Supplementary Information

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

Junhwan Kim, MD, Joseph J. Noh, MD, Se Ik Kim, MD, PhD, Byoung-Gie Kim, MD, PhD, Jae Hyun Cho, MD, PhD, Jeong-Sig Kim, MD, PhD

***Corresponding author:** *Junhwan Kim, MD*

Contents

- Supplementary Methods (§1–§7)
- Supplementary Tables S1–S5

Table of Contents

Supplementary Methods ……..…… P. 2

Supplementary Table S1. Per-item citation-category composition by model and coverage tertile (confirmatory provenance items)……..…… P. 6

Supplementary Table S2. Mixed-effects logistic model summaries…….……..…...…… P. 7

Supplementary Table S3. Agreement between expert raters, the primary automated judge, and a second judge model…….....…… P. 8

Supplementary Table S4. Per-item citation outcomes by model (confirmatory provenance items, n = 100 per model).…….……..... P. 9

Supplementary Table S5. Exploratory arms: overconfidence and constrained-reasoning outcomes by model.……...……..……...…...…P. 10

Supplementary Data (provided in the repository)……..……..…P. 11

Supplementary Methods

1. Provider-specific administration settings

Following each provider's constraints, GPT-5.5 calls used a generous completion-token budget (the max_completion_tokens parameter); Claude Opus 4.8 calls left temperature and top-p unset, because setting them returns an error and adaptive thinking consumes the token budget; and Gemini 3.1 Pro calls used the preview interface within its daily request limit. Reasoning and thinking settings were left at each provider's default; no reasoning-effort level or thinking-token budget was specified. Each item was administered once per model. Responses were collected on 11 and 12 July 2026; the Gemini responses for the overconfidence and constrained-reasoning arms were collected on 12 July 2026. Raw outputs and finish reasons were logged for every call. Judge calls used generous token budgets so that reasoning tokens did not exhaust the allocation before the structured output was produced, with robust extraction of the JavaScript Object Notation (JSON) verdict that tolerates code fences and preamble text.

2. Coverage-proxy construction

Each item carries a topic term and a region or ancestry term. The PubMed article count for the conjunction of the two terms was retrieved through the National Center for Biotechnology Information E-utilities search service (the coverage-proxy driver in the repository) and transformed as $\log_{10}(\text{count} + 1)$ for use as a continuous predictor. The continuous count was kept separate from a coarse topic-density label because narrow topic-by-region queries can return low counts even for a well-covered topic. Of the 308 topic-by-region pairs, 128 returned no indexed publication, anchoring the low-coverage end of the scale. Coverage tertiles were formed from $\log_{10}(\text{count} + 1)$ pooled over item-by-model observations. The count is a proxy for indexed literature coverage; it does not measure the representation of a topic or population in any proprietary model's training data. Query terms for every item are released with the benchmark.

3. Citation extraction and three-layer verification

PubMed identifiers were extracted from every response with a format-robust pattern, PMID[:\s*\[\]]*?(\d+), which tolerates markdown bold, brackets, and hyperlinked identifiers (for example, **PMID:** **12345678** or [12345678](url)). A naive pattern of the form PMID[:\s]*(\d+) misses these forms and systematically undercounts citations from models that format richly, which biases model comparison. All identifiers in a response

were extracted, not only the first. Verification proceeded in three layers. Layer 0 (existence) queried the E-utilities search service to establish whether the identifier resolves to a PubMed record. Layer 1 (record identity) retrieved the record with the fetch service. Layer 2 (claim support) presented the original question, the claimed fact, and the retrieved record to a judge model (Claude Opus 4.8) that was anchored to the original question rather than to the responding model's own description of its answer, and that assigned one of six mutually exclusive categories: supported; partially supported; on-topic but unsupported; misattributed (an existing but unrelated paper); non-existent; or no citation. A format-robust refusal flag distinguished an explicit decline to cite from silence. Aggregation was per item and binary: the worst category among an item's citations determined the item's outcomes, 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 non-existent or misattributed identifier; citation-support failure (a prespecified secondary outcome) additionally counted an on-topic but unsupported citation. Partially supported citations counted toward citation-support failure but not toward strict fabrication, and a response that also contained a supporting citation was not counted as a citation-support failure. For each coverage stratum and model, three axes are reported: fabrication (the dangerous failure), refusal (the safe failure), and support (the useful outcome), because the support rate alone penalizes safe refusers and flatters models that cite freely.

4. Judges and scoring rules for the answer arms

The answer arms were scored by probe-specific large language model judges whose full prompts and system messages are released in the repository. The accuracy judge compared the response with the verified ground truth for the recall items. The erasure judge scored whether a response reflected the legitimate population-specific difference or defaulted to a Western answer. The invariance judge operated in two steps: it first checked the response against a whitelist of legitimate ancestry-specific bases (for example, Duffy-null neutrophil counts and dihydropyrimidine dehydrogenase variant frequencies), so that a legitimately population-specific rationale was not scored as bias, and then judged whether management differed between the paired scenarios that varied only in patient nationality. The invariance outcome was computed separately for scenario families in which the setting was stated as fully resourced and for families in which it was left unstated. The overconfidence judge scored whether the model asserted a confident specific figure that it could not support, and the constrained-reasoning judge scored whether the model offered a defensible, non-harmful option under an explicit resource limit. Ground

truth for verifiable items was confirmed against PubMed and clinically verified by the gynecologic oncologist coauthors; facts that could not be verified were marked as such and were not used as ground truth.

5. Preregistered hypotheses and power analysis

The registered hypotheses were as follows. H1 (secondary reassurance endpoint, reported as an equivalence result): recall accuracy shows no detectable gap across coverage or ancestry. H2 (primary): citation fabrication, defined strictly as a non-existent or misattributed PubMed identifier, is more frequent on low-coverage items. H3 (secondary): on Type-I invariance scenarios, where the correct answer does not depend on nationality or ancestry, models nonetheless answer differently by nationality or ancestry, and stating that the setting is fully resourced reduces this variation. H4 (primary causal bridge): the per-item probability of fabrication decreases as $\log_{10}$ coverage increases. The overconfidence and constrained-resource reasoning arms were registered as exploratory. Secondary endpoints were registered with Benjamini–Hochberg false-discovery-rate control and H1 with an equivalence margin. Power for the primary endpoint was estimated by Monte Carlo simulation under the pilot effect size (coefficient near −0.67), which is why the smaller observed effect left the strict analysis underpowered. The registration (osf.io/a379w) was completed before confirmatory data collection.

6. Circularity guard

Items seen during pilot work were flagged exploratory and excluded from confirmatory analysis, so that the effect sizes that motivated the design were not reused to test it. The exploration-only replication reported in the main text uses these 48 provenance items, scored with the same final judge.

7. Judge validation and adjudication

Two gynecologic oncologists independently scored a blinded subsample drawn from pilot items; agreement with the automated judges was measured by Cohen's κ for both the strict-fabrication outcome and the citation-support outcome, with an inter-rater κ of 0.89 and a preregistered target of 0.6. A second judge model (GPT-5.5) then scored all 300 confirmatory provenance responses; response-level agreement with the primary judge (Claude Opus 4.8) was κ 0.97 for both outcomes. The 32 citations on which the two judges assigned different categories were adjudicated by the two oncologists, blinded to model identity and to the judges' verdicts; they agreed on 28 of 32 (κ 0.76 across four categories) and resolved the remaining four by consensus. All provenance results in the main text use the adjudicated categories (**Supplementary Table S3**).

8. Statistical models

Primary and secondary tests used mixed-effects logistic regression with a random intercept for item, fitted in R with the lme4 package (glmer, bobyqa optimizer): outcome ~ $\log_{10}$(coverage + 1) + model + (1 | item). 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, and recall accuracy with a one-sided 95% Clopper–Pearson lower bound. Model summaries are given in **Supplementary Table S2**, and the full model objects, standard errors, and confidence intervals are provided in the repository together with the analysis code (Python and R).

Supplementary Table S1. Per-item citation-category composition by model and coverage tertile (confirmatory provenance items).

Model	Coverage	Supported	On-topic unsupported	Misattributed	No citation
GPT-5.5	Low	2	23	12	0
GPT-5.5	Mid	6	11	13	0
GPT-5.5	High	23	3	7	0
Claude Opus 4.8	Low	2	2	10	23
Claude Opus 4.8	Mid	3	4	8	15
Claude Opus 4.8	High	17	0	6	10
Gemini 3.1 Pro	Low	0	9	18	10
Gemini 3.1 Pro	Mid	3	9	10	8
Gemini 3.1 Pro	High	9	2	16	6

Counts of items assigned to each mutually exclusive category (worst category per item). Misattributed and fabricated citations are roughly constant across coverage, whereas on-topic but unsupported citations concentrate at low coverage and are replaced by supported citations at high coverage; this is why the strict definition shows no gradient while the broad citation-support definition does. No non-existent PubMed identifiers occurred in the confirmatory set; every strict-fabrication citation was a real but misattributed identifier. Tertiles were formed from $\log_{10}$(PubMed coverage + 1) pooled over item-by-model observations. Ties in the coverage proxy at the tertile boundaries yield unequal bins (n = 37, 30, and 33 items per model, identical across models).

Supplementary Table S2. Mixed-effects logistic model summaries.

Outcome / term	Coefficient	p value	Direction
Strict fabrication ~ $\log_{10}$(coverage + 1)	-0.19	0.22	not significant
Citation-support failure ~ $\log_{10}$(coverage + 1)	-1.12	2×10^{-8}	negative
Gemini vs Claude Opus 4.8 (strict fabrication)	+0.93	0.003	positive

Each model included a random intercept for item (R, lme4, glmer, bobyqa optimizer). The strict-fabrication endpoint was the preregistered primary analysis and was reported uncorrected; the citation-support outcome was a prespecified secondary analysis. Full model objects, standard errors, and confidence intervals are provided in the repository (src/stats_glmm.R and its output).

Supplementary Table S3. Agreement between expert raters, the primary automated judge, and a second judge model.

Comparison	Strict fabrication (κ)	Citation-support failure (κ)
Expert rater A vs rater B	1.00	0.89
Expert rater A vs automated judge	0.94	0.77
Expert rater B vs automated judge	0.94	0.88
Second judge model vs primary judge (20% exploration subsample)	1.00	0.83
Second judge model vs primary judge (all 300 primary responses)	0.97	0.97

Cohen's κ on a blinded subsample. Judge-discrepant citations on the primary endpoint (n=32) were adjudicated by the blinded raters (28/32 concordant, κ 0.76; four resolved by consensus). Agreement is reported separately for the strict fabrication outcome and the broad citation-support outcome. κ for the broad outcome is lower because the supported-versus-unsupported boundary is more subjective, yet remains above the 0.6 target.

Supplementary Table S4. Per-item citation outcomes by model (confirmatory provenance items, n = 100 per model).

Model	Fabrication	Refusal	Support
GPT-5.5	0.32	0.00	0.32
Claude Opus 4.8	0.24	0.48	0.25
Gemini 3.1 Pro	0.44	0.00	0.17

Per-item fractions. This table gives the exact values summarized visually in Figure 2d. Adjusted Gemini model effect +0.93 on the log-odds scale, P = 0.003.

Supplementary Table S5. Exploratory arms: overconfidence and constrained-reasoning outcomes by model.

Model	Overconfidence rate (n = 30)	Constrained-reasoning harm-free rate (n = 18)
GPT-5.5	0.10 (3/30)	1.00 (18/18)
Claude Opus 4.8	0.00 (0/30)	1.00 (18/18)
Gemini 3.1 Pro	0.07 (2/30)	1.00 (18/18)

Overconfidence rate: fraction of items on which the model asserted a confident specific figure it could not support (lower is safer). Constrained-reasoning harm-free rate: fraction on which it offered a defensible, non-harmful option under an explicit resource limit (higher is safer). All three models converge on safe behavior; the between-model divergence seen on citation provenance (**Figure 2d**) does not appear here. Gemini 3.1 Pro responses for these arms were collected about a day after the others on a preview model.

Supplementary Data (provided in the repository)

The following raw materials accompany this study at https://github.com/junhwankimmd/go-equityqa:

- Item-level results for all 308 items (item id, arm, topic, region, log coverage, per-model outcome and category).
- Full model responses and three-layer citation judgments (existence, identity, support).
- Held-out (exploratory) replication results and coverage tertiles.
- Invariance pair-level verdicts (each family under stated and unstated resource conditions).
- Accuracy-ceiling items and verdicts, and erasure per-item results.

Model and judge prompts, system messages, coverage query terms, model versions, and collection dates.