A Framework to Assess Clinical Safety and Hallucination Rates of LLMs for Medical Text Summarisation

Authors: Dr Elham Asgari[1,2], Miss Nina Montaña-Brown[1], Dr Magda Dubois[1], Mr Saleh Khalil[1], Miss Jasmine Balloch[1], Dr Dominic Pimenta[1]

Affiliation: 1-Tortus AI, London, UK, 2- Guy's and St Thomas NHS Trust, London, UK

Corresponding author: asgelham@gmail.com

Abstract

The integration of large language models (LLMs) into healthcare settings holds great promise for improving clinical workflow efficiency and enhancing patient care, with the potential to automate tasks such as text summarisation during consultations. The fidelity between LLM outputs and ground truth information is therefore paramount in healthcare, as errors in medical summary generation can lead to miscommunication between patients and clinicians, leading to incorrect diagnosis and treatment decisions and compromising patient safety. LLMs are well-known to produce a variety of errors. Currently, there is no established clinical framework for assessing the safety and accuracy of LLM-generated medical text.

We have developed a new approach to: a) categorise LLM errors within the clinical documentation context, b) establish clinical safety metrics for the live usage phase, and c) suggest a framework named CREOLA for assessing the safety risk for errors. We present clinical error metrics over 18 different LLM experimental configurations for the clinical note generation task, consisting of 12,999 clinician-annotated sentences. We illustrate the utility of using our platform CREOLA for iteration over LLM architectures with two experiments. Overall, we find our best-performing experiments outperform previously reported model error rates in the note generation literature, and additionally outperform human annotators. Our suggested framework can be used to assess the accuracy and safety of LLM output in the clinical context.

Introduction

One of the most appealing applications of LLMs in healthcare is for administrative tasks [1]. Clinicians devote a substantial amount of time to documentation [2], and prolonged interaction with electronic health records, where clinical documentation is logged, has been demonstrated to raise cognitive load and lead to burnout [3]. In fact, the use of LLMs for clinical documentation, especially clinical note generation [4] or consultation summarization [5, 6], is an active area of research.

However, LLMs are known to produce errors in many settings, from document summarization [7], to general reasoning tasks as well as more clinically relevant tasks [8]. These errors can be categorised as "hallucinations" [9]: known as an event where LLMs generate information that is not present in the input data, or omissions: the event where LLMs miss relevant information from the original document. Errors in clinical documentation generation can lead to inaccurate recording and communication of facts [10, 11]. Inaccuracies in the document summarisation task can introduce misleading details [8] into transcribed conversations or summaries, potentially delaying diagnoses [12] and causing unnecessary patient anxiety.

The problem of hallucinations poses a significant challenge to date [1, 13]. The occurrence of hallucinations has previously been attributed to the data quality during model training [14, 15], the type of model training methodology [16] and prompting strategies [17].

Recent work has established that hallucination may be an intrinsic, theoretical property of all LLMs [9]. Consequently, there is a growing body of work focused on the technical evaluation of LLM accuracy and the detection and mitigation of hallucinations in LLMs [18]. However, the prevalence, causation, and evaluation of hallucinations in a clinical context, as well as their subsequent impact on clinical safety, remains an open question.

Background

Clinical Note Quality Evaluation Frameworks

Clinical documentation can be variable in quality [19, 20], and studies estimate that human-generated clinical notes have, on average, at least 1 error and 4 omissions [21]. Given the increased usage of LLMs for clinical documentation [22, 23], several methods have been proposed for evaluating clinical documentation generated using LLMs.

Relevant clinical evaluation frameworks typically include categorising clinical errors for downstream analysis. Typically, these differ from traditional natural language processing (NLP) taxonomies [16], which have separated hallucination types into distinct categories, for example, into "intrinsic" and "extrinsic" [24], "factuality" and "faithfulness" [16], "factual mirage" and "silver lining" [25] errors. The differences between general and clinical taxonomies arise from the necessity of increased granularity of clinical error types, which are not captured by the broader, general methods.

For example, Tierney et al. [26] propose using a modified version of the Physician Documentation Quality Instrument-9, accounting for hallucinations and bias, while Abacha et al. [23] propose

evaluating clinical note quality using automated metrics. However, these relevant clinical categorisations have not assessed the implications of the mistakes for downstream harm.

LLM Evaluation Methods

Despite the exponential growth in benchmarks for model reasoning abilities [27], the evaluation of LLMs on clinical tasks has typically been carried out via "question-answering" (QA) benchmarks [5, 8, 28]. These tasks assess models' accuracy over various clinical questions, typically derived from licensing exams. While these methods offer insights into factual knowledge and reasoning abilities of LLMs, they do not assess clinical or medical capabilities such as medical text summarisation.

Multiple benchmarks have been proposed to evaluate model summarisation capabilities in the biomedical domain, including over biomedical literature [29–32], medical forum conversations [33], and radiology reports [22, 34]. However, these benchmarks do not capture the nuances of patient-facing clinical interactions, where LLM-documentation holds most promise.

Recently, Umapathi et al. [35] have assessed models' tendency towards hallucination. They reported that LLMs were significantly variable in their accuracy depending on the prompts used. However, the MedHALT benchmark is limited to assessing LLM's reasoning capabilities over the medical domain in a QA format. Most relevantly, Moramarco et al. [21] benchmark BART models on the PriMock dataset and find that they produce 3.9 errors and 6.6 omissions on average per note. However, they did not assess the model's impact or human errors on patient safety as part of their study.

This study aims to contribute to the ongoing effort to ensure clinical safety in using AI by 1) introducing a clinical LLM error categorisation, with a particular emphasis on quantifying the clinical impact of hallucinations, 2) presenting a clinician-in-the-loop methodology to assess the use of LLMs in clinical practice, whilst safeguarding downstream users from potential drawbacks, and 3) sharing insights from the usage of the proposed methodology in clinical practice. Our objective is to promote the efficient, reliable, and confident use of LLMs for clinical documentation, thus supporting healthcare providers in delivering high-quality care and overall reducing the administrative workload for clinicians.

Methodology

We propose a multi-component framework to evaluate the hallucinatory potential of different information flows for LLM-assisted document summarisation. At the core of our framework is

the concept of 'clinician-in-the-loop'. Given their expertise, clinicians are uniquely positioned to identify clinical errors made by the models, making their involvement essential.

Our framework consists of four distinct components: 1) a clinically and technically-informed error taxonomy to assess LLM outputs, 2) an experimental structure to comprehensively compare outputs within our LLM document generation pipeline, 3) a clinical safety framework to assess potential harms of errors in LLM outputs, and 4) an encompassing graphical user interface (GUI) to perform and assess all previous steps. The framework design and connections between components are summarised in Fig. 1.

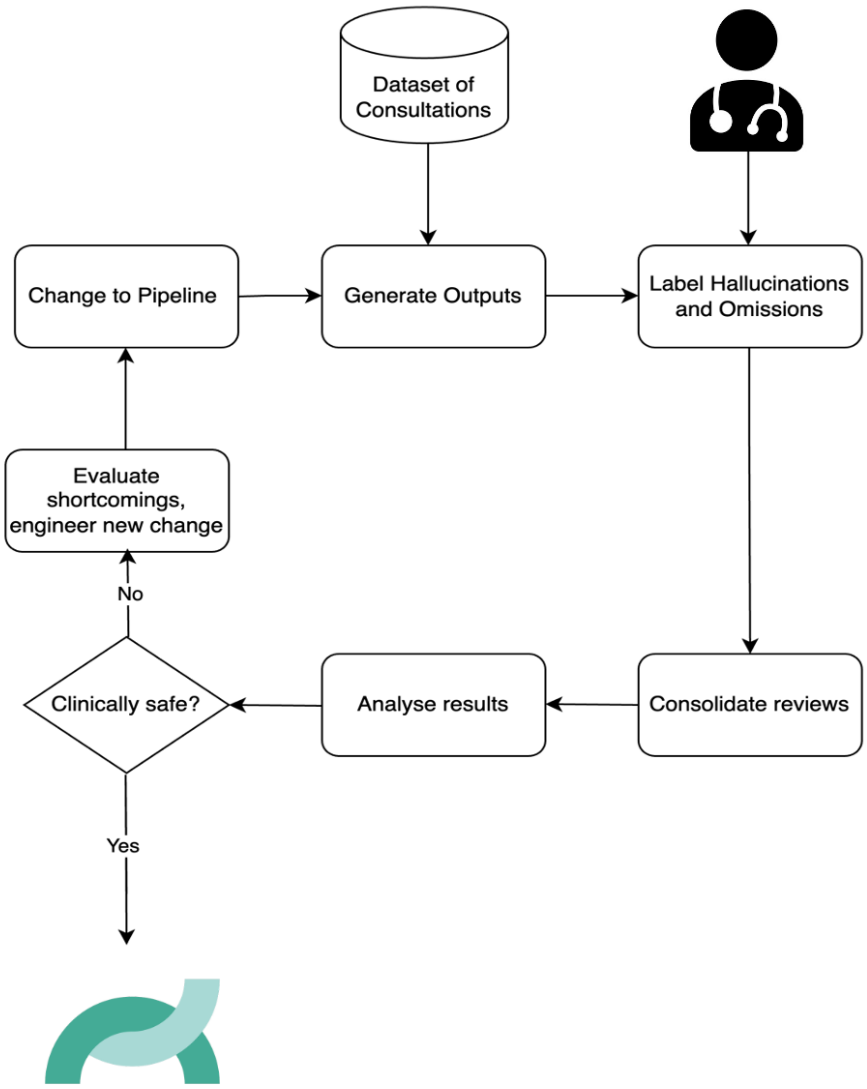


Figure 1: Our workflow for the assessment of LLM output using CREOLA platform

Hallucination and Omission Taxonomy

We follow the conventional AI literature and taxonomise LLM errors into two types 1) hallucinations, which are instances of text unsupported by the associated clinical documentation, and 2) omissions [21], which are instances where relevant details are missed in the supporting evidence. Furthermore, inspired by protocols in medical device certifications [36, 37], we categorise errors as either 'major' or 'minor', where major indicates an impact on the diagnosis or the management of the patient.

To make our categories more granular, we propose to divide hallucinations into four categories: 1) fabrication, occurring when the model produced information that was not available in the text, 2) negation, occurring when the model output negates a clinically relevant fact, 3) causality, occurring when a model speculates the cause of a given condition without explicit support from the text, and 4) contextual, occurring when the model mixes topics otherwise not related in the given context.

In the case of omissions, we further divide them into: 1) current issues, occurring when details about the current presentation were omitted, 2) PMFS (past medical history, medication history, family and social history), occurring when details about the past medical history, medications including allergies, family and social history, including drinking and smoking, were omitted, and 3) information and plan: when discussions and explanations of the condition and management plans were omitted. Examples of each of the sub-categories are provided in the Supplementary Materials.

Experimental Structure and Annotation Protocol

Here, we define a process to assess how model parameters affect the model outputs and clinical safety. To do this, we define "experiments", which are parametrised by: 1) the number of data points processed by the LLM, 2) the type of data the LLM will ingest, 3) the model configuration (type of model, random seed, temperature,...), 4) the prompt used to obtain an LLM output, and 5) the number of clinicians which must review the data point for clinical errors.

To achieve a meaningful clinical comparison of efficacy and safety in a data-driven way, our framework relies on the definition of a 'baseline' experiment against which to compare results. The baseline must have the same data points as the proposed experiment. To attribute an experiment's results to a specific change, we aim only to change one parameter from the baseline experiment configuration.

Given an experiment configuration, we extract model outputs from the input data and store the results in a database with the associated experiment metadata. We task annotators to classify whether given sub-sections of the output contain hallucinations or omissions according to our taxonomy, and explain in free text the reason for classification. The annotators were volunteer doctors who were paid £5 per note for annotations. Recognising the subjectivity inherent in annotation, we require annotation by at least two clinicians for each input-output pair. This step is followed by a consolidation step, i.e. a detailed review by our internal team of senior clinicians, ensuring a consistent evaluation of all annotations.

Clinical Safety Assessment

Recognising that safety assessment is a crucial part of using any medical technology, we designed a safety evaluation framework of the LLM outputs based on the framework used for evaluating a medical device [36, 37]. Overall, this assessment involves estimating the likelihood of an error happening (Table 1) in conjunction with the potential impact of the error on the clinical outcome if it does occur. Table 2 shows the classification of the level of harm, and Table 3 presents the estimation of risk based on the likelihood and consequences of an event.

Likelihood Category	Interpretation
Very high	Certain or almost certain; highly likely to occur
High	Not certain but very possible: reasonably expected to occur in the majority of cases
Medium	Possible
Low	Could occur but in the great majority of occasions will not
Very low	Negligible or nearly negligible possibility of occurring

Table 1. The likelihood of a hazard occurring

Consequence classification	Interpretation	Number of patients affected
Catastrophic	Death	Multiple

	Permanent life-changing incapacity and any condition for which the prognosis is death or permanent life-changing incapacity, severe injury or severe incapacity from which recovery is not expected in the short term	Multiple
Major	Death	Single
	Permanent life-changing incapacity and any condition for which the prognosis is death or permanent life-changing incapacity, severe injury or severe incapacity from which recovery is not expected in the short term	Single
	Severe injury or severe incapacity from which recovery is expected in the short term	Multiple
	Severe psychological trauma	Multiple
Considerable	Severe injury or severe incapacity from which recovery is expected in the short term	Single
	Severe psychological trauma	Single
	Minor injury or injuries from which recovery is not expected in the short term	Multiple
	Significant psychological trauma	Multiple
Significant	Minor injury or injuries from which recovery is not expected in the short term	Single
	Significant psychological trauma	Single
	Minor injury from which recovery is expected in the short term	Multiple
	Minor psychological upset; inconvenience	Multiple

Minor	Minor injury from which recovery is expected in the short term; Minor psychological upset; inconvenience; any negligible severity	Single

Table 2: Guidance for assessing the level of harm

Likelihood	Very high	3	4	4	5	5
	High	2	3	3	4	5
	Medium	2	2	3	3	4
	Low	1	2	2	3	4
	Very low	1	1	2	2	3
		Minor	Significant	Considerable	Major	Catastrophic
		Consequence				

Table 3. Risk estimation based on the likelihood and consequence of harm occurrence

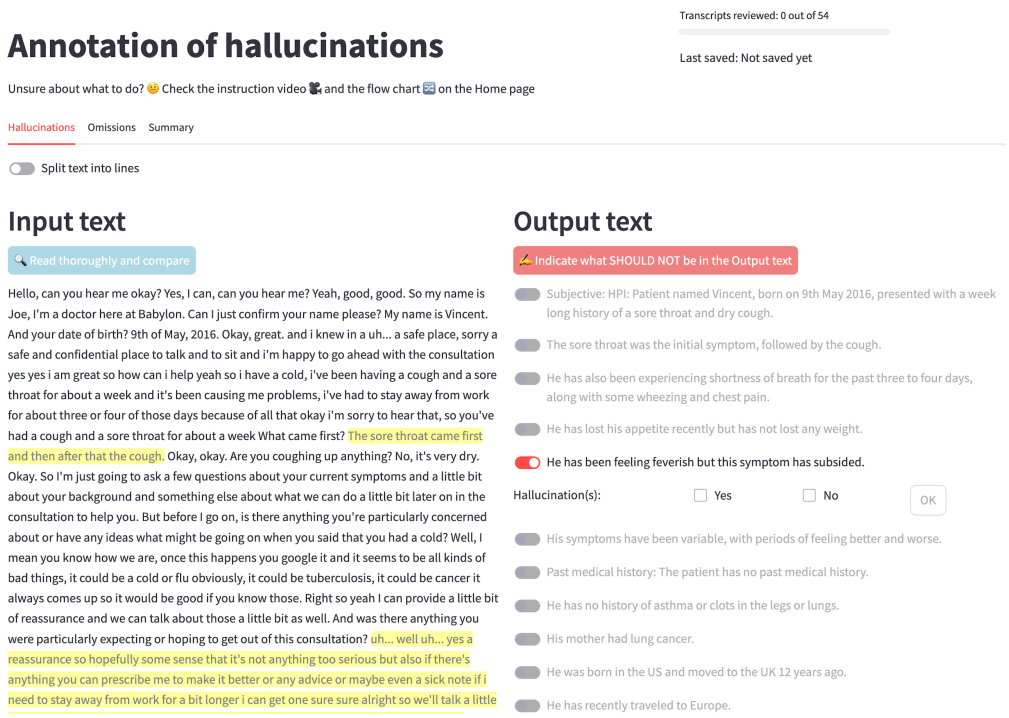
To ensure consistency in calculating the likelihood of the occurrence of hallucinations and omissions in each experiment, we describe a percentage-based metric of the presence of hallucinations and omissions across experiments, described in Table 4.

	Per 25 examples	Possibility
Very High	22.5	90%
High	15	60%
Medium	7	10-60%
Low	2.5	10%
Very Low	0.5	1%

Table 4. Calculating the likelihood of an error occurring in the text output

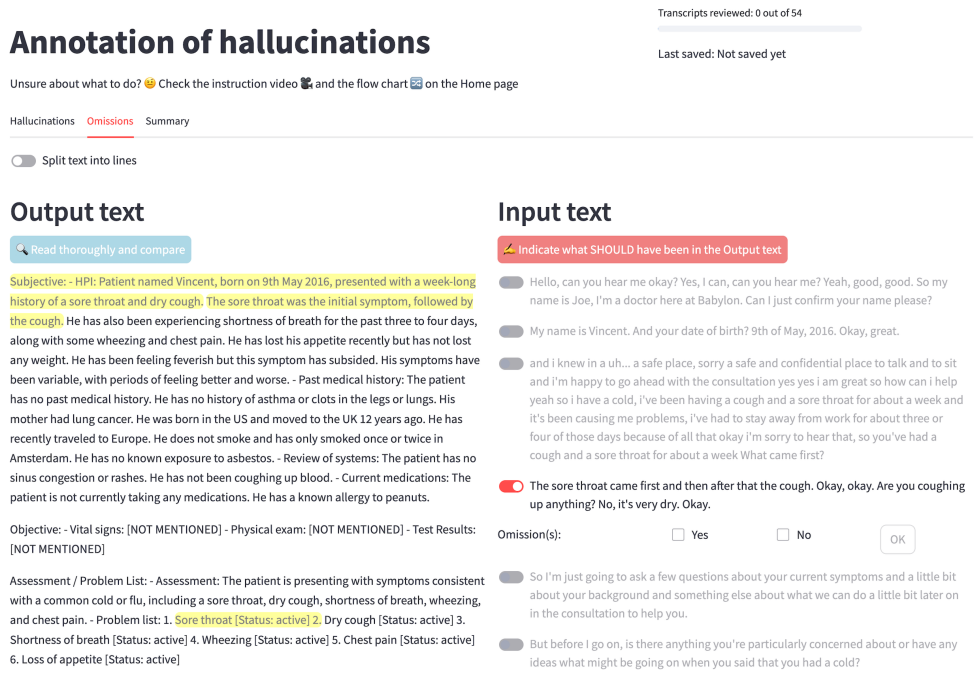
CREOLA: Clinical Review of LLMs and AI

We combine the experiment design, hallucination and omission taxonomy, and clinical safety evaluation in a platform we denote CREOLA, short for Clinical Review of LLMs and AI[1]. The platform is used to identify resultant changes in generated clinical documentation arising from changes to processes in LLM architecture. As illustrated in the "experimental structure", these changes could involve - but are not limited to - the type of model used or prompts used to obtain outputs. The platform was hosted as a Streamlit web application (https://creola.tortus.ai/); the annotation UI is displayed in Fig. 2. Further detail on the annotation flow is provided in the Supplementary Materials.



A

[1] pays tribute to Creola Katherine Johnson [46], a pioneering human computer at NASA. Just as human computers were integral to the safe landing of Apollo moon missions, clinicians play a vital role in safely integrating AI technologies into clinical practice



B

Figure 2: The annotation UI used to identify hallucinations (A) and omissions (B), as well as categorise them into major and minor categories. To facilitate review, the closest sentence matches (highlighted in yellow) for each portion of text under review were extracted from the counterpart document. In the case of hallucinations, portions of text in the note were compared to the consultation (A), whereas for omissions, portions of the transcript were compared against the note (B).

Annotator recruitment

As outlined in the Methods section, our framework requires annotators to review model outputs. Clinicians are uniquely skilled in critically assessing the veracity of clinical facts in the text. Therefore, we ask clinicians to annotate errors for our experiments. Annotators could register to contribute to the annotation through the CREOLA platform. To ensure annotators had a good understanding of the process, one-to-one tuition was initially provided by the study team. As the number of annotators grew, a short online course was developed to explain the annotation process, followed by a questionnaire to ensure a comprehensive understanding of the material. The annotators were only able to participate if they completed the questionnaire correctly. The annotators could contact the study teams with any questions through the CREOLA platform in order to ensure any problems in the platform were dealt with promptly.

Experiment Results

Our proposed platform enables clinicians to identify relevant errors in an effective and low-code manner. Overall, this system has enabled us to quantify changes in our model configurations and prioritise issues to ensure the safe integration of AI-generated summaries into clinical practice.

Experimental parameters

The focus of this study is to systematically evaluate the impact of prompting techniques and changes on clinical note generation from primary care consultations. As we propose to analyse the impact of prompting techniques on clinical safety in clinical notes, we use OpenAI's GPT-4 (GPT-4-32k-0613) as the LLM for all our experiments, with seed=210, temperature=0, and top-p=0.95. By default, we use 25 unique data points, consisting of primary care consultation transcripts from the PriMock dataset [38], and require two reviewers per data point. In other words, we propose to analyse how prompt changes affect the hallucinatory potential for consultation summarisation. We additionally identify the specific sections of the notes where the hallucinations occur. To investigate this, we manually assess the location of the occurrence of hallucinations in the main history, examination, discussion, symptoms assessment, and plan. The consolidation was performed by a senior clinician with over 20 years of clinical experience.

Safety Analysis

Out of the 25 consultations in 18 experiments, 450 medical notes, comprising 12,999 sentences, were generated. Of these, 191 sentences had hallucinations (1.47%), of which 44% were major (impacting patient diagnosis and management).

Of the hallucinations, 82 (43%) were fabricated, 56 (30%) were negations, 33 (17%) were contextual, and 20 (10%) were related to causality. Most hallucinations occurred in the assessment and plan section of the notes, as shown in Figure 3. Examples of the various hallucination types are available in supplementary materials.

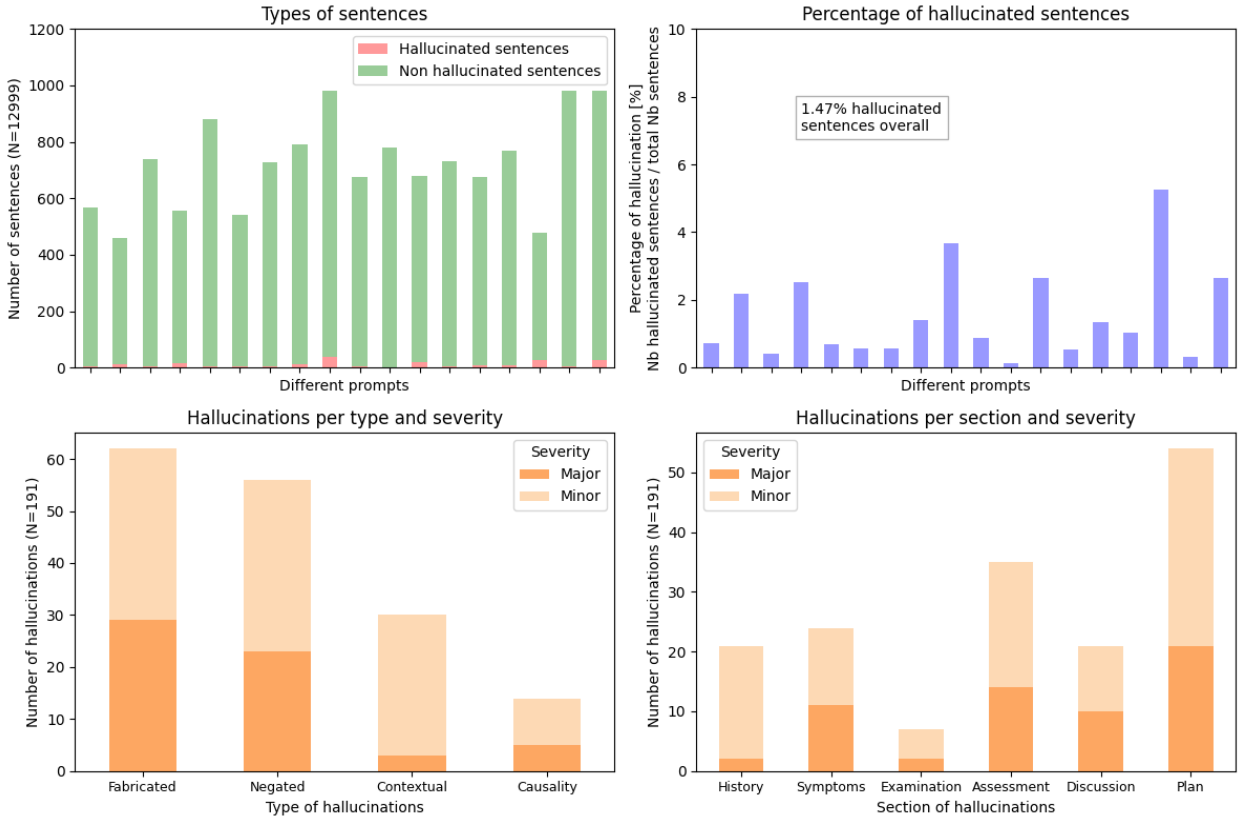


Figure 3. Incidence of hallucinations, their type and section of the note they appear in

In terms of omissions, out of 450 notes and 49,590 sentences from our transcripts, 1,712 sentences were omitted (3.45%), 286 (0.5%) of which were classified as major and 1,426 (2.95%) as minor. Figure 4 shows the number of omissions based on different prompts.

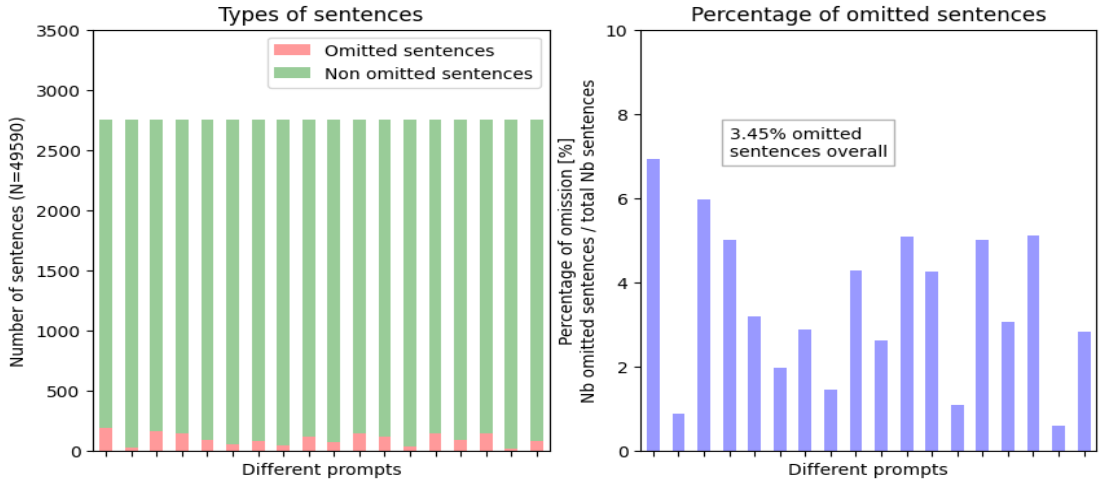


Figure 4: Number of omitted sentences and their percentage in the output based on different prompts

We evaluated the risk severity by using the provided framework for all identified hallucinations, as depicted in Figure 5. We also determined their type and likely occurrence within the note. Most of the hallucinations were found to be fabrications and primarily appeared in the planning section of the clinic note, as shown in Figure 6. The most concerning hallucinations were of the negation type, where the output contradicted what was said during the consultation. These hallucinations were mostly found in the planning section of the letters. Examples of these hallucinations are included in the supplementary materials.

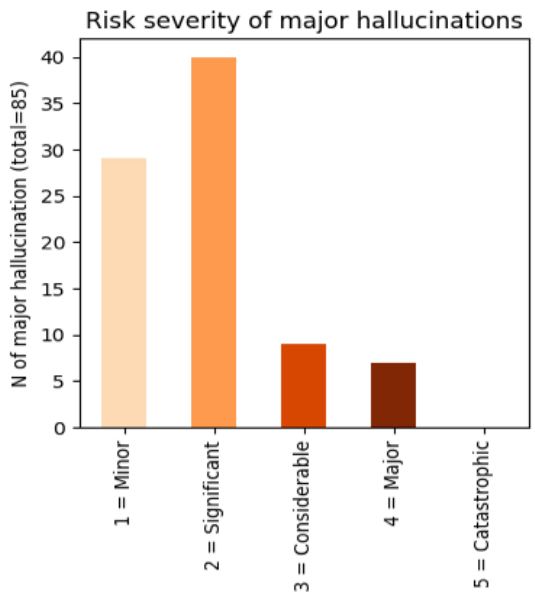


Figure 5: Severity of risk in major hallucinations

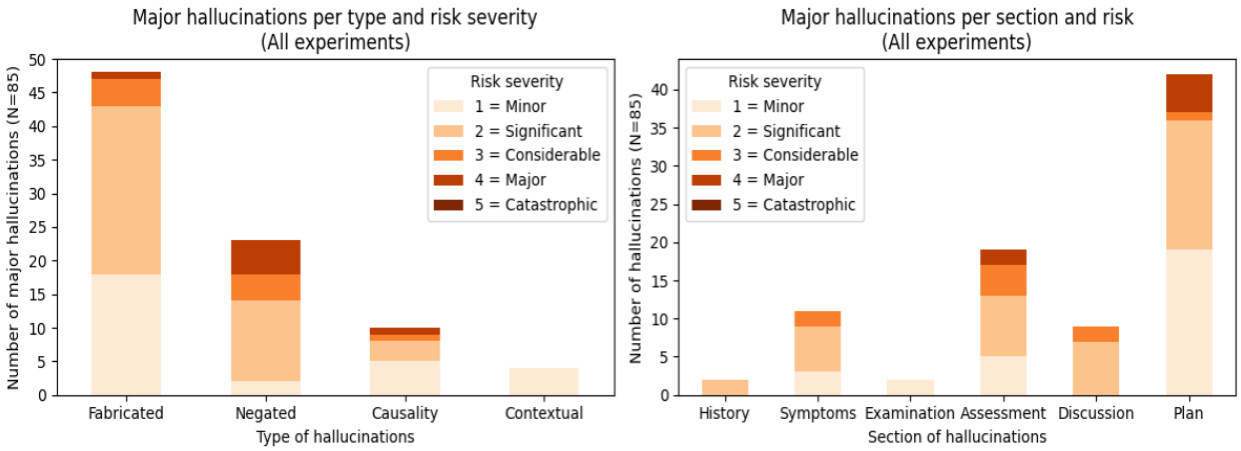


Figure 6: Hallucination risk assessment. The left box shows the type of hallucination, and the box on the right shows in which section of the clinical notes they occurred.

We conducted the risk assessment on the omissions and looked at where they had mostly occurred (Figure 7).

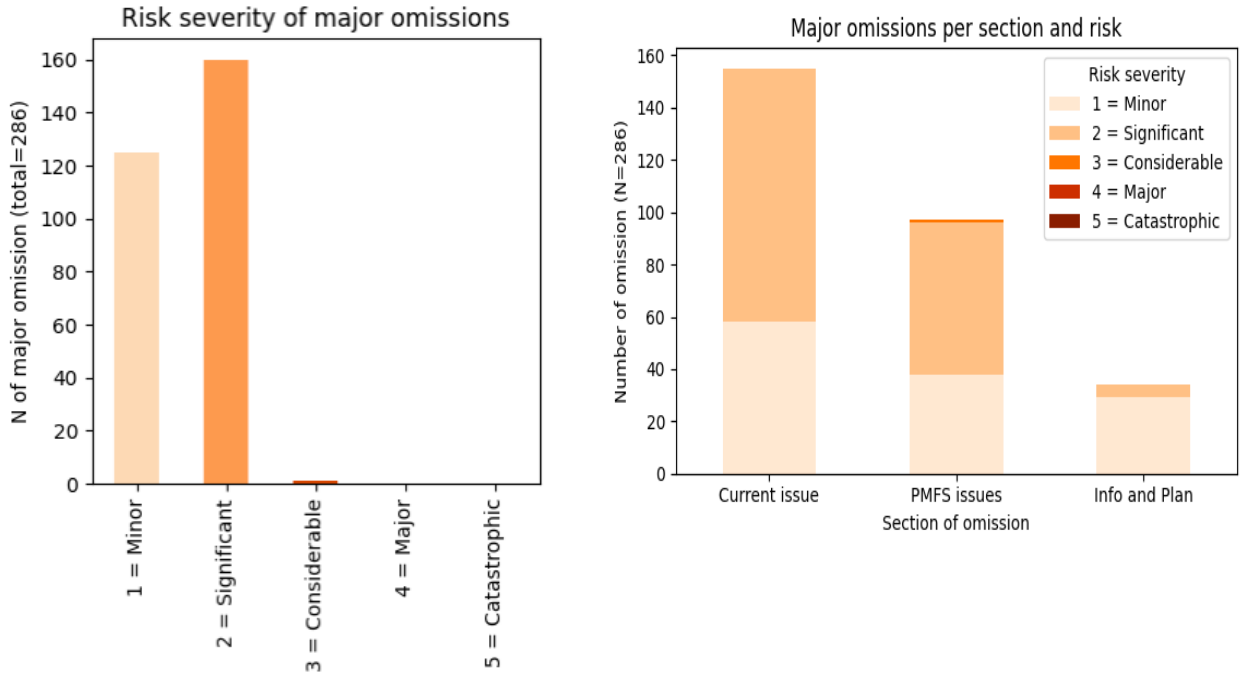


Figure 7: Severity risk for major omissions (left box) and section of the clinical note where most of the omissions occurred (right box)

We then examined the two best-performing experiments with the fewest hallucinations and omissions (experiments 8 and 11). Figure 8 illustrates the risk of hallucination in these experiments.

Experiment 11 did not have any major omissions, and the risk assessment of major omissions for experiment 8 is shown in Figure 9.

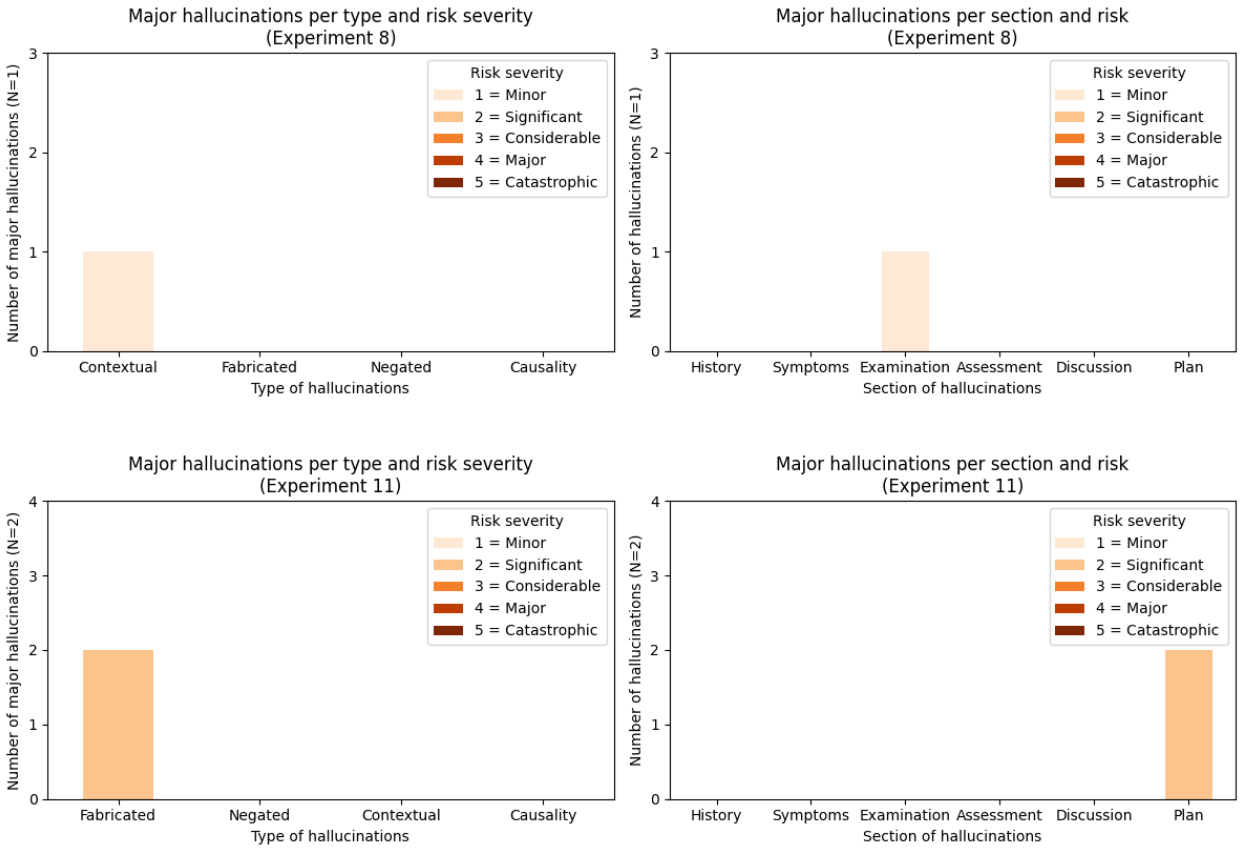


Figure 8: Risk assessment of hallucinations in experiments 8 and 11

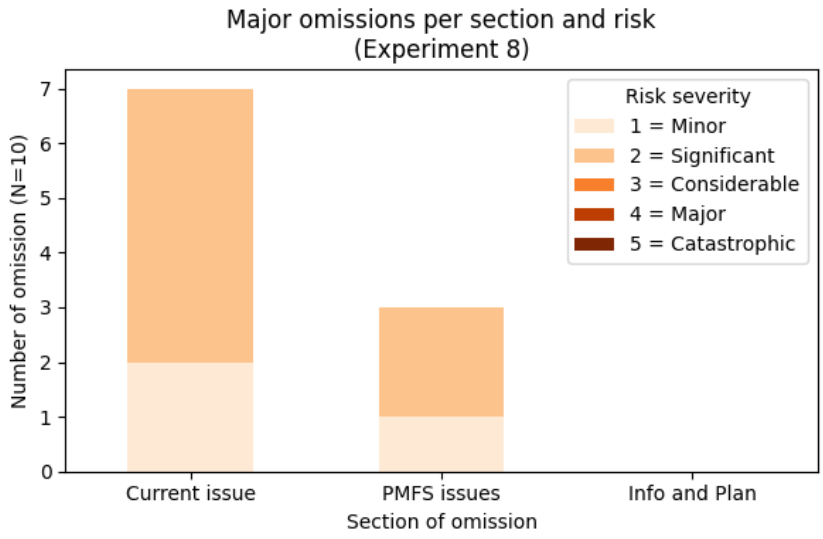


Figure 9: Risk assessment for major omissions in experiment 8 and the section in the clinical notes it has occurred

We present two examples of using the CREOLA framework to assess the overall volume of hallucinations and omissions across experimental iterations in Figs. 10 and 11.

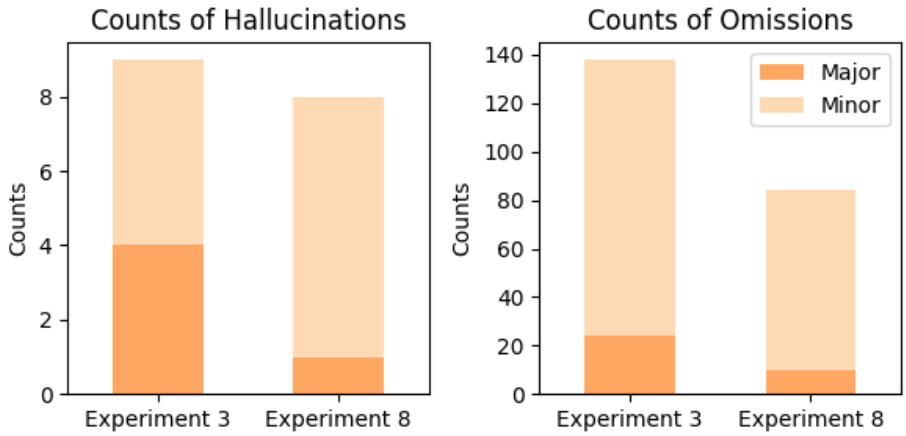


Figure 10: Comparison of hallucination and omission counts between two experiments, assessing differences in prompt engineering effect on quality of outputs. For prompt changes, see Fig. A in the Supplementary Materials.

In Fig. 10, we present an example of how we used CREOLA to iteratively assess how different prompt changes affected hallucinatory potential. We illustrate prompt changes in the Supplementary Materials; in summary, it consisted of modifying the instructions to the LLM. The proposed changes reduced major (24 to 10) and minor omission (114 to 74) counts, major hallucination counts were additionally reduced from 4 to 1. However, one should note that minor hallucinations increased from 5 to 7.

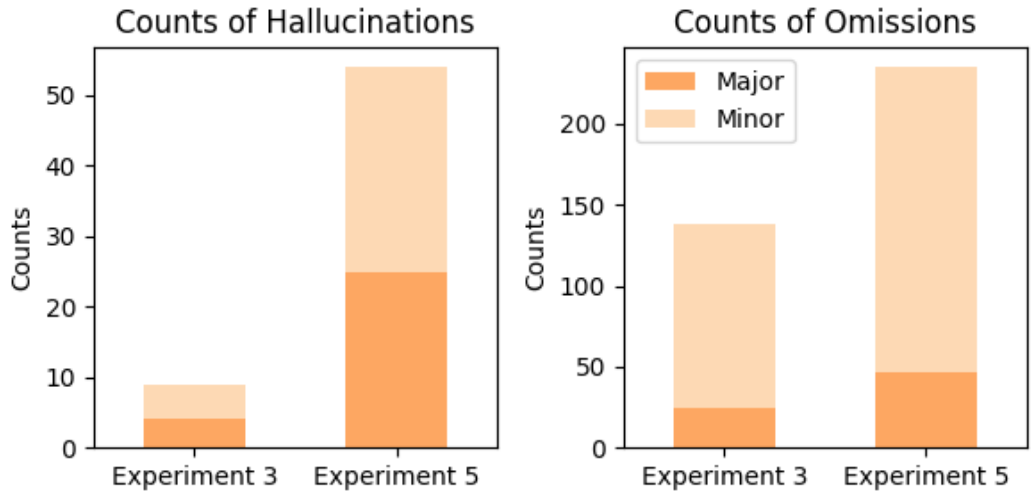


Figure 11: Comparison of hallucination and omission counts between two experiments, assessing the efficacy of a data-extraction intermediate step versus a normal note-generation step.

In contrast to Fig. 10, in Fig. 11 we observe how a proposed change majorly increases the volume of both hallucinations and omissions. The proposed experimental change was to assess whether extracting clinical facts as intermediary outputs from the original transcript would lead to fewer hallucinations than directly producing a note. This architectural modification increased major hallucinations from 4 to 25, minor hallucinations from 5 to 29, major omissions from 24 to 47, and minor omissions from 114 to 188. This result precluded the new change from being evaluated for clinical safety, as the increase in hallucinations and omissions was considered too large to be considered useful.

Discussion

LLMs are well known to produce errors in their outputs, typically known as hallucinations, and furthermore, models can produce hallucinations with high degrees of confidence that the output is correct [39]. Combined, these characteristics can be particularly harmful in a healthcare setting. Whilst several approaches have been proposed for detecting, classifying, and mitigating errors in LLM outputs, we find that they are insufficiently granular to be clinically relevant, typically bypassing assessment of the downstream impacts on patient care.

Our experimental results overall show that omissions are more likely than hallucinations; However, hallucinations are more likely to lead to downstream harm than omissions, as quantified by the likelihood of the major category amongst both error types. Overall, the hallucination and omission rates are low relative to the volume of text generated. The rate of hallucination versus omissions in our results is in line with previously reported error and omission rates [21], which find omission is more likely for human and model-generated notes. This can be attributed to the nature of the summarisation task, which requires the ablation of certain data from the original text to make it a concise, relevant, and useful summary artefact.

Furthermore, experiment 8 achieves 1 major hallucination and 10 major omissions, whilst experiment 11 achieves 2 major hallucinations and 0 major omissions over 25 notes. These results are highly encouraging, as our framework has resulted in fewer errors per note than those reported in the literature; Moramarco et al. [21] reported 3.9 errors and 6.6 omissions achieved per note as produced by a BART model and 1 error and 4 omissions per note. Although this rate is subject to change depending on the text and experiment, our results suggest that we can achieve state-of-the-art, sub-human clinical error rates by carefully engineering and subsequently validating LLMs to produce safe outputs.

Finally, we showcase two examples of the assessment of error changes between iterations in Figs. 10 and 11, illustrating how the CREOLA platform can be used to safely oversee changes in AI pipelines from a clinical perspective. By being able to observe decreases in clinical errors, we are able to make confident deployment decisions on novel LLM architectures. Additionally,

increases in clinical errors resulting from experimentation in CREOLA can validate or discredit architectures in a safe sandbox environment, reducing or eliminating the number of deployments that result in increased errors and thus reducing the potential for downstream harm.

Our study is limited in several ways. Firstly, the sample size of medical transcripts used was relatively small; the sample size was chosen to balance the trade-off of annotation volume required for the comparison of different experiments against sample size and number of experiments performed. Additionally, we only evaluated one LLM (GPT-4), selected due to its established performance in text summarization at the time of our experiments. Furthermore, our experiments use a direct prompting scheme. Newer methods such as (but not limited to) Retrieval-Augmented Generation (RAG) [40], Chain of Thought (CoT) [41], or the use of knowledge graphs [42] have recently been used to enhance the performance of LLMs. For example, by equipping LLMs with domain-specific knowledge, RAG enables the models to generate more precise and pertinent results [43, 44], whilst CoT generally enhances model reasoning abilities. A straightforward extension of this work is using this framework over different experimental configurations, such as using different models or prompting techniques, and comparing the impact on reported performance to clinical safety metrics.

Finally, using human annotators to evaluate large amounts of data is expensive and unsustainable. In the long run, the automated evaluation of model output [45] is a consequential future direction which will enable the scalable assessment of a larger volume of information, with clinicians remaining in the loop by "supervising" evaluator models via the inspection of a sub-sample of the outputs.

Conclusion

In this work, we present the CREOLA framework for clinical safety assessment of LLMs in clinical documentation scenarios. Using CREOLA, we analyse the impact of prompting techniques on the safety of LLM outputs. Our iterative modification process allows us to reach new low hallucination and omission rates - our best-performing experiments outperform previously reported model and human error rates - facilitating confident deployment of our solutions to end clinical users. Additionally, CREOLA provides a sandbox environment which buffers users and patients from harm in the case that iteration leads to higher clinical error rates. We envision CREOLA can serve as a governance and clinical safety assessment template for various organisations and further empower clinicians to become key stakeholders in the deployment of LLMs in clinical practice.

References

1. Clusmann, J. et al.: The future landscape of large language models in medicine. Communications Medicine. 3, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1
2. Becker, G. et al.: Four minutes for a patient, twenty seconds for a relative - an observational study at a university hospital. BMC Health Serv Res. 10, 94 (2010). https://doi.org/10.1186/1472-6963-10-94
3. Asgari, E. et al.: Impact of Electronic Health Record Use on Cognitive Load and Burnout Among Clinicians: Narrative Review. JMIR Med Inform. 12, e55499 (2024). https://doi.org/10.2196/55499
4. Ali, S.R., Dobbs, T.D., Hutchings, H.A., Whitaker, I.S.: Using ChatGPT to write patient clinic letters. Lancet Digit Health. 5, e179–e181 (2023). https://doi.org/10.1016/S2589-7500(23)00048-1
5. Patel, S.B., Lam, K.: ChatGPT: the future of discharge summaries? Lancet Digit Health. 5, e107–e108 (2023). https://doi.org/10.1016/S2589-7500(23)00021-3
6. Van Veen, D. et al.: Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 30, 1134–1142 (2024). https://doi.org/10.1038/s41591-024-02855-5
7. Zhang, T. et al.: Benchmarking Large Language Models for News Summarization. Trans Assoc Comput Linguist. 12, 39–57 (2024). https://doi.org/10.1162/tacl_a_00632
8. Ahmad, M.A., Yaramis, I., Roy, T.D.: Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. (2023)
9. Xu, Z., Jain, S., Kankanhalli, M.: Hallucination is Inevitable: An Innate Limitation of Large Language Models. (2024)
10. Kripalani, S. et al.: Deficits in Communication and Information Transfer Between Hospital-Based and Primary Care Physicians. JAMA. 297, 831 (2007). https://doi.org/10.1001/jama.297.8.831
11. Adane, K., Gizachew, M., Kendie, S.: The role of medical data in efficient patient care delivery: a review. Risk Manag Healthc Policy. Volume 12, 67–73 (2019). https://doi.org/10.2147/RMHP.S179259
12. Schiff, G.D.: Diagnostic Error in Medicine. Arch Intern Med. 169, 1881 (2009). https://doi.org/10.1001/archinternmed.2009.333
13. Kumar, S., Balachandran, V., Njoo, L., Anastasopoulos, A., Tsvetkov, Y.: Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 3299–3321. Association for Computational Linguistics, Stroudsburg, PA, USA (2023)
14. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623. ACM, New York, NY, USA (2021)
15. Rando, J., Tramèr, F.: Universal Jailbreak Backdoors from Poisoned Human Feedback. (2023)
16. Huang, L. et al.: A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. (2023)

17. Rawte, V. et al.: Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness. (2023)
18. S.M Towhidul Islam Tonmoy, S.M.M.Z.V.J.A.R.V.R.A.C.A. Das: A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. Computer Science, Computation and Language. (2024)
19. Overhage, J.M., Qeadan, F., Choi, E.H.E., Vos, D., Kroth, P.J.: Explaining Variability in Electronic Health Record Effort in Primary Care Ambulatory Encounters. Appl Clin Inform. 15, 212–219 (2024). https://doi.org/10.1055/s-0044-1782228
20. Shahbodaghi, A., Moghaddasi, H., Asadi, F., Hosseini, A.: Documentation Errors and Deficiencies in Medical Records: A Systematic Review. J Health Manag. 26, 351–368 (2024). https://doi.org/10.1177/09720634241229545
21. Moramarco, F. et al.: Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5739–5754. Association for Computational Linguistics, Stroudsburg, PA, USA (2022)
22. Demner-Fushman, D. et al.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association. 23, 304–310 (2016). https://doi.org/10.1093/jamia/ocv080
23. Abacha, A. Ben, Yim, W., Michalopoulos, G., Lin, T.: An Investigation of Evaluation Metrics for Automated Medical Note Generation. (2023)
24. Ji, Z. et al.: Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 55, 1–38 (2023). https://doi.org/10.1145/3571730
25. Huang, Y., Tang, K., Chen, M., Wang, B.: A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry. (2024)
26. Tierney, A.A. et al.: Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal. 5, (2024). https://doi.org/10.1056/CAT.23.0404
27. Minaee, S. et al.: Large Language Models: A Survey. (2024)
28. Evans, O. et al.: Truthful AI: Developing and governing AI that does not lie. (2021)
29. Cohan, A. et al.: A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. (2018)
30. Gupta, V., Bharti, P., Nokhiz, P., Karnick, H.: SumPubMed: Summarization Dataset of PubMed Scientific Articles. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop. pp. 292–303. Association for Computational Linguistics, Stroudsburg, PA, USA (2021)
31. Luo, Z., Xie, Q., Ananiadou, S.: CitationSum: Citation-aware Graph Contrastive Learning for Scientific Paper Summarization. (2023)
32. DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B., Wang, L.L.: MS2: Multi-Document Summarization of Medical Studies. (2021)
33. Song, Y., Tian, Y., Wang, N., Xia, F.: Summarizing Medical Conversations via Identifying Important Utterances. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 717–729. International Committee on Computational Linguistics, Stroudsburg, PA, USA (2020)

34. Johnson, A.E.W. et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data. 6, 317 (2019). https://doi.org/10.1038/s41597-019-0322-0
35. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Med-HALT: Medical Domain Hallucination Test for Large Language Models. (2023)
36. International Organization for Standardization: https://www.iso.org/standard/59752.html.
37. International Organization for Standardization: https://www.iso.org/standard/72704.html.
38. Papadopoulos Korfiatis, A. and M.F. and S.R. and S.A.: PriMock57: A Dataset of Primary Care Mock Consultations. https://github.com/babylonhealth/primock57. (2022)
39. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature. 630, 625–630 (2024). https://doi.org/10.1038/s41586-024-07421-0
40. Lewis, P. et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. (2020)
41. Wei, J. et al.: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. (2022)
42. Jia, M., Duan, J., Song, Y., Wang, J.: medIKAL: Integrating Knowledge Graphs as Assistants of LLMs for Enhanced Clinical Diagnosis on EMRs. (2024)
43. Wang, Y., Ma, X., Chen, W.: Augmenting Black-box LLMs with Medical Textbooks for Clinical Question Answering. (2023)
44. Gilbert, S., Kather, J.N., Hogan, A.: Augmented non-hallucinating large language models as medical information curators. NPJ Digit Med. 7, 100 (2024). https://doi.org/10.1038/s41746-024-01081-0
45. Desmond, M., Ashktorab, Z., Pan, Q., Dugan, C., Johnson, J.M.: EvaluLLM: LLM assisted evaluation of generative outputs. In: Companion Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 30–32. ACM, New York, NY, USA (2024)
46. https://en.wikipedia.org/wiki/Katherine_Johnson: Katherine Johnson

Supplementary materials

Examples of hallucinations in the medical output text:

Fabricated Facts/Negations

Facts (diagnostics, plan, communication, …) that are completely fabricated (were never mentioned)

Type	Reason	Sentence
Minor	clinical assessment not explicitly mentioned but accurate	Impression or clinical assessment: Chronic cough with sputum production and exertional dyspnea, possibly exacerbated by smoking, with a history of asthma.
Major	This is not in the input text shown	Follow up: Urgent referral to emergency department for further evaluation and management.
Minor	Not in the transcript	Treatment plan: Rest, continue with Advil as needed for pain relief.
Minor	Only neck mentioned, no other joints	Pertinent review of systems mentioned: No stiffness or loss of motion in joints.

Negated Events

Clinical events (e.g., symptoms, negations, surgeries, diagnostics) that were mentioned but not recorded

Type	Reason	Sentence
Major	Patient reports tiredness and feeling hot	No other significant symptoms.

Minor	Has had attendance due to peanut allergy	She has not had any surgeries or hospitalisations.

Context Conflation

Topics that were mentioned but mixed up in the note.

Type	**Reason**	**Sentence**
Minor	Exposure to mold and dust is at home and asbestos exposure in living and workplace.	Family/Social History: Works in crowded, poorly conditioned workplace with potential mold, dust, and asbestos exposure.
Minor	The cough preceded the shortness of breath and has been over three weeks.	History of presenting complaint: Shortness of breath and coughing started two weeks ago, worsening over time.
Minor	The past medical history is epilepsy. The missed medication is related to the current presentation.	Past Medical History: Epilepsy, seizure disorder, infrequent seizures; missed medication dose.

Assumed causality

Made-up causality between two things that were mentioned independently

Type	**Reason**	**Sentence**
Major	I can't see in the text that the physician ascribes the pain as 'stress related'.	Impression or clinical assessment: Chest pain of unknown origin, possibly stress related.
Major	Shouldn't make associations without being explicitly mentioned in the transcript	Impression or clinical assessment: Possible upper respiratory tract infection or COPD exacerbation, considering significant smoking history and environmental exposures.

Minor	No mention of smoking exacerbating dyspnoea	Impression or clinical assessment: Chronic cough with sputum production and exertional dyspnea, possibly exacerbated by smoking, with a history of asthma.

Examples of omissions in the medical output text and the section it occurred:

Current Issue

Type	**Reason**	**Sentence**
Major	Palpitations not mentioned. This points the diagnosis towards Pulmonary embolism/infective cause rather than cardiac	Yes, since this morning, it's felt like my heart is racing. Okay. Have you noticed any skin changes at all? Any rashes? No rashes. Okay.
Minor	Blood in sputum important	And has there been any kind of green, slimy mucus in your cough at all? Uh, no, it's been a dry cough. Dry cough. Okay. And have you noticed at all coughing up blood?
Minor	Need to mention if any change in smell/taste as may indicate COVID infection	Uh, no, I haven't noticed that. What about your sense of taste? Are you being, are you able to taste things? Yeah, I think so. Okay.

Information and Plan

Type	**Reason**	**Sentence**
Major	Advice: if the pregnancy test is positive, call straight back to organise be seen. Important safety net	So if you can do that today and if it's positive we need to see you straight away so give us a call back. Okay. Have you got any other questions? Right.
Minor	Explanation of shortness of breath important to include to gauge severity of this	Oh yeah, it's like when I'm doing walking, and I do exercise now and then, so it hasn't been... Super bad, like it hasn't stopped me from doing any of those things, but it's been pretty unusual and I’m a little bit worried.

Minor	Discussing diagnosis	So yeah, based on what we discussed today, it seems that you may have an upper respiratory tract infection.

PMFS Issues

Type	**Reason**	**Sentence**
Major	No allergies is an important negation to record	Okay. Right. Any allergies? Nothing. Yeah. Nothing. No allergies at all. Right here. Fine.
Major	Need to mention smoking history as this is important in someone presenting with chest pain. Is a risk factor for myocardial infarction etc	And do you use any recreational drugs like marijuana? Never. Okay. And do you smoke? Yes, I smoke about two packs of cigarettes per day.
Minor	Smoking status in an asthmatic should be recorded	No, I don't smoke. Do you drink any alcohol regularly? Yeah, one glass. How often is it? One glass of what? One glass of wine. Every... how often did you say?

Prompt comparison

EXPERIMENT 3	EXPERIMENT 8
You are a highly accurate medical office assistant drafting documentation for a physician. Every decision you take is life or death and must be 100% accurate. DO NOT ADD any content that isn't specifically mentioned IN THE TRANSCRIPT. From the attached transcript generate a clinical note based on the below template format for the physician to review, include all the relevant information and do not include any information that isn't explicitly mentioned in the transcript. If nothing is mentioned just return [NOT MENTIONED]. It is vital that all the information in the note is as accurate as possible. Avoid repeating the same information in different sections where possible. Write the note from the perspective of the physician. DO NOT add associate or relate causes for medical conditions unless explicitly specified by the Physician. See below for a template to outline the structure of the output and style preferences to follow.	You are a medical office assistant drafting documentation for a physician. DO NOT ADD any content that isn't specifically mentioned IN THE TRANSCRIPT. From the attached transcript generate a SOAP note based on the below template format for the physician to review, include all the relevant information and do not include any information that isn't explicitly mentioned in the transcript. If nothing is mentioned just return [NOT MENTIONED]. It is VITAL that all the information in the note is as accurate as possible. Avoid repeating the same information in different sections where possible. Write the note from the perspective of the physician. Only include any section of the template if there is information from the transcript, otherwise omit it.
Template: Referral Reason / reason for appointment: History - Allergies - Medications - History of presenting complaint - Past Medical History - Family/Social History - Sensitive information Observations: - Examination findings - Investigation results - Impression or clinical assessment Plan - Planned investigations - Follow up - New prescribed medication or therapies - Communication, reassurance & patient understanding of care - Actions for referrer/GP Clarity - Explained medical terms	Template for Clinical SOAP Note Format: Subjective: - HPI: [include here any mentioned symptoms, chronological narrative of patients complaints, information obtained from other sources(always identify source if not the patient).] - Past medical history. [include here all of the patients past conditions, treatments and encounters, also include relevant social history here including smoking, alcohol, drug use and occupation/travel history] - Review of systems [include here any additional symptoms in other organs that is relevant to the initial presentation] - Current medications [list medicines out each on a seperate line, in a standard format where the information is mentioned: [DRUG NAME][DRUG DOSE][DRUG FREQUENCY][INDICATION]] Objective: - Vital signs [including any mentioned blood pressure, pulse rate, oxygen saturation, temperature] - Physical exam [the examination findings from the physical exam, if mentioned] - Test Results [include in this section any lab test results or imaging reports] Assessment / Problem List: - Assessment: [A one sentence description of the patient and major problem as described by the physician, including the diagnosis the physician has identified] - Problem list: [A numerical list of clinical problems arising from this encounter and active ongoing medical problems the patient has. Present each problem as [Condition][Status:active/suspected/confirmed/past/unknown], list each problem on a separate line, leave status as unknown if not mentioned in the transcript] Plan:[include here any management plan mentioned in the transcript, including patient education, prescriptions, tests, referrals or other plans.] Follow-up: [include here any plan mentioned to see the patient again, or to be discharged.]
Style preferences: - Write from the perspective of the physician (first person) - Be ultra concise - Use bullet points and broken sentences	Please adhere to the following style guidelines: - Write from the perspective of the physician (first person) - Be ultra-concise - Be ultra-precise, do not use generalising terms - Be highly detailed - Include ALL important negations in the relevant sections (e.g. the patient has no fever) the clinician has elicited as well as all positive findings. - Use bullet points and single words, not sentences. - Always list medications in a list in the following format for each one: medicine, dose, frequency, indication - Always document if drug allergies are present or not - Examination findings always refer to a physical exam, only include signs here, not symptoms - Preserve quantities if mentioned in the text

Figure A: Prompt changes proposed that led to the decreased hallucinations and omissions illustrated in Figure 11. Highlighted sections show where the prompt has changed