Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Assessing Completeness of Clinical Histories Accompanying Imaging Orders Using Adapted Open-Source and Closed-Source Large Language Models

David B. Larson, MD, MBA[1,2] • *Arogya Koirala, MS*[1,2] • *Lina Y. Cheuy, PhD*[1,2] • *Magdalini Paschali, PhD*[1,2] • *Dave Van Veen, PhD*[3] • *Hye Sun Na*[1,2] • *Matthew B. Petterson, MD*[1] • *Zhongnan Fang, PhD*[1,2] • *Akshay S. Chaudhari, PhD*[1,2,4]

Author affiliations, funding, and conflicts of interest are listed at the end of this article.

Radiology 2025; 314(2):e241051 • https://doi.org/10.1148/radiol.241051 • Content code: AI

Background: Incomplete clinical histories are a well-known problem in radiology. Previous dedicated quality improvement efforts focusing on reproducible assessments of the completeness of free-text clinical histories have relied on tedious manual analysis.

Purpose: To adapt and evaluate open-source and closed-source large language models (LLMs) for their ability to automatically extract clinical history elements within imaging orders and to use the best-performing adapted open-source model to assess the completeness of a large sample of clinical histories as a benchmark for clinical practice.

Materials and Methods: This retrospective single-site study used previously extracted information accompanying CT, MRI, US, and radiography orders from August 2020 to May 2022 at an adult and pediatric emergency department of a 613-bed tertiary academic medical center. Two open-source (Llama 2–7B [Meta], Mistral-7B [Mistral AI]) and one closed-source (GPT-4 Turbo [OpenAI]) LLMs were adapted using prompt engineering, in-context learning, and fine-tuning (open-source only) to extract the elements "past medical history," "what," "when," "where," and "clinical concern" from clinical histories. Model performance, interreader agreement using Cohen κ (none to slight, 0.01–0.20; fair, 0.21–0.40; moderate, 0.41–0.60; substantial, 0.61–0.80; almost perfect, 0.81–1.00), and semantic similarity between the models and the adjudicated manual annotations of two board-certified radiologists with 16 and 3 years of postfellowship experience, respectively, were assessed using accuracy, Cohen κ, and BERTScore, an LLM metric that quantifies how well two pieces of text convey the same meaning; 95% CIs were also calculated. The best-performing open-source model was then used to assess completeness on a large dataset of unannotated clinical histories.

Results: A total of 50 186 clinical histories were included (794 training, 150 validation, 300 initial testing, 48 942 real-world application). Of the two open-source models, Mistral-7B outperformed Llama 2–7B in assessing completeness and was further fine-tuned. Both Mistral-7B and GPT-4 Turbo showed substantial overall agreement with radiologists (mean κ, 0.73 [95% CI: 0.67, 0.78] to 0.77 [95% CI: 0.71, 0.82]) and adjudicated annotations (mean BERTScore, 0.96 [95% CI: 0.96, 0.97] for both models; $P = .38$). Mistral-7B also rivaled GPT-4 Turbo in performance (weighted overall mean accuracy, 91% [95% CI: 89, 93] vs 92% [95% CI: 90, 94]; $P = .31$) despite being a smaller model. Using Mistral-7B, 26.2% (12 803 of 48 942) of unannotated clinical histories were found to contain all five elements.

Conclusion: An easily deployable fine-tuned open-source LLM (Mistral-7B), rivaling GPT-4 Turbo in performance, could effectively extract clinical history elements with substantial agreement with radiologists and produce a benchmark for completeness of a large sample of clinical histories. The model and code will be fully open-sourced.

Supplemental material is available for this article.

The clinical history accompanying an imaging order has long been recognized as important for helping the radiologist understand the clinical context needed to provide an accurate and relevant diagnosis (1). Clinical histories that are accurate, relevant, complete, and concise have been shown to help improve key aspects of the radiology report, including accuracy of detection and characterization of abnormalities (2,3). A brief clinical history directly provided by the clinician communicates the clinician's specific thought process, including the most relevant background information and clinical question to be answered.

Despite their usefulness and recognition that the provision of informative clinical histories results in more accurate and contextually relevant radiology reports, clinical histories provided to interpreting radiologists are notoriously incomplete or unhelpful—a problem that has plagued radiology since the specialty's inception (4,5). Although this has been shown to be amenable to improvement with dedicated quality improvement efforts, challenges with sustained improvements and broad adoption of such measures have remained, leaving a gap for an easily adoptable solution (6–11).

Such efforts depend on reproducible assessments of the completeness of clinical histories to help facilitate improvement. Currently, this depends on manual review, which is time-consuming and expensive. However, recent advances in large language models (LLMs) may enable this quality assessment activity to be automated. Recent studies (12,13) have shown how fine-tuning LLMs with human annotations and prompt engineering can help align model outputs with the expected outputs in a variety of tasks, including tasks for radiology. LLMs have even outperformed human experts in summarizing clinical text across multiple tasks (12,14–18).

To determine the ability of an LLM to assess the completeness of clinical histories accompanying imaging orders, this study aimed to *(a)* adapt open-source and closed-source (proprietary) LLMs to assess the completeness of clinical histories based on five key elements: "past medical history," "what," "when," "where,"

Abbreviation

LLM = large language model

Summary

A fine-tuned, open-source large language model (Mistral-7B; Mistral AI) effectively extracted clinical history elements from imaging orders with substantial agreement with radiologists and rivaled a closed-source model (GPT-4 Turbo; OpenAI).

Key Results

- Three adapted large language models, two open-source (Llama 2–7B [Meta] and Mistral-7B [Mistral AI]) and one closed-source (GPT-4-Turbo [OpenAI]), extracted "past medical history," "what," "where," "when," and "clinical concern" from clinical histories accompanying emergency department imaging orders (CT, MRI, US, radiography).
- After in-context learning, Mistral outperformed Llama and was further fine-tuned; both fine-tuned Mistral and GPT (after prompt engineering) showed substantial overall agreement with radiologists (mean κ, 0.73 [95% CI: 0.67, 0.78] to 0.77 [95% CI: 0.71, 0.82]) and adjudicated annotations (mean BERTScore, 0.96 [95% CI: 0.96, 0.97] for both models; $P = .38$), with Mistral rivaling GPT in accuracy (weighted mean, 91% [95% CI: 89, 93] vs 92% [95% CI: 90, 94]; $P = .31$).
- Using Mistral, 26.2% (12 803 of 48 942) of emergency department clinical histories were found to contain all five elements.

and "clinical concern"; *(b)* evaluate model performance in extracting each key element of the clinical history; and *(c)* use the best-performing adapted open-source LLM to assess the completeness of a sample of approximately 50 000 clinical history entries from the emergency department of a large academic medical center to serve as a benchmark for clinical practice.

Materials and Methods

Clinical Context

This retrospective single-site study used previously extracted information accompanying CT, MRI, US, and radiography orders from August 2020 to May 2022 at an adult and pediatric emergency department of a 613-bed tertiary academic medical center (Stanford Medical Center). This information was previously extracted to evaluate the clinical histories provided from all ordering providers for all types of examinations. To evenly weight the representation of all ordering clinicians, a sampling approach was used to reduce the dataset size. The clinical histories for this time period were first separated into radiography orders and cross-sectional imaging orders (CT, MRI, US). Then, clinical histories accompanying one radiograph order and one cross-sectional imaging order from each authorized ordering clinician were extracted for use in the study. Because all data were de-identified, the study did not constitute human subjects research, and institutional review board oversight was waived.

In 2018, the imaging order process at the study institution introduced a prepopulated smart-phrase comment field with prompted text: "CLINICAL HISTORY: what, when, where, concern for. (eg, 2 days RLQ pain, concern for appendicitis)" (Fig 1). However, ordering clinicians can choose to delete phrases, provide other information (including no information), and/or select other comment fields, which are often automatically populated with "See Comments," referring to the previous comment field. Thus, the included clinical history fields varied based on ordering clinician preference, which raises the question of whether the five clinical history elements can still be extracted in an automated fashion on a large scale.

Datasets and Study Design

Training and validation sets.—De-identified clinical history entries were extracted from the radiology information system. Nonduplicated entries by date were selected for model adaptation in a training set and validation set (Table 1). These entries were manually annotated by two board-certified radiologists (D.B.L. and M.B.P., with 16 and 3 years of postfellowship experience, respectively), with one radiologist (M.B.P.) annotating 80% (756 of 944) of the entries due to radiologist availability. Five elements, based on the study by Hawkins et al (6) and modified by Bor et al (19), were extracted from each entry as follows: *(a)* relevant past medical history; *(b)* "what" (nature of the symptoms, description of injury, or cause for clinical concern); *(c)* "when" (duration of symptoms or time of injury); *(d)* "where" (focal site of pain or abnormality, if applicable); and *(e)* "clinical concern" (the ordering provider's concern, working differential diagnosis, or main clinical question).

Clinical History: Relevant PMH ***. Presents with *** for a duration of ***. Specific location of issue (if applicable): ***. Concern for ***.

Figure 1: The prepopulated comment field accompanying imaging orders at the emergency department of the study institution, where "***" represents required text.

Table 1: Modality Distribution of the Datasets

Dataset	Clinical History Entries	CT	MRI	US	Radiography
Model adaptation	944	284 (30)	21 (2.2)	129 (14)	510 (54)
Training	794	245 (31)	20 (2.5)	100 (13)	429 (54)
Validation	150	39 (26)	1 (0.7)	29 (19)	81 (54)
Initial testing	300	75 (25)	75 (25)	75 (25)	75 (25)
Real-world application	48 942	15 255 (31.2)	872 (1.8)	5687 (11.6)	27 128 (55.4)

Note.—Data in parentheses are percentages. The first nonduplicated entries ($n = 944$) by date were selected for model adaptation including prompt engineering, in-context learning, and fine-tuning, and their combinations for training and validation. A separate modality-balanced evaluation dataset from the most-recent clinical history entries was then used for initial testing to assess model performance, interreader agreement, and semantic similarity between the models and the radiologist's manual annotations. The best-performing open-source model was applied on the remaining unannotated de-identified clinical history entries (representing a real-world application).

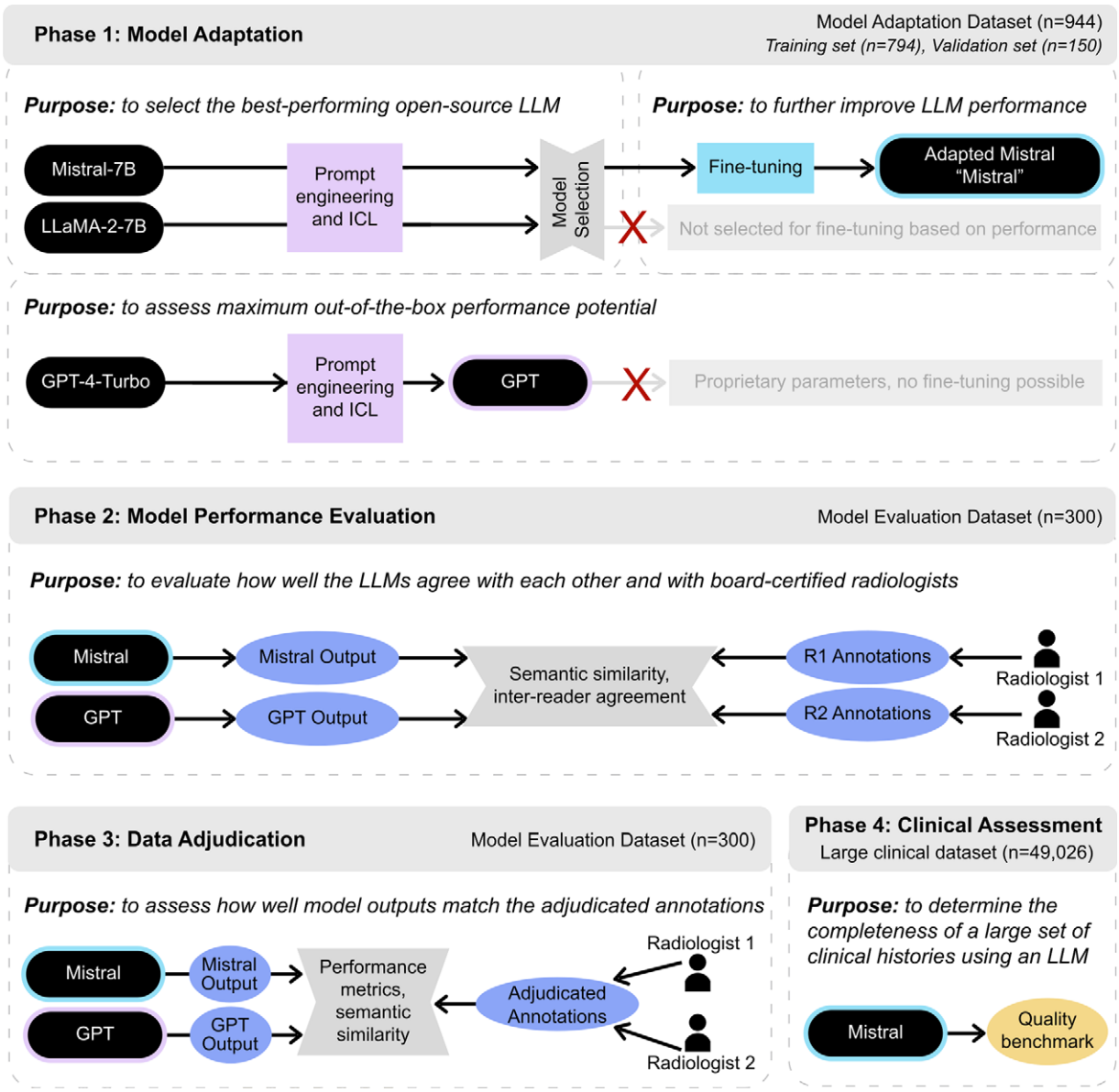

Figure 2: Schematic shows the overall study design, including the goal of each phase, which models were developed with which model adaptation strategy, and the datasets used in each phase of the study. A total of 365 097 previously extracted clinical histories (123 553 for CT, 7401 for MRI, 37 146 for US, 196 987 for radiography) were separated into radiography orders and cross-sectional imaging orders (CT, MRI, US). Then, clinical histories accompanying one radiography order and one cross-sectional imaging order from each authorized ordering clinician were extracted for use in the study, resulting in 50 186 clinical histories for this study (794 training, 150 validation, 300 initial testing, 48 942 unannotated clinical histories representing a real-world application). ICL = in-context learning, LLM = large language model, R1 = radiologist 1, R2 = radiologist 2.

Initial test set.—A separate modality-balanced evaluation dataset (initial test set; n = 75 each for CT, MRI, US, and radiography) was curated from the most recent clinical history entries, as temporal separation helps assess real-world performance and robustness against data shifts (eg, in patient demographics or clinical workflows).

These three datasets (training, validation, and initial testing) were used for different phases of the study (Fig 2).

Phase 1: Model Adaptation

All LLMs require adaptation for specific tasks. Prompt engineering, in-context learning, and fine-tuning and their combinations were explored using the training and validation sets. First, prompt engineering and in-context learning were performed for two open-source models, Llama 2–7B (Meta) (20) and Mistral-7B (Mistral AI) (21), which were considered state of the art at the time of this writing; open-source models are advantageous in cost, compute requirements, and ability to be deployed locally (22,23). The best-performing open-source model was then selected for fine-tuning to further improve performance. One larger closed-source model, GPT-4 Turbo (OpenAI) (24), with proven performance was included to understand the maximum out-of-the-box performance potential for this task (25). Since the GPT-4 Turbo parameters are not publicly disclosed, GPT-4 Turbo only underwent prompt engineering and in-context learning, as fine-tuning was not possible.

Additional model adaptation details, including prompt structure design, in-context learning process, and fine-tuning parameters are available in Appendix S1. Model details can be found in Table 2. The model code can be found on the Stanford AIDE Lab github repository: *https://github.com/stanfordaide*.

Table 2: Details of the Models Evaluated

Model	Type	Developer	No. of Parameters	Maximum Input Sequence Length	Access Date
GPT-4 Turbo	Closed source	OpenAI	Not publicized	128 000	January 24–26, 2024
Llama 2–7B	Open source	Meta	7 billion	4096	January 24, 2024
Mistral-7B	Open source	Mistral AI	7 billion	4096	January 24, 2024

Phase 2: Agreement Evaluation

To evaluate model agreement with radiologists, the best-performing open-source model and GPT-4 Turbo were applied to the initial test set, which the two radiologists independently annotated using Microsoft Excel (version 2018). Because the data were already de-identified, no further blinding was needed. Consensus agreements were obtained in phase 3 of the study through adjudication, as described in the next section. Each clinical history element was first categorized as "included" or "not included" for interreader agreement analysis. If no localization was needed for "where," such as for cough or fever, "not applicable" was noted. "No relevant past medical history" was noted when the clinical history specifically indicated no past medical history relevant to the current imaging study. Weighted and unweighted overall values were computed for all elements regardless of label using radiologist-assigned weights based on each element's perceived importance for clinical diagnosis (past medical history, 0.15; what, 0.4; when, 0.1; where, 0.15; and clinical concern, 0.2).

Then, the clinical history elements (past medical history, what, where, when, and clinical concern) were extracted. Semantic similarity between model outputs and radiologist annotations was quantified using BERTScore, an LLM metric that quantifies how well two pieces of text convey the same meaning (26). Unweighted and weighted overall values were computed only for elements labeled as "included."

Phase 3: Annotation Adjudication

To simplify model performance evaluations, the two radiologists adjudicated their manual annotations (n = 300) and discussed any disagreements to reach consensus agreements. The performance metrics of sensitivity, specificity, precision, and accuracy, along with BERTScore, were computed using model outputs and the adjudicated annotations (from the initial test set). All metric definitions, adjudication, and consensus agreement details are provided in Appendix S1.

Phase 4: Assessment of Clinical Performance Using a Real-world Application

To determine the percentage of complete clinical histories at the study institution, the best-performing open-source model was applied on the remaining unannotated de-identified clinical history entries (representing a real-world application). To assess the clinical history entries, completeness was defined as the presence of all five clinical history elements (past medical history, what, where, when, and clinical concern).

Statistical Analysis

Statistical analyses were performed (Z.F.) using Python version 3.8 *(https://www.python.org/downloads/release/python-380/)* and statsmodels version 0.14.1 *(https://www.statsmodels.org/stable/index.html)*. The 95% CIs were calculated by bootstrapping, and differences between the models and radiologists were assessed through either bootstrapping hypothesis testing (sensitivity, specificity, precision, accuracy, and interreader agreement) or t tests (BERTScore), with a P value threshold of .05. Interreader agreement was evaluated using Cohen κ (no to slight agreement, 0.01–0.20; fair agreement, 0.21–0.40; moderate agreement, 0.41–0.60; substantial agreement, 0.61–0.80; almost perfect agreement, 0.81–1.00) (23).

Results

Dataset Characteristics

After using the described sampling strategy on the 365 097 previously extracted clinical histories (123 553 for CT; 7401 for MRI; 37 146 for US; 196 987 for radiography), a total of 50 186 clinical histories were included (794 training, 150 validation, 300 initial testing, and 48 942 unannotated clinical histories representing a real-world application) (Table 1, Fig 2).

Model Selection and Performance

With prompt engineering alone, without in-context input-output pairs (in-context learning examples), the model outputs were inconsistent and unstructured, and the model could "hallucinate" information (Fig S1). Experimenting with different prompt structures showed that the combination of a prefix, a task description, in-context learning examples, and the clinical history to be examined yielded the most consistent results (Fig 3).

The performance of all models improved when the number of in-context input-output pairs increased, with 16 pairs leading to the best performance (Fig S2). Thus, only models that underwent prompt engineering and in-context learning with 16 in-context input-output pairs were assessed. After prompt engineering and in-context learning, Mistral-7B performed better than Llama 2–7B (mean overall accuracy, 90% vs 79%; mean overall BERTScore, 0.95 vs 0.92) (Fig S2; Tables S1, S2). After Mistral-7B was further fine-tuned, its performance improved to a mean overall accuracy of 91% and mean overall BERTScore of 0.96.

Agreement between Models and Radiologists

In the interreader agreement analysis, there was substantial overall agreement between radiologists 1 and 2 (mean κ, 0.76 [95% CI: 0.70, 0.81]). Both the fine-tuned Mistral-7B model and GPT-4 Turbo also showed substantial overall agreement with radiologist 1 (mean κ, 0.73 for both [95% CI: 0.66, 0.79]) and radiologist 2 (mean κ, 0.75 [95% CI: 0.69, 0.79] and 0.77 [95% CI: 0.71, 0.82]), indicating strong similarity between the models' outputs and radiologist annotations (Fig 4A; Table S3, unweighted agreement values are shown in Table S4).

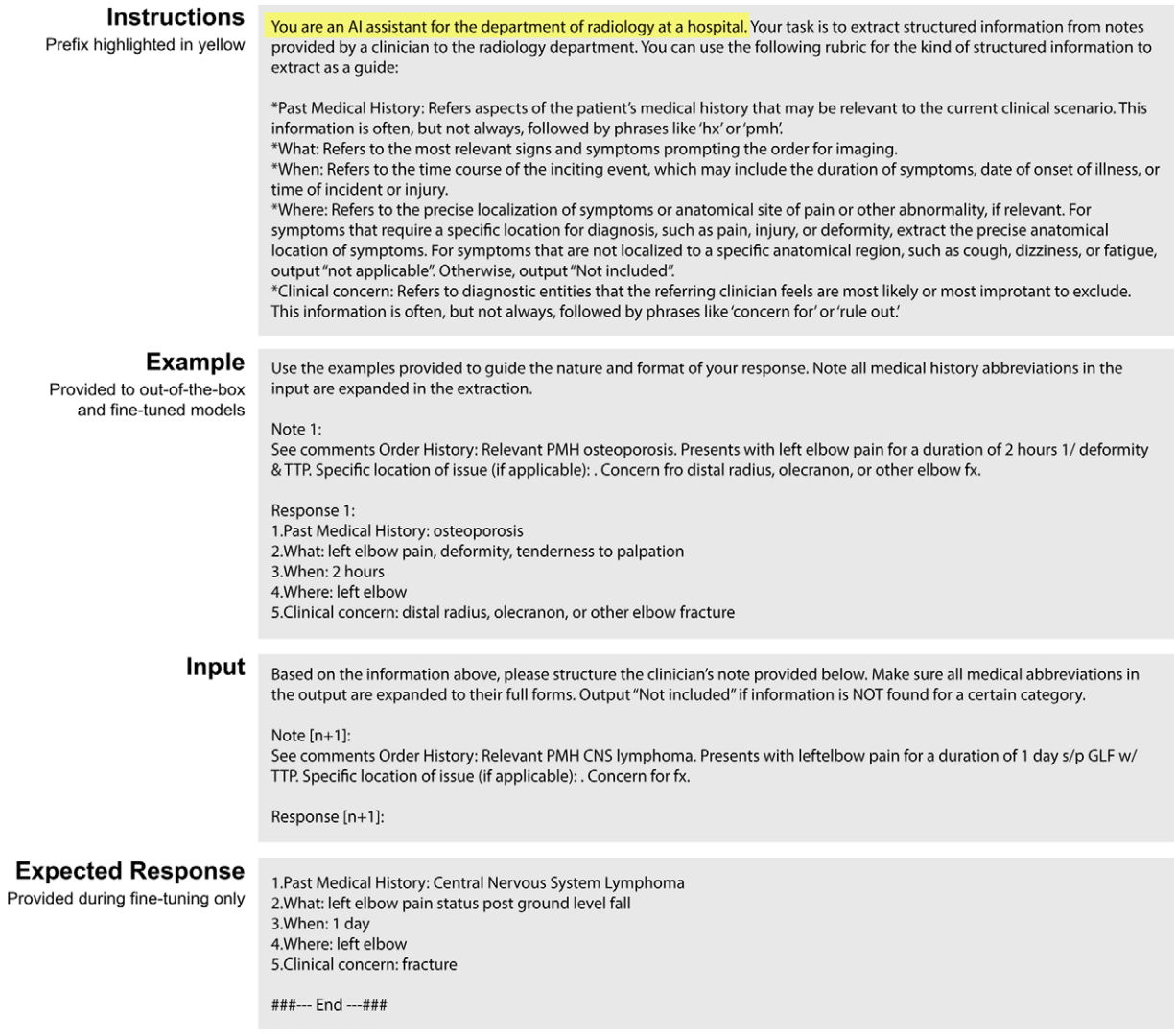

Figure 3: Example structure of a prompt for model training, showing the initial prefix, instructions, and task, with the addition of in-context examples and response depending on which adaptation methods were applied. Misspellings in this image are intentional. CNS = central nervous system, PMH = past medical history.

At the element-specific level, both models showed substantial agreement with radiologist annotations for four out of five elements ("past medical history," "what," "when," and "clinical concern," with mean κ ranging from 0.65 [95% CI: 0.51, 0.77] to 0.94 [95% CI: 0.89, 0.97]), and moderate agreement for "where" (mean κ, 0.50 [95% CI: 0.40, 0.58] to 0.63 [95% CI: 0.55, 0.71]). This was consistent with the agreement observed between radiologist 1 and radiologist 2 for the five elements (Fig 4B). The two elements with the greatest agreement both between models and radiologists and between radiologists were "when" (mean κ, 0.88 [95% CI: 0.83, 0.94] to 0.94 [95% CI: 0.89, 0.97]) and "clinical concern" (mean κ, 0.85 [95% CI: 0.77, 0.91] to 0.93 [95% CI: 0.87, 0.97]), suggesting that these elements are less subjective; in contrast, the agreement values (mean κ, 0.50 [95% CI: 0.41, 0.58] to 0.63 [95% CI: 0.55, 0.71]) for "where" suggest it is more subjective.

BERTScores were also highest for the two elements "when" (mean values, 0.96 [95% CI: 0.96, 0.97] to 0.99 [95% CI: 0.98, 0.99]) and "clinical concern" (mean values, 0.97 [95% CI: 0.97, 0.98] to 0.99 [95% CI: 0.98, 0.99]), echoing greater semantic similarity for these elements (Table S5). Overall, both the fine-tuned Mistral-7B model and GPT-4 Turbo had equal or greater mean BERTScores with radiologists 1 and 2 than the two radiologists had between themselves (Fig 4C). For the individual elements, Mistral-7B, GPT-4 Turbo, and the radiologists achieved mean BERTScores ranging from 0.95 (95% CI: 0.94, 0.96) to 0.99 (95% CI: 0.98, 0.99) (*P* value range, .10–.95) (Fig 4D). Both models had better agreement with radiologist 2 (mean Cohen κ: Mistral-7B, 0.75 [95% CI: 0.69, 0.79]; GPT-4 Turbo, 0.77 [95% CI: 0.71, 0.82]) than radiologist 1 (mean Cohen κ, 0.73 for both models [95% CI: 0.66, 0.79]), even exceeding interradiologist agreement in semantic meaning (mean Cohen κ, 0.76 [95% CI: 0.70, 0.81]) in most cases (*P* value range, .22–.99).

Annotation Adjudication and Model Performance

During data adjudication, disagreements were seen for all five clinical history elements (Table S6). Overall, the fine-tuned Mistral-7B model and GPT-4 Turbo demonstrated no evidence of statistical differences in specificity (mean, 82% [95% CI: 76, 88] vs 76% [95% CI: 69, 83]; *P* = .07), accuracy (mean, 91% [95% CI: 89, 93] vs 92% [95% CI: 90, 94]; *P* = .31), or BERTScore (mean, 0.96 [95% CI: 0.95, 0.97] vs 0.96 [95% CI: 0.95, 0.97]; *P* = .64) (Table 3). Mistral-7B also showed higher precision than GPT-4

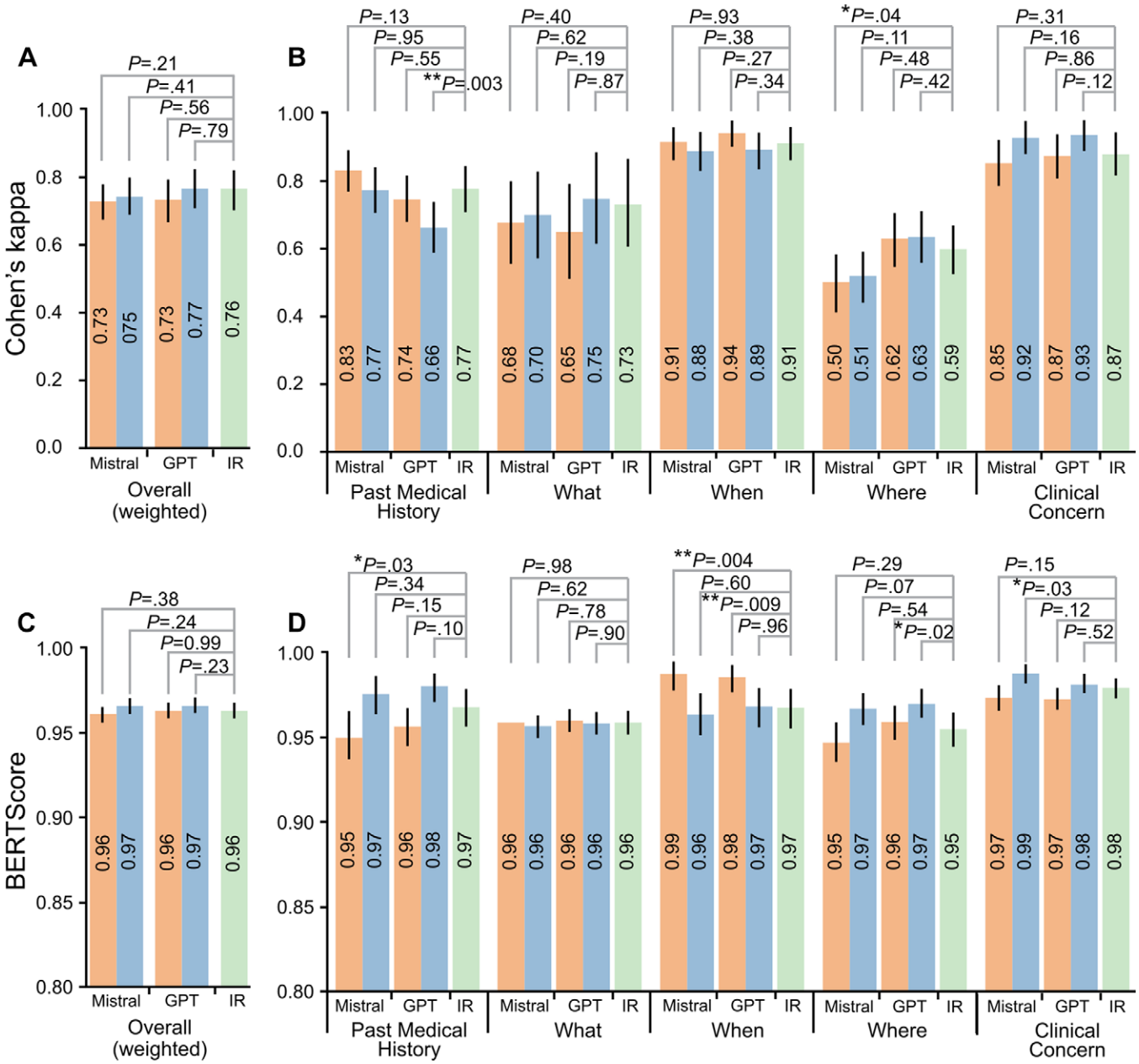

Figure 4: Interreader agreement between the models and radiologists as well as between radiologists 1 and 2 in terms of **(A)** weighted overall agreement and **(B)** for each individual clinical history element. BERTScore was used to assess **(C)** weighted overall semantic similarity between the models and readers and between the two radiologists, as well as **(D)** for each element. Mistral refers to fine-tuned Mistral-7B, GPT refers to GPT-4 Turbo after prompt engineering and in-context learning with 16 in-context input-output pairs, and IR refers to interreader agreement between the two radiologists. Weighted overall metrics were calculated using the following weights: past medical history, 0.15; what, 0.4; when, 0.1; where, 0.15; and clinical concern, 0.2. * *P* value ≤ .05, ** *P* value ≤ .01. Error bars show 95% CIs estimated by bootstrapping. Differences between the models and radiologists were assessed through either bootstrapping hypothesis testing (Cohen κ) or *t* tests (BERTScore). Unweighted overall metrics can be found in Figure S3.

Turbo (mean, 95% [95% CI: 94, 96] vs 93% [95% CI: 91, 95]; *P* = .003), while GPT-4 Turbo demonstrated higher sensitivity (mean 95% [95% CI: 94, 96] vs 92% [95% CI: 90, 94]; *P* = .001).

Clinical History Quality Evaluation

Because open-source models can be deployed locally, helping keep patient information secure, the locally adapted fine-tuned Mistral-7B model was applied to evaluate the quality of a large dataset containing 48 942 unannotated clinical history entries. Of a possible 244 710 elements, 153 934 were present, leading to a nonweighted average inclusion rate of all relevant clinical history elements of 62.9%.

When considering each clinical history as a collective entity, 49.8% (24 374 of 48 942), 91.5% (44 798 of 48 942), 48.5% (23 729 of 48 942), 78.5% (38 461 of 48 942), and 68.0% (33 295 of 48 942) contained the elements "past medical history," "what," "where," "when," and "clinical concern," respectively (Table 4). Overall, 26.2% (12 803 of 48 942) were complete, meaning that they contained all five elements, and 40.2% (19 680 of 48 942) contained the three elements with the highest weights ("past medical history," "what," "clinical concern"). Using the previously described weights, the weighted average inclusion rate of all relevant clinical history elements was 74.3% (36 376 of 48 942).

Table 3: Performance Metrics for Adapted Models, Mistral-7B and GPT-4-Turbo

	Mistral-7B		GPT-4 Turbo		
Element and Metric	Value	95% CI	Value	95% CI	*P* Value*
Past medical history					
Sensitivity (%)	89 (122/137)	83, 95	93 (128/137)	89, 97	.09
Specificity (%)	92 (150/163)	87, 95	81 (132/163)	75, 87	.003
Precision (%)	90 (122/135)	85, 95	81 (128/159)	75, 87	.003
Accuracy (%)	91 (272/300)	87, 95	87 (260/300)	83, 91	.06
BERTScore	0.95	0.94, 0.96	0.96	0.95, 0.97	.39
What					
Sensitivity (%)	94 (250/265)	91, 97	97 (258/265)	95, 99	.02
Specificity (%)	80 (28/35)	66, 94	63 (22/35)	46, 80	.06
Precision (%)	97 (250/257)	95, 99	95 (258/271)	92, 98	.08
Accuracy (%)	93 (278/300)	90, 96	93 (258/300)	90, 96	.68
BERTScore	0.96	0.95, 0.97	0.96	95, 97	.77
When					
Sensitivity (%)	93 (103/113)	88, 98	97 (110/113)	94, 100	.11
Specificity (%)	97 (182/187)	94, 100	97 (181/187)	94, 100	.80
Precision (%)	95 (105/110)	91, 99	95 (110/116)	91, 99	.84
Accuracy (%)	96 (287/300)	94, 98	97 (291/300)	95, 99	.41
BERTScore	0.99	0.98, 1.00	0.98	0.97, 0.99	.67
Where					
Sensitivity (%)	85 (218/257)	80, 90	88 (227/257)	84, 92	.16
Specificity (%)	49 (21/43)	34, 64	67 (29/43)	52, 82	.02
Precision (%)	91 (218/240)	87, 95	94 (227/241)	91, 97	.01
Accuracy (%)	80 (239/300)	75, 75	85 (256/300)	81, 89	.02
BERTScore	0.95	0.94, 0.96	0.96	0.95, 0.97	.10
Clinical concern					
Sensitivity (%)	94 (217/232)	91, 97	96 (223/232)	93, 99	.04
Specificity (%)	97 (66/68)	93, 100	93 (63/68)	86, 100	.10
Precision (%)	99 (217/219)	98, 100	98 (223/228)	96, 100	.11
Accuracy (%)	94 (283/300)	91, 97	95 (286/300)	93, 97	.42
BERTScore	0.97	0.96, 0.98	0.97	0.97, 0.97	.95
Overall (weighted)					
Sensitivity (%)	92	90, 94	95	94, 96	.001
Specificity (%)	82	76, 88	76	69, 83	.07
Precision (%)	95	94, 96	93	91, 95	.003
Accuracy (%)	91	89, 93	92	90, 94	.29
BERTScore	0.96	0.95, 0.97	0.96	0.95, 0.97	.47

Note.—Unless otherwise indicated, data are percentages and data in parentheses are numerators/denominators. Differences between the models and radiologists were assessed through either bootstrapping hypothesis testing (sensitivity, specificity, precision, accuracy, and Cohen κ) or *t* tests (BERTScore). The 95% CIs were calculated by bootstrapping. Weighted overall metrics were calculated using the following weights: past medical history, 0.15; what, 0.4; when, 0.1; where, 0.15; and clinical concern, 0.2. Both large language models, open-source (Mistral-7B) and proprietary (GPT-4 Turbo), underwent prompt engineering and in-context learning. As proprietary models (or closed-source models) cannot be fine-tuned, GPT underwent only prompt engineering and in-context learning. Mistral was further fine-tuned. BERTScore is an LLM metric that quantifies how well two pieces of text convey the same meaning. Accuracy = (true-positives + true-negatives)/(true-positives + true-negatives + false-positives + false-negatives); precision = (true-positives)/(true-positives + false-positives); sensitivity = (true-positives)/(true-positives + false-negatives); and specificity = (true-negatives)/(true-negatives + false-positives).

* $P < .05$ indicated statistically significant difference.

Example Clinical History Outputs

The fine-tuned Mistral-7B model generally performed well in extracting each element from the provided clinical history entries. In most cases, the model correctly identified each of the five clinical history elements, noting their presence or absence and generating text to indicate what was provided for each element in the original clinical history (Fig 5A). However, in some cases, the model was affected by the inherent subjectivity of identifying the five elements, especially for clinical histories in which the clinicians themselves did not have clear consensus, suggesting that some clinical history entries may contain a higher degree of inherent subjectivity than others. For example, in Figure 5B, given the history "POD 10 from craniotomy and mass resection, now with HA, N/V concern for edema, mass, bleed," Mistral-7B responded with "10 days ago" as the label for the "when" element. In contrast, for "when," radiologist 1 said "Now" and radiologist 2 said this information was "Not included." In this case, it is not fully

Table 4: Composition of Elements Observed in 48 942 Clinical Histories Overall and by Modality

Modality and Inclusion	Past Medical History	What	When	Where	Clinical Concern
All modalities					
Included	20 939	44 798	23 729	25 785	33 295
Not applicable	…	…	…	12 676	…
No relevant past medical history	3435	…	…	…	…
Not included	24 568	4144	25 213	10 481	15 647
Percent included (%)*	49.8	91.5	48.5	78.5	68.0
CT					
Included	7872	14 242	8765	8717	11 211
Not applicable	…	…	…	3795	…
No relevant past medical history	593	…	…	…	…
Not included	6790	1013	6490	2743	4044
Percent included (%)*	55.5	93.4	57.5	82.0	73.5
MRI					
Included	506	795	495	460	642
Not applicable	…	…	…	229	…
No relevant past medical history	76	…	…	…	…
Not included	290	77	377	183	230
Percent included (%)*	66.7	91.2	56.8	79.0	73.6
US					
Included	2828	5256	3599	4264	4750
Not applicable	…	…	…	575	…
No relevant past medical history	700	…	…	…	…
Not included	2159	431	2088	848	937
Percent included (%)*	62.0	92.4	63.3	85.1	83.5
Radiography					
Included	9733	24 505	10 870	12 344	16 692
Not applicable	…	…	…	8077	…
No relevant past medical history	2066	…	…	…	…
Not included	15 329	2623	16 258	6707	10 436
Percent included (%)*	43.5	90.3	40.1	75.3	61.5

* Elements considered not applicable and not relevant were counted as "included" elements.

clear when the most recent symptoms began, indicating that the confusion is likely more related to the subtle ambiguity of the history than a problem with the model. Remarkably, Mistral-7B was successful when provided with a clinical history (Fig 5C) that contained many clinical acronyms, likely due to its large pretraining corpora. These examples illustrate that Mistral-7B performed in a manner similar to expert human raters in assessing clinical histories.

Discussion

Incomplete clinical histories are a common frustration among radiologists, and previous improvement efforts have relied on tedious manual analysis. This study illustrates the feasibility of automatically assessing the completeness of clinical histories using large language models (LLMs). Two open-source (Llama 2–7B and Mistral-7B) and one closed-source (GPT-4 Turbo) LLMs were adapted using prompt engineering, in-context learning, and fine-tuning (open-source only) to extract the elements "past medical history," "what," "when," "where," and "clinical concern" from clinical histories. After model adaptation, the fine-tuned Mistral-7B showed no evidence of a difference in accuracy compared with GPT-4 Turbo (91% [95% CI: 89, 93] vs 92% [95% CI: 90, 94]; $P = .31$) despite being a substantially smaller model. Both Mistral-7B and GPT-4 Turbo showed substantial overall agreement with radiologists (mean Cohen κ, 0.73 [95% CI: 0.67, 0.78] to 0.77 [95% CI: 0.71, 0.82]) and adjudicated manual annotations by the radiologists (mean BERTScore, 0.96 [95% CI: 0.96, 0.97] for both models; $P = .38$). Using Mistral-7B, 26.2% (12 803 of 48 942) of unannotated clinical histories were found to contain all five elements.

These results are promising because smaller models require fewer computing resources, facilitating their deployment. Additionally, open-source models are unaffected by unpublicized model updates and can be fully deployed locally, avoiding some clinical data privacy concerns with external servers required by proprietary (or closed-source) models. Thus, the fine-tuned Mistral-7B can feasibly automate similar improvement efforts and may improve patient care, as ensuring that clinical histories contain the most relevant information can help radiologists better understand the clinical context of the imaging order.

Both models had better agreement with radiologist 2 (mean Cohen κ, 0.75 [95% CI: 0.69, 0.79] for Mistral-7B and 0.77 [95% CI: 0.71, 0.82] for GPT-4 Turbo) than radiologist 1 (mean Cohen κ, 0.73 for both models [95% CI: 0.66, 0.79]). Because radiologist 2 annotated more clinical history entries due to differences in availability (ratio of 4:1 with radiologist 1), it is unsurprising that stronger agreement was seen between

A **Input**

hx of htn, hld, here w/ epigastricx pain, ?sob, eval pna, effusion

Output

Element	Fine-tuned Mistral	Reader 1	Reader 2
Past Medical History	hypertension, hyperlipidemia	Hypertension, hyperlipidemia	hypertension, hyperlipidemia
What	epigastric pain, possible shortness of breath	Epigastric pain, questionable shortness of breath	epigastric pain, possible shortness of breath
When	Not included	Not included	Not included
Where	epigastric	epigastric	epigastric
Clinical Concern	pneumonia, effusion	Pneumonia, effusion	pneumonia, effusion

B **Input**

POD 10 from craniotomy and mass resection, now with HA, N/V concern for edema, mass, bleed

Output

Element	Fine-tuned Mistral	Reader 1	Reader 2
Past Medical History	Not included	Craniotomy and mass resection	Not included
What	headache, nausea, vomiting	Headache, nausea, vomiting	post operative day 10 from craniotomy and mass resection, headache, nausea, vomiting
When	10 days after surgery	Now	Not included
Where	head	Head	head
Clinical Concern	edema, mass, bleeding	Edema, mass, bleed	edema, mass, bleed

C **Input**

See comments Order History: Relevant PMH CAD s/p PCI, HTN, TIIDM. Presents with chest pain, marked HTN for a duration of 1 day. Specific location of issue (if applicable): . Concern for dissection.

Output

Element	Fine-tuned Mistral	Reader 1	Reader 2
Past Medical History	coronary artery disease status post percuaneous coronary intervention, hypertension, type II diabetes mellitus	Coronary artery disease status post percutaneous intervention, hypertension, type II diabetes mellitus	coronary artery disease status post percutaneous coronary intervention, hypertension, diabetes mellitus type 2
What	chest pain, marked hypertension	Chest pain, marked hypertension	chest pain, marked hypertension
When	1 day	1 day	1 day
Where	chest	Chest	chest
Clinical Concern	dissection	Dissection	dissection

Figure 5: Example input clinical history entries with the outputs generated by the fine-tuned Mistral-7B model and the two radiologists (reader 1 and reader 2). Misspellings in this image are intentional. **(A)** An example clinical history where the model and readers were all in agreement and correctly annotated each element. **(B)** An example clinical history shows a high degree of subjectivity that led to inconsistent labeling between the two human readers and the model for certain elements, such as "when." **(C)** An example of a difficult clinical history with many clinical abbreviations that expert radiologists would be expected to recognize and that the model also correctly annotated. The example **(C)** also shows the default clinical history entry "See comments" field that is included in some of the orders from the emergency department of the study institution.

the models and radiologist 2. Nonetheless, comparable agreement was observed between the two models and radiologists 1 and 2 and between the two radiologists, illustrating that the model versus radiologist discrepancies rival interradiologist discrepancies in extracting clinical history elements. The slight discrepancy between radiologists (mean Cohen κ, 0.76) shows the inherently subjective nature of the task.

These discrepancies motivated an additional adjudication step to create a set of reference annotations to evaluate model performance. Using the adjudicated annotations, the fine-tuned Mistral-7B model generally showed higher precision (overall mean, 95% [95% CI: 95, 96] vs 93% [95% CI: 91, 95]; P = .003), but lower sensitivity (overall mean, 92% [95% CI: 90, 94] vs 0.95% [95% CI: 94, 96]; P = .001) than GPT-4 Turbo. These results suggest that there can be more confidence in Mistral-7B model outputs, as GPT-4 Turbo may result in a higher false-positive rate. Additionally, the similar accuracy and BERTScores between the two models, along with overall strong performance, suggest feasible local deployment of the fine-tuned Mistral-7B model at scale.

Assessment of the large dataset of 48 942 clinical histories revealed that 26.2% (12 803 of 48 942) were considered fully "complete" (contained all five elements). However, since not all elements are necessarily equally valuable, the weighted mean inclusion rate may be more representative of overall "completeness" (74.3% in this study). In comparison, using the three categories Hawkins et al (6) considered ("what," "when," "where"), 37.0% of clinical history entries in this study would be considered complete, compared with the 38% and 92% seen by Hawkins et al before and after a dedicated improvement project, respectively. Bor et al (19) considered four categories ("what," "when," "where," "clinical concern"),

reporting completeness rates of 16% and 52% before and after a dedicated improvement project, respectively, compared with 33.6% in this study. Notably, these comparison studies illustrated that the problem of inadequate clinical histories is amenable to improvement, though both relied on manual auditing. Thus, similar improvements may be more readily achievable at additional sites by automating the task and minimizing some of the associated time costs.

Given the ubiquity of incomplete clinical histories accompanying imaging orders, some believe that artificial intelligence (AI) tools should directly extract data from the electronic medical record (12). However, the brief clinical history accompanying imaging orders represents the ordering clinician's assessment of the clinical context and its most relevant aspects, articulating a specific clinical question to radiologists in a way that electronic medical record data often cannot. Thus, while AI summarizations of clinical histories may be helpful to radiologists, they do not necessarily obviate the need for succinct communication directly from the ordering clinician.

Despite its successful aspects, this study had limitations. First, this study used single-site data. Use of multi-institutional datasets will be required to expand the relevant training data, assess cross-site performance, and enhance generalizability (27). Second, incomplete data and missing data remain an issue that may affect generalizability. Additional analyses using more comprehensive data will further improve model robustness and validate performance when the model is applied to datasets with different patterns of missing clinical history elements. Moreover, because the model will be made publicly available, conducting additional analyses, such as robustness analyses, will become easier and more pervasive. Third, the prompt structure used during model development was manually designed for this specific task, and its generalizability to other institutions should be explored. Fourth, although inclusion rates were reported for the clinical histories accompanying imaging studies ordered from the emergency department at the study site, the representativeness of this rate remains unclear until other sites can be studied using standardized assessment methods, including validation of the five clinical history elements used. Finally, while similar key clinical history elements have been previously used, there is a need to further study, refine, and accept this definition as a complete clinical history.

In summary, a fine-tuned, open-source large language model (Mistral-7B) effectively extracted clinical history elements accompanying imaging orders (CT, MRI, US, radiography) with substantial agreement with radiologists. The fine-tuned Mistral-7B model was also competitive with a closed-source model, GPT-4 Turbo, enabling high-quality local adaptation and deployment to assess the completeness of clinical histories. Both models underwent prompt engineering and in-context learning, although Mistral-7B underwent further fine-tuning. Future studies should explore the impact of order prompts on the ability of large language models to assess clinical history completeness and quality at other sites. This study also highlights the potential of artificial intelligence for quality assessments of text, which could have broad applications in quality assurance and improvement in radiology. Therefore, the model and code from this study are being made freely available for public use.

Deputy Editor: Linda Moy
Scientific Editor: Sarah Atzen

Author affiliations:

[1] Department of Radiology, Stanford University School of Medicine, 453 Quarry Rd, MC 5659, Stanford, CA 94304

[2] AI Development and Evaluation Laboratory, Stanford University School of Medicine, Palo Alto, Calif

[3] Department of Electrical Engineering, Stanford University, Stanford, Calif

[4] Department of Biomedical Data Science, Stanford University, Stanford, Calif

Received April 15, 2024; revision requested April 30; final revision received January 2, 2025; accepted January 7.

Address correspondence to: D.B.L. (email: david.larson@stanford.edu).

Funding: Authors declared no funding for this work.

Acknowledgment: Microsoft provided Azure OpenAI credits for partial experiments in this project via the Accelerate Foundation Models Research (AFMR) program. Mistral-7B by Mistral AI, Llama-7B by Meta, and GPT-4 Turbo by OpenAI *(https://chat.openai.com/)* were used to generate results for this study.

Author contributions: Guarantors of integrity of entire study, **D.B.L., A.K., L.Y.C., H.S.N., Z.F., A.S.C.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **D.B.L., A.K., L.Y.C., M.P., D.V.V., Z.F.**; clinical studies, **D.B.L., D.V.V., Z.F.**; experimental studies, **D.B.L., A.K., M.P., D.V.V., Z.F., A.S.C.**; statistical analysis, **D.B.L., A.K., Z.F., A.S.C.**; and manuscript editing, **D.B.L., A.K., L.Y.C., M.P., D.V.V., H.S.N., Z.F., A.S.C.**

Data sharing: Data generated or analyzed during the study are available from the corresponding author by request.

Disclosures of conflicts of interest: D.B.L. Member of the Board of Chancellors of the American College of Radiology and Board of Trustees of the American Board of Radiology, shareholder in Bunkerhill Health; receives research funding from the Gordon and Betty Moore Foundation. **A.K.** No relevant relationships. **L.Y.C.** No relevant relationships. **M.P.** No relevant relationships. **D.V.V.** No relevant relationships. **H.S.N.** No relevant relationships. **M.B.P.** No relevant relationships. **Z.F.** No relevant relationships. **A.S.C.** Grants to institution from the NIH (R01HL167974, R01AR077604, R01EB002524, R01AR079431, and P41EB027060) and contracts (75N92020C00008 and 75N92020C00021), ARPA-H, GE HealthCare, and Philips; royalties or licenses from LVIS though Stanford; consulting fees from Elucid Bioimaging, Patient Square Capital, and Chondrometrics; payment from Genentech for expert testimony; patents planned, issued or pending from GE HealthCare through Stanford; participation on an Advisory Board for Brain Key and Chondrometrics GmbH; stock or stock options in LVIS, Subtle Medical, and Brain Key; Microsoft Azure OpenAI credits.

References

1. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to Create a Great Radiology Report. RadioGraphics 2020;40(6):1658–1670.
2. Yapp KE, Brennan P, Ekpo E. The Effect of Clinical History on Diagnostic Imaging Interpretation - A Systematic Review. Acad Radiol 2022;29(2):255–266.
3. Hattori S, Yokota H, Takada T, et al. Impact of clinical information on CT diagnosis by radiologist and subsequent clinical management by physician in acute abdominal pain. Eur Radiol 2021;31(8):5454–5463.
4. Ihuhua P, Pitcher RD. Is the devil in the detail? The quality and clinical impact of information provided on requests for non-trauma emergency abdominal CT scans. Acta Radiol 2016;57(10):1217–1222.
5. Finger A, Harris M, Nishimura E, Yoon H-C. Inadequate Clinical Indications in Computed Tomography Chest and Abdomen/Pelvis Scans. Perm J 2018;22(4):18-017.
6. Hawkins CM, Anton CG, Bankes WM, et al. Improving the availability of clinical history accompanying radiographic examinations in a large pediatric radiology department. AJR Am J Roentgenol 2014;202(4):790–796.
7. Alkasab TK, Alkasab JR, Abujudeh HH. Effects of a computerized provider order entry system on clinical histories provided in emergency department radiology requisitions. J Am Coll Radiol 2009;6(3):194–200.
8. Ganeshan D, Duong PT, Probyn L, et al. Structured Reporting in Radiology. Acad Radiol 2018;25(1):66–73.

9. Gunderman RB, Phillips MD, Cohen MD. Improving clinical histories on radiology requisitions. Acad Radiol 2001;8(4):299–303.
10. Schuster DM, Gale ME. The malady of incomplete, inadequate, and inaccurate radiology requisition histories: a computerized treatment. AJR Am J Roentgenol 1996;167(4):855–859.
11. Cohen MD, Alam K. Radiology clinical synopsis: a simple solution for obtaining an adequate clinical history for the accurate reporting of imaging studies on patients in intensive care units. Pediatr Radiol 2005;35(9):918–922.
12. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30(4):1134–1142.
13 Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 22). New Orleans, La, November 28 to December 9, 2022. Curran Associates, 2022; 27730–27744.
14. Zakka C, Shad R, Chaurasia A, et al. Almanac - Retrieval-Augmented Language Models for Clinical Medicine. NEJM AI 2024;1(2):10.1056/aioa2300068.
15. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI Feedback. arXiv 2212.08073 [preprint] https://arxiv.org/abs/2212.08073. Posted December 15, 2022. Accessed February 2, 2024.
16. Zhou Y, Muresanu AI, Han Z, et al. Large Language Models Are Human-Level Prompt Engineers. arXiv 2211.01910 [preprint] https://arxiv.org/abs/2211.01910. Posted November 3, 2022. Accessed February 2, 2024.
17. Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In: CHI EA '21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama Japan, May 2021. ACM, 2021; 1–7.
18. Van Veen D, Van Uden C, Attias M, et al. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, Canada, July 2023. Association for Computational Linguistics, 2023; 449–460.
19. Bor DS, Sharpe RE Jr, Payán G Jr, et al. Sustained Increases in the Quality of Imaging Histories Provided by Ordering Providers Resulting From Multidisciplinary Collaboration and Systems Design. AJR Am J Roentgenol 2019;213(5):1023–1028.
20. Llama 2–7B. Meta. https://www.llama.com/llama2/. Accessed January 24, 2024.
21. Mistral-7B. Mistral AI. https://mistral.ai/news/announcing-mistral-7b/. Accessed January 24, 2024.
22. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv 2307.09288 [preprint] https://arxiv.org/abs/2307.09288. Posted July 18, 2023. Accessed February 9, 2024.
23. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. arXiv 2310.06825 [preprint] https://arxiv.org/abs/2310.06825. Posted October 10, 2023. Accessed February 9, 2024.
24. GPT-4. OpenAI. https://platform.openai.com/docs/models#gpt-4o. Accessed January 24–26, 2024.
25. Liang P, Bommasani R, Lee T, et al. Holistic Evaluation of Language Models. arXiv 2211.09110 [preprint] https://arxiv.org/abs/2211.09110. Posted November 16, 2022. Accessed February 9, 2024.
26. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. arXiv 1904.09675 [preprint] https://arxiv.org/abs/1904.09675. Posted April 21, 2019. Accessed February 9, 2024.
27. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA 2023;330(9):866–869.