Source-Grounded Prompting as a Mitigation for Quotation Fabrication and Unsupported Factual Claims in AI History Tutors

A Pre-Registered, Mixed-Effects Evaluation Across Well- and Under-Documented U.S. History Topics

Research Question:

To what extent does source-grounded prompting reduce fabricated quotations and unsupported factual claims in AI tutoring responses to U.S. history questions, and does its effectiveness vary between well-documented and under-documented topics?

Submitted by:
RAJ GADGE
raj.r.gadge@gmail.com

Category:
Social Sciences and Education

1. Research Problem and Significance

AI tutoring tools are now common in U.S. classrooms and student study routines, but early controlled evidence is concentrated in STEM and remains limited for K–12 humanities subjects such as history (Kestin et al., 2025). The failure mode of greatest concern for history is quotation fabrication. Plausible-sounding but invented quotations have already produced documented harms: a California Court of Appeal sanctioned an attorney in September 2025 after finding 21 of 23 case quotations in his brief were AI-generated (CalMatters, 2025), and a Massachusetts federal case over AI use in an AP U.S. History assignment raised questions about how schools should penalize AI-generated coursework and nonexistent citations (Toppo, 2024). For a student, citing a quotation Frederick Douglass never produced amounts to learning a fictional past, with consequences for historical understanding well beyond a single missed grade.

This risk is unevenly distributed. Training corpora over-represent dominant political narratives and under-represent marginalized communities (Bender et al., 2021), a pattern with deep historiographic roots: Trouillot (1995) theorized *archival silence* as the systematic absence of marginalized voices from the historical record, and Hartman (2008) demonstrated how that silence forces historians of slavery toward speculation in the absence of adequate sources. Where archives are sparse, models have the least basis for grounding and the greatest incentive to fall back on familiar but unsupported narratives. Source-grounded prompting (instructing the model to answer only from a provided document set) is the leading recommendation issued to students; whether its benefit holds across topic types remains untested in published research.

2. Literature Review and Identified Gap

The hallucination literature has converged on shared methods. Maynez et al. (2020) distinguish *intrinsic* errors (which contradict the source) from *extrinsic* errors (which introduce unverifiable content); Ji et al. (2023) survey approximately 200 mitigation studies. Atomic-claim methods such as FActScore (Min et al., 2023) decompose long outputs into individually verifiable units, and the Attributable to Identified Sources framework (AIS; Rashkin et al., 2023) operationalizes "supported by sources" as a binary judgment. Retrieval-augmented generation (RAG; Lewis et al., 2020) and source-grounded prompting are the dominant mitigations.

Education-specific evidence remains limited and mixed. Kestin et al. (2025), in a Harvard physics RCT, found that prompts enriched with worked solutions improved AI-tutor performance over default settings. Liu et al. (2024) demonstrated a key limitation of grounding: model accuracy on retrieved evidence follows a U-shape, with material in the middle of long contexts systematically ignored, a pattern known as the *lost-in-the-middle* effect. Grounding also cannot compensate for material absent from the corpus.

Three gaps remain. First, no published study has held model and prompt fixed while varying topic documentation density. Second, there is little peer-reviewed factuality comparison of the new tutoring and learning modes in non-STEM domains. Third, no work connects the historiographic literature on archival silence (Trouillot, 1995; Hartman, 2008) to empirical research on LLM coverage bias. This proposal addresses all three.

3. Research Question and Hypotheses

Central question. To what extent does source-grounded prompting reduce fabricated quotations and unsupported factual claims in AI tutoring responses to U.S. history questions, and does its effectiveness vary between well-documented and under-documented topics?

H1 (main effect on quotations). Source-grounded prompting reduces the rate of fabricated quotations relative to baseline prompting (predicted odds ratio < 1).

H2 (main effect on factual claims). Source-grounded prompting reduces the rate of unsupported factual claims relative to baseline (predicted OR < 1).

H3 (moderation by documentation density). The reduction predicted in H1 and H2 is smaller for under-documented topics than for well-documented ones (predicted positive Condition ×

Documentation interaction, OR > 1) because, even with matched packet size, under-documented topics are more likely to rely on indirect, fragmentary, or institutionally mediated evidence.

4. Methodology

Item bank. A 2 (prompt: baseline vs. source-grounded) × 2 (topic: well- vs. under-documented) within-items design uses 80 short-answer items adapted from publicly released AP U.S. History exam questions and scoring information (College Board, 2025), NAEP, and state standards. Documentation density is classified *externally* by two history teachers using a rubric covering volume of digitized primary sources, archival accessibility (drawing on Trouillot, 1995, and Hartman, 2008), and peer-reviewed scholarly coverage. A third rater resolves disagreements (target Cohen's $\kappa \geq .75$; Cohen, 1960). Items are matched across documentation categories by historical period, question type, cognitive demand, and expected answer length, to reduce confounding between documentation density and topic difficulty. Under-documented items draw from histories of enslaved persons, Indigenous diplomacy, and immigrant labor.

Source-packet control. To prevent confounding documentation density with packet size, each item receives a packet of five documents within a fixed token budget, identical across conditions. Where under-documented topics cannot supply five high-quality primary sources, peer-reviewed scholarly excerpts complete the packet. Packet token count is logged for every item and entered as a covariate in analysis, so any moderation effect is interpretable as documentation density rather than packet volume.

Models and prompts. Three frontier LLMs (the leading model from OpenAI, Google, and Anthropic) are accessed through stable API endpoints, with exact model versions, dates, decoding parameters, and prompts logged. Commercial tutoring modes (ChatGPT Study Mode, Gemini Guided Learning, Claude Learning Mode) motivate the study, but the main experiment tests reproducible model behavior under fixed prompts; comparison of those product modes is treated as exploratory. Both prompts equally invite quotation, to prevent the grounded condition from inflating quotation count as a measurement artifact. Baseline: "Answer this U.S. history question in 150–200 words. Include one direct quotation if relevant." Source-grounded: "Using only the attached sources, answer this U.S. history question in 150–200 words. Include one direct quotation if the provided sources support one. If the sources do not support a quotation or answer, say so." Each item runs three times per model per condition, yielding 1,440 responses; order is randomized within model.

Outcome and analysis. Following FActScore atomic decomposition (Min et al., 2023) and AIS attribution rubrics (Rashkin et al., 2023), two trained coders, blinded to model identity, item classification, and study hypotheses (with condition labels removed where possible), score (a) *fabricated quotation rate* (unverifiable quotations / total); (b) *unsupported claim rate*; and (c) *appropriate abstention*. An 80-response pilot calibrates the rubric. Mixed-effects logistic regression with fixed effects for condition, documentation, and their interaction, packet size as covariate, and random intercepts for item and model, tests H1–H3. Power estimates assuming a baseline fabrication rate of approximately 25% versus 10% under grounding (ICC ≈ .20, target ≥ .80 for main effects) are treated as provisional and will be updated after the pilot using simulated mixed-effects models. Materials and analysis script are pre-registered on the Open Science Framework before data collection.

5. Significance, Conclusion, and Feasibility

This project would provide one of the first empirical comparisons of factuality across the new tutoring-mode generation, and would be among the first to operationalize Trouillot's (1995) and Hartman's (2008) archival-silence theory as a quantitative LLM moderator. If H1 and H2 hold and H3 does not, source-grounded prompting can be confidently recommended as a uniform classroom mitigation; if H3 also holds, the corrective lies in better source coverage for marginalized histories rather than in further prompt engineering. Scope is bounded for a single high-school researcher: six months, approximately $300 in API credits and $800 in coder compensation, with a stratified-subsample fallback pre-registered should atomic coding prove slower than projected.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.

CalMatters. (2025, September 22). California issues historic fine over lawyer's ChatGPT fabrications. *CalMatters*. https://calmatters.org/economy/technology/2025/09/chatgpt-lawyer-fine-ai-regulation/

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

College Board. (2025). *AP United States History exam questions and scoring information*. AP Central. https://apcentral.collegeboard.org/courses/ap-united-states-history/exam

Hartman, S. (2008). Venus in two acts. *Small Axe, 12*(2), 1–14.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.

Kestin, G., Miller, K., Klales, A., et al. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, 17458.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of ACL 2020* (pp. 1906–1919).

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. In *Proceedings of EMNLP 2023* (pp. 12076–12100).

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2023). Measuring attribution in natural language generation models. *Computational Linguistics, 49*(4), 777–840.

Toppo, G. (2024, October 31). Could Massachusetts AI cheating case push schools to refocus on learning? *The 74*. https://www.the74million.org/article/could-massachusetts-ai-cheating-case/

Trouillot, M.-R. (1995). *Silencing the past: Power and the production of history*. Beacon Press.