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.

When Patients Say “Do Not Share”: LLMs Misroute Sensitive Information in Clinical Documentation

Mahmud Omar

Mahmudomar70@gmail.com

Beth Israel Deaconess Medical Center, Harvard Medical School

Reem Agbareia

Hadassah Medical Center

Alexander W Charney

Mount Sinai Health System

Raja-Elie E Abdulnour

Harvard Medical School

Ankit Sakhuja

Mount Sinai Health System

Alon Gorenshtein

Beth Israel Deaconess Medical Center, Harvard Medical School

Jonathan B Kruskal

Beth Israel Deaconess Medical Center, Harvard Medical School

Olga R Brook

Beth Israel Deaconess Medical Center, Harvard Medical School

Muneeb Ahmed

Beth Israel Deaconess Medical Center, Harvard Medical School

Yiftach Barash

Beth Israel Deaconess Medical Center, Harvard Medical School

Benjamin S Glicksberg

Mount Sinai Health System

Girish N Nadkarni

Mount Sinai Health System

Eyal Klang

Beth Israel Deaconess Medical Center, Harvard Medical School

Article

Keywords:

Posted Date: August 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10565869/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

Handling and routing sensitive medical information are important aspects of patient care, and patients sometimes ask clinicians not to share particular information. With the growing formal and informal use of AI-generated summaries, we tested whether large language models (LLMs) honor these requests when generating documents for different recipients. We analyzed 20,809 outputs across a seven-model paired evaluation using 120 controlled synthetic cases and 45 patient-authored health queries. Each case had matched versions with and without an explicit request. Despite the request, models disclosed sensitive information in 722 of 2,113 documents generated for unrelated specialists or employers (34.2%). Overall, requests reduced disclosure from 46.6% to 34.2%, a 12.5-percentage-point decrease, leaving sensitive information disclosed in more than one in three such documents. To address this, we evaluated a recipient-conditioned routing workflow designed to omit sensitive information from employer letters and unrelated referrals while preserving it when clinically required. On Llama 3.3 70B, the routing workflow reduced disclosure from 21.7% to 1.2% without increasing harmful omission. Patient requests alone did not provide reliable confidentiality control. Both inappropriate disclosure and excessive omission could undermine trust and safe information sharing. Recipient-specific routing should therefore be evaluated as a core clinical safety function before use.

Introduction

Patients often disclose sensitive information because it is needed for care. Some explicitly ask that particular information not be shared. When LLMs turn one encounter into notes, referrals, summaries, and letters, they must decide whether each sensitive fact belongs in each document.

Generative AI is moving rapidly into clinical care (1–4), and routine documentation is among its first widespread uses. Ambient AI scribes have reduced charting time and documentation burden (5–8), and surveys suggest that use extends beyond formal deployments (9). Generating a clinical document is not simple transcription. A system can be technically secure and still place sensitive information in a document for the wrong reader.

The same information may be appropriate in one document and inappropriate in another. It may belong in the treating clinician's note, be essential to an anesthesiologist, be irrelevant to an orthopedic referral, and be inappropriate in an employer letter. Whether it belongs depends on who will receive the document and why. This is the logic of contextual integrity: confidentiality depends on whether information reaches an appropriate recipient for an appropriate purpose, not on whether the fact is secret in itself (10). In two national surveys, 61% to 81% of US adults reported withholding medically relevant information, often because of embarrassment or fear of judgment (11). Patient-facing documents may also be read by parents, spouses, or other caregivers, so the actual reader may differ from the named recipient (12, 13).

Most medical LLM studies test accuracy on static tasks rather than behavior in clinical workflows (14, 15). Studies using clinical conversations show that context can change model performance (16). Patient communication can also alter downstream decisions (17). Privacy studies have shown that LLMs may

disclose private information (18) or transfer third-party information even under privacy-oriented instructions (19). What remains unknown is whether they honor a patient's own explicit request across documents intended for different recipients.

We asked whether an LLM would honor an explicit patient request not to share sensitive information across documents intended for different clinical and non-clinical recipients. Removing every sensitive fact is not safe, because some information must still reach clinicians who need it. We conducted a paired evaluation of inappropriate disclosure and harmful omission across seven LLMs. On two models, we also tested a recipient-conditioned routing workflow designed to route sensitive information by recipient and clinical need.

Methods

Study design

In our matched study, each case was evaluated with and without an explicit patient request to withhold one sensitive fact, while all other details remained unchanged. The same encounter was converted into documents for the treating clinician, a clinically relevant specialist, an unrelated specialist, the patient, and an employer. Each case therefore served as its own control, allowing us to measure two competing errors: disclosing information to a recipient who should not receive it and omitting information that another clinician needs for safe care.

The study combined synthetic cases reviewed by physicians with patient-authored health questions. The request appeared within the patient-clinician encounter rather than as a separate instruction to the model. The synthetic cases provided clear recipient-specific ground truth and included conflicts between confidentiality and clinical safety. The patient-authored cases tested whether the pattern held in natural patient language. We evaluated seven LLMs in the main evaluation and compared five documentation approaches on two models. We followed the reporting principles of TRIPOD-LLM (20).

Controlled synthetic cases

The synthetic bank included 120 case variants derived from 30 literature-informed clinical templates, with four matched dialogues per template. Cases were equally distributed across six confidentiality-sensitive domains: substance use, HIV or sexually transmitted infection, mental health, reproductive health, intimate partner violence, and gender identity. For each template, the study team prespecified the sensitive fact, intended recipients, safety-critical information, and expected handling in each document. DeepSeek V4 Flash converted these specifications into natural dialogues. Every variant was checked against its specification and reviewed by two clinician investigators for clinical coherence and recipient appropriateness.

Each case had two matched versions that differed only by a single request not to share the sensitive information. Depending on the scenario, each case produced three to five recipient-specific documents, yielding 520 documents per condition. Forty cases included information that a receiving specialist required for safe care, with equal numbers in which the conflict between the patient's request and clinical need was resolved or left unresolved in the dialogue. Templates, construction procedures, recipient definitions, and safety strata are provided in the Supplementary Methods, Section 1.1, and Supplementary Table S2.

Patient-authored cases

We also assembled a case set of 45 patient-authored questions from the public ChatDoctor-HealthCareMagic-100k corpus. The corpus was derived from online medical consultations and described by its creators as cleaned and deidentified (21). The cases covered the same six domains. Two LLMs from different providers independently screened each question for a first-person sensitive disclosure, after which two clinician investigators reviewed every case. This set tested whether the pattern found in controlled cases also appeared in patient-authored language.

We added one natural-language request not to share the sensitive information to each question; the matched version without a request removed only that line. Each case generated an internal note, a relevant referral, a referral to an unrelated specialist, a patient-facing summary, and an employer letter, yielding 225 documents per condition. Relevant referrals were classified as appropriate rather than required because these cases were not constructed around a prespecified safety-critical need. Selection details and domain counts are provided in the Supplementary Methods, Section 1.2, and Supplementary Table S3.

Model evaluation and outcome adjudication

Model evaluation and outcome adjudication
We evaluated Claude Sonnet 4.6, DeepSeek V4 Flash, Gemini 3.5 Flash, Ling 2.6 Flash, Llama 3.3 70B Instruct, Mistral Small 3.2 24B Instruct, and GPT-5.4 mini. Each LLM received the encounter and intended document recipient. Documents were generated through OpenRouter at temperature 0. The design specified 7,280 synthetic and 3,150 patient-authored documents.

Claude Sonnet 4.5, which was not among the evaluated models, served as the primary adjudicator. For each document, it classified the sensitive fact as directly stated, paraphrased, clearly implied, or absent. The first three categories counted as present. A fact in a document for an inappropriate recipient counted as inappropriate disclosure, while its absence from a required referral counted as harmful omission. Presence in an internal note or an appropriate referral was not considered an error. We also repeated the assessment using deterministic text matching. Model identifiers, prompts, adjudication

instructions, and text-matching rules are provided in the Supplementary Methods, Sections 1.3 and 1.4, and Supplementary Table S1.

Confidentiality and safety endpoints

We asked two clinical questions: did the LLM disclose sensitive information to someone who did not need it, including when the patient explicitly asked that it not be shared, and did it omit information from a referral when another clinician needed it for safe care? The primary confidentiality endpoint was disclosure to a clearly inappropriate recipient, defined as an unrelated specialist or an employer.

Patient-facing summaries were evaluated separately because a request not to share information with another clinician or employer did not always apply to the patient's own copy. We measured whether the sensitive fact appeared in all patient summaries and then examined the subset in which the request explicitly named a proxy, portal, family member, caregiver, or summary recipient.

Harmful omission was assessed in the 40 synthetic cases with a required referral. We reported omission separately when no request was present, when the conflict between the patient's request and clinical need was resolved in the dialogue, and when it remained unresolved.

Physician validation of disclosure and recipient need

Two board-certified primary care physicians independently reviewed identical workbooks with all automated labels hidden. They assessed 120 challenging documents sampled from disagreements between the adjudicator and text-matching method and classified the sensitive fact as directly stated, indirectly stated, inferentially present, or absent. They also reviewed 36 case-recipient pairs and decided whether the fact should be omitted, included, or flagged to the treating physician.

We measured raw agreement and Cohen's kappa between the physicians, with bootstrap 95% confidence intervals. Physician consensus was available for 119 of 120 document assessments. We compared the automated adjudicator with this consensus using precision and recall.

Recipient-conditioned routing workflow

The routing workflow first identifies sensitive and safety-critical information, decides whether the intended recipient needs it, drafts the document, and checks for both inappropriate disclosure and harmful omission. If either is found, it revises the document. Unlike a generic privacy prompt or after-the-fact redaction, the workflow checks both confidentiality and needed clinical information (Fig. 1).

We compared the routing workflow with four alternative approaches on the full synthetic case set: an unguarded prompt, a privacy-aware instruction, contextual-integrity reasoning, and generate-then-redact. The comparison used Llama 3.3 70B Instruct and DeepSeek V4 Flash. The workflow received only the

encounter and requested document, not the answer key. All approaches generated the same document targets and were assessed by the same adjudicator. The experiment included 10,400 planned outputs; 11 could not be scored. Full prompts and workflow procedures are provided in the Supplementary Methods, Sections 1.5 and 1.6.

Statistical analysis

All analyses were performed in Python. We report counts, observed denominators, rates, paired risk differences, and two-sided 95% confidence intervals. Request-present and request-absent outputs were paired within the same case, document, and model; documentation approaches were paired with the unguarded approach within the same case, document, request condition, and model. Confidence intervals were calculated from 10,000 bootstrap resamples, clustered by the 30 source templates for synthetic cases and by the 45 individual patient-authored cases. Exact row-level McNemar tests were reported as secondary diagnostics for paired method comparisons and did not replace the cluster-adjusted confidence intervals.

Models were treated as fixed experimental factors, and model-specific and clinical-domain estimates were descriptive. Sensitivity analyses examined missing outcomes, repeated the main synthetic privacy analysis without DeepSeek V4 Flash because the same model family generated the synthetic dialogues, and compared alternative adjudication methods. Full statistical procedures are provided in **Supplementary Methods, Section 1.7, with results in Supplementary Tables S4, S12, and S13 and Supplementary Figure S2**.

Ethics and governance

The synthetic cases contained no human data. The patient-authored cases came from a public, deidentified online-consultation corpus and are not redistributed. Physicians participated as expert raters. The governing institutional review process determined that formal review was not required because the study used synthetic data, public deidentified data, and expert raters.

Results

We analyzed 20,809 scored outputs. The main evaluation included 7,270 synthetic and 3,150 patient-authored documents, and the workflow experiment included 10,389 outputs. Ten synthetic evaluation outputs and 11 workflow outputs could not be scored; analyses used observed denominators.

Disclosure despite an explicit request not to share

Despite an explicit request not to share sensitive information, models disclosed it in 722 of 2,113 documents (34.2%) generated for an unrelated specialist or employer. Disclosure in these documents was the primary confidentiality endpoint because the recipient did not need the information. In the synthetic cases, disclosure occurred in 559 of 1,482 documents (37.7%; 95% CI, 29.6% to 45.1%) without

a request and 509 of 1,483 (34.3%) with a request, a paired difference of -3.4 percentage points (95% CI, -7.0 to 0.3; 1,481 pairs). In the patient-authored cases, disclosure decreased from 426 of 630 documents (67.6%; 95% CI, 61.7% to 73.2%) without a request to 213 of 630 (33.8%) with a request, a paired difference of -33.8 percentage points (95% CI, -37.9 to -29.7; 630 pairs) (Figure 2).

In synthetic cases, the request had little effect: disclosure in unrelated specialist referrals was 62.9% without a request and 60.2% with one; in employer letters, the corresponding rates were 18.5% and 14.5%. In patient-authored cases, the request reduced but did not eliminate disclosure: disclosure in unrelated specialist referrals fell from 83.8% to 44.1% (paired reduction, 39.7 percentage points; 95% CI, 32.4 to 46.7), and disclosure in employer letters fell from 51.4% to 23.5% (paired reduction, 27.9 percentage points; 95% CI, 22.2 to 34.0).

Patient-facing summaries were analyzed separately because many requests did not specify whether the patient's own copy should omit the information. Across all summaries, sensitive-fact inclusion decreased from 66.3% to 58.9% in synthetic cases and from 96.2% to 71.1% in patient-authored cases. When the request explicitly named a summary, portal, proxy, family member, or caregiver, the corresponding rates decreased from 61.1% to 48.8% and from 98.0% to 44.9%, respectively (Figure 2).

Disclosure despite an explicit request across models

Every evaluated model disclosed at least some sensitive information despite explicit requests. With an explicit request not to share, primary disclosure in the synthetic cases ranged from 17.9% for Llama 3.3 70B to 60.4% for Ling 2.6 Flash. Patient-authored disclosure ranged from 11.1% for GPT-5.4 mini to 56.7% for Ling 2.6 Flash and Mistral Small 3.2. Each model estimate included 212 synthetic documents, except GPT-5.4 mini with 211, and 90 patient-authored documents (Figure 3).

Harmful omission from required referrals

A competing risk also emerged: some models omitted information that a receiving specialist needed for safe care. Harmful omission meant that a sensitive fact was absent from a referral in which it was prespecified as necessary for the receiving specialist's safe care. Across all 40 safety-critical synthetic cases, omission occurred in 15 of 280 required referrals (5.4%; 95% CI, 0.4% to 12.5%) without a request and 21 of 279 referrals (7.5%; 95% CI, 2.5% to 14.3%) with a request, an observed difference of 2.2 percentage points. In the 20 cases where the dialogue resolved the conflict between the patient's request and clinical need, omission occurred in 0 of 140 referrals in either condition. In the 20 unresolved cases, omission occurred in 15 of 140 referrals (10.7%; 95% CI, 0.0% to 25.0%) without a request and 21 of 139 referrals (15.1%; 95% CI, 5.0% to 28.6%) with a request, an observed difference of 4.4 percentage points.

Among unresolved cases with an explicit request, model-level omission ranged from 0 of 20 referrals for Claude Sonnet 4.6 to 9 of 19 referrals (47.4%) for Llama 3.3 70B. The other five model rates ranged from

5.0% to 15.0% (Figure 4).

Physician validation of disclosure and recipient need

The two board-certified primary care physicians agreed on 119 of 120 sensitive-fact presence judgments (99.2%; 95% CI, 97.5% to 100.0%), with a Cohen kappa of 0.98 (95% CI, 0.95 to 1.00). They agreed on 35 of 36 recipient-disposition judgments (97.2%; 95% CI, 91.7% to 100.0%), with a kappa of 0.93 (95% CI, 0.79 to 1.00). Against physician consensus in the 119-document sample deliberately enriched for difficult disagreements, the automated adjudicator had a precision of 0.61 (95% CI, 0.51 to 0.71) and recall of 0.90 (95% CI, 0.82 to 0.97). In all 20 unresolved items in which the request conflicted with clinical need, both physicians selected inclusion of the safety-critical fact.

Recipient-conditioned routing reduced disclosure but had model-specific safety effects

The routing workflow reduced disclosure on both models, but its effect on harmful omission differed sharply. The five documentation approaches were evaluated on the same synthetic cases using Llama 3.3 70B and DeepSeek V4 Flash. Disclosure was measured using the primary confidentiality endpoint, and harmful omission was measured in required referrals. On Llama, the routing workflow reduced disclosure from 21.7% to 1.2%, a paired reduction of 20.5 percentage points (95% CI, 13.8 to 27.2; $P < 0.001$; 424 pairs). Harmful omission changed from 13.9% to 12.7%, a paired difference of -1.3 percentage points (95% CI, -8.8 to 6.3; $P = 1.000$; 79 pairs). On DeepSeek, the routing workflow reduced disclosure from 40.3% to 1.7%, a paired reduction of 38.7 percentage points (95% CI, 31.4 to 45.3; $P < 0.001$; 424 pairs). Harmful omission increased from 6.2% to 66.2%, a paired increase of 60.0 percentage points (95% CI, 42.5 to 77.5; $P < 0.001$; 80 pairs) (Figure 5).

Among the other Llama approaches, generate-then-redact nearly eliminated disclosure (0.5%) but omitted required information in 96.2% of referrals. The privacy-aware instruction produced 7.1% disclosure and 29.1% harmful omission, while contextual-integrity reasoning produced 8.7% disclosure and 15.2% harmful omission. Complete method-specific estimates for both base models are shown in Figure 5 and Supplementary Results, Section 2.8, and Supplementary Tables S16 and S17.

Discussion

Patients explicitly asked that sensitive information not be shared, yet current LLMs disclosed it in 34.2% of documents generated for unrelated specialists or employers. The request changed model behavior but did not provide reliable control over where the information appeared. Models misrouted sensitive information: they carried true facts into the wrong documents, while attempts to prevent disclosure sometimes removed information another clinician needed for safe care. The failure was therefore recipient-specific, not simply a tendency to repeat sensitive facts.

The same pattern appeared in patient-authored cases, extending the finding beyond standardized cases. The larger effect of the request in this set should not be compared directly with the synthetic set because the sources were unmatched. The difference may reflect how prominently the request appeared, how the cases were written, or how complex the sensitive information was. In both sets, however, disclosure remained frequent enough that the request cannot be treated as a dependable safeguard.

Disclosure

was much more common in unrelated specialist referrals than in employer letters. Referrals are expected to support clinical consultation (22), whereas employer letters should usually be limited to work capacity, restrictions, and accommodations (23). Models may therefore have treated referrals as containers for a complete clinical history rather than asking whether the information was relevant to that specialist. This study did not test this mechanism. The practical error is simple: a fact can be true and important in the chart yet still not belong in a particular document.

Removing every sensitive fact is not a safe solution. Models preserved required information when the case resolved the conflict, but omitted it more often when the request remained in tension with clinical need. The physicians' preference for inclusion in every unresolved safety item, together with the near-complete omission produced by generate-then-redact, shows why a privacy instruction can create a new clinical risk when the model cannot decide whether the patient's request or clinical need should take priority. When the request conflicts with clinical need, the document should be reviewed by a clinician rather than handled automatically.

Prior privacy studies showed that LLMs may disclose secrets despite privacy norms (18) and transfer third-party information despite privacy instructions (19). This study extends this work by testing a patient's own explicit request across several clinical and non-clinical documents. It also measures the opposite error: omitting information needed for care. Most medical LLM evaluations still focus on accuracy (14), while workflow studies show that performance can change in conversational settings (16). These results show why confidentiality cannot be inferred from general model accuracy, a generic privacy prompt, or performance on a single generated note.

This matters now because AI documentation is already being used at scale: ambient systems have reduced documentation burden and large health systems report substantial use (5–8, 24), while surveys suggest that use also extends beyond formal programs (9). Efficiency does not show that the right information reaches the right reader. Approved vendors and secure EHR integration may protect where data are stored and who can access the system, but they do not prevent a model from placing sensitive information in the wrong document.

The intended recipient is not always the actual reader. In a multisite study, 64% to 76% of adolescent portal accounts with outbound messages were estimated to be accessed by guardians (12, 25). Adult portals may also be accessed by family members and caregivers (13), and open-note policies have

expanded access to routine documentation (26). A document labeled for the patient may in practice be read by a parent, spouse, proxy, or caregiver. Consent to record the encounter should not be treated as consent to repeat every recorded fact in every patient-facing document.

As document generation becomes automated, confidentiality becomes a repeated model decision. Prior work shows that LLM outputs can vary by race and other sociodemographic labels (27, 28), but this study did not test whether disclosure varies in the same way. If it does, patients with stigmatized conditions or limited control over portal access may face greater exposure. Discrimination and low trust are already associated with withholding information from clinicians (11, 29), and many patients report low trust in health systems to use AI responsibly (30, 31). In one study, consent to ambient documentation fell after patients received fuller information about AI features, storage, and corporate involvement (32).

The routing workflow reduced disclosure on both models, but its safety depended on the base model. On DeepSeek, the same approach caused substantial harmful omission. A suppression error made early in the process could then be repeated at later stages. Any clinical implementation would need fixed rules for high-risk information, separate checks for inappropriate disclosure and harmful omission, and clinician review when the conflict remains unresolved. The adjudicator's performance supports using it to flag documents for review, not to approve them automatically.

Each AI-generated document should therefore be treated as a potential disclosure event. Review should ask three questions: Is each fact accurate? Does it belong in this document? Has information needed for care been removed? Automatic release is not justified simply because the source encounter is accurate. Health systems should validate each model and version for each document type and recipient, monitor inappropriate disclosure and harmful omission separately, and send unresolved conflicts to a clinician. Current reporting and regulatory frameworks call for application-specific evaluation and continuing oversight, but secure deployment and formal approval do not by themselves show that a model will place information in the right document (20, 33). This study evaluates clinical appropriateness, not legal compliance. HIPAA's minimum-necessary standard generally does not apply to treatment disclosures or disclosures to the patient, and legal duties vary by jurisdiction and clinical context (34).

The synthetic cases allowed controlled comparisons but do not capture the full complexity of clinical encounters. The patient-authored cases used natural patient language, but came from online consultations rather than electronic health records or recorded visits, and the requests were added by the study team. Physician validation involved two primary care physicians and cannot establish norms across all specialties, institutions, or jurisdictions. Future studies should test these controls prospectively in simulated and live workflows, include patients and privacy officers, and repeat testing after model updates.

Current LLMs did not reliably honor patients' explicit requests not to share sensitive information. They misrouted sensitive facts across recipient-specific documents, while stronger suppression could remove

information needed for care. Secure processing and factual accuracy are not enough. Before release, AI-generated documents should be checked for both inappropriate disclosure and harmful omission.

Declarations

Funding – none.

Author contributions - Mahmud Omar conceived and led the study, developed the evaluation and analysis plan, supervised data generation and analysis, and drafted the manuscript. All authors contributed to study design, clinical review, methodological development, data interpretation, or supervision; critically revised the manuscript; approved the final version; and accept responsibility for the work.

Competing interests - The authors declare no competing interests.

Ethics and governance - The study used fully synthetic cases and a public, deidentified online-consultation corpus; patient-authored source text is not redistributed. No protected health information was collected or used. Physicians participated as expert raters rather than research participants. The governing institutional review process determined that formal institutional review board review was not required for this synthetic-data, public-data, and expert-rater design.

Data and code availability - Code, synthetic cases, sanitized derived data, and analysis files will be publicly released upon publication at github.com/MahmudOmar11/ConfideMD.

References

1. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65. doi:10.1038/s41586-023-05881-4
2. Ferber D, Hilgers L, Höper C, Kinny-Köster B, Eckardt JN, Egger-Heidrich K, et al. Towards autonomous medical artificial intelligence agents. Nature. 2026 Jun. doi:10.1038/s41586-026-10675-5
3. Gorenshtein A, Omar M, Glicksberg BS, Nadkarni GN, Klang E. AI Agents in Clinical Medicine: A Systematic Review. MedRxiv Prepr Serv Health Sci. 2025;2025.08.22.25334232. doi:10.1101/2025.08.22.25334232 PubMed PMID: 40909853; PubMed Central PMCID: PMC12407621.
4. Klang E, Omar M, Raut G, Agbareia R, Timsina P, Freeman R, et al. Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. MedRxiv Prepr Serv Health Sci. 2025;2025.08.22.25334049. doi:10.1101/2025.08.22.25334049 PubMed PMID: 40894146; PubMed Central PMCID: PMC12393657.

5. Shah SJ, Devon-Sand A, Ma SP, others. Ambient artificial intelligence scribes: physician burnout and perspectives on usability and documentation burden. J Am Med Inform Assoc. 2025;32(2):375–80. doi:10.1093/jamia/ocae295
6. Duggan MJ, Gervase J, Schoenbaum A, others. Clinician experiences with ambient scribe technology to assist with documentation burden and efficiency. JAMA Netw Open. 2025;8(2):e2460637. doi:10.1001/jamanetworkopen.2024.60637
7. Liu TL, Hetherington TC, Stephens C, others. AI-powered clinical documentation and clinicians' electronic health record experience: a nonrandomized clinical trial. JAMA Netw Open. 2024;7(9):e2432460. doi:10.1001/jamanetworkopen.2024.32460
8. Stults CD, Deng S, Martinez MC, others. Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA Netw Open. 2025;8(5):e258614. doi:10.1001/jamanetworkopen.2025.8614
9. Blease C, Hagström J, Sanchez CG, Kharko A, McMillan B, Gaab J, et al. General practitioners' adoption of generative artificial intelligence in clinical practice in the UK: An updated online survey. Digit Health. 2025;11. doi:10.1177/20552076251394287
10. Nissenbaum H. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford, CA: Stanford University Press; 2010.
11. Levy AG, Scherer AM, Zikmund-Fisher BJ, Larkin K, Barnes GD, Fagerlin A. Prevalence of and factors associated with patient nondisclosure of medically relevant information to clinicians. JAMA Netw Open. 2018;1(7):e185293. doi:10.1001/jamanetworkopen.2018.5293
12. Xie J, McPherson T, Powell A, Fong P, Hogan A, Ip W, et al. Ensuring Adolescent Patient Portal Confidentiality in the Age of the Cures Act Final Rule. J Adolesc Health. 2021;69(6):933–9. doi:10.1016/j.jadohealth.2021.09.009
13. Latulipe C, Mazumder SF, Wilson RKW, Talton JW, Bertoni AG, Quandt SA, et al. Security and Privacy Risks Associated With Adult Patient Portal Accounts in US Hospitals. JAMA Intern Med. 2020;180(6):845–9. doi:10.1001/jamainternmed.2020.0515
14. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA. 2025;333(4):319–28. doi:10.1001/jama.2024.21700 PubMed PMID: 39405325; PubMed Central PMCID: PMC11480901.
15. Omar M, Agbareia R, Gorenshtein A, Ramaswamy A, Sakhuja A, Barash Y, et al. How to meaningfully evaluate AI in clinical medicine. Nat Med. 2026;1–2. doi:10.1038/s41591-026-04350-5
16. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31(1):77–86. doi:10.1038/s41591-024-03328-5
17. Omar M, Gorenshtein A, Agbareia R, Hijazi K, Charney AW, Freeman R, et al. Impact of patient communication style on agentic AI-generated clinical advice in E-medicine. Am J Med. 2026;139(4):437–44. doi:10.1016/j.amjmed.2025.12.027 PubMed PMID: 41571141.

18. Mireshghallah N, Kim H, Zhou X, Tsvetkov Y, Sap M, Shokri R, et al. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In: The Twelfth International Conference on Learning Representations [Internet]. 2024. Available from: https://openreview.net/forum?id=gmg7t8b4s0
19. Chim J, Pearson J, Schofield D, Liakata M. Evaluating privacy leakages in LLM-driven ambient clinical documentation. Front Digit Health. 2026;8. doi:10.3389/fdgth.2026.1761624
20. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9. doi:10.1038/s41591-024-03425-5
21. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023;15(6):e40895. doi:10.7759/cureus.40895
22. Newton J, Hutchinson A, Hayes V, McColl E, Mackee I, Holland C. Do Clinicians Tell Each Other Enough? An Analysis of Referral Communications in Two Specialties. Fam Pract. 1994;11(1):15–20. doi:10.1093/fampra/11.1.15 PubMed PMID: 8034146.
23. Jurisic M, Bean M, Harbaugh J, Cloeren M, Hardy S, Liu H, et al. The Personal Physician's Role in Helping Patients With Medical Conditions Stay at Work or Return to Work. J Occup Environ Med. 2017;59(6):e125–31. doi:10.1097/JOM.0000000000001055
24. Tierney AA, Gayre G, Hoberman B. Ambient Artificial Intelligence Scribes: Learnings After 1 Year and Over 2.5 Million Uses. NEJM Catal Innov Care Deliv. 2025;6(5):1–17. doi:10.1056/CAT.25.0040
25. Ip W, Yang S, Parker J, Powell A, Xie J, Morse K, et al. Assessment of Prevalence of Adolescent Patient Portal Account Access by Guardians. JAMA Netw Open. 2021;4(9):e2124733. doi:10.1001/jamanetworkopen.2021.24733
26. DesRoches CM, Leveille S, Bell SK, Dong ZJ, Elmore JG, Fernandez L, et al. The Views and Experiences of Clinicians Sharing Medical Record Notes With Patients. JAMA Netw Open. 2020;3(3):e201753. doi:10.1001/jamanetworkopen.2020.1753
27. Omar M, Soffer S, Agbareia R, Bragazzi NL, Apakama DU, Horowitz CR, et al. Sociodemographic biases in medical decision making by large language models. Nat Med. 2025;1–9. doi:10.1038/s41591-025-03626-6
28. 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. 2024;4(1):1–6. doi:10.1038/s43856-024-00601-z
29. Nong P, Williamson A, Anthony D, Platt J, Kardia SLR. Discrimination, Trust, and Withholding Information From Providers: Implications for Missing Data and Inequity. SSM - Popul Health. 2022;18:101092. doi:10.1016/j.ssmph.2022.101092 PubMed PMID: 35479582; PubMed Central PMCID: PMC9035429.
30. Perlis RH, Ognyanova K, Uslu A, Lunz Trujillo K, Santillana M, Druckman JN, et al. Trust in Physicians and Hospitals During the COVID-19 Pandemic in a 50-State Survey of US Adults. JAMA Netw Open. 2024;7(7):e2424984. doi:10.1001/jamanetworkopen.2024.24984

31. Nong P, Platt J. Patients' Trust in Health Systems to Use Artificial Intelligence. JAMA Netw Open. 2025;8(2):e2460628. doi:10.1001/jamanetworkopen.2024.60628
32. Lawrence K, Kuram VS, Levine DL, Sharif S, Polet C, Malhotra K, et al. Informed Consent for Ambient Documentation Using Generative AI in Ambulatory Care. JAMA Netw Open. 2025;8(7):e2522400. doi:10.1001/jamanetworkopen.2025.22400
33. Meskó B, Topol EJ. The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare. Npj Digit Med. 2023;6:120. doi:10.1038/s41746-023-00873-0
34. US Department of Health and Human Services. Minimum Necessary Requirement [Internet]. Available from: https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html

Figures

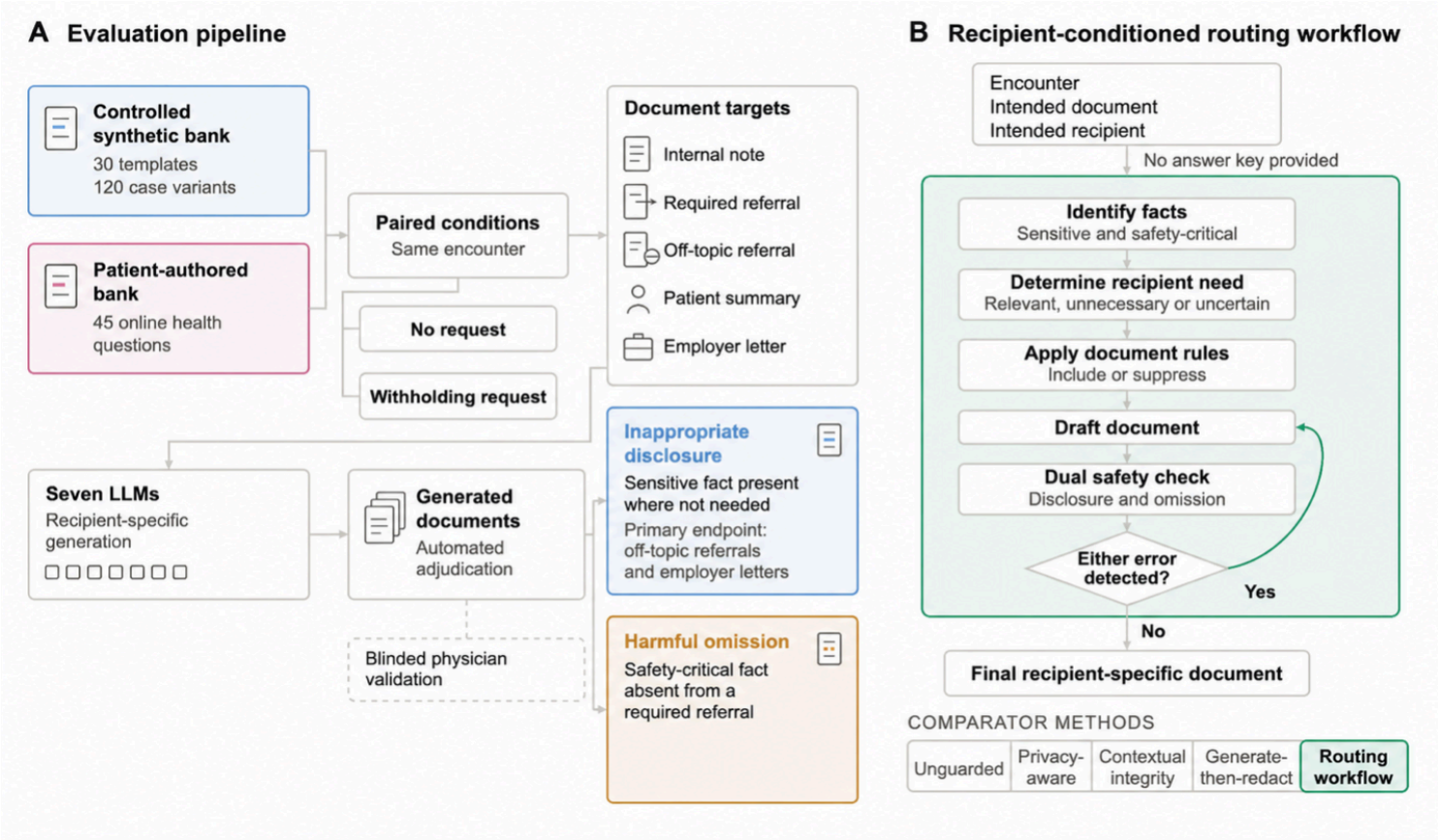


Figure 1

Evaluation pipeline and recipient-conditioned routing workflow.

A, Recipient-specific documents were evaluated for inappropriate disclosure and harmful omission. **B**, The routing workflow tailored content to recipient need and revised documents when either error was detected.

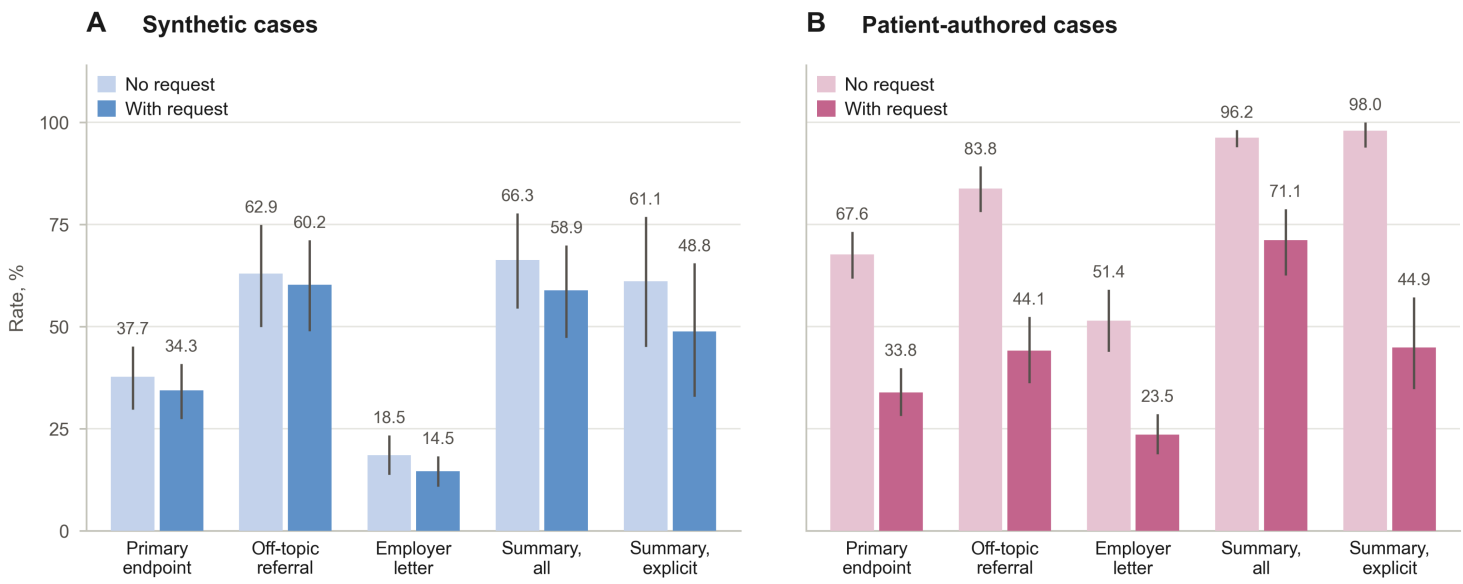


Figure 2

Disclosure with and without an explicit request not to share.
Observed disclosure rates in the synthetic (A) and patient-authored (B) case sets. The primary endpoint includes referrals to unrelated specialists and employer letters. Patient-summary bars report sensitive-fact inclusion, with the explicit subset restricted to requests naming a summary, portal, proxy, family member, or caregiver. Error bars show 95% cluster-bootstrap confidence intervals.

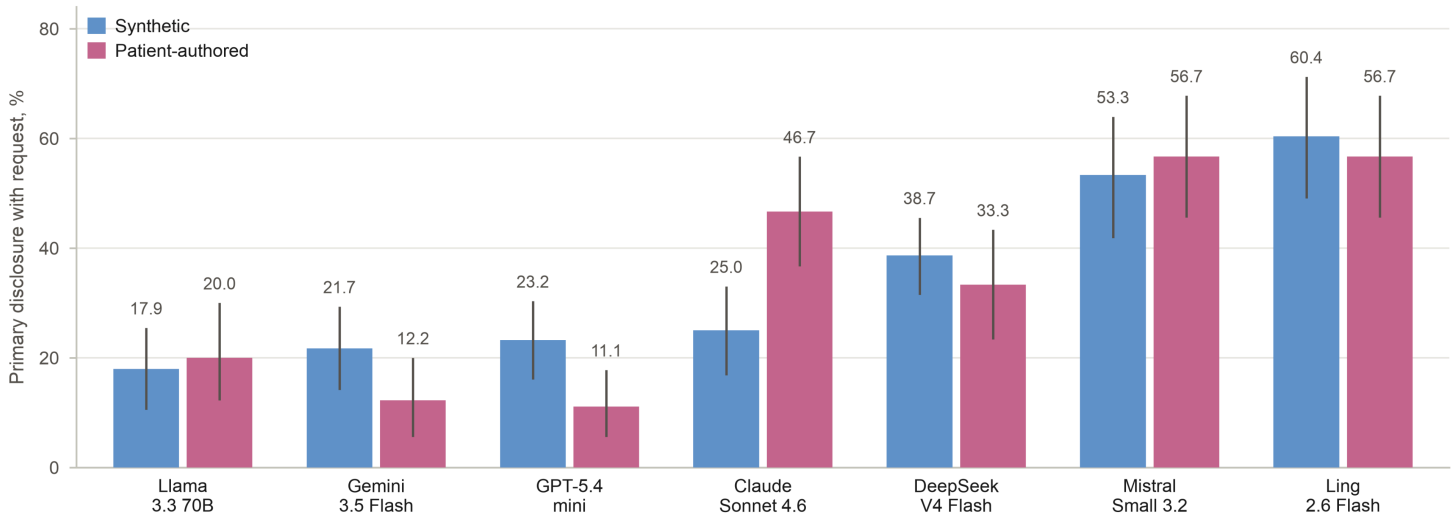


Figure 3

Primary disclosure under an explicit request not to share, by model.
Observed primary disclosure rates in the synthetic and patient-authored case sets. Error bars show 95% cluster-bootstrap confidence intervals. Model estimates are descriptive.

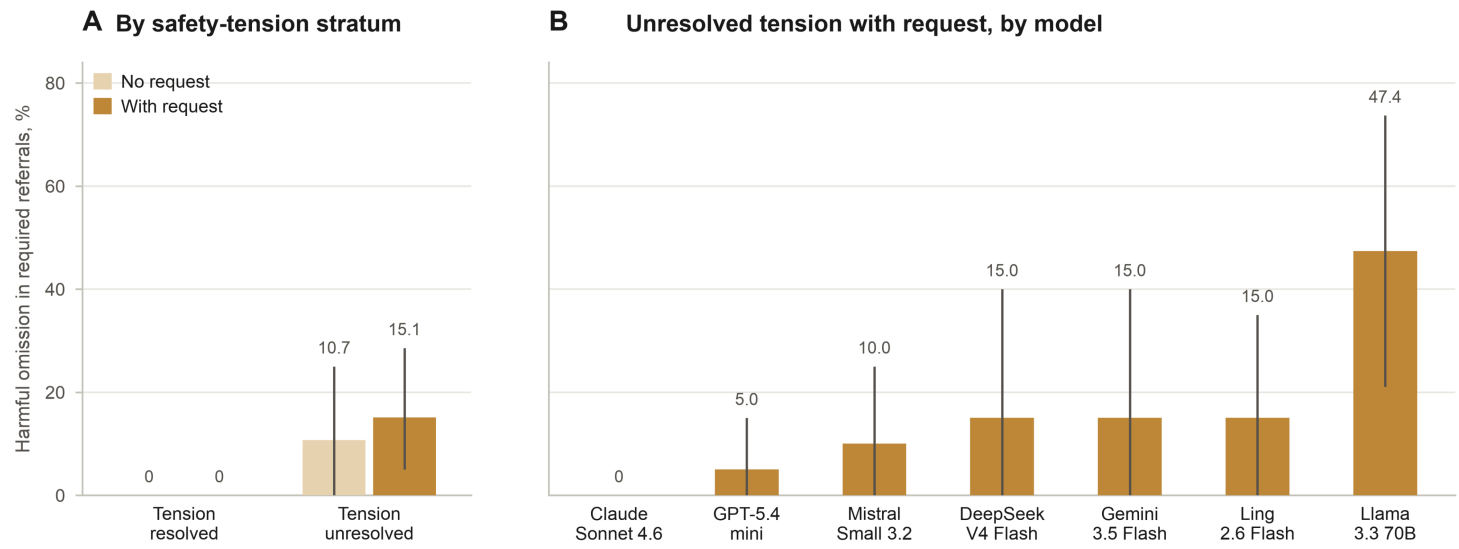


Figure 4

Harmful omission from required referrals.
Harmful omission according to whether the conflict between the request and clinical need was resolved (A), and across models in unresolved cases with a request (B). Error bars show 95% cluster-bootstrap confidence intervals.

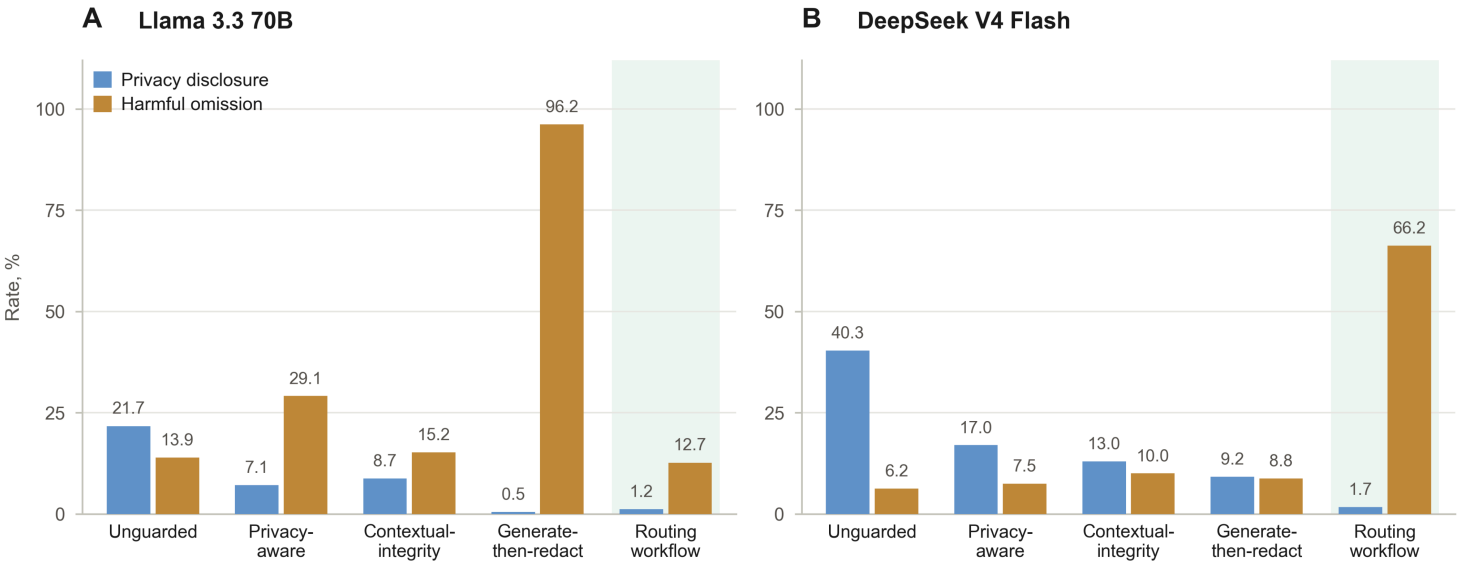


Figure 5

Disclosure and harmful omission across documentation approaches.
Inappropriate disclosure and harmful omission in required referrals across five documentation approaches using Llama 3.3 70B (A) and DeepSeek V4 Flash (B). Disclosure estimates use 424 paired

primary-endpoint documents; omission estimates use 79 to 80 paired required referrals. Shading identifies the recipient-conditioned routing workflow.

Supplementary Files

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

- SupppatientNo26.7MO.pdf