplied in the prompt rather than independent verification. A retrieval-augmented judge with access to a live citation database is identified as a priority for future work.

True citation metadata for general knowledge. True citations for FEVER-sourced claims use author, venue, and year metadata back-filled from other domain citation pools, since Wikipedia articles lack structured academic citation records. These entries are flagged as `citation_matches_claim = False`, and results for this condition should be interpreted accordingly.

Compute and resource constraints. All experiments were conducted on a single university-owned GPU with 50GB VRAM, constraining full-dataset evaluation to open-source models in the 3B–8B range. API-based models were evaluated on the balanced 15K subset and are not directly comparable in scale to locally evaluated models.

Prestige tier operationalization. Venue prestige tiers should be interpreted as controlled experimental variables rather than definitive measures of authority or quality.

Author demographic proxies. Country-coded surnames are a coarse proxy for perceived demographic identity and do not capture finer-grained signals such as name familiarity or intersectional combinations. Effects absent at regional scale may be present at finer granularity.

Template coverage. The 40 templates do not cover web-native or informal citation formats such as hyperlinks, social media references, or conversational citations.

Future Work

Three directions follow directly from this work. First, mechanistic analysis via hidden state or attention inspection would clarify why citation signals disrupt true-claim processing and why the suppression/amplification split across model families emerges. Second, the benchmark provides a testbed for evaluating prompt-based, fine-tuning-based, or architectural mitigations against citation deference. Third, expanded demographic analysis at finer granularity would give a more complete picture of identity effects on epistemic authority judgments. Evaluation of larger frontier models is a further natural extension.

Impact Statement

This paper introduces AuthorityBench, a benchmark for studying how citation-based authority signals influence epistemic behavior in large language models. We identify and quantify a failure mode in which citation presence degrades factual accuracy, including on claims models would otherwise answer correctly. The implications are most direct for retrieval-augmented generation systems in high-stakes domains: users and developers should not assume that grounding outputs in cited sources improves reliability. We hope this benchmark accelerates development of models that treat citations as evidence rather than authority. All datasets and evaluation code are available at: https://github.com/floating-reeds/AuthorityBench. The benchmark does not involve human subjects, personally identifiable information, or content that poses direct harm risk.

References

Boothe, A. Popular names by country dataset. `https://github.com/sigpwned/popular-names-by-country-dataset`, 2023. Accessed: 2024.

Dassen, M., Kotula, R., Murray, K., Yates, A., Lawrie, D., Kayi, E., Mayfield, J., and Duh, K. FACTUM: Mechanistic detection of citation hallucination in long-form RAG, 2026. URL `https://arxiv.org/abs/2601.05866`.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL `https://aclanthology.org/2023.emnlp-main.398/`.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL `https://doi.org/10.1145/3703155`.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W., and Lu, X. PubMedQA: A dataset for biomedical research question answering. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL `https://aclanthology.org/D19-1259/`.

Kotek, H., Dockum, R., and Sun, D. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, pp. 12–24, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701139. doi: 10.1145/3582269.3615599. URL `https://doi.org/10.1145/3582269.3615599`.

Li, J., Cheng, X., Zhao, X., Nie, J.-Y., and Wen, J.-R. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.397. URL `https://aclanthology.org/2023.emnlp-main.397/`.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL `https://aclanthology.org/2022.acl-long.229/`.

Mao, N., Kaushik, V., Shivkumar, S., Sharafoleslami, P., Zhu, K., and Dev, S. Visualizing and benchmarking LLM factual hallucination tendencies via internal state analysis and clustering. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 289–298, Mumbai, India, 2025. Association for Computational Linguistics. URL `https://aclanthology.org/2025.ijcnlp-srw.24/`.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Pal, A., Umapathi, L. K., and Sankarasubbu, M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Pataranutaporn, P., Powdthavee, N., and Maes, P. Algorithmic inheritance: Surname bias in AI decisions reinforces intergenerational inequality, 2025. URL https://arxiv.org/abs/2501.19407.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Schuster, J., Gautam, V., and Markert, K. Whose facts win? LLM source preferences under knowledge conflicts, 2026. URL https://arxiv.org/abs/2601.03746.

SCImago Research Group. SCImago journal rankings. https://www.scimagojr.com, 2024. Accessed: 2024.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. FEVER: a large-scale dataset for fact extraction and VERification. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Vectara. Hughes hallucination evaluation model (HHEM) leaderboard. https://github.com/vectara/hallucination-leaderboard, 2024. Accessed: 2026-04-21.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or fiction: Verifying scientific claims. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Washington and Lee University School of Law. Washington & Lee law journal rankings. https://managementtools4.wlu.edu/LawJournals/, 2024. Accessed: 2024.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. In Derczynski, L., Xu, W., Ritter, A., and Baldwin, T. (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Wilson, K. and Caliskan, A. *Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval*, pp. 1578–1590. AAAI Press, 2025.

Xie, J., Zhang, K., Chen, J., Lou, R., and Su, Y. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Xu, R., Qi, Z., Guo, Z., Wang, C., Wang, H., Zhang, Y., and Xu, W. Knowledge conflicts for LLMs: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models, 2024b. URL https://arxiv.org/abs/2401.11817.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pp. 159–168, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3462757.3466088. URL https://doi.org/10.1145/3462757.3466088.