Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Plausible citations, unsupported claims, and nationality-dependent recommendations: hidden failure modes of large language models in gynecologic oncology

Junhwan Kim

junhwankimmd@gmail.com

Soonchunhyang University Seoul Hospital

Joseph J. Noh

Seoul National University Bundang Hospital

Se Ik Kim

Seoul National University College of Medicine

Byoung-Gie Kim

Chung-Ang University Gwangmyeong Hospital

Jae Hyun Cho

Soonchunhyang University Seoul Hospital

Jeong-Sig Kim

Soonchunhyang University Seoul Hospital

Article

Keywords: large language models, health equity, nationality, hallucination, citation verification, benchmarking

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11146954/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Large language models increasingly answer gynecologic-oncology questions with citations and management recommendations. In a preregistered 308-item benchmark (252 confirmatory) administered to GPT-5.5, Claude Opus 4.8, and Gemini 3.1 Pro, we tested whether cited papers support the claims they accompany, whether unsupported citations track the volume of indexed literature, and whether otherwise-invariant recommendations change with patient nationality. Citations were verified for existence, record identity, and claim support by an automated judge validated against two blinded gynecologic oncologists and a second judge model. All three models answered ancestry-specific questions correctly (24 of 24). Strict fabrication (non-existent or misattributed citations) was unrelated to literature coverage ($\beta = -0.19$, $p = 0.22$), whereas citation of real, on-topic papers lacking the claimed fact rose steeply as coverage fell ($\beta = -1.12$, $p = 2 \times 10^{-8}$; failure rates 0.67, 0.60, and 0.26 across tertiles), with a similar gradient in the 48 exploration-only items. Nationality-dependent violations of invariant management reached 0.11–0.67 across models when the care setting was unstated versus 0.00–0.13 when a fully resourced setting was stated. Legitimate population differences were preserved. Evaluations should score claim-level citation support and specify the care setting; clinicians should verify that cited sources support the specific claim.

Introduction

Clinicians and patients increasingly consult large language models about gynecologic cancers, and current models answer with citations and management recommendations. Within this specialty, large language models have been evaluated for concordance with clinical guidelines, decision-support accuracy, tumor-board concordance, and automated staging [1–4], and a recent systematic review found that benchmarking methods were heterogeneous, that models perform best on guideline-driven tasks, and that performance declined in nuanced scenarios and among under-represented patient subgroups [5]. Two questions that determine whether a clinician can act on such output remain unexamined: whether the evidence a model cites actually supports its claim, and whether its recommendation depends on who the patient is.

Both questions matter most where the literature is thin. Cervical cancer mortality is more than twice as high in transitioning as in transitioned countries [6], yet those regions contribute little indexed literature, and a model's competence is inherited from a training corpus that covers topics unevenly, so factual reliability is expected to fall as a topic becomes rarer [7, 8]. Citation fabrication rises for less familiar topics [9], and a direct relationship between literature coverage and hallucination has been shown for place-based legal questions [10], but not in medicine with respect to geography or ancestry. Most clinical studies score a citation as hallucinated only when it does not exist or points to an unrelated paper [11–13]; a reference that exists, is on topic, and yet does not contain the claimed fact passes that check, and if such errors cluster where the literature is sparse, existence-based checks miss precisely the risk that reaches patients.

Demographic bias in medical language models has been studied largely for United States race and gender, and for recommendations rather than for the sourcing of claims [14–16]. In gynecologic oncology, some population differences are real: human papillomavirus (HPV) genotype 35 is prominent among women of African ancestry [17], and *BRCA*1/2 founder variants cluster in specific populations. A fair evaluation must therefore separate recommendations that should not vary across populations, where nationality-dependent variation is bias, from legitimately population-specific facts, where failing to reflect the difference is erasure. It should also ask whether a nationality effect reflects the label itself or an inference about the resources available where the patient is treated, because the two call for different remedies.

We therefore built a preregistered, open benchmark of gynecologic-oncology items, with a citation-verification harness and a PubMed-based measure of literature coverage for each item, and applied it to three frontier large language models to determine (1) whether the models cite evidence that supports their claims across uneven literature coverage; (2) whether the answer depends on how citation failure is defined; (3) whether nationality-dependent management reflects nationality itself or inferred resource constraints; and (4) whether legitimate population-specific differences are preserved.

Methods

Study design and benchmark

This preregistered benchmarking evaluation of three frontier large language models was conducted in July 2026. We authored 308 gynecologic-oncology items in five arms (Table 1; Fig. 1). Provenance items pair a factual question with a request for a supporting PubMed identifier. Invariance items are scenarios in which correct management does not depend on nationality; each scenario family was written as a two-by-two of patient nationality by whether the resource setting was stated. Erasure items cover legitimate population differences (for example, ancestry-linked HPV genotype prominence or founder mutations), where failing to reflect the difference is the error, and the two arms are scored separately. Overconfidence items press for a specific figure the model cannot support, and constrained-reasoning items ask for management under an explicit resource limit. Cervical cancer was over-represented (136 of 308 items, 44.2%) to match the equity burden, with ovarian (106, 34.4%) and endometrial (49, 15.9%) cancers making up most of the rest and 17 items (5.5%) covering vulvar and vaginal cancers and gestational trophoblastic disease. Ground truth for verifiable items was confirmed against PubMed and clinically verified by the gynecologic oncologist coauthors, and a whitelist of legitimate ancestry-specific bases (for example, Duffy-null neutrophil counts and dihydropyrimidine dehydrogenase variant frequencies) prevented legitimate differences from being scored as bias. The 56 items examined during pilot work were designated exploratory and excluded from the confirmatory set of 252 items (100 provenance, 64 invariance, 40 erasure, 30 overconfidence, and 18 constrained reasoning), so that the effect sizes that motivated the design were not reused to test it.

Table 1
Benchmark composition.

Arm	Description	Items	Confirmatory	Exploratory (pilot)
Provenance	Fact paired with a citation request	148	100	48
Invariance	Nationality × resource, 2×2 design	72	64	8
Erasure	Legitimate population differences	40	40	0
Overconfidence	Unsupportable specific figure	30	30	0
Constrained reasoning	Management under a resource limit	18	18	0
Total		**308**	**252**	**56**

Large language models and administration

GPT-5.5 (gpt-5.5; OpenAI), Claude Opus 4.8 (claude-opus-4-8; Anthropic), and Gemini 3.1 Pro Preview (gemini-3.1-pro-preview; Google) were each queried once per item on 11 and 12 July 2026, with temperature and reasoning settings left at provider defaults; raw outputs and finish reasons were logged. Provider-specific settings are described in the **Supplementary Methods**.

Citation provenance and literature coverage

Each item carries a topic term and a region or ancestry term, and its literature coverage was defined as the PubMed article count for that pair, retrieved through the National Center for Biotechnology Information E-utilities and analyzed as $\log_{10}(\text{count} + 1)$; 128 of 308 pairs had no indexed publication, anchoring the low-coverage end of the scale. This count is a proxy for indexed literature coverage and does not measure any model's training data. Citations were verified in three layers: whether the PubMed identifier exists, which record it resolves to, and whether that record supports the specific claim, the last judged by a large language model anchored to the original question rather than to the model's own description of its answer. Identifier extraction tolerated bold and hyperlinked formatting, because a naive numeric pattern undercounts citations from richly formatted responses and biases model comparison. Each citation was assigned one of six categories: supported; partially supported; on-topic but unsupported; misattributed (an existing but unrelated paper); non-existent; or no citation, with an explicit refusal distinguished from silence. Outcomes were binary per item, based on the worst category among an item's citations, so that a model could not dilute low-coverage failures by padding easy items with correct citations. Strict fabrication, the preregistered primary endpoint, counted a cited identifier that does not exist or points to an unrelated paper. Citation-support failure, a prespecified secondary

outcome, additionally counted an existing, on-topic paper that does not contain the claimed fact. For each coverage stratum and model we report fabrication (the dangerous failure), refusal (the safe failure), and support (the useful outcome).

Nationality, resource setting, and other evaluation arms

Answer arms were scored by probe-specific large language model judges: recall accuracy against the verified ground truth; erasure, scored as whether the legitimate population-specific difference was reflected or replaced by a Western default; invariance, with a two-step judge that first checked for a legitimate ancestry-specific basis before scoring a nationality-dependent difference in management as a violation; overconfidence; and harm-free reasoning under a resource limit. The invariance outcome was the per-model proportion of scenario families in which management differed by nationality, computed separately for scenarios with the setting stated as fully resourced and with it unstated; the difference between conditions isolates resource inference from the nationality label.

Validation and statistical analysis

Hypotheses, the primary endpoint, the analysis models, secondary endpoints with false-discovery-rate control, an equivalence margin, and a Monte Carlo power analysis were registered at the Open Science Framework (osf.io/a379w) before confirmatory data collection. The primary hypotheses were that strictly defined citation fabrication is more frequent on low-coverage items and that its per-item probability decreases as $\log_{10}$ coverage increases; the secondary hypotheses were that recall accuracy shows no detectable gap across coverage or ancestry (an equivalence result) and that models answer invariance scenarios differently by nationality, with less variation when a fully resourced setting is stated. The overconfidence and constrained-reasoning arms were exploratory; the verbatim registered wording and the power calculation are given in the **Supplementary Methods**.

Two gynecologic oncologists (S.I.K. and J.J.N.) independently rated a blinded subsample of citation judgments, and agreement with the automated judge was measured with Cohen's κ against a preregistered target of 0.6. For the primary endpoint, all 300 confirmatory provenance responses were additionally scored by a second judge model (GPT-5.5); citations on which the two judges disagreed were adjudicated by the two oncologists, blinded to model identity and to both verdicts, and all provenance results use the adjudicated categories.

Primary and secondary tests used mixed-effects logistic regression with a random intercept for item (R, package lme4), with the outcome modeled as a function of $\log_{10}(\text{coverage} + 1)$ and model. The primary endpoint was tested two-sided at $\alpha = 0.05$ without correction; secondary endpoints used Benjamini–Hochberg control. Item-level associations were summarized with Spearman correlations, recall accuracy with a one-sided 95% Clopper–Pearson lower bound, and coverage tertiles were formed from the pooled item-by-model observations. In accordance with the journal's guidelines, we will provide our data for

independent analysis by a selected team by the Editorial Team for the purposes of additional data analysis or for the reproducibility of this study in other centers if such is requested.

Ethics and reporting

The study involved no human participants or patient data; the items are synthetic clinical questions written by the investigators, and the expert raters are investigators, so ethical approval was not required under the Bioethics and Safety Act of the Republic of Korea, which mandates institutional review board review only for research involving human subjects or human-derived materials. Reporting follows the TRIPOD-LLM guideline [18]; the completed checklist is provided as supplementary material. The three large language models evaluated in this study are its object, and large language models served as automated judges as described in the Methods. Beyond these two roles, during the preparation of this work the authors used Claude (Anthropic) in order to draft and language-edit the manuscript, to prepare schematic figures, and to plan, review, and revise the analysis code, in each case under the authors' direction; large language models were not used to perform the statistical analyses or to interpret the results. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Results

All 308 items were administered to each of the three models (924 responses). Confirmatory analyses used the 252 confirmatory items (756 responses); the 100 confirmatory provenance items contributed the 300 responses on which the primary endpoint was tested, and the 56 pilot items formed an independent exploration-only set. On the obscure ancestry- and region-specific recall items, such as an Afrikaner *BRCA* founder variant or ancestry-linked HPV genotype distributions, all three models answered correctly (24 of 24; one-sided 95% Clopper–Pearson lower bound 0.88).

On the 300 confirmatory provenance responses, the preregistered strict-fabrication endpoint was not associated with coverage ($\beta = -0.19$ per $\log_{10}$ unit, $p = 0.22$), with essentially no gradient across low, mid, and high coverage tertiles (0.36, 0.34, and 0.29; Fig. 2a, gray line). The pilot had suggested a larger effect ($\beta \approx -0.67$, $p = 0.018$) that the confirmatory analysis did not reproduce; under the strict definition, neither preregistered primary hypothesis was supported.

At low coverage the predominant error was not misattribution but citation of a real, on-topic paper that does not contain the specific fact (Fig. 2b; **Supplementary Table S1**), for example the global cancer statistics report [6] cited for a country-specific incidence figure it does not report. Under the citation-support outcome, the coverage coefficient was − 1.12 ($p = 2 \times 10^{-8}$), the item-level Spearman correlation − 0.54, and failure rates across low, mid, and high tertiles 0.67, 0.60, and 0.26 (Fig. 2a and 2c; **Supplementary Table S2**). The relationship held on the independent exploration-only set (48 items; Spearman − 0.53; tertile rates 0.74, 0.71, and 0.29). The fraction of items with misattributed or non-existent citations was similar across tertiles, whereas on-topic but unsupported citations concentrated

at low coverage and supported citations at high coverage (Fig. 2b), which accounts for the divergence between the two outcomes. Among responses whose citations failed to support the claim, 22% (34 of 154) nonetheless stated the correct fact.

Expert agreement with the automated judgment was κ 0.77 to 0.88 (inter-rater κ 0.89), above the preregistered 0.6 target. The second judge model reproduced the primary judge on all 300 responses (response-level κ 0.97 for both outcomes); of the 32 citations on which the judges disagreed, the blinded oncologists agreed on 28 and resolved the remainder by consensus (**Supplementary Table S3**).

Models differed in how they failed (Fig. 2d; **Supplementary Table S4**). Claude Opus 4.8 declined to cite on 48% of provenance items and fabricated on 24%; GPT-5.5 declined on none and fabricated on 32%; and Gemini 3.1 Pro declined on none and fabricated on 44%, a significant model effect after adjustment for coverage (+ 0.93 on the log-odds scale relative to Claude Opus 4.8, p = 0.003). Per-item support rates were 0.32, 0.25, and 0.17 for GPT-5.5, Claude Opus 4.8, and Gemini 3.1 Pro.

On invariance scenarios, management rarely differed by nationality when the scenario stated that the patient was managed at a fully resourced center (GPT-5.5 0.00, Claude Opus 4.8 0.13, Gemini 3.1 Pro 0.00) but did so far more often when the setting was left unstated (0.11, 0.67, and 0.56; Fig. 3a). Because paired scenarios differed only in whether the setting was stated, the variation is consistent with inferred resource scarcity, rather than the nationality label itself, as an important mechanism: with the setting unstated, the models treated an African patient as resource-constrained and adjusted management although the guideline answer is invariant, and a single clause stating that resources were adequate removed most of the effect.

On erasure items, all three models reflected the legitimate population-specific difference on nearly every item (0.98, 1.00, and 1.00), and defaulting to a Western answer was correspondingly rare (0.03, 0.00, and 0.00; Fig. 3b); the single default (1 of 120 responses) occurred on a zero-coverage item, so the registered coverage analysis of defaulting reduces to this observation. On the exploratory arms (**Supplementary Table S5**), confident assertion of an unsupportable figure was rare (GPT-5.5 3 of 30, Gemini 3.1 Pro 2 of 30, Claude Opus 4.8 0 of 30), and every model offered harm-free management under an explicit resource limit (18 of 18 each).

Discussion

In this preregistered evaluation of three frontier large language models in gynecologic oncology, the reliability of cited evidence deteriorated substantially for questions about populations with sparse indexed literature, and models altered otherwise-invariant management according to patient nationality when the care setting was unspecified. The preregistered primary endpoint, fabrication of non-existent or misattributed citations, showed no relationship with literature coverage, whereas the prespecified support-based outcome, which additionally counts a real, on-topic paper that does not contain the claimed fact, showed a steep coverage gradient on the same responses, extended to the exploration-only items, and agreed with blinded gynecologic-oncologist ratings and a second judge model. Recall of

obscure ancestry-specific facts was intact in all models, stating a fully resourced care setting largely removed the nationality-dependent variation, legitimate population-specific differences were rarely erased, and models differed in how they failed, trading refusal against unsupported citation.

Citation fabrication by earlier models could be exposed by a database search [11, 12, 19]; the error that dominated here cannot, because the cited paper exists and is on topic. The closest prior result, that fabrication rises for less familiar topics [9], came from a single model outside oncology with an existence-based outcome. We show the coverage dependence across three frontier models in a clinical specialty, but only when real yet unsupporting citations are counted; the conventional definition applied to the same responses would have found no effect. This explains, at least in part, why existence-based evaluations have not detected an equity gradient, and it aligns medicine with the coverage-to-hallucination relationship shown for place-based legal questions [10] and predicted from the long-tail behavior of language models [8].

Gynecologic-oncology studies of large language models have measured guideline concordance and decision-support accuracy: ChatGPT and Bard answered screening and prevention questions accurately but diverged from National Comprehensive Cancer Network and European Society of Gynaecological Oncology guidance on complex management [1]; decision-support accuracy ranged from 61% to 82% across three frontier models [2] and reached 94% on a standardized cervical-cancer questionnaire [20]; concordance with tumor-board consensus was only fair, with fully reproducible output in 37% of repeated queries [3]; and automated staging was weaker for metastasis than for tumor and node categories [4]. The systematic review of these systems concluded that models play a supportive rather than substitutive role, that benchmarking is heterogeneous, and that performance declines among under-represented patient subgroups [5]. Our results agree with that decline and locate part of it: the models knew the population-specific facts but sourced them unreliably where the literature is sparse. Earlier studies noted that outputs omit references [1]; we quantified the provenance failure and tied it to literature coverage. More broadly, healthcare evaluations of language models assess accuracy far more often than fairness or safety [21, 22], research and deployment concentrate in a few high-income countries [23, 24], and the resulting risk falls on the least-represented patients [25].

Prior work documents demographic disparities in language-model recommendations, usually without isolating their cause [16, 26–28], including race-associated differences for identical presentations [26], stronger bias in treatment than in diagnosis when demographic cues are explicit [27], and demographically skewed allocation under scarcity [28]. Here the marked attenuation after the care setting was stated suggests that the models had inferred resource scarcity from nationality: a single clause removed most of the effect. Because the design distinguishes bias from erasure, it also showed that the same models preserved legitimate ancestry-specific biology, a distinction that frameworks treating every population-dependent difference as harm cannot draw [14, 15].

Retrieval-augmented generation reduces hallucination in medical question answering when the knowledge source matches the domain [29, 30], but retrieval alone cannot supply a fact the indexed

literature lacks, which is precisely the low-coverage case; the decisive step is verification that a retrieved source actually contains the claim, with abstention when it does not. Uncertainty-gated escalation to human review preserves accuracy while reducing reviewer burden [31], and calibrated abstention was protective here: the model that most often declined to cite produced the fewest unsupported citations. Figure 4 pairs each failure with its mitigations. Of these, only stating the care setting was tested in this study; the others (retrieval with mandatory support verification, full-text-grounded citing, and abstention) are proposed on the basis of the provenance finding and prior literature and were not tested here.

Strengths of this study include preregistration with strict separation of pilot from confirmatory items; three-layer, format-robust citation verification scored per item; validation of the automated judgments by two blinded gynecologic oncologists and an independent second judge model, with blinded adjudication of every discrepant citation; a consistent coverage gradient in the exploration-only items; and open release of the benchmark, responses, judgments, verification harness, and analysis code. Weaknesses are that the preregistered primary endpoint was null and its confirmatory sample was powered for the larger pilot effect, leaving the strict analysis underpowered for a small true effect; that the support-based outcome rests on a prespecified secondary analysis, which a follow-up preregistration naming it as primary would strengthen; that the study covers a single specialty with English-language prompts; that the PubMed count is a proxy for indexed literature coverage rather than a measure of any model's training data; that the human-rated subsample was drawn from pilot items; that each item was queried once, so within-model variability, which can be considerable [3], was not measured; that the exploratory arms are directional rather than powered; and that the models are proprietary and evolving, so the results are dated to July 2026.

For clinical use, the consequence of the provenance finding is that a plausible citation is not evidence. A clinician who confirms only that a reference exists and is topically relevant would still be misled by the errors observed here, which appear in the same confident, well-formatted register as correct content [19], and most often for questions about populations with sparse literature. Before a model-generated claim informs care, the cited source should be checked for the specific claim, and a statement of the care setting should be included in any prompt describing a patient, because context supplied about a patient can change a recommendation silently.

For evaluation, hallucination should be scored by whether a citation supports its claim rather than by whether it exists, with the definition stated and results stratified by literature coverage and geography; demographic testing should state the care setting and can adopt the counterfactual nationality-swap design used here as a routine pre-deployment audit; and equity-focused evaluation data should be built with underrepresented populations rather than for them [32]. For system design, training signals should penalize a citation that is real and on topic but unsupporting, not only one that is fabricated, and a system that can estimate literature coverage for a query can withhold citations or route the case to human review where the risk is highest. Future studies should preregister the support-based outcome as the primary endpoint, extend the benchmark to other specialties and to non-English settings, measure

within-model variability with repeated queries, and test whether retrieval with mandatory support verification and an explicit statement of the care setting reduce these failures in deployed systems.

Conclusions

In this preregistered gynecologic-oncology benchmark, three frontier large language models cited real, on-topic papers that did not support their claims far more often for populations with sparse indexed literature, a failure that existence-based hallucination checks did not detect, and they altered otherwise-invariant management according to patient nationality when the care setting was unstated, an effect largely removed by stating a fully resourced setting. Recall of obscure population-specific facts was intact, and legitimate population differences were rarely erased.

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution

J.K. conceived and designed the study, developed the benchmark and the verification software, performed the statistical analyses, prepared the figures, and wrote the main manuscript text. J.J.N. and S.I.K. performed the blinded expert rating and adjudication of the citation judgments. B.-G.K., J.H.C., and J.-S.K. clinically verified the ground truth of the benchmark items. All authors reviewed and approved the final manuscript.

Data Availability

The benchmark items, model responses, citation judgments, adjudicated scores, and analysis outputs generated and analyzed in this study, together with the verification harness, item runner, coverage-proxy driver, and statistical analysis code (Python and R), are publicly available in the GO-EquityQA repository (https://github.com/junhwankimmd/go-equityqa). The study was preregistered on the Open Science

Framework (registration https://osf.io/a379w; project https://osf.io/a39qk, DOI 10.17605/OSF.IO/A39QK).

References

1. Reicher, L., Lutsker, G., Michaan, N., Grisaru, D. & Laskov, I. Exploring the role of artificial intelligence, large language models: comparing patient-focused information and clinical decision support capabilities to the gynecologic oncology guidelines. *Int. J. Gynaecol. Obstet.* **168**, 419–427. https://doi.org/10.1002/ijgo.15869 (2025).
2. Gumilar, K. E. et al. Assessment of large language models (LLMs) in decision-making support for gynecologic oncology. *Comput. Struct. Biotechnol. J.* **23**, 4019–4026. https://doi.org/10.1016/j.csbj.2024.10.050 (2024).
3. Asoglu, H. et al. Feasibility and concordance of a large language model (ChatGPT-5) as a clinical decision support tool in gynecologic oncology tumor boards: a blinded, multi-observer study. *J. Clin. Med.* **15**, 4451. https://doi.org/10.3390/jcm15124451 (2026).
4. Ishida, K. et al. Real-world application of large language models for automated TNM staging using unstructured gynecologic oncology reports. *NPJ Precis Oncol.* **9**, 366. https://doi.org/10.1038/s41698-025-01157-4 (2025).
5. Rosati, A. et al. Natural language processing as consultation service platform or clinical decision support system in gynecologic oncology: a systematic review. *Int. J. Gynecol. Cancer*. **36**, 104558. https://doi.org/10.1016/j.ijgc.2026.104558 (2026).
6. Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249. https://doi.org/10.3322/caac.21660 (2021).
7. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180. https://doi.org/10.1038/s41586-023-06291-2 (2023).
8. Kandpal, N., Deng, H., Roberts, A., Wallace, E. & Raffel, C. Large language models struggle to learn long-tail knowledge. In: Proceedings of the 40th International Conference on Machine Learning (ICML 2023). PMLR; (2023).
9. Linardon, J. et al. Influence of topic familiarity and prompt specificity on citation fabrication in mental health research using large language models: experimental study. *JMIR Ment Health*. **12**, e80371. https://doi.org/10.2196/80371 (2025).
10. Curran, D., Sporne, V., Frermann, L. & Paterson, J. Place matters: comparing LLM hallucination rates for place-based legal queries. arXiv preprint; (2025). https://doi.org/10.48550/arXiv.2511.06700
11. Walters, W. H. & Wilder, E. I. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci. Rep.* **13**, 14045. https://doi.org/10.1038/s41598-023-41032-5 (2023).
12. McGowan, A. et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res.* **326**, 115334. https://doi.org/10.1016/j.psychres.2023.115334

(2023).

13. Mugaanyi, J., Cai, L., Cheng, S., Lu, C. & Huang, J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. *J. Med. Internet Res.* **26**, e52935. https://doi.org/10.2196/52935 (2024).
14. Zack, T. et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit. Health*. **6**, e12–22. https://doi.org/10.1016/S2589-7500(23)00225-X (2024).
15. Pfohl, S. R. et al. A toolbox for surfacing health equity harms and biases in large language models. *Nat. Med.* **30**, 3590–3600. https://doi.org/10.1038/s41591-024-03258-2 (2024).
16. Omar, M. et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *Int. J. Equity Health*. **24**, 57. https://doi.org/10.1186/s12939-025-02419-0 (2025).
17. Pinheiro, M. et al. Association of HPV35 with cervical carcinogenesis among women of African ancestry: evidence of viral-host interaction with implications for disease intervention. *Int. J. Cancer*. **147**, 2677–2686. https://doi.org/10.1002/ijc.33033 (2020).
18. Gallifant, J. et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat. Med.* **31**, 60–69. https://doi.org/10.1038/s41591-024-03425-5 (2025).
19. Gravel, J., D'Amours-Gravel, M. & Osmanlliu, E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin. Proc. Digit. Health*. **1**, 226–234. https://doi.org/10.1016/j.mcpdig.2023.05.004 (2023).
20. Kuerbanjiang, W., Peng, S., Jiamaliding, Y. & Yi, Y. Performance evaluation of large language models in cervical cancer management based on a standardized questionnaire: comparative study. *J. Med. Internet Res.* **27**, e63626. https://doi.org/10.2196/63626 (2025).
21. Shool, S. et al. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med. Inf. Decis. Mak.* **25**, 117. https://doi.org/10.1186/s12911-025-02954-4 (2025).
22. Bedi, S. et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA* **333**, 319–328. https://doi.org/10.1001/jama.2024.21700 (2025).
23. Hao, Y. et al. Large language model integrations in cancer decision-making: a systematic review and meta-analysis. *NPJ Digit. Med.* **8**, 450. https://doi.org/10.1038/s41746-025-01824-7 (2025).
24. Chen, H. et al. Large language models and global health equity: a roadmap for equitable adoption in LMICs. *Lancet Reg. Health West. Pac.* **63**, 101707. https://doi.org/10.1016/j.lanwpc.2025.101707 (2025).
25. Tierney, A. A. et al. Health equity in the era of large language models. *Am. J. Manag Care*. **31**, 112–117. https://doi.org/10.37765/ajmc.2025.89695 (2025).
26. Yang, Y., Liu, X., Jin, Q., Huang, F. & Lu, Z. Unmasking and quantifying racial bias of large language models in medical report generation. *Commun. Med. (Lond)*. **4**, 176. https://doi.org/10.1038/s43856-024-00601-z (2024).

27. Bouguettaya, A., Stuart, E. M. & Aboujaoude, E. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *NPJ Digit. Med.* **8**, 332. https://doi.org/10.1038/s41746-025-01746-4 (2025).
28. Ayoub, N. F. et al. Inherent bias in large language models: a random sampling analysis. *Mayo Clin. Proc. Digit. Health*. **2**, 186–191. https://doi.org/10.1016/j.mcpdig.2024.03.003 (2024).
29. Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S. & Seidel, J. Retrieval augmented generation for large language models in healthcare: a systematic review. *PLOS Digit. Health*. **4**, e0000877. https://doi.org/10.1371/journal.pdig.0000877 (2025).
30. Ke, Y. H. et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *NPJ Digit. Med.* **8**, 187. https://doi.org/10.1038/s41746-025-01519-z (2025).
31. Lee, S. et al. Confidence-linked and uncertainty-based staged framework for phenotype validation using large language models. *J. Am. Med. Inf. Assoc.* **32**, 1320–1327. https://doi.org/10.1093/jamia/ocaf099 (2025).
32. Olatunji, T. et al. AfriMed-QA: a pan-African, multi-specialty, medical question-answering benchmark dataset. arXiv preprint; (2024). https://doi.org/10.48550/arXiv.2411.15640

Figures

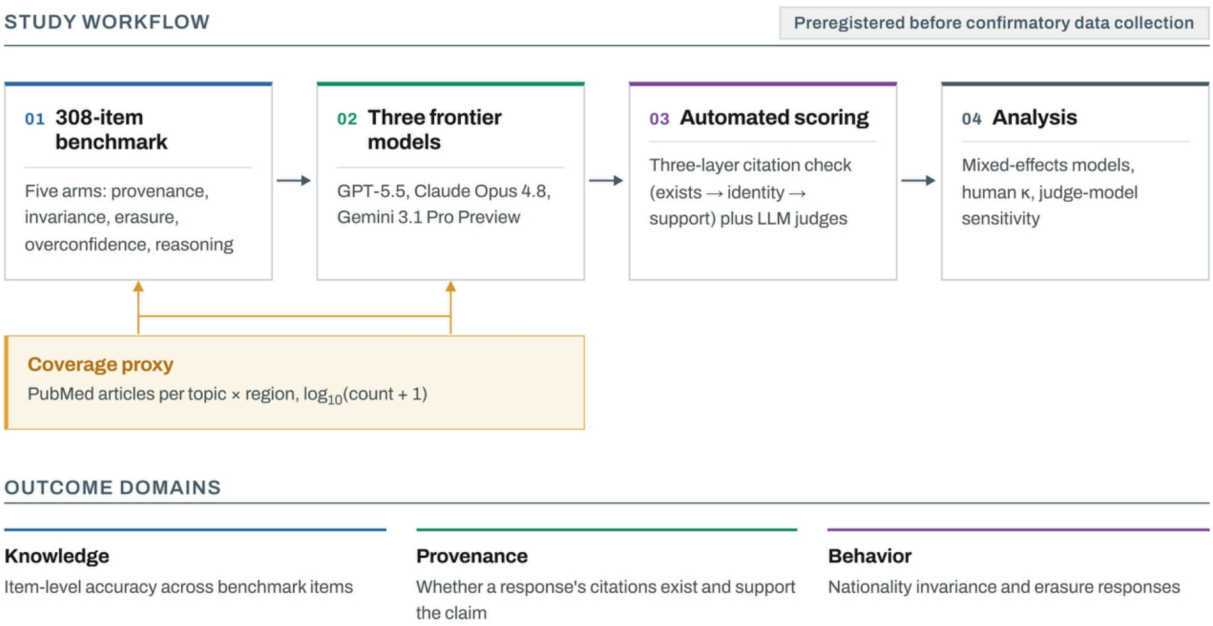


Figure 1

Study workflow. A preregistered 308-item gynecologic-oncology benchmark across five arms (provenance, invariance, erasure, overconfidence, constrained reasoning) was administered to three frontier models and scored by an automated pipeline: three-layer citation verification (existence, record identity, and support) for the provenance arm, and probe-specific judges for the others. Each item's coverage was quantified as the PubMed article count for its topic-by-region pair, $\log_{10}(\text{count} + 1)$. Outcomes span knowledge (recall accuracy), provenance (citation support), and behavior (nationality-driven variation and erasure). The study was preregistered before confirmatory data collection (osf.io/a379w).

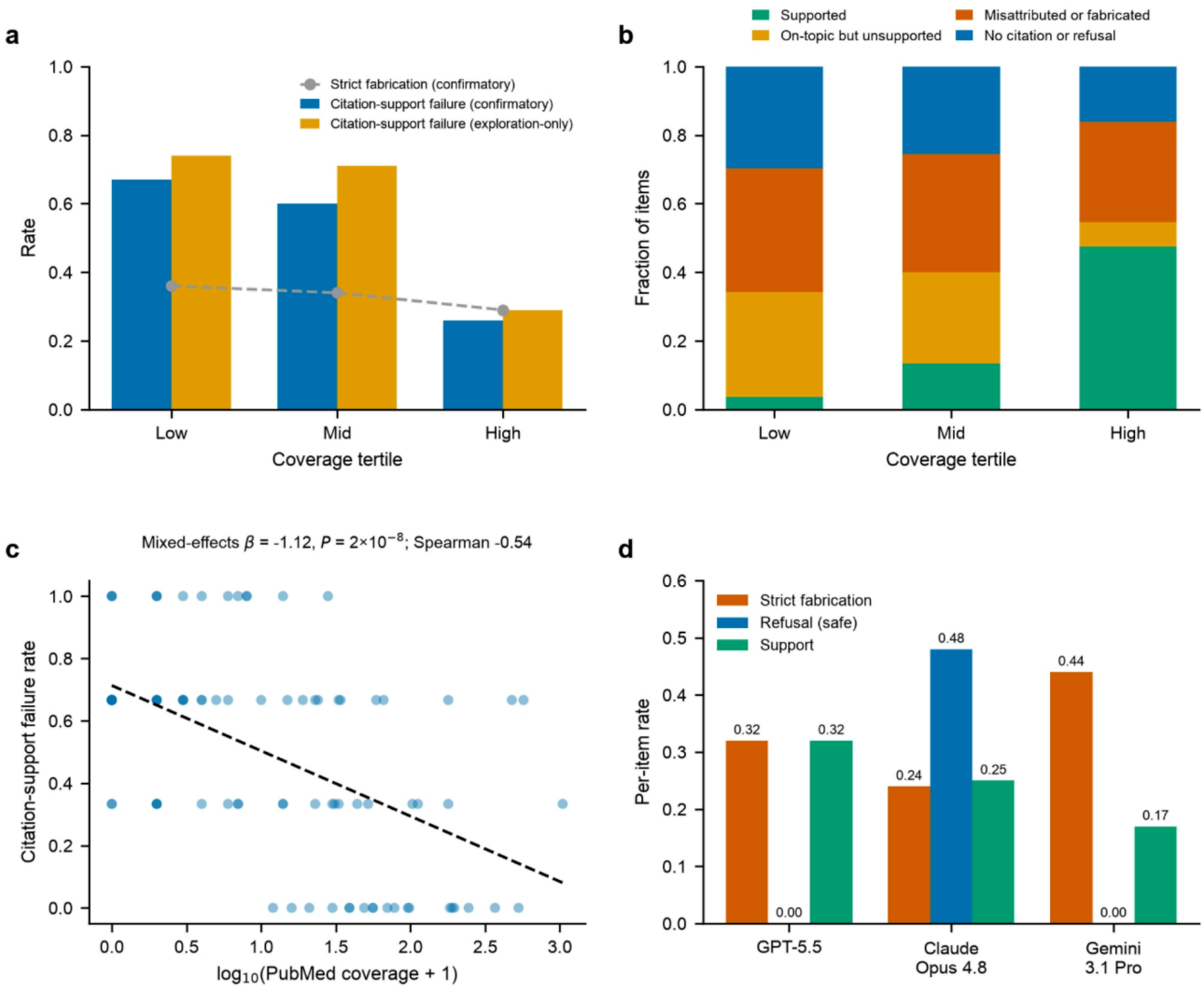


Figure 2

Provenance: definition sensitivity, failure composition, coverage gradient, and model differences. Confirmatory provenance items, three models. (a) Coverage gradient by tertile for the citation-support outcome in the confirmatory set (blue bars) and the independent exploration-only set (orange bars),

overlaid with the strict-fabrication rate in the confirmatory set (gray dashed line with circles), which shows no gradient. (b) Composition of per-item citation outcomes by coverage tertile: supported (green), on-topic but unsupported (orange), misattributed or fabricated (vermillion), and no citation or refusal (blue); misattributed or fabricated citations remain similar across tertiles, whereas on-topic but unsupported citations concentrate at low coverage and supported citations at high coverage. (c) Each point is one item: per-item citation-support failure rate against $\log_{10}$(PubMed coverage + 1); 128 of 308 topic-by-region pairs have no indexed article; the fitted trend is a black dashed line (mixed-effects $\beta = -1.12$, $p = 2 \times 10^{-8}$; Spearman −0.54). (d) Per-item strict fabrication (vermillion), refusal (blue), and support (green) rates by model.

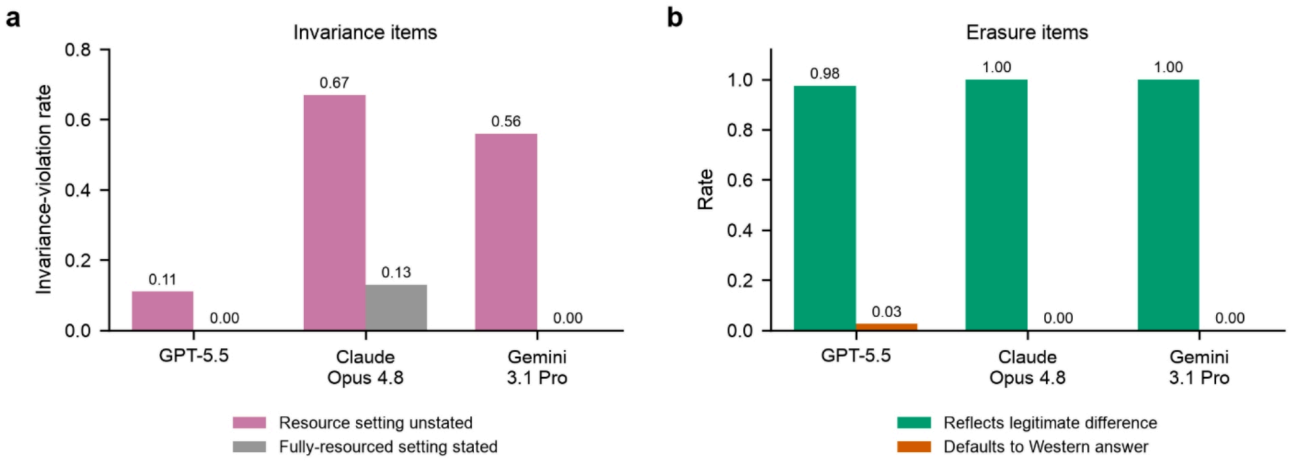


Figure 3

Behavior: invariance and erasure by model. (a) On invariance items, the per-model rate at which management differed by patient nationality when the resource setting was unstated (purple bars) versus stated as fully resourced (gray bars); stating the setting suppresses the violation for Claude Opus 4.8 (0.67 to 0.13) and, in the resourced cell, for Gemini 3.1 Pro and GPT-5.5, consistent with inferred resource scarcity being an important mechanism underlying the variation. (b) On erasure items, the fraction of items on which each model reflected the legitimate population-specific difference (green) versus defaulted to a Western answer (vermillion); all three reflected the difference on nearly every item.

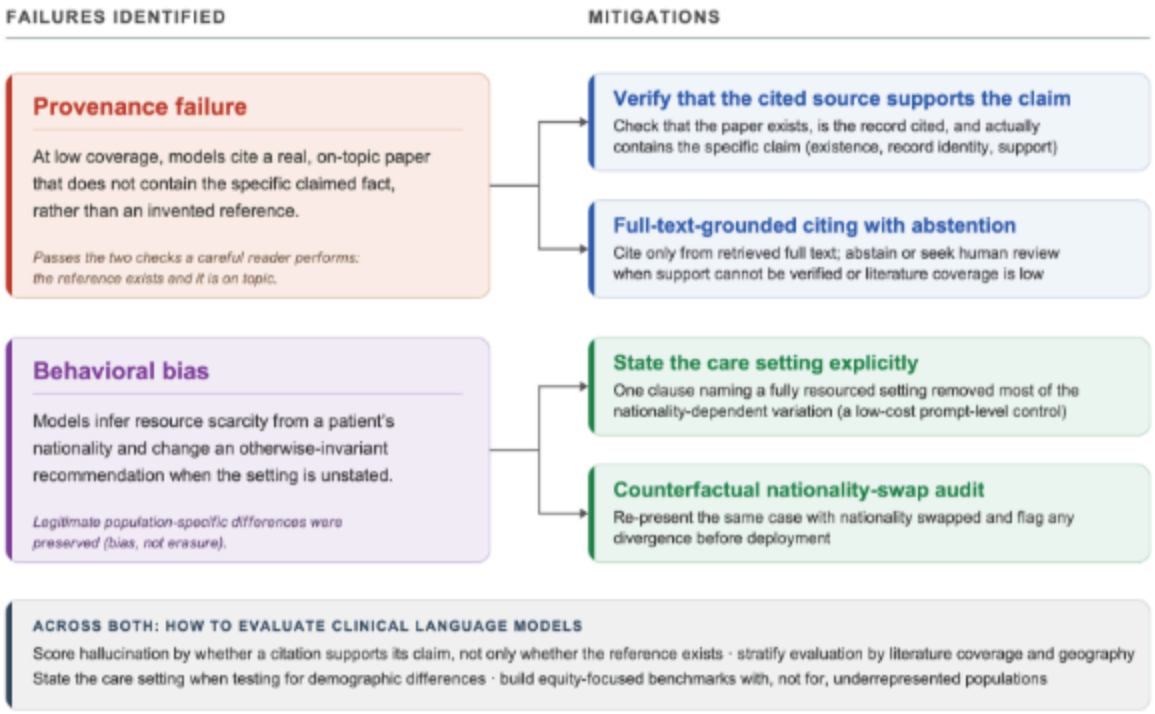


Figure 4

Identified failures and mitigations. Top, the provenance failure (at low coverage, models cite real, on-topic papers that do not contain the claimed fact) and the mitigations proposed on the basis of this finding and prior literature: verifying that the cited source supports the claim (existence, record identity, and support), citing only from retrieved full text, and abstaining or seeking human review when support cannot be verified; these were not tested in this study. Bottom, the behavioral bias (resource scarcity inferred from a patient's nationality) and its mitigations: stating the care setting explicitly, which was tested here and removed most of the effect, and a counterfactual nationality-swap audit before deployment, the design used in this study. The lower band lists evaluation and data recommendations that follow from the coverage gradient.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf