




# RadSearch, a Semantic Search Model for Accurate Radiology Report Retrieval with Large Language Model Integration

Cody H. Savage, MD<sup>1,2</sup> • Gunvant Chaudhari, MD<sup>3</sup> • Andrew D. Smith, MD, PhD<sup>2,4</sup> • Jae Ho Sohn, MD, MS<sup>3</sup>

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

See also the editorial by Yasaka and Abe in this issue.

Radiology 2025; 315(1):e240686 • <https://doi.org/10.1148/radiol.240686> • Content codes:   

**Background:** Current radiology report search tools are limited to keyword searches, which lack semantic understanding of underlying clinical conditions and are prone to false positives. Semantic search models address this issue, but their development requires scalable methods for generating radiology-specific training data.

**Purpose:** To develop a scalable method for training semantic search models for radiology reports and to evaluate a model, RadSearch, trained using this method.

**Materials and Methods:** In this retrospective study, a scalable method for generating training examples for semantic search was applied to CT and MRI reports generated between December 2021 and January 2022, and was used to train the model RadSearch. RadSearch performance was evaluated using four internal test sets (including one subset) and one external test set from another large tertiary medical center, including chest, abdomen, and head CT reports generated between December 2015 and June 2023. Performance was evaluated for findings-to-impression matching, retrieving reports with the same examination type, retrieving reports relevant to free-text queries, and improving the ability of a large language model (LLM) (Llama 3.1 8B Instruct) to provide accurate diagnoses from report finding descriptions. RadSearch performance was compared with that of other embedding models specialized for symmetric (All MPNet Base) and asymmetric (MS MARCO DistilBERT Base) semantic search and a state-of-the-art semantic search model (GTE-large). A reference set of 100 diagnoses with common radiologic descriptions was used for the LLM evaluation. Findings-to-impression matching and free-text query accuracy  $P$  values were calculated using  $\chi^2$  and McNemar tests.

**Results:** The training set included 16690 reports; the internal test sets included 13 598, 6178, and 9954 reports; and the external test set included 13958 reports. For simulated free-text clinical queries, RadSearch successfully retrieved reports containing the specified findings for 83.0% (498 of 600) of reports and matching location for 89.8% (521 of 580) of reports, outperforming GTE-large, with performance at 65.7% (394 of 600;  $P < .001$ ) and 58.8% (341 of 580;  $P < .001$ ), respectively. For 100 report finding descriptions, the baseline accuracy of Llama 3.1 8B Instruct in providing the correct diagnosis without any embedding model search assistance was 30% (30 of 100), improving to 61% (61 of 100) with RadSearch integration ( $P < .001$ ), which outperformed GTE-large integration (47% [47 of 100];  $P = .03$ ).

**Conclusion:** A semantic search model trained with scalable methods achieved state-of-the-art performance in retrieving reports with relevant findings and improved LLM diagnostic accuracy.

© RSNA, 2025

Supplemental material is available for this article.

Structured information retrieval from medical documents remains one of the greatest challenges for search algorithms within the medical domain. Despite advances in artificial intelligence and search technologies, the majority of medical images and radiology reports are indexed using manually assigned keywords or metadata generated during image acquisition. Conventional search mechanisms within radiology information systems predominantly use keyword-based search, which is limited by poor specificity, susceptibility to misclassification, and lack of semantic understanding (1). These limitations can create problems for both radiologists and researchers. For radiologists, these limitations increase the time and energy investment in locating reports with similar imaging findings for comparison. For researchers, current search methods often lack the specificity required, necessitating manual curation of datasets, which is time-consuming and introduces the potential for bias (2–7).

Advances in large language model (LLM) technologies like retrieval-augmented generation may overcome these limitations. Retrieval-augmented generation connects LLMs to embedding models that search databases for similar items based on user input, providing contextual information for responses (8). In

radiology, Rau et al (9) demonstrated that retrieval-augmented generation improved the adherence of an LLM to the American College of Radiology Appropriateness Criteria, with the LLM even outperforming radiologists. Despite the critical role of embedding models, they have received little attention in training radiology-specific models (10). A major contributing factor is their dependence on manual annotations by radiologists to generate training data, which is time-consuming and prevents the creation of the large, diverse datasets needed to train these models. Scalable approaches that reduce dependence on expert annotation are essential to keep pace with artificial intelligence development and power the future search capabilities of LLMs.

The objective of this study was to develop a scalable method for training a domain-specific embedding model for radiology report semantic search. The search performance of the resulting model, termed RadSearch, was evaluated using several search metrics. The primary outcome metrics were retrieving reports with similar report findings using free-text queries and enhancing LLM diagnostic accuracy. Figure 1 illustrates RadSearch as an educational tool for radiology residents, with and without LLM integration. Term definitions are provided in Table S1.

## Abbreviations

LLM = large language model, mAP = mean average precision, UAB = University of Alabama at Birmingham, UCSF = University of California, San Francisco

## Summary

RadSearch, a specialized semantic search model for radiology reports trained using a scalable method, achieved state-of-the-art performance in retrieving relevant radiology reports for queried report findings, improving the diagnostic accuracy of a large language model.

## Key Results

- This retrospective study introduces RadSearch, a specialized semantic search model.
- Tasked with retrieving radiology reports with similar findings to simulated free-text clinical queries, RadSearch retrieved reports containing the queried report finding for 83.0% (498 of 600) of reports and location for 89.8% (521 of 580), outperforming a state-of-the-art semantic search model (GTE-large) with performance at 65.7% (394 of 600;  $P < .001$ ) and 58.8% (341 of 580;  $P < .001$ ), respectively.
- Asked to identify the most likely diagnosis for a given report finding description, the performance of a large language model (LLM) (Llama 3.1 8B Instruct) without any search assistance was 30% (30 of 100), increasing to 61% (61 of 100) with RadSearch assistance ( $P < .001$ ), outperforming the LLM with GTE-large assistance (47% [47 of 100];  $P = .03$ ).

## Materials and Methods

### Datasets

This retrospective study was approved by the institutional review board of the University of Alabama at Birmingham (UAB) and the University of California, San Francisco (UCSF) and was compliant with the Health Insurance Portability and Accountability Act. The requirement for written informed consent was waived due to the retrospective nature of the study. Six datasets (datasets A–E plus dataset B subset) consisting of radiology reports from two separate, large tertiary medical centers (UAB and UCSF) were used for this study (Fig 2). Reports with duplicate findings sections or reports where the findings section or impression section was missing were excluded. The findings and impression sections were extracted from reports in each dataset using a Python (version 3.12; Python Software Foundation) parser script customized separately for the format of reports at UAB and UCSF. Additional information about each of the datasets, the report exclusion process, and the parser script is provided in Appendix S1.

### Model Architecture and Training

RadSearch used a Siamese network architecture with RadBERT-RoBERTa-4m as its weight initialization and was fine-tuned for semantic search using a contrastive learning approach (11,12). The contrastive learning method, illustrated in Figure 3, used the findings and impression sections from individual full radiology reports to create positive and negative training pairs. Positive pairs were pairings of a findings section with its own corresponding impression section, and negative pairs were pairings of the findings section with the impression section of a different report. For example, for a positive pair, the findings section of report A would be paired with the impression section of report A. For a negative pair, the findings section of report A would be paired

with the impression section of report B. For negative pairs, this process was repeated for all unique impression sections, thereby creating multiple negative pairs for every positive pair. The training examples for RadSearch were created from dataset A (training set) using this contrastive learning method. Additional details of the architecture of RadSearch and the contrastive learning method are provided in Appendix S1. Training and evaluation source code is provided at <https://github.com/csavages/RadSearch>.

The comparator embedding models—All MPNet Base (13), MS MARCO DistilBERT Base (14), and GTE-large (15)—were not fine-tuned using training examples from dataset A. These models were previously fine-tuned for semantic search using large, generalized, manually curated query-and-answer pair datasets (eg, 500 000 Bing search queries and relevant text passages from web sources retrieved using the queries). The rationale for fine-tuning only RadSearch with training examples from dataset A was to isolate the impact of the fine-tuning method and to compare a model trained using this fine-tuning method (RadSearch) with a diverse set of existing high-performing semantic search models.

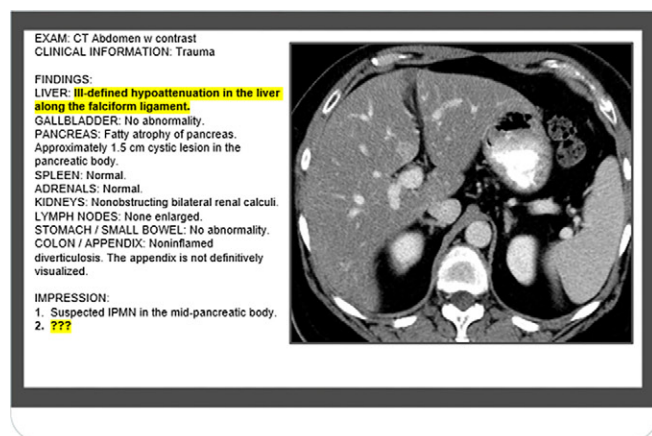
### Model Evaluation

The search performance of RadSearch was evaluated using a multistage approach that included four search metrics: matching the findings section of a report with its respective impression section, matching reports with the same examination type, matching reports with similar report findings using free-text queries, and improving LLM diagnostic accuracy. The report findings-to-impression matching metric and the examination type matching metric were also measured for an embedding model fine-tuned for symmetric semantic search (All MPNet Base) (13), an embedding model fine-tuned for asymmetric semantic search (MS MARCO DistilBERT Base) (14), and GTE-large (15). GTE-large, a state-of-the-art semantic search model, was selected because it was one of the highest-performing sentence transformer models on the Massive Text Embedding Benchmark Leaderboard at the time that the study evaluations were performed (May 2024) (16). *State-of-the-art* was defined as being within the top 10 ranks on the Massive Text Embedding Benchmark Leaderboard.

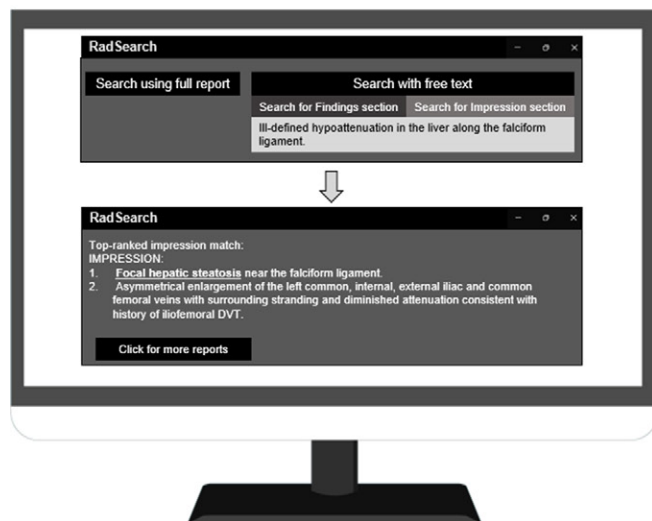
### Report findings section-to-impression section matching.—

As illustrated in Figure 4, the report findings section-to-impression section matching metric was evaluated using each findings section as a query to the embedding models to retrieve the top five most similar impression sections. This evaluation was performed in dataset B ( $n = 13958$ ) and dataset E ( $n = 13958$ ), one of the internal test sets and the external test set, respectively. The embedding models generated similarity rankings by calculating the cosine similarity scores between the input query and each impression section in the dataset being searched. Higher cosine similarity scores corresponded to higher ranks. If the correct impression section (ie, the impression corresponding to the queried findings section of the full report) was among the top five retrieved, it was considered a match for that report. Top five ranking was considered to be the most clinically relevant, as prior studies have shown that most users view only the top 5–10 search results for a given query (17). Given that dataset E included only contrast-enhanced chest CT reports, a subset analysis of only the

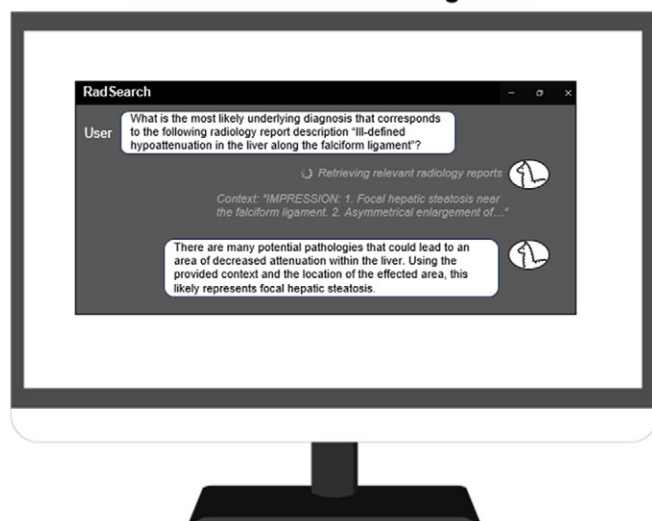
contrast-enhanced chest CT reports from dataset B ( $n = 2490$ ) was performed to provide more direct evaluation of performance differences between the UAB and UCSF datasets. Additional information is provided in Appendix S1.



### RadSearch Alone



### RadSearch with LLM Integration

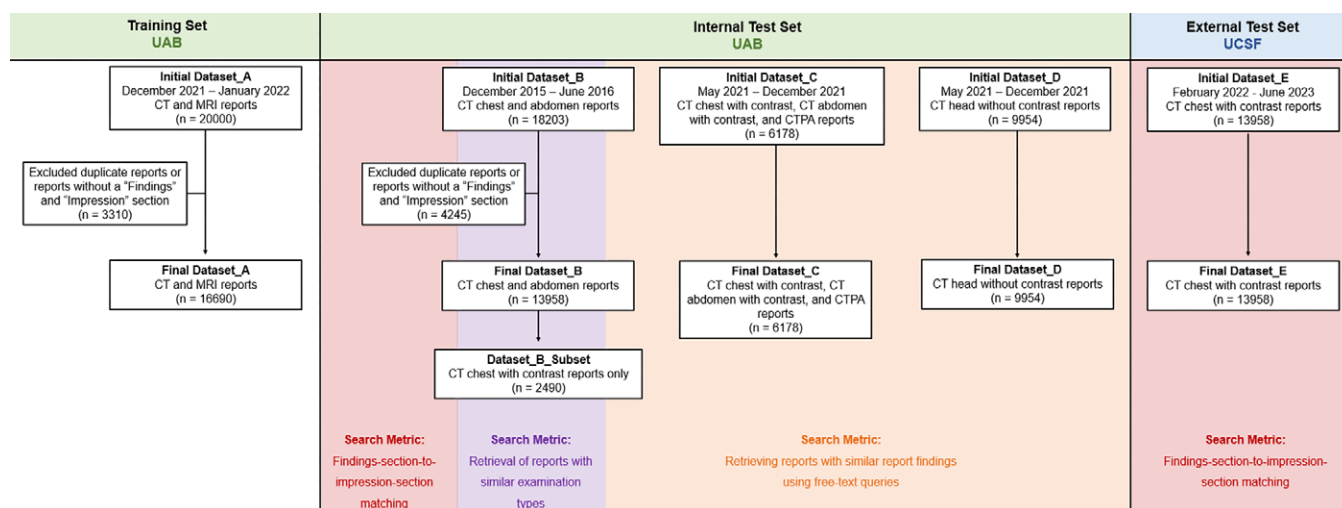


**Matching reports with the same examination type.**—Models were evaluated on their ability to match reports of the same examination type, to provide a more general quantitative measure of their ability to retrieve similar reports (assuming that text in reports of the same examination type have more similar imaging findings and anatomic structures than different examination types). Dataset B was used for this evaluation, and mean average precision (mAP) was calculated from report rankings (18). mAP is a ranking quality metric that considers the number of relevant reports retrieved in response to a query and their position in the list of total ranks. Here, the number of relevant reports was the number of chest reports in dataset B, and the total number of ranks was the total number of reports in dataset B. mAP can range from 0% to 100%, with higher percentages indicating higher precision. The metric is visualized in Figure 5, and further details are provided in Appendix S1.

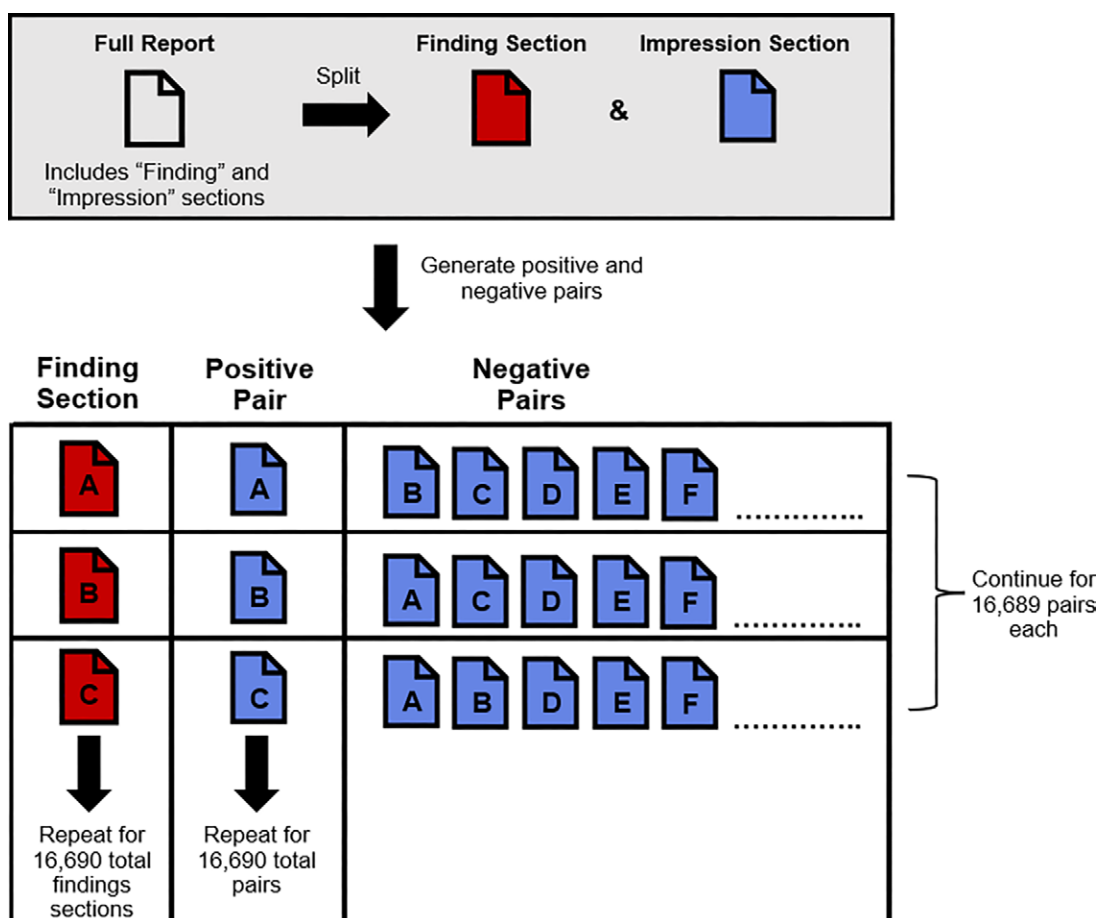
**Retrieving reports with similar report findings using free-text queries.**—To better evaluate model performance on free-text queries, simulated free-text clinical queries were generated, and RadSearch and GTE-large were asked to retrieve reports similar to these queries, for six general report finding categories: aneurysms, pulmonary emboli, intracranial hemorrhage, cholecystitis, pancreatic neoplasms, and spinal fractures (10 free-text sentences per category, for a total of 60 free-text sentences; Table S2). Details regarding the category selection process are described in Appendix S1. Top 10 ranking was used for this metric to allow for greater stratification of the results between models, given the smaller number of queries evaluated overall (compared with the findings-to-impression matching metric). This query approach simulated how RadSearch might be used in a real-world clinical setting and was a primary outcome metric.

The free-text sentences were entered as queries to RadSearch and GTE-large, and the top 10 most similar full reports (ie, findings section and impression section) were retrieved for each query. Dataset B was used for free-text queries of aneurysms, cholecystitis, pancreatic cancer, and spinal fractures. Two other internal test sets, datasets C and D, were used for free-text queries of pulmonary emboli and intracranial hemorrhage, respectively. The

**Figure 1:** Illustration of how RadSearch could be used as a diagnostic support tool. In this example, the interpreting radiologist comes across an unfamiliar imaging finding for which they do not know the underlying diagnosis (highlighted text in top image). The radiologist can enter their description of the finding into the RadSearch widget text box and choose to either (a) search and retrieve reports with similar findings sections, to see other potential ways that that imaging finding may have been described (eg, “focal fatty infiltration along the falciform ligament”), or (b) search and retrieve impressions that are associated with that imaging finding description, to identify potential diagnoses (eg, “focal hepatic steatosis”) (middle image). Of note, only one retrieved impression is shown in this example, but the number can be adjusted to any number ranging from one to the total number of reports in the dataset being searched. By using semantic search, RadSearch is able to retrieve reports with similar underlying meaning, even if they share no common keywords (eg, retrieving “focal hepatic steatosis” for “ill-defined hypodensity along the falciform ligament”). RadSearch can also be integrated with a large language model (LLM), such as Llama 3.1 (bottom image). Here, RadSearch retrieves relevant radiology reports from a database, which are then used by the LLM as additional context when responding to the user’s query. DVT = deep vein thrombosis, IPMN = intraductal papillary mucinous neoplasm, w = with.

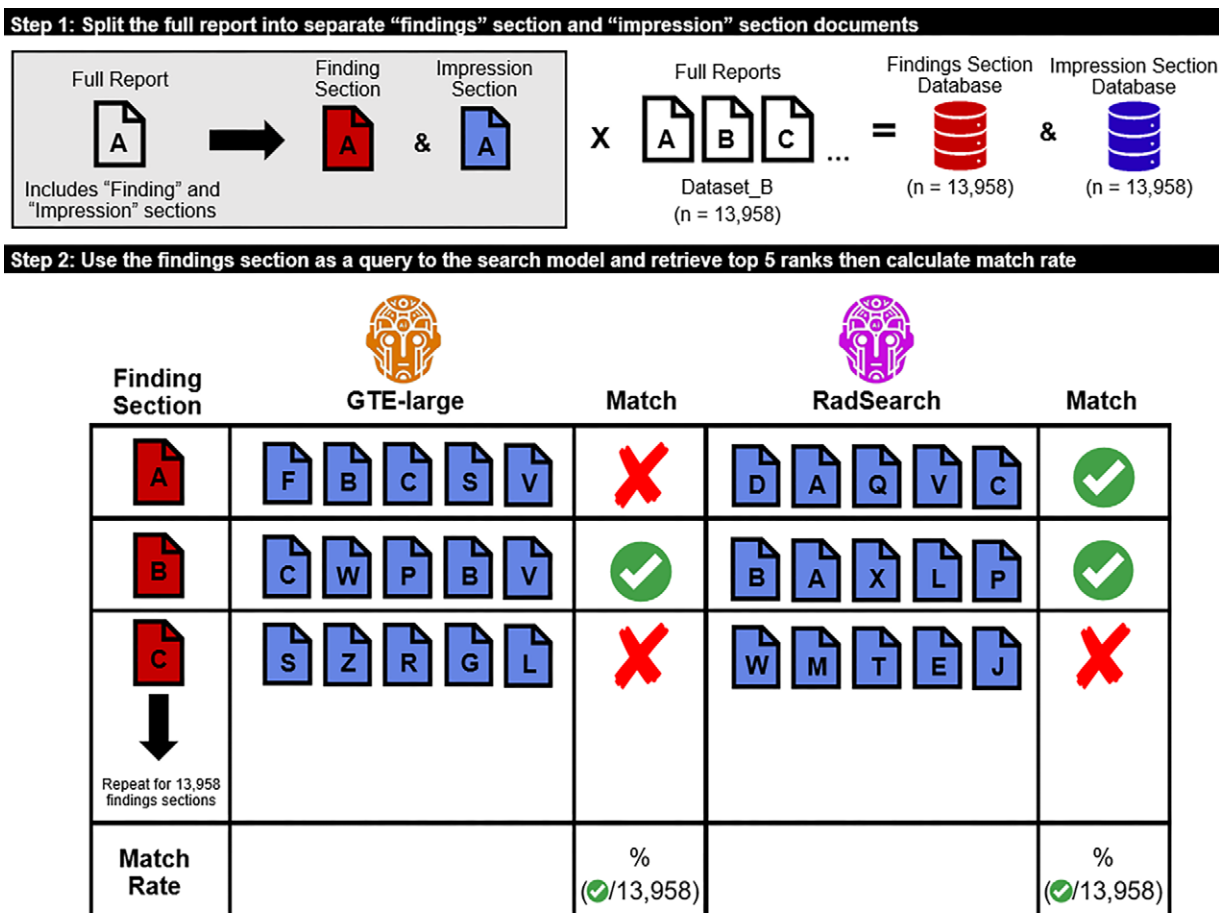


**Figure 2:** Flowchart detailing dataset selection and processing workflow. The study involved six datasets (A, B, B subset, C, D, and E) and involved two large tertiary medical centers. Datasets A–D were obtained from the University of Alabama at Birmingham (UAB) (green), and dataset E was obtained from the University of California, San Francisco (UCSF) (blue). Duplicate reports and reports missing a findings section or impression section were excluded from dataset A (the training set) and dataset B. Datasets B–D served as internal test sets, and dataset E served as an external test set. These datasets were used to evaluate three search performance metrics: (a) findings section-to-impression section matching (red shading, datasets B and E), (b) examination type matching (purple shading, dataset B), and (c) free-text similar report retrieval (orange shading, datasets B–D). Dataset B was used for free-text queries of aneurysm, cholecystitis, pancreatic cancer, and spinal fractures, while datasets C and D were used for free-text queries of pulmonary emboli and intracranial hemorrhage, respectively. CTPA = CT pulmonary angiogram.



**Figure 3:** A contrastive learning approach was used to train RadSearch to understand the underlying semantics of radiology reports. In this approach, each item in a training dataset is paired with at least one corresponding semantically similar item (positive pair) and one semantically dissimilar item (negative pair). For RadSearch, each item in the training dataset had one positive pair and 16,689 negative pairs (number of training dataset items minus one) (with the exception described below). To create these pairs, the full reports were first split into two documents corresponding to the findings section and the impression section, generating 16,690 findings sections and 16,690 impression sections. For each findings section, a positive pair was created by pairing it with its original impression section (from the full report); 16,689 negative pairs were generated for each findings section by pairing it with the impression section from a different radiology report. For reports that had an impression section that was identical to the impression section of another report, the number of negative pairs was equal to 16,689 minus the number of other reports with an identical impression section.



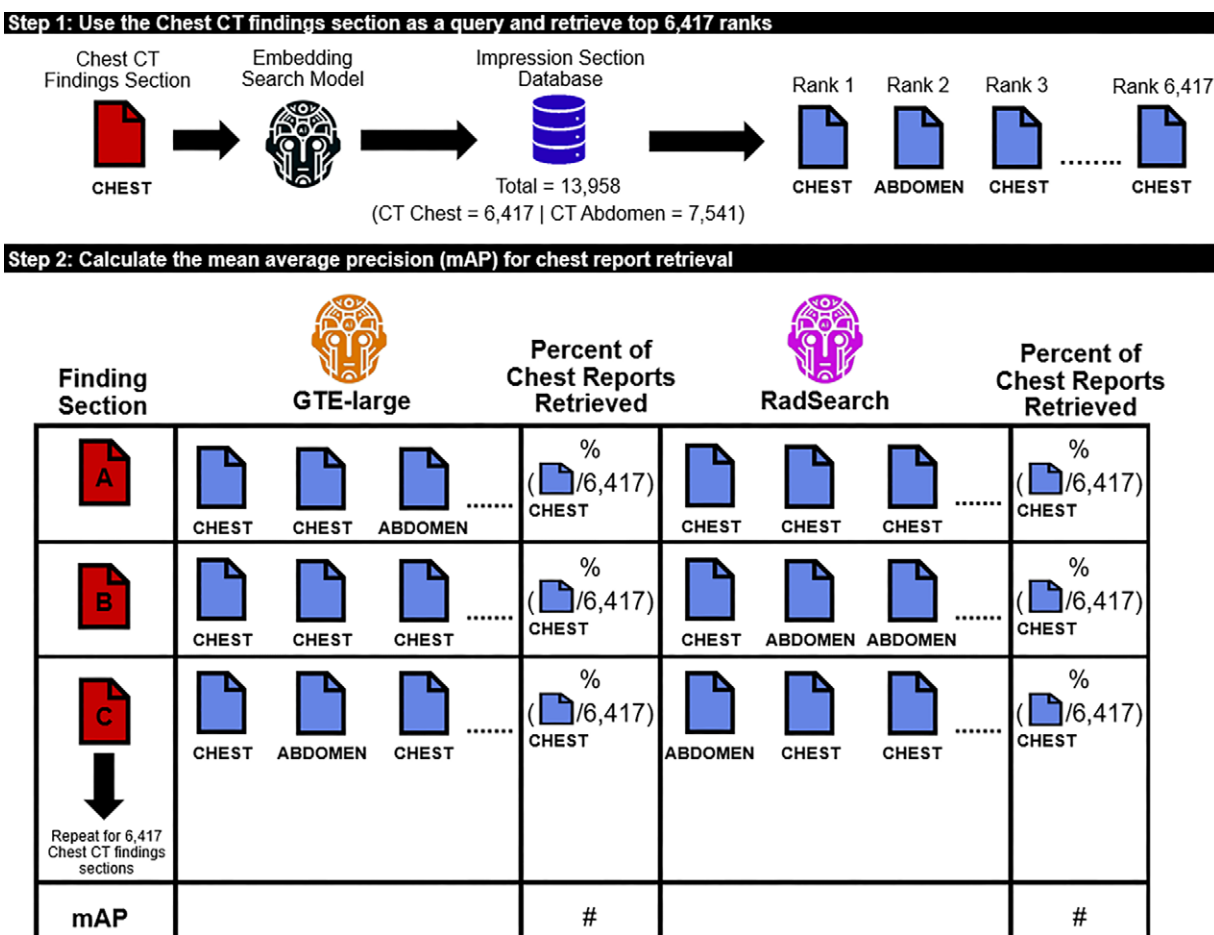


**Figure 4:** Diagram of the findings-to-impression matching metric for evaluating embedding model performance. In step 1, the 13,958 full radiology reports from an internal test set (dataset B), each of which includes a findings section (red) and an impression section (blue), are split, generating two separate documents for each full radiology report. This step produces a findings section database and an impression section database. Step 2 shows findings sections A, B, and C used as an input query to the embedding models GTE-large (15) and RadSearch. The models search the impression section database and retrieve the top five–ranked impression sections for being the most semantically similar to the queried findings section. GTE-large did not retrieve impression section A within its top five–ranked impression sections for findings section A and, therefore, was not counted as having a match (red X). RadSearch did retrieve impression section A within its top five–ranked impression sections for findings section A and was counted as having a match (green checkmark). This process was repeated for all 13,958 findings sections, and the percentage of matches (match rate) was calculated for GTE-large and RadSearch. Findings-to-impression matching was likewise evaluated for All MPNet Base (13) and MS MARCO DistilBERT Base (14) in database B, and for all four models in the external database (database E).

top 10–ranked reports from RadSearch and GTE-large for each query (1200 total reports) were evaluated by an intern radiology resident (C.H.S.) for their similarity to the query in terms of the presence of the general report finding (eg, aneurysm), the location of the general report finding (eg, *aortic* aneurysm), and the unique characteristics of the general report finding (eg, *partially thrombosed* aneurysm). Each retrieved report was independently manually reviewed by the resident, and the resident was blinded to which model retrieved the report. Two of the 60 queries lacked a location and eight queries lacked a unique characteristic, resulting in a total of 580 reports reviewed for similar location and 520 reports reviewed for similar unique characteristic. Additional details regarding the evaluation process are provided in Appendix S1.

**Evaluating LLM diagnostic performance using report finding descriptions without and with embedding model search assistance.**—To evaluate whether use of RadSearch could improve the ability of an LLM to generate an accurate diagnosis, the LLM Llama 3.1 8B Instruct (8 billion parameter version) (19)

was given a report finding description (eg, “Hepatic mass with nodular, discontinuous, peripheral enhancement on venous phase and progressive centripetal ‘fill-in’ on delayed images”) and prompted to provide the most likely diagnosis (eg, hepatic hemangioma). In total, 100 report finding descriptions were evaluated. This process was then repeated separately with RadSearch, where RadSearch calculated the cosine similarity score between the report finding description and each of the full reports from dataset B, an internal test set. The full report with the highest cosine similarity score was retrieved and added to the input prompt for the LLM, to provide additional context and assist the LLM in generating a diagnosis. This process was also separately performed using GTE-large the same way as RadSearch. The LLM without embedding model assistance provided the baseline LLM diagnostic performance. The LLM was not trained or altered, to ensure that any observed performance differences would be attributable to the embedding models rather than changes in the LLM. Additional details regarding the prompt format, reference standard process, and LLM settings are provided in Appendix S1.



**Figure 5:** Diagram of the examination type matching metric for evaluating embedding model performance. In step 1, the findings section (red) of a chest CT radiology report is given to the model as a search query for which it searches the impression section database of 13 958 impression sections (blue) and retrieves the top 6417-ranked impression sections. This number of impression sections was chosen to mirror the number of impression sections from the database that were from chest CT reports, to allow for the embedding model to potentially achieve a 100% chest report retrieval rate. In step 2, this process was repeated for all 6417 chest CT findings sections for GTE-large (15) and RadSearch. For each findings section, the percentage of the top 6417-ranked impressions sections that were from chest CT reports was calculated. The mean average precision (mAP) was then measured by calculating the mean percentage of chest impression sections retrieved for the 6417 chest CT findings section queries for each database. Examination type matching was likewise evaluated for All MPNet Base (13) and MS MARCO DistilBERT Base (14).

### Statistical Analysis

The match rate was calculated for the findings-to-impression matching metric, and was defined as the percentage of reports where the findings section was matched with its corresponding impression section within the top five-ranked impression sections. A pairwise comparison of match rates between All MPNet Base, MS MARCO DistilBERT Base, GTE-large, and RadSearch for reports from dataset B (UAB) was performed using the McNemar test with Bonferroni correction for multiple comparisons, adjusting the significance level to  $P \leq .008$  ( $\alpha = .05/6$ ). A separate pairwise comparison was performed for reports from dataset E (UCSF) (Table 1). Match rates were compared within models for contrast-enhanced chest CT reports from UAB (dataset B subset) versus UCSF (dataset E subset) using the  $\chi^2$  test (Table 2). The mAP of RadSearch was compared with the mAP of All MPNet Base, MS MARCO DistilBERT Base, and GTE-large using repeated-measures analysis of variance. A Bonferroni correction for multiple comparisons was applied, adjusting the significance level to  $P \leq .017$  ( $\alpha = .05/3$ ). The 95% CIs for mAP values were calculated using the Student *t*-distribution. For the free-text query analysis,

the reports retrieved by a model were assigned a score of 1 (present) or 0 (not present) individually for the general category, location, and unique characteristic. The binary results for the general category were compared between RadSearch and GTE-large using the McNemar test. This process was repeated for location and unique characteristic. A Bonferroni correction for multiple comparisons (general category, location, and unique characteristic) was applied, adjusting the significance level to  $P \leq .017$  ( $\alpha = .05/3$ ). The 95% CIs for proportions were calculated using the Wilson method. The diagnostic accuracy of the LLM with RadSearch integration was compared separately with that of the LLM alone and the LLM with GTE-large integration using McNemar tests. Statistical tests were performed using SciPy (version 1.13.0).

## Results

### Dataset Characteristics

The report exclusion process is depicted in Figure 2. The final training set included 16 690 reports (initial set,  $n = 20\,000$ ; excluded,  $n = 3310$ ). The final sample sizes for the internal test

**Table 1: Findings-to-Impression Matching Performance**

Measure	Dataset B*	Dataset E†
Match rate‡		
All MPNet Base	8.7 (1214/13 958)	4.0 (558/13 958)
MS MARCO DistilBERT Base	8.7 (1214/13 958)	9.7 (1354/13 958)
GTE-large	24.0 (3350/13 958)	18.7 (2610/13 958)
RadSearch	52.0 (7258/13 958)	39.3 (5485/13 958)
<i>P</i> value for comparison§		
All MPNet Base vs MS MARCO DistilBERT Base	.980	<.001
All MPNet Base vs GTE-large	<.001	<.001
All MPNet Base vs RadSearch	<.001	<.001
MS MARCO DistilBERT Base vs GTE-large	<.001	<.001
MS MARCO DistilBERT Base vs RadSearch	<.001	<.001
GTE-large vs RadSearch	<.001	<.001

Note.—The findings section of a report was used as a query to a model, which then searched a dataset of 13 958 impression sections and retrieved its top five–ranked impression sections for being the most semantically similar to the queried findings section. If the actual impression section of the queried findings section was present within those five retrieved impression sections, the outcome was defined as a “match” for that report. This process was repeated for each of the findings sections for dataset B and dataset E (Fig 4). The percentage of queried findings sections for which the model retrieved the actual impression section of the report (match rate) was calculated for All MPNet Base (13), MS MARCO DistilBERT Base (14), GTE-large (15), and RadSearch.

\* Internal test set of chest and abdomen CT reports from the University of Alabama at Birmingham ( $n = 13\,958$ ).

† External test set of contrast-enhanced chest CT reports from the University of California, San Francisco ( $n = 13\,958$ ).

‡ Data are percentages, with numbers of reports in parentheses.

§ Pairwise comparison of the match rate was performed using the McNemar test with Bonferroni correction for multiple comparisons, adjusting the significance level to  $P \leq .008$  ( $\alpha = .05/6$ ).

**Table 2: Findings-to-Impression Matching Performance on Contrast-Enhanced Chest CT Reports from Two Institutions**

Model	Match Rate for Dataset B Subset (UAB)*	Match Rate for Dataset E Subset (UCSF)†	<i>P</i> Value
All MPNet Base	11.9 (296/2490)	9.7 (242/2490)	.13
MS MARCO DistilBERT Base	12.3 (306/2490)	17.0 (423/2490)	.90
GTE-large	26.4 (657/2490)	29.5 (735/2490)	.79
RadSearch	46.0 (1145/2490)	51.7 (1287/2490)	.93

Note.—Data are percentages, with numbers of reports in parentheses. The percentage of reports where the findings section was matched with its correct impression section within the top five–ranked impression sections retrieved by the model (match rate) for each model is shown. The performance of All MPNet Base (13), MS MARCO DistilBERT Base (14), GTE-large (15), and RadSearch was compared between dataset B and dataset E using the  $\chi^2$  test. UAB = University of Alabama at Birmingham, UCSF = University of California, San Francisco.

\* Internal test subset of contrast-enhanced chest CT reports from UAB ( $n = 2490$ ).

† External test subset of an equal number of randomly selected contrast-enhanced chest CT reports from UCSF ( $n = 2490$ ) to compare to the internal test set.

sets were 13 958 for dataset B (initial set,  $n = 18\,203$ ; excluded,  $n = 4245$ ), 6178 for dataset C (no exclusions), and 9954 reports for dataset D (no exclusions). Dataset B subset included only contrast-enhanced chest CT reports ( $n = 2490$ ) from dataset B after exclusions. The final external test set included 13 958 reports (no exclusions).

### Report Findings-to-Impression Matching

When the findings section was used as the query and the model was tasked with retrieving the top five most similar impression sections in dataset B (internal test set with chest and abdomen CT reports), RadSearch retrieved the impression section corresponding to the findings section for 52.0% (7258 of 13 958) of queried reports, outperforming GTE-large, at 24.0% (3350 of 13 958); All MPNet Base, at 8.7% (1214 of 13 958); and MS MARCO DistilBERT Base, at 8.7% (1214 of 13 958) (all  $P < .001$ ) (Table 1). Likewise, in an external test set of

contrast-enhanced chest CT reports (dataset E), RadSearch retrieved the corresponding impression section for 39.3% (5485 of 13 958) of queried reports, outperforming GTE-large, at 18.7% (2610 of 13 958); All MPNet Base, at 4.0% (558 of 13 958); and MS MARCO DistilBERT Base, at 9.7% (1354 of 13 958) (all  $P < .001$ ).

To evaluate differences in performance across institutions, performance was compared for a subset of only contrast-enhanced chest CT reports from dataset B and an equal number of randomly selected reports from dataset E. There was no evidence of a difference in RadSearch performance in findings-to-impression matching across institutions (dataset B: 46.0% [1145 of 2490]; dataset E: 51.7% [1287 of 2490];  $P = 0.93$ ) (Table 2).

### Matching Reports with the Same Examination Type

For retrieving impression sections with the same examination type (CT chest) as findings section queries, RadSearch had an

**Table 3: Mean Average Precision for Examination Type Matching**

Model	Mean Average Precision*	P Value†
All MPNet Base	47.0 (46.5, 47.4)	<.001
MS MARCO DistilBERT Base	31.0 (30.8, 31.2)	<.001
GTE-large	52.0 (51.5, 52.5)	<.001
RadSearch	48.6 (48.2, 49.0)	...

Note.—The mean average precision of the four models (All MPNet Base [13], MS MARCO DistilBERT Base [14], GTE-large [15], and RadSearch) in matching reports with the same examination type are shown. The findings sections of the chest CT examinations in dataset B (contrast and noncontrast,  $n = 6417$ ) were used as a query to retrieve the impression sections of the chest and abdomen CT reports of dataset B ( $n = 13958$ ), and models were tasked with ranking the most similar impressions. To give the model a chance to achieve 100% retrieval precision, the model ranked the top 6417 most similar impression sections (equal to the total number of chest CT examinations in the dataset). The percentage of the 6417 chest CT reports that were present within the top 6417-ranked impression sections was calculated separately for each of the 6417 queries (ie, findings sections of CT chest reports), and these percentages were used to calculate the mean average precision of each models' rankings (Fig 5). Mean average precision ranges from 0% to 100%, with higher percentages indicating higher precision; 95% CIs were calculated using the Student  $t$ -distribution.

\* Data are percentages, with 95% CIs in parentheses.

† P value is for the comparison of mean average precision between each model and RadSearch using repeated-measures analysis of variance. A Bonferroni correction for multiple comparisons was applied, adjusting the significance level to  $P \leq .017$  ( $\alpha = .05/3$ ).

mAP of 48.6% (20 012 454 of 41 177 889), greater than the mAP of 47.0% (19 353 608 of 41 177 889) for All MPNet Base and 31.0% (12 765 146 of 41 177 889) for MS MARCO DistilBERT Base (both  $P < .001$ ) (Table 3). However, GTE-large outperformed RadSearch, with an mAP of 52.0% (21 412 502 of 41 177 889;  $P < .001$ ).

### Retrieving Reports Similar to Free-Text Queries

The performance of RadSearch and GTE-large in retrieving report findings similar to free-text queries is shown in Table 4. RadSearch retrieved results for a single query in less than 1 second for all queries. Overall, of the reports retrieved by RadSearch, 83.0% (498 of 600) matched the general report finding category of the query, 89.8% (521 of 580) matched the location of the query, and 50.6% (263 of 520) matched the unique characteristic of the query, outperforming GTE-large for general report finding category (65.7% [394 of 600];  $P < .001$ ) and location (58.8% [341 of 580];  $P < .001$ ). GTE-large had greater accuracy than RadSearch in retrieving matching reports for the general report finding of intracranial hemorrhage (92.0% [92 of 100] vs 66.0% [66 of 100];  $P < .001$ ) and for the unique characteristic of spinal fractures (63.3% [57 of 90] vs 30.0% [27 of 90];  $P < .001$ ). RadSearch outperformed GTE-large for the remaining five general report finding categories ( $P$  value range, <.001 to .004) and for location for all six categories ( $P$  value range, <.001 to .003).

### Diagnostic Performance of LLM without and with Embedding Model Search Integration

When the LLM was prompted alone without any embedding model, it correctly identified the most likely diagnosis for 30% (30 of 100) of the queries based on the given report finding descriptions. However, when the LLM was prompted with added context from RadSearch, its accuracy improved to 61% (61 of 100;  $P < .001$ ). This performance also surpassed that of the LLM when prompted with added context from GTE-large, which achieved an accuracy of 47% (47 of 100;  $P = .03$ ).

### Discussion

Semantic search models hold promise to improve radiology report search tool capabilities but lack scalable approaches to generate training data that capture the breadth of radiology-domain knowledge. Thus, in the present study, we developed a scalable approach for adapting a semantic search model for radiology report retrieval and trained a model, RadSearch, using this approach. The main study findings are as follows: First, in a simulated real-world search task with free-text queries, RadSearch retrieved reports containing the queried report finding for 83.0% (498 of 600) of reports and the finding location for 89.8% (521 of 580) of reports, outperforming a state-of-the-art semantic search model (GTE-large), at 65.7% (394 of 600;  $P < .001$ ) and 58.8% (341 of 580;  $P < .001$ ), respectively. Second, the performance of a large language model (LLM, Llama 3.1 8B Instruct) in providing a correct diagnosis for a given report finding description doubled from a baseline of 30% (30 of 100) without embedding model search integration to 61% (61 of 100) with RadSearch integration ( $P < .001$ ), outperforming the LLM with GTE-large integration, at 47% (47 of 100;  $P = .03$ ).

Shi et al (20) previously explored automated annotation methods for training semantic search models for radiology report retrieval and, in agreement with our results, noted improved performance in retrieving sentences related to aneurysms and pulmonary emboli compared to the baseline embedding model (Sentence-BERT). Likewise, their model achieved mAP values of 46% and 47% for retrieving queried chest radiograph findings in the Indiana Network for Patient Care chest radiograph dataset (21) and National Institutes of Health ChestX-ray8 dataset (22), respectively. While RadSearch did achieve an mAP of 48.6% on the examination type matching task, performance cannot be directly compared between these models because of differences in what text was used for queries and the number of retrievable items for a given query. Another important distinction is that Shi et al used lexicon-driven concept detection to develop query-sentence positive and negative pairs for training and evaluation. In contrast, in our study, a findings section coupled with its corresponding impression section was considered a positive pair, and a findings section coupled with the impression section of another report was considered a negative pair. This difference means that the model of Shi et al was trained for sentence-level retrieval, while RadSearch is more suited for retrieval at the report section level. An added benefit of our training approach is that it requires very little preprocessing (ie, only extracting the findings sections and impression sections from reports). In comparison, the approach of Shi et al required an already established chest radiograph finding lexicon (23) to generate labeled data. Such lexicons are not available for most modalities and their associated imaging findings (eg, tendinopathies of the knee at MRI), which limits the overall translatability to other modalities or examination types.



**Table 4: Report Retrieval Performance for Free-Text Queries**

Finding	Percent Correct for GTE-large	Percent Correct for RadSearch	<i>P</i> Value*
<b>Aneurysm</b>			
General category	60.0 (60/100) [50.2, 69.1]	84.0 (84/100) [75.6, 89.9]	<.001
Location	32.2 (29/90) [23.5, 42.4]	85.6 (77/90) [76.8, 91.4]	<.001
Unique characteristic	25.0 (20/80) [16.8, 35.5]	37.5 (30/80) [27.7, 48.5]	.11
<b>Pulmonary embolism</b>			
General category	75.0 (75/100) [65.7, 82.5]	92.0 (92/100) [85.0, 95.9]	.003
Location	61.0 (61/100) [51.2, 70.0]	89.0 (89/100) [81.4, 93.7]	<.001
Unique characteristic	38.9 (35/90) [29.5, 49.2]	42.2 (38/90) [32.5, 52.5]	.77
<b>Intracranial hemorrhage</b>			
General category	92.0 (92/100) [85.0, 95.9]	66.0 (66/100) [56.3, 74.5]	<.001
Location	74.4 (67/90) [64.6, 82.3]	95.6 (86/90) [89.1, 98.3]	<.001
Unique characteristic	76.3 (61/80) [65.9, 84.2]	61.3 (49/80) [50.3, 71.2]	.74
<b>Cholecystitis</b>			
General category	42.0 (42/100) [32.8, 51.8]	83.0 (83/100) [74.5, 89.1]	<.001
Location	64.0 (64/100) [54.2, 72.7]	100.0 (100/100) [96.3, 100.0]	<.001
Unique characteristic	38.9 (35/90) [29.5, 49.2]	63.3 (57/90) [53.0, 72.6]	.004
<b>Pancreatic cancer</b>			
General category	57.0 (57/100) [47.2, 66.3]	88.0 (88/100) [80.2, 93.0]	<.001
Location	58.0 (58/100) [48.2, 67.2]	79.0 (79/100) [70.0, 85.8]	.003
Unique characteristic	42.2 (38/90) [32.5, 52.5]	68.9 (62/90) [58.7, 77.5]	.002
<b>Spinal fracture</b>			
General category	68.0 (68/100) [58.3, 76.3]	85.0 (85/100) [76.7, 90.7]	.004
Location	62.0 (62/100) [52.2, 70.9]	90.0 (90/100) [82.6, 94.5]	<.001
Unique characteristic	63.3 (57/90) [53.0, 72.6]	30.0 (27/90) [21.5, 40.1]	<.001
<b>All</b>			
General category	65.7 (394/600) [61.8, 69.4]	83.0 (498/600) [79.8, 85.8]	<.001
Location	58.8 (341/580) [54.7, 62.7]	89.8 (521/580) [87.1, 92.0]	<.001
Unique characteristic	47.3 (246/520) [43.1, 51.6]	50.6 (263/520) [46.3, 54.9]	.33

Note.—Data are percentages, with numbers of reports in parentheses and 95% CIs in brackets. Ten free-text queries were created for each of the six general report finding categories (aneurysm, pulmonary embolism, intracranial hemorrhage, cholecystitis, pancreatic cancer, and spinal fracture), for a total of 60 free-text queries. GTE-large (15) and RadSearch retrieved the top 10 most similar full reports (findings and impression sections combined) for each query, resulting in 100 retrieved reports for each finding category. Performance was evaluated by determining whether the retrieved reports contained the correct general category, location, and unique characteristic mentioned in the free-text query. Two of the 60 queries lacked a location and eight queries lacked a unique characteristic, resulting in a total of 580 reports reviewed for location and 520 reports reviewed for unique characteristic. Aneurysm, cholecystitis, pancreatic cancer, and spinal fracture were evaluated using dataset B ( $n = 13\,958$  reports) as the queried database. Pulmonary embolism and intracranial hemorrhage were evaluated using dataset C ( $n = 6178$  reports) and dataset D ( $n = 9954$  reports), respectively, as the queried database. Datasets B–D were all internal test sets. Definitions for general category, location, and unique characteristic are detailed in Appendix S1. The reports retrieved by RadSearch and GTE-large were assigned a score of 1 (present) or 0 (not present) individually for the general category, location, and unique characteristic. 95% CIs were calculated using the Wilson method.

\* *P* value is for comparison between GTE-large and RadSearch using the McNemar test. A Bonferroni correction for multiple comparisons (general category, location, and unique characteristic) was applied, adjusting the significance level to  $P \leq .017$  ( $\alpha = .05/3$ ).

The implications of this study are multifaceted, spanning both clinical and research applications. RadSearch can be directly integrated into LLMs using retrieval-augmented generation to improve their ability to answer questions or perform tasks that require searching through radiology report databases. This integration could increase the specificity and reproducibility of LLM responses by providing relevant context for a query to the LLM before the LLM generates a response. For example, RadSearch integration could improve LLM performance in question answering, as supported by our results demonstrating improved accuracy in generating the most likely diagnosis. RadSearch could also be used as a diagnostic decision support tool.

Researchers could also potentially use RadSearch to streamline dataset curation. Current radiology information systems use keyword-based retrieval mechanisms and lack semantic search

capabilities, which limits the diversity and comprehensiveness of the retrieved dataset. RadSearch, by leveraging semantic understanding, can identify and group similar radiology reports based on nuanced textual descriptors, thereby capturing a broader and more diverse range of clinical scenarios. For instance, by matching reports for a specific imaging finding, RadSearch can facilitate the retrieval of medical images that represent a wide array of disease presentations for that finding. This diversity is critical for training robust artificial intelligence models, as they must be able to adapt to the variability and complexity of real-world clinical environments. Furthermore, this diversity is more likely to capture inherent disease variability—influenced by factors such as comorbidities, age, sex, and imaging modalities used—that may be missed when a dataset is curated using user-defined keywords or manual classification of radiology reports.

While these results are promising, limitations exist. First, RadSearch's errors often occurred when it was "distracted" by unintended findings in a query, especially when matching reports for a single target finding. A query describing findings seen in acute cholecystitis might retrieve reports that match the examination type and related findings like fat stranding, even if the gallbladder is normal. This issue may stem from using training data where the findings section includes many unique imaging findings besides acute cholecystitis, causing the model to associate these imaging findings with cholecystitis. Training RadSearch on a sentence-level dataset pairing specific findings with their corresponding impression sentences could improve specificity, but creating such a dataset requires substantial manual effort and is not scalable. Second, our study is limited by its retrospective design, which prevents direct real-world evaluation of RadSearch use by residents or attending radiologists. Given that a real-world evaluation would require addressing interoperability challenges, compliance with institutional information technology policies, and the creation of a complete software package tailored for deployment in real-world settings (eg, with user interfaces and error handling), it was considered beyond the scope of the current study.

In conclusion, we developed a scalable method to train a semantic search model for the radiology domain and evaluated the resulting model (RadSearch) on several retrieval tasks. RadSearch demonstrated high accuracy in retrieving radiology reports with similar imaging findings for simulated free-text clinical queries (outperforming a state-of-the-art semantic search model) and improved the ability of a large language model to provide accurate diagnoses. Further evaluation in a real-world clinical setting is needed to fully understand the utility and limitations of the model and will be performed in a separate future study.

**Deputy Editor:** Linda Moy

**Scientific Editor:** Sarah Atzen

#### Author affiliations:

<sup>1</sup> Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, Md

<sup>2</sup> Department of Radiology, University of Alabama at Birmingham Heersink School of Medicine, Birmingham, Ala

<sup>3</sup> Center for Intelligent Imaging, Department of Radiology and Biomedical Imaging, University of California, San Francisco, 505 Parnassus Ave, San Francisco, CA 94143

<sup>4</sup> Department of Diagnostic Imaging, St Jude Children's Research Hospital, Memphis, Tenn  
Received March 12, 2024; revision requested April 16; final revision received February 25, 2025; accepted February 28.

**Address correspondence to:** J.H.S. (email: jacho.sohn@ucsf.edu).

**Funding:** Supported by the Department of Radiology and Biomedical Imaging, University of California, San Francisco.

**Author contributions:** Guarantors of integrity of entire study, C.H.S., J.H.S.; 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, C.H.S., J.H.S.; clinical studies, C.H.S., J.H.S.; experimental studies, C.H.S., G.C., J.H.S.; statistical analysis, C.H.S., J.H.S.; and manuscript editing, all authors

**Disclosures of conflicts of interest:** C.H.S. On the *Radiology: Artificial Intelligence* Trainee Editorial Board. G.C. On the *Radiology: Artificial Intelligence* Trainee Editorial Board. A.D.S. Associate editor for *Radiology*. J.H.S. Early career consultant to the editor for *Radiology*.

## References

1. Goldschmidt DE, Krishnamoorthy M. Comparing keyword search to semantic search: a case study in solving crossword puzzles using the Google API. *Softw Pract Exper* 2008;38(4):417–445.
2. Beheshtian E, Putman K, Santomartino SM, Parekh VS, Yi PH. Generalizability and bias in a deep learning pediatric bone age prediction model using hand radiographs. *Radiology* 2023;306(2):e220505.
3. Bachina P, Garin SP, Kulkarni P, et al. Coarse race and ethnicity labels mask granular underdiagnosis disparities in deep learning models for chest radiograph diagnosis. *Radiology* 2023;309(2):e231693.
4. Rouzrokh P, Khosravi B, Faghani S, et al. Mitigating bias in radiology machine learning: 1. data handling. *Radiol Artif Intell* 2022;4(5):e210290.
5. Zhang K, Khosravi B, Vahdati S, et al. Mitigating bias in radiology machine learning: 2. model development. *Radiol Artif Intell* 2022;4(5):e220010.
6. Glocker B, Jones C, Roschewitz M, Winzeck S. Risk of bias in chest radiography deep learning foundation models. *Radiol Artif Intell* 2023;5(6):e230060.
7. Brady AP, Allen B, Chong J, et al. Developing, purchasing, implementing and monitoring AI tools in radiology: practical considerations. A multi-society statement from the ACR, CAR, ESR, RANZCR and RSNA. *Radiol Artif Intell* 2024;6(1):e230513.
8. Chen TL, Emerling M, Chaudhari GR, et al. Domain specific word embeddings for natural language processing in radiology. *J Biomed Inform* 2021;113:103665.
9. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology* 2023;308(1):e230970.
10. Liu Z, Zhong A, Li Y, et al. Radiology-GPT: a large language model for radiology. *arXiv 2306.08666* [preprint] <https://arxiv.org/abs/2306.08666>. Posted June 14, 2023. Updated March 19, 2024. Accessed April 25, 2024.
11. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. *arXiv 1908.10084* [preprint] <https://arxiv.org/abs/1908.10084>. Posted August 27, 2019. Accessed January 29, 2024.
12. Yan A, McAuley J, Lu X, et al. RadBERT: adapting transformer-based language models to radiology. *Radiol Artif Intell* 2022;4(4):e210258.
13. sentence-transformers/all-mpnet-base-v2. Hugging Face. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>. Accessed October 20, 2023.
14. sentence-transformers/msmarco-distilbert-base-v4. Hugging Face. <https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v4>. Accessed October 20, 2023.
15. Alibaba-NLP/gte-large-en-v1.5. Hugging Face. <https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5>. Accessed May 1, 2024.
16. Muennighoff N, Tazi N, Magne L, Reimers N. MTEB: Massive Text Embedding Benchmark. *arXiv 2210.07316* [preprint] <https://arxiv.org/abs/2210.07316>. Posted October 13, 2022. Updated March 19, 2023. Accessed May 1, 2024.
17. Dean B. We analyzed 4 million Google search results. Here's what we learned about organic CTR. *Backlinko*. <https://backlinko.com/google-ctr-stats>. Published 2019. Updated March 4, 2025. Accessed March 22, 2025.
18. Su W, Yuan Y, Zhu M. Threshold-free evaluation of medical tests for classification and prediction: average precision versus area under the ROC curve. *arXiv 1310.5103* [preprint] <https://arxiv.org/abs/1310.5103>. Posted October 18, 2013. Accessed August 20, 2024.
19. Dubey A, Jauhri A, Pandey A, et al. The Llama 3 herd of models. *arXiv 2407.21783* [preprint] <https://arxiv.org/abs/2407.21783>. Posted July 31, 2024. Updated November 23, 2024. Accessed March 22, 2025.
20. Shi L, Syeda-Mahmood T, Baldwin T. Improving neural models for radiology report retrieval with lexicon-based automated annotation. In: Carpuat M, de Marneffe MC, Meza Ruiz IV, editors. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, 2022; 3457–3463.
21. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc* 2016;23(2):304–310.
22. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers, 2017; 3462–3471.
23. Syeda-Mahmood T, Wong KCL, Wu JT, Jadhav A, Boyko O. Extracting and learning fine-grained labels from chest radiographs. *arXiv 2011.09517* [preprint] <https://arxiv.org/abs/2011.09517>. Posted November 18, 2020. Accessed January 29, 2024.