Journal of the American Medical Informatics Association, 2024, **31(11)**, 2660–2667
https://doi.org/10.1093/jamia/ocae223
Advance access publication 23 August 2024
Research and Applications

Research and Applications

Generating colloquial radiology reports with large language models

Cynthia Crystal Tang, BS[†,1], Supriya Nagesh, PhD[†,2], David A. Fussell, MD[1,*], Justin Glavis-Bloom, MD[1], Nina Mishra, PhD[2], Charles Li, MD[1], Gillean Cortes, DO[1], Robert Hill, MD[1], Jasmine Zhao, MD[1], Angellica Gordon, MD[1], Joshua Wright, MD[1], Hayden Troutt, MPH[1], Rod Tarrago, MD[3], Daniel S. Chow, MD[1]

[1]Department of Radiological Sciences, University of California, Irvine, Irvine, CA 92868, United States, [2]Amazon Web Services, East Palo Alto, CA 94303, United States, [3]Amazon Web Services, Seattle, WA 98121, United States

*Corresponding author: David Fussell, MD, Department of Radiological Sciences, University of California, Irvine Medical Center, 101 The City Drive South, Orange, CA 92868, United States (fusselld@hs.uci.edu)

[†]First and second author contributed equally.

Abstract

Objectives: Patients are increasingly being given direct access to their medical records. However, radiology reports are written for clinicians and typically contain medical jargon, which can be confusing. One solution is for radiologists to provide a "colloquial" version that is accessible to the layperson. Because manually generating these colloquial translations would represent a significant burden for radiologists, a way to automatically produce accurate, accessible patient-facing reports is desired. We propose a novel method to produce colloquial translations of radiology reports by providing specialized prompts to a large language model (LLM).

Materials and Methods: Our method automatically extracts and defines medical terms and includes their definitions in the LLM prompt. Using our method and a naive strategy, translations were generated at 4 different reading levels for 100 de-identified neuroradiology reports from an academic medical center. Translations were evaluated by a panel of radiologists for accuracy, likability, harm potential, and readability.

Results: Our approach translated the Findings and Impression sections at the 8th-grade level with accuracies of 88% and 93%, respectively. Across all grade levels, our approach was 20% more accurate than the baseline method. Overall, translations were more readable than the original reports, as evaluated using standard readability indices.

Conclusion: We find that our translations at the eighth-grade level strike an optimal balance between accuracy and readability. Notably, this corresponds to nationally recognized recommendations for patient-facing health communication. We believe that using this approach to draft patient-accessible reports will benefit patients without significantly increasing the burden on radiologists.

Key words: prompt engineering; radiology; large language model; machine learning; natural language processing.

Introduction

Healthcare organizations are increasingly sharing patient information via online portals to the electronic health record (EHR). While much of the content in an EHR is intended for use by clinicians, patient portals are designed to be a shared resource allowing clinicians and patients to communicate to advance care. As a result, patients have faster and easier access to clinical content including radiology reports. These reports have historically been written to communicate with other health professionals using medical terminology that may not be readily understandable by patients and families.[1] For example, 1 group found that standard radiology reports were written at greater than a 13th-grade reading level.[2] Because many organizations release results rapidly to patients,[3,4] a patient may read a radiology report before the healthcare professional who ordered the study has had an opportunity to review it and communicate with them. Lacking immediate guidance on the content and context of the results, patients turn to readily available but unverified tools such as web search engines, social media, and chatbots.[5] In response, legislation, such as Senate Bill No. 1419 (SB 1419) in California, has been passed to prevent immediate electronic disclosure for certain conditions, such as imaging studies that reveal new or recurrent malignancy. Healthcare organizations are thus challenged to balance improving patient access to clinical content with the reality that patients may have difficulty comprehending that content.

Some researchers have argued that the frequency of worry experienced by patients is low and that most patients prefer to receive results immediately.[6] Some have suggested that radiologists should generate a separate "patient-accessible" summary, though existing high clinical volumes makes adding further tasks unappealing to radiologists.[7] Several organizations have experimented with traditional artificial intelligence in an attempt to improve the readability of radiology reports. Potential approaches are to add glossaries, add

Received: April 22, 2024; **Revised:** August 5, 2024; **Editorial Decision:** August 7, 2024; **Accepted:** August 8, 2024

definitions via hyperlinks, or to automatically replace information with standard concept names, though these approaches may lack specificity to individual patient contexts.[8–10]

More recently, there has been interest in leveraging large language models (LLMs) and generative Artificial Intelligence (Gen AI) to bridge gaps in physician-patient communication and simplify medical terminology for patients.[11–15] Much of the work to date has focused on single models,[16] which involved a small number of reports,[17] including the use of synthetic reports and various scoring methods.[18] Common problems in employing LLMs include "hallucinations," in which false information is presented as truth, and the omission of important information, which may interfere with comprehension or lead patients to inaccurate conclusions.

In this study, we describe a novel method of prompt engineering for colloquial translation of radiology reports that overcomes many of the limitations inherent in a naive prompting strategy. We compare the accuracy, likability, and harm potential of translations generated by our method to those generated by a naive approach and measure readability of the translations using standardized scales. Finally, to verify quality and identify hallucinations, we describe a novel approach for mapping individual clauses from colloquial translations back to the portion of the original report from which they are derived.

Methods

Data description

In this study, we used 100 de-identified, clinically-relevant neuroradiology reports from an academic medical center. Each radiology report contained Indications, Findings, and Impression sections that were used in this work. Real reports were used instead of synthetic data as they better capture clinical contexts that could otherwise be overlooked, unrepresented, or inaccurately interpreted.[19] Synthetic data may not fully encapsulate the spectrum of patient examinations due to their inherit constraints.[20]

Problem statement

Given a radiology report, our goal was to produce a colloquial translation. We posited that a good colloquial translation is concise, easy to understand, accurate, and comments primarily on the abnormal aspects in the report.

Our solution was designed to meet the following requirements: (1) Summarize the original report without clinical jargon; (2) Produce a translation that has readability at the given education level; and (3) Produce a granular attribution mapping from the translation to the original report.

Responsible use of patient data was critical. Consequently, we chose a cloud-based LLM service, Amazon Bedrock (Amazon, Seattle, WA), and an LLM on that provider (Claude v1.3, Anthropic, San Francisco, CA) that guaranteed that it would not further train an LLM on our private data. Moreover, a readability index was computed locally with our own implementation of measures such as Flesch-Kincaid.[21]

Baseline solution: Directly prompting an LLM

As a first attempt at a solution, we provided a prompt to an LLM instructing it to return a translation that meets the above requirements. To produce desirable translations, we provided few-shot examples, or demonstrations of reports and the corresponding translations, within the LLM prompt. See Figure S1 for an example of a prompt to produce a colloquial translation given the original report, the desired education level, and few-shot examples. In this work, we used 5 original reports and translations produced by a radiologist as few-shot examples.

Our solution: Medical knowledge-based prompting

With the baseline method of directly prompting an LLM, we noticed that the LLM produced certain inaccuracies in translations; for example, hematoma was translated to "blood clot." We hypothesized that pointed questions such as "What does hematoma mean in a head CT scan?" to an LLM would result in more accurate responses when compared to the long-form prompt used in the baseline method. Since LLMs such as Claude are designed to understand and generate natural responses, pointed questions may reduce errors in long-form LLM responses.[22]

We implemented several additional enhancements to improve our results (Figure 1), including:

Medical entity extraction

The first step was to obtain a list of medical diagnoses from the reports using a medical entity extractor.[23] See Figure 2 for an example report and extracted entities. For example, if the original report reads as "There is a stable 10 mm left subdural hematoma," our goal in this step is to extract the phrase left subdural hematoma.

Generating colloquial definitions

Our next step was to generate definitions in simple terms for the different entities identified. We generated these definitions through a pointed question given as a prompt to an LLM. See Figure S2 for the prompt used to generate a definition in simple terms for a given entity. Given a list of medical entities from the previous step, we prompted the LLM to generate the definition for each of them. Figure 2 illustrates this step for entities extracted from the example report.

Final translation

The final step in generating the translation was providing the different entities and their definitions in addition to the original report in the prompt to the LLM. Figure S3 illustrates the prompt in the final step to generate a translation.

Results

We generated translations of 100 neuroradiology reports using 4 different combinations in the prompt: (1) baseline (few-shot prompting only); (2) baseline with inclusion of study indications in the prompt; (3) medical knowledge (MK)-based prompt; (4) MK plus inclusion of study indications in the prompt. We assessed the translated radiology reports for accuracy, physician likability, readability, and harm potential. Assessments of accuracy, harm potential, and physician likability for Findings and Impression were performed by radiologists.

Accuracy

Each report's translation was labeled as 0 (inaccurate) or 1 (accurate) by a radiologist. We computed the percentage accuracy across the 100 reports. The accuracy of the translations of the Findings and Impression sections is illustrated in Table 1 and in Figures 3A and B. We found that the Findings

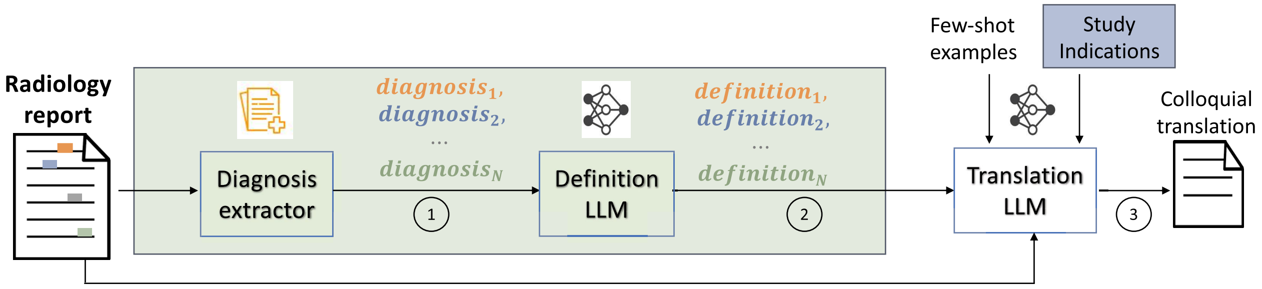


Figure 1. Our translation pipeline: medical knowledge (MK)-based prompting.

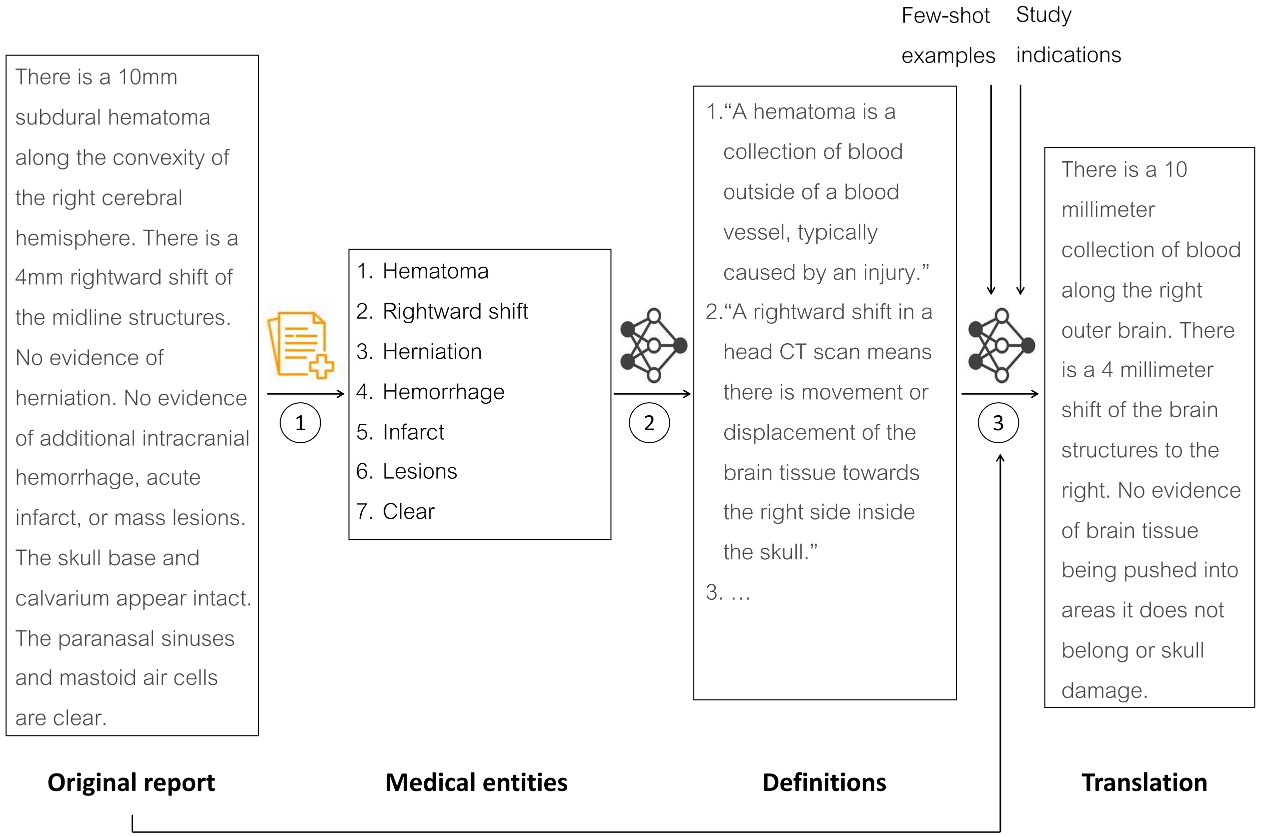

There is a 10mm subdural hematoma along the convexity of the right cerebral hemisphere. There is a 4mm rightward shift of the midline structures. No evidence of herniation. No evidence of additional intracranial hemorrhage, acute infarct, or mass lesions. The skull base and calvarium appear intact. The paranasal sinuses and mastoid air cells are clear.
1
1. Hematoma
2. Rightward shift
3. Herniation
4. Hemorrhage
5. Infarct
6. Lesions
7. Clear
2
1."A hematoma is a collection of blood outside of a blood vessel, typically caused by an injury."
2."A rightward shift in a head CT scan means there is movement or displacement of the brain tissue towards the right side inside the skull."
3. …
Few-shot examples
Study indications
3
There is a 10 millimeter collection of blood along the right outer brain. There is a 4 millimeter shift of the brain structures to the right. No evidence of brain tissue being pushed into areas it does not belong or skull damage.
Original report
Medical entities
Definitions
Translation

Figure 2. Illustration of the steps involved for 1 example report.

Table 1. Accuracy of MK-based prompting with indications and baseline methods.

	Grade	Baseline	MK + Indications	*P*-value
Findings	Fifth	60.0%	83.0%	**.0006**
	Eighth	65.0%	88.0%	**.0002**
	12th	68.0%	87.0%	**.0023**
	College	68.7%	87.0%	**.0033**
Impression	Fifth	82.0%	90.0%	.32
	Eighth	83.0%	93.0%	**.05**
	12th	84.0%	94.0%	**.04**
	College	91.0%	98.0%	1

Bold indicates significance with $P <= .05$.

accuracy improved as we increased the grade level of the translation. Across the different methods, we found that the baseline method (few-shot prompting only) achieved an accuracy of 60%-68%, depending on targeted grade level. Including the study indications in the prompt increased the accuracy consistently across all the grade levels. We found a global notable improvement by using MK-based prompting along with study indications. For instance, the accuracy for the college grade level improved from 68.7% to 87.0%.

We see the same pattern for Impression sections in terms of the accuracy improving with an increase in grade level, and the accuracy improving with MK-based prompting along with indications (Figure 3B). We observe that the accuracy of translating the shorter Impression is substantially better than the longer Findings, with the best accuracy of Impression at college level being 98% compared to 87% for Findings. This may be related to the LLM having a longer text to translate.

Physician likability

Radiologists rated each of the translations on a 5-point Likert scale for likability. Figures 3C and D illustrate the average likability across all the reports at different grade levels on the Findings and Impression. We observe that, in the case of both

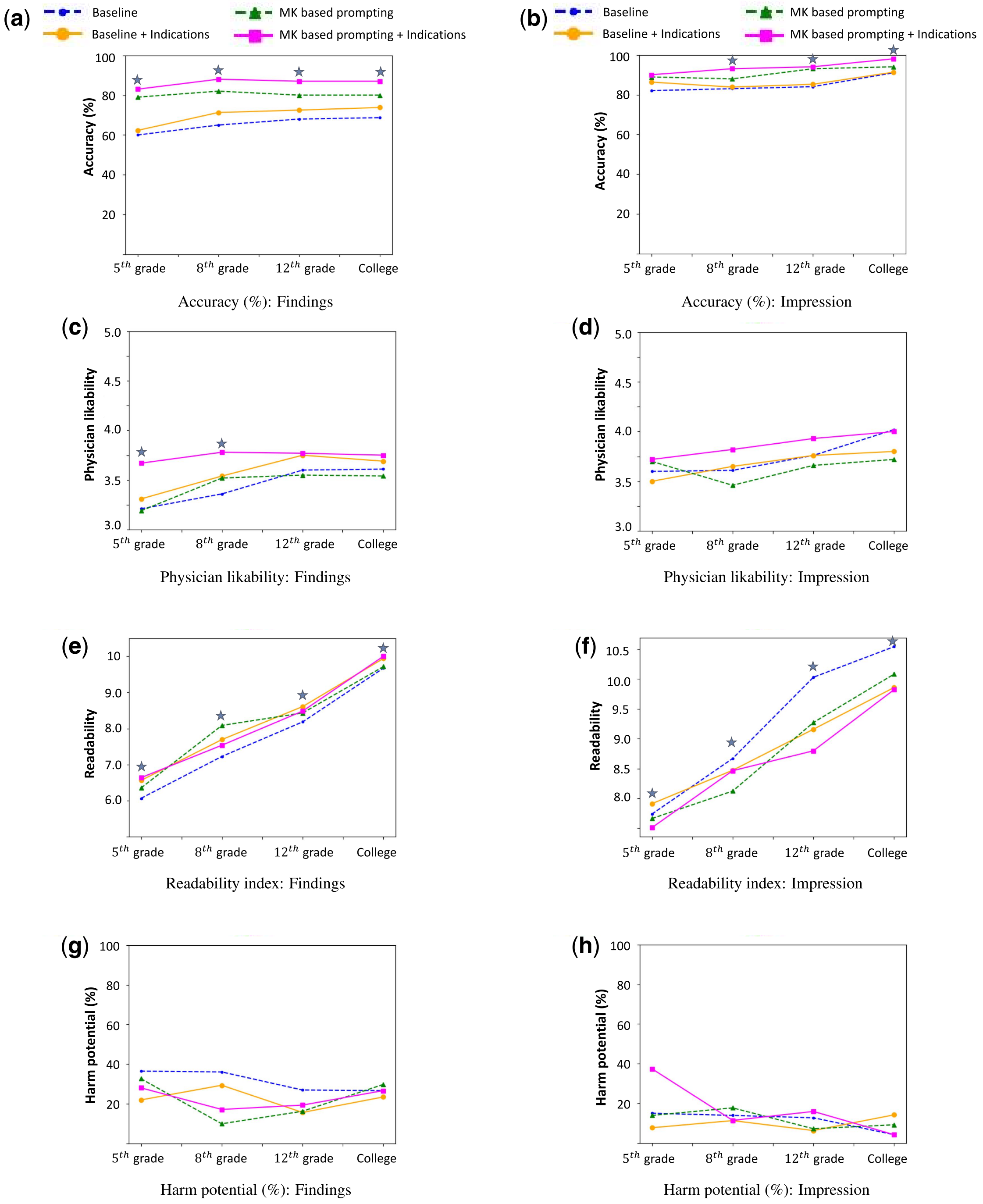


Figure 3. (A and B): Accuracy of translations generated by each prompting method at 4 different education levels. (C and D): Physician likability of translations generated by each prompting method at 4 different education levels. Higher is better. (E and F): Readability index of translations generated by each prompting method at 4 different education levels. Lower readability index corresponds to simpler language. (G and H): Harm potential of translations generated by each prompting method at 4 different education levels. Lower is better. In all panels, asterisk indicates statistical significance for MK + Indications model vs baseline ($P < .05$).

Findings and Impression, the MK-based prompting consistently scored higher than the other methods. Physicians also tended to prefer the reports written at a higher grade level, such as 12th grade or college level.

Readability

We assessed readability index (RI) by computing the mean of 5 standard readability scores: Flesch-Kincaid,[21] Gunning-Fog,[24] SMOG,[25] Coleman-Liau,[26] and Automated Readability Index.[27] The RI is an estimate of the grade level required for comprehension, such that a lower RI represents more readable text. Mean RI of Findings sections from the original reports was 14.06 and that of Impression sections was 14.31. RI of translations produced by the baseline and MK methods ranged from 6.64 to 10.00 depending on target grade level (Figures 3E and 3F); all translations had significantly lower RI, or were more readable, than original reports ($P < .001$; Table S2).

Harm potential

Each translation was evaluated by a radiologist for potential harm it could cause and labeled as either 0 (not harmful) or 1 (harmful). A harmful translation is one that might exclude an important aspect of a patient's report, or one that might include incorrect information. The average harm potential of the translated Findings and Impression sections is shown in Figures 3G and H. Lower average harm potential is better. In our study, translations of Impression sections had lower average harm potential than those of Findings sections, consistent with the results for accuracy. Among the different methods, we found lower harm potential when we included MK prompting. Importantly, by adding medical knowledge to the prompt, we saw moderate decreases in harm potential; for example, at the eighth-grade level, harm potential of Findings section translations decreased from 36% to 10%. Nevertheless, some harm potential is present in translations across all the methods. This is significant as these translations are meant to be shared with patients.

Discussion

In this study, we evaluated the use of a general purpose LLM to generate colloquial translations of neuroradiology reports. Our findings indicate that, as compared with a naive method, careful prompt engineering can improve the accuracy and readability of translations while reducing their harm potential. Our findings are consistent with previous work showing that careful prompt engineering improved LLM accuracy compared to baseline.[28,29] A downside is that there exists a potential for oversimplification of radiology reports. This, in turn, creates increased risk of inaccuracy through hallucinations or information omission.[30]

Readability vs accuracy

We observe that readability, or ease of reading, and accuracy ratings were inversely related; that is, as higher grade levels were targeted, translations became more accurate at the cost of being less readable. However, gains in accuracy above the eighth-grade level were marginal, suggesting that targeting this level may offer the best combination of accuracy and readability for patients. Notably, the eighth-grade level corresponds to nationally recognized recommendations for patient-facing public health communication.[31]

Improved accuracy with better prompting

As expected, the inclusion of contextual information, like clinical indications, further improved overall accuracy of radiology reports simplified by the LLM. Clinical indications are concise and directly relevant to the input of radiology reports; including them in the prompt helps focus the LLM and limit distractions from irrelevant information.[32] This corroborates findings of LLMs producing more relevant output when provided with contextual information, as explored across other domains.[33]

Translation accuracy: Findings vs impression

The Impression section of a radiology report is essentially a short summary of the more detailed Findings section. In our study, translations of the more concise Impression sections had greater average accuracy; this association has been substantiated in other, non-radiology healthcare-related domains.[33] In addition to being more accurate, Impression section translations had significantly lower word counts, which further contributes to improved patient comprehension.[34]

Detecting the quality of translation through attribution

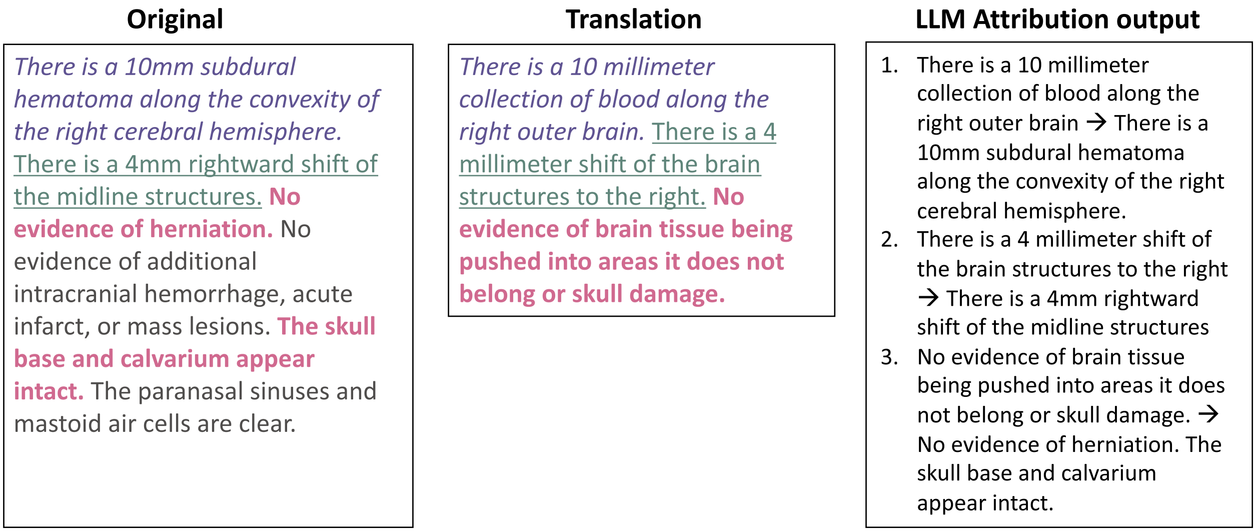
Despite reductions in harm potential using our method, relatively high rates of potential harm persisted, particularly in translating the Findings (Figure 3G). This precludes immediate use of this tool in a clinical setting unless other "harm checks" are included. Potential harm in a translation could be caused by incorrect translation of sentences in the report, omission of important aspects of the original report, or addition of new concepts to the translation (hallucination).

In this study, we undertook preliminary work to detect sources of harm and thus reduce the risk of patients receiving inaccurate translations. The objective is to automatically attribute each sentence in the translation to one in the original report (Figures 4 and 5). This is done by providing the colloquial translation and the original report to an LLM that is prompted to produce pairs of sentences from the translation and original report. The prompt used here is shown in Figure S4.

First, attributions help identify any incorrect translations of sentences. In Figure 4, the attribution outputs produced by the LLM are used to format-match the translation and original report. For example, "There is a 10 millimeter collection of blood along the right outer brain" is attributed to "There is a 10 mm subdural hematoma along the convexity of the right cerebral hemisphere" in the original report. This is a way to check for the first source of harm: incorrect translations. Second, attributions help with detecting the omission of parts of the original report. In Figure 4, we see sentences in the original report such as "No evidence of additional intracranial hemorrhage, acute infarct, or mass lesions," which are not attributed. This indicates that these parts of the original report are not included in the translation. While the omitted sentences in this example only describe normal findings, a scenario where a patient's tumor size increased but is not included in the translation would be detected in this manner and potentially corrected. Finally, we used attribution to detect the addition of new concepts (hallucinations) to the translation. Figure 5 shows an example where the original Findings reads "Please see impression." The LLM is only given the Findings to produce the translation. However, the generated translation here includes multiple new concepts or hallucinations. From the attribution output, we find that none of the sentences in the translation can be attributed back to the Findings. This is one way to detect and screen out potential hallucinations.

We believe that attribution is potentially a useful way for an end user to quickly identify any inaccuracies in the translation. This initial work on attribution for detecting inaccuracies can serve as a basis for future work.

Our study had several limitations. First, the number of radiology reports included was limited to 100 reports total as each report required extensive time to review by radiologists. Given the multiple iterations of reading grade levels and translations before and after prompt-engineering and context information, each report required multiple evaluations. We considered possible effects, including reader fatigue, to reduce confounding variables. Future studies may include a greater number of reports and increase the number of readers. Second, this study evaluated just 1 LLM, Claude v1.3. Ideally, multiple LLMs would have been compared. However, there are practical



Original

There is a 10mm subdural hematoma along the convexity of the right cerebral hemisphere. There is a 4mm rightward shift of the midline structures. **No evidence of herniation.** No evidence of additional intracranial hemorrhage, acute infarct, or mass lesions. **The skull base and calvarium appear intact.** The paranasal sinuses and mastoid air cells are clear.

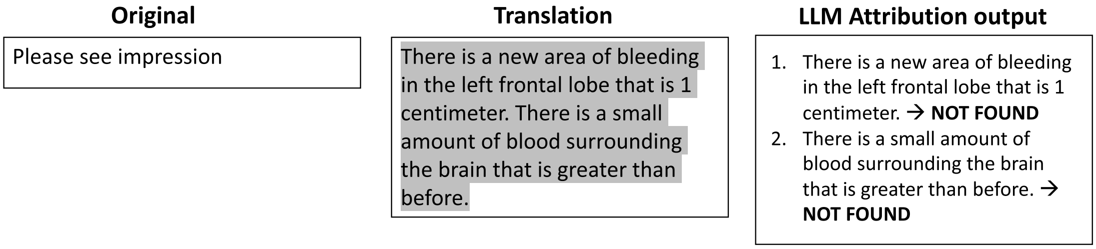
Translation

There is a 10 millimeter collection of blood along the right outer brain. There is a 4 millimeter shift of the brain structures to the right. **No evidence of brain tissue being pushed into areas it does not belong or skull damage.**

LLM Attribution output

1. There is a 10 millimeter collection of blood along the right outer brain → There is a 10mm subdural hematoma along the convexity of the right cerebral hemisphere.
2. There is a 4 millimeter shift of the brain structures to the right → There is a 4mm rightward shift of the midline structures
3. No evidence of brain tissue being pushed into areas it does not belong or skull damage. → No evidence of herniation. The skull base and calvarium appear intact.

Figure 4. Example of attribution when the translation is accurate. The attribution is shown by color coding the sentences, for example, the text in the translation is attributed to the text in the original report. Note that there is text in the original report that is not used for attribution, meaning that it is not present in the translation.



Original

Please see impression

Translation

There is a new area of bleeding in the left frontal lobe that is 1 centimeter. There is a small amount of blood surrounding the brain that is greater than before.

LLM Attribution output

1. There is a new area of bleeding in the left frontal lobe that is 1 centimeter. → **NOT FOUND**
2. There is a small amount of blood surrounding the brain that is greater than before. → **NOT FOUND**

Figure 5. Example of attribution when there is hallucination in the translation. The attribution output for each of the sentences in the translation says "NOT FOUND."

constraints to evaluating additional models due to the substantial radiologist time required—1600 reports evaluated for 1 LLM. Moreover, a different LLM, which was not trained on private data, could have been chosen to conduct the study, such as open source LLMs or GPT 4. Claude v1.3 was chosen because past work suggests it is competitive. According to 2 studies,[38,39] Claude v1.3 ranks second or third when compared to numerous other LLMs. In another work,[40] the authors propose a method to reduce inference time in which Claude v1.3 performs better than vanilla-GPT 4. Future work is needed to compare other open source and GPT models. Next, the included radiology reports were limited to CT scans of the head. Future studies may include greater variety of imaging types across other clinical domains. Finally, our study specifies only textual data in the form of clinical indications and radiology reports; future work may incorporate actual imaging data to aid colloquial translation. We emphasize the capabilities of text in LLMs, although future studies may investigate potential improvement by incorporating other image-based deep learning tools.

Conclusion

Pilot studies of the patient experience reading and comprehending radiology reports have found a preference for simplified or "translated" layman reports.[35] Proposed workflow solutions suggest that radiologists directly modify writing style of reports to favor readability. However, this conflicts with widely recognized radiology guidelines emphasizing the importance of precise anatomic and radiologic terminology for medical clarity. The results of this study indicate great potential for LLMs to generate the colloquial reports preferred by patients without disrupting workflows for clinical stakeholders, including radiologists, referring physicians, and insurers.[36] More work must be done to investigate the use of LLMs to translate content from medical to non-medical language, since current regulations apply to machine translation from 1 language to another, rather than from medical to colloquial language.[37] Further study of tools such as attribution is also needed to foster trust from clinicians and patients who are concerned about LLM-generated hallucinations.

Author contributions

Cynthia Crystal Tang, Supriya Nagesh, David Fussell, Justin Glavis-Bloom, Nina Mishra, Charles Li, Robert Hill, and Daniel S. Chow all contributed to the main manuscript text. David Fussell performed statistical analysis. Gillean Cortes, Robert Hill, Jasmine Zhao, Angellica Gordon, and Joshua Wright contributed to report analysis. All authors reviewed the manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

The study received no outside funding.

Conflicts of interest

Authors S.N., N.M., and R.T. are employed by Amazon Web Services.

Data availability

Patient data analyzed as part of this study are not publicly available to protect participant privacy. Data generated by physician raters areavailable by request.

Ethics approval

The study was approved by the UC Irvine institutional review board.

References

1. Trofimova A, Vey BL, Safdar NM, et al. Radiology report readability: an opportunity to improve patient communication. *J Am Coll Radiol*. 2018;15(8):1182-1184.
2. Patil S, Yacoub JH, Geng X, et al. Radiology reporting in the era of patient-centered care: how can we improve readability? *J Digit Imaging*. 2021;34(2):367-373.
3. Mehan WA, Brink JA, Hirsch JA. 21st century Cures Act: patient-facing implications of information blocking. *J Am Coll Radiol*. 2021;18(7):1012-1016.
4. Johnson AJ, Easterling D, Nelson R, et al. Access to radiologic reports via a patient portal: clinical simulations to investigate patient preferences. *J Am Coll Radiol*. 2012;9(4):256-263.
5. Alarifi M, Patrick T, Jabour A, et al. Understanding patient needs and gaps in radiology reports through online discussion forum analysis. *Insights Imaging*. 2021;12(1):50-59.
6. Steitz BD, Turer RW, Lin CT, et al. Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Netw Open*. 2023;6(3):e233572.
7. Amin K, Khosla P, Doshi R, et al. Focus: big data: artificial intelligence to improve patient understanding of radiology reports. *Yale J Biol Med*. 2023;96(3):407-417.
8. Cook TS, Oh SC, Kahn CE. Patients' use and evaluation of an online system to annotate radiology reports with lay language definitions. *Acad Radiol*. 2017;24(9):1169-1174.
9. Oh SC, Cook TS, Kahn CE. PORTER: a prototype system for patient-oriented radiology reporting. *J Digit Imaging*. 2016;29 (4):450-454.
10. Qenam B, Kim TY, Carroll MJ, et al. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *J Med Internet Res*. 2017;19(12):e8536.
11. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nature Medicine*. 2023;29(8):1930-1940.
12. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Medicine*. 2023;6(1):120-126.
13. Tippareddy C, Jiang S, Bera K, et al. Radiology reading room for the future: harnessing the power of large language models like ChatGPT. *Curr Probl Diagn Radiol*. 2023; in press.
14. Doshi-Velez F, Kim B. 2017. Towards a rigorous science of interpretable machine learning. arXiv, arXiv:1702.08608, preprint: not peer reviewed.
15. Pons E, Braun LMM, Hunink MGM, et al. Natural language processing in radiology: a systematic review. *Radiology*. 2016;279 (2):329-343.
16. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art*. 2023;6(1):9.
17. Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. *Cureus*. 2023;15(12):e50881.
18. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. 2023;34(5):2817-2825.
19. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLOS Digit Health*. 2023;2(1):e0000082.
20. Rothrock SG, Rothrock AN, Swetland SB, et al. Quality, trustworthiness, readability, and accuracy of medical information regarding common pediatric emergency medicine-related complaints on the web. *J Emerg Med*. 2019;57(4):469-477.
21. Kincaid J, Fishburne R, Rogers R, et al. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) for navy enlisted personnel. Accessed January 3, 2024, https://apps.dtic.mil/sti/citations/ADA006655.
22. Dhuliawala S, Komeili M, Xu J, et al. 2023. Chain-of-verification reduces hallucination in large language models. arXiv, arXiv:230911495, preprint: not peer reviewed.
23. Bhatia P, Celikkaya B, Khalilia M, et al. Comprehend medical: A named entity recognition and relationship extraction web service. In: *Proceedings – 18th IEEE International Conference on Machine Learning and Applications*, IEEE ICMLA 2019, Boca Raton, FL, USA. 2019:1844-1851.
24. Gunning. *The technique of clear writing*. Rev ed. McGraw-Hill; 1968.
25. McLaughlin GH. SMOG grading-a new readability formula. *J Read*. 1969;12:639-646.
26. Coleman M, Liau TL. A computer readability formula designed for machine scoring. *J Appl Psychol*. 1975;60(2):283-284.
27. Smith EA, Senter RJ. Automated readability index. *Amrl Tr*. 1967 May:1-14.
28. Wang J, Shi E, Yu S, et al. 2023. Prompt engineering for healthcare: methodologies and applications. arXiv, arXiv:230414670, preprint: not peer reviewed.
29. Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*. 2023;55(9):1-35.
30. Olthof AW, van Ooijen PMA, Cornelissen LJ. Deep learning-based natural language processing in radiology: the impact of report complexity, disease prevalence, dataset size, and algorithm type on model performance. *J Med Syst*. 2021;45(10):91.
31. Rooney MK, Santiago G, Perni S, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J Patient Exp*. 2021;8:2374373521998847.
32. Shi F, Chen X, Misra K, et al. *Large Language Models Can Be Easily Distracted by Irrelevant Context*. In: *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 31210-31227.
33. Davis R, Eppler M, Ayo-Ajibola O, et al. Evaluating the effectiveness of artificial intelligence-powered large language models application in disseminating appropriate and readable health information in urology. *J Urol*. 2023;210(4):688-694.
34. Martin-Carreras T, Cook TS, Kahn CE. Readability of radiology reports: implications for patient-centered care. *Clin Imaging*. 2019;54:116-120.
35. Cabarrus M, Naeger DM, Rybkin A, et al. Patients prefer results from the ordering provider and access to their radiology reports. *J Am Coll Radiol*. 2015;12(6):556-562.
36. Hall FM. The radiology report of the future. *Radiology*. 2009;251 (2):313-316.
37. Youdelman M, Turner W, Coursolle A, et al. *Questions and Answers on the 2022 Proposed Rule Addressing Nondiscrimination Protections under the ACA's Section 1557*. National Health Law Program; 2022.

38. Fu Y, Ou L, Chen M, Wan Y, Peng H, Khot T. 2023. Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance. arXiv, arXiv:2305.17306, preprint: not peer reviewed.
39. Sun L, Han Y, Zhao Z, et al. Scieval: a multi-level large language model evaluation benchmark for scientific research. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2024;38(17):19053-19061.
40. Jiang H, Wu Q, Lin CY, Yang Y, Qiu L. 2023. Llmlingua: Compressing prompts for accelerated inference of large language models. arXiv, arXiv:2310.05736, preprint: not peer reviewed.

Journal of the American Medical Informatics Association, 2024, 31, 2660–2667
https://doi.org/10.1093/jamia/ocae223
Research and Applications