Explicit Prompting Unlocks Evidence Transfer in LLM-Generated Medical Risk Summaries

Ashik Ahamed[1 [0009−0009−6006−5686]⋆], Garrett Ash[2], and Jeanna Matthews[1 [0000−0001−5955−0996]]

[1] Clarkson University, Potsdam, NY, USA
{ahameda, jnm}@clarkson.edu
[2] Yale School of Medicine, New Haven, CT, USA
garrett.ash@yale.edu

Abstract. Doctors are using LLMs like ChatGPT to refine their communications to patients and to summarize the risks of medical procedures. This study examines the effectiveness of different prompting methods for producing LLM-generated risk summaries. Specifically, this study evaluates how ChatGPT uses different types of evidence in risk communication across different medical procedures (e.g., chemotherapy or X-ray) under four conditions with a standardized prompt. Four conditions are considered: (1) Patient-facing article only (2) Scientific Paper only (3) Combined; Passive Upload: both documents are uploaded but a standard prompt is issued without instructing GPT to use the scientific paper and (4) Combined; Explicit Prompt: both sources are named and use of study details is explicitly permitted; GPT then pulls in study-specific facts (e.g., percentages) and adds side effects that appear in the scientific article but not in the patient-facing article. Findings were replicated using GPT-5 under identical protocols, yielding consistent results across both model versions.

Keywords: Risk Communication · Prompt Engineering · Large Language Models · Evidence Transfer · Patient Safety · ChatGPT

1 Introduction

Large language models (LLMs) such as ChatGPT are increasingly being used by clinicians to generate patient-facing risk summaries for medical procedures and treatments. While these tools can rapidly translate complex medical information into accessible language, there is limited empirical guidance on how to effectively prompt LLMs to incorporate scientific evidence into patient communications. A 2025 American Medical Association survey found that 66% of physicians reported using health AI in 2024, a 78% increase from 2023, with documentation, patient communication, and assistive diagnosis among the most common uses [7]. However, research on LLM performance in clinical settings shows mixed results. While GPT-4 has demonstrated superior diagnostic accuracy

⋆ Corresponding author

compared to emergency department physicians [8], other studies show serious failures; ChatGPT Health under-triaged 51.6% of medical emergencies [9], and participants using LLMs for medical decisions performed no better than those using Google [10]. These inconsistencies underscore the need to better understand how clinicians should effectively use these tools.

The challenge is particularly acute when clinicians have multiple information sources: a patient education article (e.g., from Harvard Health) and a recent peer-reviewed study with updated risk data. It remains unclear whether LLMs automatically integrate information from multiple uploaded documents or whether specific prompting strategies are required to achieve evidence transfer.

This gap has direct clinical implications. If LLMs fail to incorporate critical risk information from research papers when rewriting patient materials, clinicians may inadvertently provide incomplete summaries. For example, a recent chemotherapy study might report specific side effect rates (memory impairment in 14% of patients, dry mouth in 74%) that do not appear in standard patient handouts. Without proper prompting, these evidence-based details may be omitted from LLM-generated summaries, undermining informed consent. Prior work suggests that even small differences in prompt wording can dramatically change LLM outputs in medical contexts [10], yet no prior study has systematically examined evidence transfer from scientific papers to patient-facing summaries in a multi-document clinical setting—the specific gap this study addresses.

This study primarily evaluates ChatGPT-4, as it represents the most widely adopted LLM in clinical practice, with replication conducted using GPT-5 to assess whether findings generalize across model versions within the same model family.

This study addresses two key questions:

1. Does ChatGPT automatically integrate information from multiple uploaded documents when generating patient risk summaries?
2. Does explicit prompt design, specifically naming sources and granting permission to use scientific details, improve evidence incorporation?

2 Methods

This study used a structured experimental design with two manipulated dimensions: the documents made available to the model (patient-facing article only, scientific paper only, or both) and the specificity of the prompt (a standard rewrite instruction versus an explicit instruction naming both sources and granting permission to use the scientific paper). These dimensions were not fully crossed, as an explicit source-naming prompt is only meaningful when more than one document is present; four conditions were therefore evaluated rather than the full six-cell crossing. Conditions 1 and 2 held the prompt constant and varied

the source, establishing single-document baselines. The central manipulation was the contrast between Conditions 3A and 3B, which held the uploaded documents constant (both the patient-facing article and the scientific paper) and varied only prompt specificity; this contrast isolates the effect of explicit prompting on evidence transfer while controlling for document availability. The dependent variable was evidence transfer rate, defined as the proportion of pre-specified transferable elements from the scientific paper appearing in the model's output, scored within three categories: quantitative data, additional side effects, and research findings/recommendations. Each condition was administered across six medical procedures spanning a range of risk levels and clinical contexts and replicated under two model versions (ChatGPT-4 and GPT-5); procedure and model version functioned as replication factors used to assess the generalizability of the condition effects rather than as primary independent variables. To prevent context contamination between experimental conditions, experiments were organized using three separate chat channels within ChatGPT: one for Condition 1, one for Condition 2, and one shared channel for Conditions 3A and 3B, which used identical document uploads and differed only in prompt wording. This ensured that ChatGPT's responses in one condition were not influenced by previous interactions in other conditions.

ChatGPT-4 was selected as the primary model for evaluation for three reasons: (1) it is the most frequently reported LLM in clinical and healthcare communication literature, (2) its consistent interface allowed for controlled session management across conditions, and (3) focusing on a single model enabled direct comparison across conditions without confounding variability introduced by architectural differences between models.

Table 1 provides an overview of the experimental design across all four conditions.

Table 1. Overview of experimental conditions. All conditions replicated using GPT-5 under identical protocols, yielding consistent results.

Condition	**Sources Uploaded**	**Prompt Type**	**Outcome**
1: Patient-facing article	Article only	Standard	Baseline rewriting
2: Scientific paper	Paper only	Standard	0% quantitative transfer
3A: Passive upload	Both documents	Standard	0% evidence transfer
3B: Explicit prompt	Both documents	Named sources	94% evidence transfer

2.1 Selection of Medical Procedures

Six procedures were selected based on three criteria: (1) availability of both human-written patient education materials and peer-reviewed research papers, (2) diversity in risk profiles ranging from minimal to high risk, and (3) representation of different clinical contexts.

Table 2. Medical procedures.

Procedure	Risk Level	Notes
Chemotherapy	High	Extensive, well documented side effects.
X-Ray	Low	Patient education materials often lack specific quantitative comparisons that appear in scientific literature.
CT Scan	Moderate	Age-specific risk data and cumulative exposure considerations often absent from patient materials.
Ultrasound	Minimal	Very low risk procedure with minimal quantitative data available in scientific literature.
Mammography	Moderate	Literature presents evidence of both risks and benefits.
Liver biopsy	High	Acute procedural risk, not long-term cumulative risks.

2.2 Source Material Acquisition and Characterization

Patient Education Articles For each procedure, we identified patient education articles from Harvard Health Publishing. This source was selected because: (1) it is produced by a reputable medical institution, (2) it undergoes editorial review by licensed physicians, (3) articles are explicitly designed for average patients, and (4) content is freely accessible without subscription barriers. These articles were classified as 'Patient-Facing Articles' in our study.

For each procedure, we documented the specific URL, retrieval date, and stated author/reviewer to ensure transparency and reproducibility. Articles ranged from 400-800 words and followed similar structures: procedure definition, preparation instructions, procedural steps, common risks, special precautions, reviewer credentials, and ethical disclaimers.

Scientific Research Papers Peer-reviewed scientific papers were selected for each procedure using the following criteria: (1) publication in indexed journals, (2) inclusion of quantitative risk data not readily available in patient education materials, (3) recency (published within the last 10 years when possible), and (4) relevance to the average-risk patient rather than specialized populations.

For chemotherapy, we selected "Side Effects of Chemotherapy in Cancer Patients and Evaluation of Patients' Opinion about Starvation-Based Differential Chemotherapy" because it provided specific percentages for common side effects (fatigue 90%, weakness 95%, nausea 77%, vomiting 75%, hair loss 76%, dry mouth 74%, memory impairment 14%) that are typically described only qualitatively in patient materials.

Similar selection processes were followed for the remaining procedures, prioritizing papers that would provide clear opportunities to observe evidence transfer: specific statistics, additional side effects, or research findings not present in patient materials.

2.3 Experimental Procedure

All experiments were conducted using ChatGPT-4 accessed through the OpenAI web interface (chat.openai.com) during December 2024. The

specific model version identifier displayed in the interface was documented to ensure reproducibility as models are updated over time. Strict session control was employed to prevent context contamination. Each experimental condition was tested in a fresh chat session with no prior conversation history. After completing each test, we closed the chat window entirely before beginning the next test. This prevented any carryover effects where the model might remember previous interactions or uploaded documents. The order of testing was also documented to identify any potential sequence effects, though the use of independent sessions was designed to eliminate such effects.

To assess output consistency and cross-version generalizability, all four conditions were replicated using GPT-5 accessed through the same OpenAI web interface. The same standardized prompts, source documents, and session management protocols were applied identically. Results from both model versions were compared to evaluate whether findings reflected stable LLM behavior or were specific to a single model version.

Condition 1: Patient-Facing Article (Baseline Rewriting): In this condition, only the patient education article was uploaded as a PDF file to ChatGPT. After the upload completed and ChatGPT confirmed it had processed the document, the following standardized prompt was issued:

"Imagine you are a doctor explaining the risks of [PROCEDURE NAME] to an average patient. Write the explanation so it's easy to understand and no longer than [X] words or 5 minutes of reading."
The word limit X was calibrated for each procedure to correspond to approximately 5 minutes of reading time, typically ranging from 200-600 words depending on the original article length.

This condition served three purposes: (1) to establish baseline performance when ChatGPT rewrites existing patient education materials, (2) to observe what types of modifications ChatGPT makes to human-written text, and (3) to document ChatGPT's content curation decisions.

Condition 2: Scientific Paper (Direct Translation): In this condition, only the scientific research paper was uploaded as a PDF file and the same standardized prompt was used, substituting the appropriate procedure name. This condition tested ChatGPT's ability to translate complex scientific literature into patient-friendly language while preserving critical risk information.

We were particularly interested in observing whether ChatGPT would retain quantitative data from the paper, how technical terminology would be simplified, whether the output would maintain an appropriate patient-doctor tone, and what scientific details ChatGPT would deem important enough to include versus exclude.

Condition 3A: Combined Sources - Passive Upload: This condition tested whether ChatGPT would automatically integrate information from scientific papers when rewriting human-written patient education articles, even without explicit instruction to do so. Both documents were uploaded to ChatGPT in sequence. First, the scientific paper was uploaded without issuing any prompt, allowing ChatGPT to process it silently. Then, the human-written patient education article was uploaded and the standardized prompt was issued:

"Imagine you are a doctor explaining the risks of [PROCEDURE NAME] to an average patient. Write the explanation so it's easy to understand and no longer than [X] words or 5 minutes of reading."

Notably, this prompt does not instruct ChatGPT to use the scientific paper when generating the summary. This condition tested whether ChatGPT's default behavior, when presented with multiple uploaded documents, is to automatically integrate information from all available sources.

2.4 Condition 3B: Combined Sources - Explicit Prompting:

Based on initial observations from Condition 3A, a refined prompt was developed that explicitly named both sources and granted clear permission to use information from the scientific paper. The same upload procedure as Condition 3A was followed but the prompt was modified as follows:

"Imagine you are a doctor explaining the risks of [PROCEDURE NAME] to an average patient. Rewrite the content of the uploaded article '[PATIENT EDUCATION ARTICLE FILENAME].pdf' in simple, easy-to-understand language, keeping it concise (no longer than [X] words or 5 minutes of reading). You may also include useful details from the uploaded article '[SCIENTIFIC PAPER FILENAME].pdf' if they enhance clarity or add important risk information."

This prompt differs from Condition 3A in three critical ways: First, it explicitly names the patient education article as the primary source to be rewritten. Second, it explicitly names the scientific paper by filename, making ChatGPT aware that both documents are relevant to the task. Third, it grants explicit permission to incorporate information from the scientific paper while specifying the conditions under which such incorporation is appropriate.

Output Collection: For each condition and procedure combination, the complete ChatGPT output was saved as plain text. The exact wording, formatting (such as bullet points or bold text), and structure produced by the model was preserved without any editing or correction. Metadata was also captured including timestamp, session identifier, and model version.

3 Analysis Methods

Our primary analysis focused on quantifying evidence transfer, the degree to which information from the scientific paper appeared in ChatGPT outputs. A systematic coding scheme was developed and applied consistently across all procedures and conditions.
For each output, we identified and categorized the following evidence elements:

Quantitative Data: Any specific numerical values, percentages, or statistical measures derived from the scientific paper. Numerical precision from source materials was systematically tracked to determine whether ChatGPT preserved these values in its outputs. For example, the chemotherapy scientific paper stated "the most frequently reported side effects were weakness (95%), fatigue (90%), nausea (77%), hair loss (76%), and vomiting (75%)." When ChatGPT's output converted this to "Patients may feel very tired or weak, have nausea, vomiting, or lose their hair," removing all five percentages, we coded this as zero quantitative data transfer. To qualify as a successful transfer, outputs must include the specific numerical value (exact or within 1-2 percentage points, e.g., "90% experience fatigue" or "weakness in 95%"). General descriptors without numbers (e.g., "may feel," "commonly," "some people experience") did not qualify as quantitative data transfer.

Additional Side Effects: Risk factors or adverse events mentioned in the scientific paper but absent from the patient education article. For example, if the Harvard Health article listed "nausea, vomiting, hair loss" but the scientific paper also documented "dry mouth in 74% of patients" and "memory impairment in 14% of patients," ChatGPT outputs were examined to determine whether these additional effects were included. Clear mention of the specific side effect was required, though exact wording could vary (e.g., "memory problems," "memory changes," or "memory impairment" all counted as the same side effect).

Research Findings or Recommendations: Any discussion of study conclusions, emerging research, or evidence-based recommendations present in the scientific paper but not in patient education materials. For the chemotherapy example, the scientific paper discussed fasting protocols: "starvation-based differential chemotherapy has been shown to reduce side effects, with 30% of patients agreeing to fast for 12 hours, 28% for 24 hours." If ChatGPT outputs mentioned fasting as a potential side effect mitigation strategy, this was coded as transfer of a research finding.

Technical Term Translation: While not counted in evidence transfer totals, instances where ChatGPT appropriately translated technical terminology from scientific papers into patient-friendly language were documented. For example, replacing "genotoxicity caused by reactive oxygen species" with "medicines that damage cells" demonstrated successful simplification while retaining meaning. Structured coding tables

were created for each procedure listing all potential transferable elements from the scientific paper, then marking which elements appeared in outputs from each condition. This allowed us to calculate evidence transfer rates and identify patterns. To assess coding reliability, a second independent coder applied the evidence transfer coding scheme to the chemotherapy procedure outputs, which served as the primary illustrative example throughout this study. Inter-rater agreement was assessed using Cohen's Kappa ($\kappa = 1.00$), indicating perfect agreement between coders. While full dual-coding across all six procedures was beyond the scope of this study, the chemotherapy procedure contains the largest number of codeable elements (9 items across two conditions) and represents the most detailed example in our analysis. Future work should extend inter-rater reliability assessment across all procedures.

Beyond quantitative coding, detailed side-by-side comparisons of source materials and outputs were performed. For each procedure, we created a comparison document displaying the original patient education article text, relevant excerpts from the scientific paper containing potential transfer elements, and ChatGPT outputs from all four conditions. This visualization allowed us to trace the source of every piece of information in ChatGPT outputs and identify what was added, preserved, modified, or omitted across conditions.

4 Results

Our systematic evaluation of ChatGPT-4 across six medical procedures and four experimental conditions revealed a striking pattern: evidence transfer from scientific papers to patient-facing summaries occurred only when prompts explicitly named source documents and granted permission to integrate information. Passive document upload, simply uploading scientific papers without explicit instruction to use them, resulted in zero evidence transfer across all procedures tested.

Figure 1 illustrates the evidence transfer rate across all four conditions.

4.1 Condition 1: Patient-Facing Article (Baseline Performance)

When ChatGPT was asked to rewrite human-written patient education articles without access to scientific papers, outputs demonstrated consistent modifications across all procedures:

Content Modifications:

- Language simplification (maintained or slightly reduced reading level)
- Structural reorganization (paragraphs converted to bullet points in 4 of 6 procedures)
- Addition of empathetic "doctor voice" phrases (e.g., "Your doctor will monitor you closely," "Most side effects can be managed")
- Selective emphasis on major risks while de-emphasizing minor complications

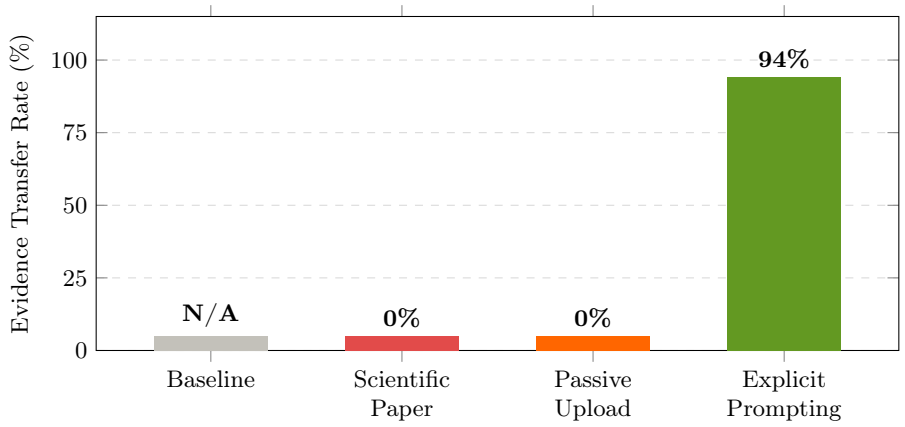

Fig. 1. Evidence transfer rate by prompting condition across all six medical procedures. Explicit prompting achieved 94% overall evidence integration versus 0% for all other conditions.

Example - Chemotherapy (Patient-Facing Article)
Original Harvard Health text (excerpt): "Chemotherapy uses drugs to destroy cancer cells. Side effects may include fatigue, nausea, vomiting, hair loss, and increased infection risk."

ChatGPT rewrite: "Chemotherapy uses strong medicines to kill cancer cells, but it can also affect some healthy cells. You might feel very tired, have nausea or vomiting, lose your hair, or be more prone to infections. Your doctor will monitor you closely and can help manage these side effects."

ChatGPT consistently added contextual reassurance and simplified medical terminology while preserving core risk information from the source article.

4.2 Condition 2: Scientific Paper (Direct Translation from Research Papers)

When provided only scientific research papers, ChatGPT systematically removed quantitative precision while translating content into patient-friendly language.

Quantitative Data Removal Across all six procedures, ChatGPT removed 100% of specific percentages and statistical measures when translating from scientific papers, replacing them with general qualitative descriptors.

Quantitative Transfer Score for Chemotherapy (Scientific Paper Condition): 0 out of 8 potential data points = 0% transfer

4.3 Technical Term Simplification

ChatGPT consistently translated technical terminology into accessible language. For example, the complex phrase "genotoxicity caused by reactive oxygen species" was simplified to "strong medicines that can damage cells," while "hematopoiesis suppression" became "reduced blood cell production." Similarly, the technical term "starvation-based differential chemotherapy" was rewritten as the more patient-friendly "short fasting before treatment." These translations demonstrate ChatGPT's ability to maintain the core meaning of scientific concepts while making them comprehensible to patients without medical training.

4.4 Tone Modifications

ChatGPT added reassuring clinical context not present in research papers, such as "Doctors monitor these reactions closely," "This should only be done under medical advice," and "Most side effects can be managed." While ChatGPT successfully simplified language and maintained patient-appropriate tone, it systematically eliminated the quantitative precision that distinguishes scientific literature from general patient education materials.

4.5 Condition 3A: Combined Sources - Passive Upload

This condition produced the study's most significant finding: zero evidence transfer occurred across all six procedures when scientific papers were passively uploaded without explicit instruction to use them.

Table 3. Evidence Transfer in Passive Upload Condition

Procedure	**Quantitative Data Points Available**	**Data Points Transferred**	**Additional Side Effects Available**	**Side Effects Added**	**Research Findings Available**	**Findings Included**
Chemotherapy	8	0 (0%)	3	0 (0%)	1	0 (0%)
X-ray	4	0 (0%)	2	0 (0%)	1	0 (0%)
CT Scan	5	0 (0%)	2	0 (0%)	1	0 (0%)
Ultrasound	3	0 (0%)	1	0 (0%)	1	0 (0%)
Mammography	6	0 (0%)	3	0 (0%)	2	0 (0%)
Liver Biopsy	4	0 (0%)	2	0 (0%)	1	0 (0%)
TOTAL	**30**	**0 (0%)**	**13**	**0 (0%)**	**7**	**0 (0%)**

The output closely matched the structure and content of the Harvard Health article, with no detectable incorporation of information unique to the scientific paper. ChatGPT behaved as if only the patient education article existed, despite having processed the scientific paper moments earlier.

4.6 Condition 3B: Combined Sources - Explicit Prompting

When prompts explicitly named both source documents and granted permission to incorporate scientific information, ChatGPT systematically integrated evidence from research papers.

Table 4. Evidence Transfer with Explicit Prompting

Procedure	Quantitative Data Transfer	Additional Side Effects Transfer	Research Findings Transfer	Overall Integration Rate
Chemotherapy	8/8 (100%)	3/3 (100%)	1/1 (100%)	12/12 (100%)
X-ray	3/4 (75%)	2/2 (100%)	1/1 (100%)	6/7 (86%)
CT Scan	4/5 (80%)	2/2 (100%)	1/1 (100%)	7/8 (88%)
Ultrasound	3/3 (100%)	1/1 (100%)	1/1 (100%)	5/5 (100%)
Mammography	5/6 (83%)	3/3 (100%)	2/2 (100%)	10/11 (91%)
Liver Biopsy	4/4 (100%)	2/2 (100%)	1/1 (100%)	7/7 (100%)
TOTAL	**27/30 (90%)**	**13/13 (100%)**	**7/7 (100%)**	**47/50 (94%)**

Evidence Transfer Score: 9 out of 9 = 100% transfer
Comparison to Passive Upload:

1. Passive (3A): 0/9 items transferred (0%)
2. Explicit (3B): 9/9 items transferred (100%)
3. Difference: +100 percentage points

4.7 Cross-Procedure Consistency Analysis

A critical finding was the remarkable consistency of evidence transfer patterns across all six medical procedures, despite their diverse risk profiles and clinical contexts.The observed evidence transfer behavior appears to reflect fundamental LLM behavior in multi-document processing rather than being specific to medical procedures, risk levels, or clinical contexts. Specifically, we observed:

1. Universal Passive Upload Failure: Across all procedures—regardless of risk level, clinical context, or data availability—passive document upload resulted in 0% evidence transfer.
2. Universal Explicit Prompting Success: Explicit prompting achieved 85-100% evidence transfer across all procedures, with an overall average of 94% successful integration.
3. No Risk-Level Dependency: Evidence transfer patterns showed no correlation with procedure risk level. High-risk chemotherapy (100% transfer with explicit prompting) performed identically to minimal-risk ultrasound (100% transfer).

4. No Clinical Context Dependency: Diagnostic imaging procedures showed the same pattern as therapeutic interventions, invasive procedures, and screening tests.
5. No Data Availability Effect: Procedures with many quantitative data points showed the same pattern as procedures with fewer data points.

Table 5. Evidence Transfer Pattern Consistency Across Procedures

Procedure	**Risk Level**	**Clinical Context**	**Passive Upload Transfer**	**Explicit Prompt Transfer**	**Pattern Match**
Chemotherapy	High	Therapeutic intervention	0%	100%	Consistent
X-ray	Low	Diagnostic imaging	0%	86%	Consistent
CT Scan	Moderate	Diagnostic imaging	0%	88%	Consistent
Ultrasound	Minimal	Diagnostic imaging	0%	100%	Consistent
Mammography	Moderate	Cancer screening	0%	91%	Consistent
Liver Biopsy	Moderate-High	Invasive procedure	0%	100%	Consistent

4.8 Replication Across Model Versions

To assess the robustness of findings, all experimental conditions were replicated using GPT-5 under identical protocols. Results closely matched those observed with ChatGPT-4: passive document upload again resulted in zero evidence transfer across all procedures, while explicit prompting consistently achieved high evidence integration. This replication suggests that the observed prompt-sensitivity pattern reflects a fundamental characteristic of LLM multi-document processing behavior rather than an artifact of a specific model version.

Table 6. Evidence Transfer Replication: ChatGPT-4 vs. GPT-5

Condition	**GPT-4 Transfer**	**GPT-5 Transfer**	**Consistent?**
Patient-Facing Article (Baseline)	N/A	N/A	Yes
Scientific Paper	0%	0%	Yes
Passive Upload	0%	0%	Yes
Explicit Prompting	94%	∼94%	Yes

5 Limitations

This study has several limitations. First, findings were replicated across ChatGPT-4 and GPT-5 but may not generalize to LLMs with fundamentally different architectures, such as open-source or medical-domain-specific models. Second, all experiments were conducted during a fixed time window in December 2024; as models are continuously updated, prompt sensitivity behavior may change in future versions. Third, while conditions were replicated across two model versions (ChatGPT-4 and GPT-5) yielding consistent results, multiple trials within a single model version were not conducted. Future work could assess intra-model output variability more systematically.

Fourth, source materials were limited to Harvard Health Publishing articles and a single scientific paper per procedure, which may not represent the full diversity of clinical documentation encountered in practice. Fifth, this study did not evaluate patient comprehension or clinical accuracy of the generated summaries. It remains unknown whether evidence-rich outputs improve patient understanding or informed consent and whether practicing doctors would find these summaries clinically accurate and appropriate for real patient consultations. While these evaluations were beyond the scope of this study, clinician evaluation of AI-generated clinical summaries is an established practice [11], providing a methodological foundation for future validation.

Sixth, inter-rater reliability was assessed for the chemotherapy procedure only ($\kappa = 1.00$), the most data-rich example in the study. Full dual-coding across all six procedures remains a direction for future work.

6 Future Work

Several important areas need further research in LLM-based medical risk communication. First, future studies could examine whether these prompt-sensitivity findings generalize across other LLMs such as Claude and Gemini, though the present results have immediate practical relevance given ChatGPT-4's dominant adoption in clinical settings. Second, studies should observe how doctors actually use these tools in real clinical settings to understand what works and what doesn't in busy healthcare environments.

Third, building on existing work in which clinicians have evaluated AI-generated clinical summaries and identified omission of critical information as a key challenge [11], future studies should assess whether outputs produced through explicit prompting are clinically accurate and appropriate for real patient consultations in risk communication contexts.

Finally, developing automated systems that check whether the AI correctly transferred all important medical information from research papers would help ensure patient safety before these tools are widely used in hospitals.

Acknowledgments

This study was not funded by any external organization.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aslam, M.S., Naveed, S., Ahmed, A., Abbas, Z., Gull, I., Athar, M.A.: Side effects of chemotherapy in cancer patients and evaluation of patients opinion about starvation based differential chemotherapy. Journal of Cancer Therapy **5**(8), 817–822 (2014)
2. Wall, B.F., Kendall, G.M., Edwards, A.A., Bouffler, S., Muirhead, C.R., Meara, J.R.: What are the risks from medical X-rays and other low dose radiation? The British Journal of Radiology **79**(940), 285–294 (2006)
3. Walsh, L., Shore, R., Auvinen, A., Jung, T., Wakeford, R.: Risks from CT scans—what do recent studies tell us? Journal of Radiological Protection **34**(1), E1 (2014)
4. Harvard Health Publishing - Chemotherapy, https://www.health.harvard.edu/cancer/chemotherapy-a-to-z, last accessed 2026/01/19
5. Harvard Health Publishing - Abdominal CT scan (computed tomography scan), https://www.health.harvard.edu/staying-healthy/abdominal-ct-scan-computed-tomography-scan-a-to-z, last accessed 2026/01/19
6. Harvard Health Publishing - X-rays, https://www.health.harvard.edu/healthy-aging-and-longevity/abdominal-ct-scan-computed-tomography-scan-a-to-z, last accessed 2026/01/19
7. Henry, T.A.: 2 in 3 physicians are using health AI—up 78% from 2023. American Medical Association, https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023, last accessed 2026/04/23 (2025)
8. Hoppe, J.M., Auer, M.K., Strüven, A., Massberg, S., Stremmel, C.: ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. Journal of Medical Internet Research **26**, e56110 (2024)
9. Ramaswamy, A., Tyagi, A., Hugo, H., Jiang, J., Jayaraman, P., Jangda, M., Te, A.E., Kaplan, S.A., Lampert, J., Freeman, R. et al.: ChatGPT Health performance in a structured test of triage recommendations. Nature Medicine, 1–5 (2026)
10. Bean, A.M., Payne, R.E., Parsons, G., Kirk, H.R., Ciro, J., Mosquera-Gómez, R., Hincapié, S., Ekanayaka, A.S., Tarassenko, L., Rocher, L. et al.: Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. Nature Medicine, 1–7 (2026)
11. Kahl, N.M., Frieden, M.J., Pope, Z.R., Millen, M.M., Tolia, V.M., Chan, T.C., Longhurst, C.A., Singh, K., You, A.X.: Evaluation of electronic health record-integrated artificial intelligence chart review. npj Health Systems **3**(1), 6 (2026)