Supplementary Materials

Data Pre-Processing Steps

Radiology reports were preprocessed by splitting each clinical note into a **context** section and an **impression** section (the final diagnostic summary provided by the radiologist) as shown in Table S1. These were extracted from the original MIMIC IV 2.2 free-form notes using regular expressions. The context was fed into the retrieval data store and prompts for fine-tuning, while the prediction target was the parsed impression text. Context and impression sections were extracted using case-insensitive regular expressions matching section headers (e.g., “IMPRESSION:”), with all preceding text assigned to context and the remaining text assigned to the prediction target. Reports lacking a valid impression header or exhibiting malformed structure were excluded, ensuring consistent supervision targets while preserving coverage in the external datastore.

Table S1: Radiology Notes Text Structure

Original radiology report (raw MIMIC note format)	**EXAMINATION:** CHEST (PA AND LAT) **INDICATION:** ___ with new onset ascites // eval for infection **TECHNIQUE:** Chest PA and lateral **COMPARISON:** None. **FINDINGS:** There is no focal consolidation, pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomediastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted. **IMPRESSION:** No acute cardiopulmonary process.
Preprocessed model input (Context)	examination: chest (pa and lat) indication: with new onset ascites eval for infection technique: chest pa and lateral comparison: none findings: there is no focal consolidation, pleural effusion or pneumothorax. bilateral nodular opacities that most likely represent nipple shadows. the cardiomediastinal silhouette is normal. clips project over the left lung, potentially within the breast. the imaged upper abdomen is unremarkable. chronic deformity of the posterior left sixth and seventh ribs are noted.
Prediction target (Impression)	impression: no acute cardiopulmonary process.

Description of Full MIMIC IV 2.2 Radiology Notes

Prior to experimentation, we explored the MIMIC-IV radiology notes table, which contains semi-structured free-text reports from imaging studies including X-ray, CT, MRI, and ultrasound, typically following protocol-specific reporting templates. After deduplication by note identifier, the dataset comprised **2,321,355 unique radiology notes (note_id)** from **237,427 distinct patients (subject_id)**. A substantial proportion of patients had longitudinal records, contributing more than one report, reflecting repeated imaging across multiple visits or admissions. Notes were predominantly full radiology reports (RR, **2,295,635; 98.9%**), with a small fraction of addenda (AR, **25,720; 1.1%**). Each note is linked to a patient identifier (*subject_id*), hospitalization identifier (*hadm_id*), and time of charting (*charttime*), enabling both patient-level aggregation and admission-level temporal analyses.

Table S2 summarizes row-level duplication across different components of radiology reports after preprocessing (see Table S4 for definitions of *context* and *impression*). Although each report is associated with a unique note identifier, textual duplication arises naturally in clinical documentation due to repeated imaging protocols, follow-up examinations, and the use of standardized reporting templates. At the level of full report text, only **33,483 of 2,321,355 notes (1.46%)** were exact duplicates, indicating that complete reports are rarely repeated verbatim. Duplication is slightly higher in the context portion of reports, with **60,633 notes (2.61%)** sharing identical contextual descriptions, reflecting reuse of protocol-driven phrasing and similar examination workflows. In contrast, diagnostic summaries (i.e. Impressions) are substantially more templated: **762,999 notes (32.9%)** share identical impression text, consistent with the frequent use of short, standardized conclusions such as “no acute cardiopulmonary process.” These patterns indicate that duplication is concentrated in short diagnostic statements and protocol-driven sections rather than in full narrative reports, supporting the use of retrieval methods that leverage shared clinical phrasing without relying on memorization of entire documents.

Table S2. Duplicate and unique text structure in radiology reports after preprocessing. See Table S1 for explanation of context and impression.

Category	Count of note_id	Count of duplicate note_ids	% of Total duplicate note_ids
Total # of notes (note_id)	2,321,355		
Raw Examinations (Full Report Text)	2,287,872	33,483	1.46%
Contexts	2,260,722	60,633	2.61%
Impressions	1,558,356	762,999	32.9%

Model Training Dataset

We used a **subsampled subset** of the MIMIC-IV radiology notes table from Table S2 to construct the external retrieval datastore used for both kNN-LM and RETOMATON experiments (Table S3). Reports consist of semi-structured free-text radiology narratives associated with imaging studies including X-ray, CT, MRI, ultrasound, and interventional procedures, typically following protocol-specific reporting templates. After preprocessing and filtering to the subset used for datastore construction and training, the dataset comprised **162,475 radiology notes** from **85,397 distinct patients (subject_ids)**. Notes in our subsample were overwhelmingly full radiology reports (RR, **162,386; 99.95%**), with a very small fraction of addenda (AR, **89; 0.05%**). Each note is linked to a patient identifier (*subject_id*), hospitalization identifier (*hadm_id*), and time of charting (*charttime*), enabling patient-level aggregation and admission-level temporal analyses.

Table S3 summarizes row-level textual duplication within the radiology subset used for fine-tuning models and constructing the retrieval datastore for the experiments in Figures 3–4 (see Table S4 for definitions of *context* and *impression*). Although each report is associated with a unique note identifier, limited duplication arises naturally from standardized reporting practices, repeated imaging protocols, and follow-up examinations. Duplication is rare at the level of full report text (**511 of 162,475 notes; 0.31%**), indicating that complete narrative reports are almost always unique. Context sections, which contain structured examination details and findings, show slightly higher duplication (**1,354 notes; 0.83%**), reflecting reuse of protocol-driven phrasing and templated descriptions. In contrast, impressions are substantially more repetitive (**30,377 notes; 18.7%**), as they are short, highly standardized diagnostic summaries that frequently reuse common clinical conclusions (e.g., “no acute cardiopulmonary process” or “no evidence of malignancy”). This distribution confirms that duplication is concentrated in brief diagnostic statements rather than in full narrative content, and supports the use of retrieval methods that reinforce common diagnostic phrasing while preserving diversity in detailed clinical descriptions.

Table S3. Duplicate and unique text structure in radiology reports after preprocessing for fine-tuning models and data store creation experiments in Figure 3. See Table S1 for explanation of context and impression.

Category	Count of note_id	Count of duplicate note_ids	% of total duplicate note_ids
Total # of notes (note_id)	162,475		
Raw Examinations (Full Report Text)	161,964	511	0.31%
Contexts	161,121	1,354	0.83%
Impressions	132,098	30,377	18.7%

Final Generations for Testing Dataset and Blind Evaluation

A held-out blind test set was constructed from the MIMIC-IV radiology notes table for final generation and evaluation of next-token retrieval models across multiple metrics. Reports consist of semi-structured free-text radiology narratives associated with imaging studies including X-ray, CT, ultrasound, and interventional procedures, typically following protocol-specific reporting templates. After filtering and preprocessing, the evaluation set comprised **500 radiology notes** from **498 distinct patients**, with only **2 patients contributing more than one report (subject_id)**, ensuring minimal patient-level leakage across samples. All notes in the evaluation set were full radiology reports (RR, **500; 100%**), with no addenda included. Each note is linked to a patient identifier (*subject_id*), hospitalization identifier (*hadm_id*), and time of charting (*charttime*), enabling patient-level and admission-level aggregation if required.

Radiology reports were preprocessed using the same pipeline as the datastore subset, by splitting each note into a **context** and impression, as shown in Table S4. **Table S4** summarizes textual duplication in the held-out blind test set used for final radiology generation and evaluation (N = 500). This split was constructed to minimize overlap at the level of model inputs, ensuring that performance reflects generalization rather than retrieval of near-duplicate contexts. Accordingly, no duplication was observed in either full report text or preprocessed context inputs (0 of 500 notes, 0%). Limited duplication remains in diagnostic summaries, with **41 notes (8.2%)** sharing identical impression text, reflecting the use of short, standardized clinical conclusions such as “no acute cardiopulmonary process” or “no acute fracture.” This distribution preserves realistic repetition in diagnostic phrasing while ensuring that all model inputs are textually distinct, supporting fair evaluation of next-token retrieval methods under clinically realistic conditions.

Table S4. Statistics for radiology notes in the blind test set (N = 500). Row-level duplication is reported separately for full report text, preprocessed context (model input), and impression (reference target). No duplication is present in full reports or contexts, while limited repetition occurs in impressions due to standardized diagnostic phrasing, enabling evaluation of retrieval-augmented generation without input-level memorization effects.

Category	Count of note_id	Count of duplicate note_ids	% of total duplicate note_ids
Total # of notes (note_id)	500		
Raw Examinations (Full Report Text)	500	0	0%
Contexts	500	0	0%
Impressions	459	41	8.2%

Given the use of standardized radiology reporting templates in the original dataset, we performed an additional check to ensure that preprocessed context texts (**defined in Table S1**) in the blind test set (Table S4) overlapped with those used to construct the datastore (Table S3). This analysis identified three test-set notes whose context text was identical to contexts present in the training datastore. **Specifically, blind test notes 10080299-RR-13 (matching training notes 14608837-RR-42, 15231816-RR-90, 18200435-RR-37, and 18191686-RR-55), 17081475-RR-11 (matching 14440624-RR-30), and 11614807-RR-56 (matching 14681464-RR-45) shared identical context text with training examples.** Given the small number of overlapping cases, we expect the impact on aggregate performance metrics to be negligible, but we report this overlap in case further analysis of individual records is required. The test set with 500 records and results reported include these 3 overlapping examples with the input datastore.

Supplementary Results

Table S5 for Figure 3 Results. Normality assessment for LLaMA metrics (Shapiro–Wilk test): Table S1. Kruskal–Wallis and Dunn's post-hoc comparisons for LLaMA (Base vs. variants). *Kruskal–Wallis non-parametric tests were used to assess overall group differences, followed by Dunn's post-hoc tests (adjusted p-values) comparing each condition against the base model.* Shapiro–Wilk tests were performed separately for each metric and condition (n = 500 per condition). All distributions violated the assumption of normality (α = 0.05), motivating the use of non-parametric statistical tests throughout.

Metric	Base	Base vs 10% R-LM	Base vs 25% R-LM	Base vs 50% R-LM	Base vs 75% R-LM	Base vs 90% R-LM	Base vs Finetuning
ROUGE-L Kruskal-Wallis **χ^2 = 265.68**	0.23 ± 0.13 (Control)	0.25 ± 0.13 (ns)	0.28 ± 0.17 (***)	0.31 ± 0.21 (****)	0.31 ± 0.23 (****)	0.31 ± 0.23 (****)	0.44 ± 0.26 (****)
Perplexity Kruskal-Wallis **χ^2 = 659.28**	11.72 ± 10.30 (Control)	6.68 ± 4.82 (****)	4.91 ± 4.37 (****)	5.68 ± 4.49 (****)	6.21 ± 5.45 (****)	7.52 ± 7.46 (****)	2.70 ± 1.58 (****)

Significance codes: ns = not significant; ** p< 0.01; *** p < 0.001; **** p < 0.0001.

Table S6. Kruskal–Wallis and Dunn's post-hoc comparisons for OLMo (Base vs. variants): Kruskal–Wallis non-parametric tests were used to assess overall group differences, followed by Dunn's post-hoc tests (adjusted p-values) comparing each condition against the base model. Shapiro–Wilk tests were performed separately for each metric and condition (n = 500 per condition). All distributions violated the assumption of normality (α = 0.05), motivating the use of non-parametric statistical tests throughout.

Metric	Base	Base vs 10% R-LM	Base vs 25% R-LM	Base vs 50% R-LM	Base vs 75% R-LM	Base vs 90% R-LM	Base vs Finetuning
ROUGE-L Kruskal-Wallis **χ^2 = 481.58**	0.19 ± 0.09 (Control)	0.20 ± 0.10 (ns)	0.24 ± 0.15 (***)	0.28 ± 0.22 (****)	0.28 ± 0.22 (****)	0.28 ± 0.22 (****)	0.44 ± 0.25 (****)
Perplexity Kruskal-Wallis **χ^2 = 860.11**	17.92 ± 18.18 (Control)	9.65 ± 8.38 (****)	8.92 ± 8.11 (****)	9.17 ± 9.03 (****)	11.09 ± 12.26 (****)	15.37 ± 19.51 (***)	2.63 ± 1.60 (****)

Significance codes: ns = not significant; ** p< 0.01; *** p < 0.001; **** p < 0.0001.

Table S7. Kruskal–Wallis and Dunn's post-hoc comparisons for Qwen (Base vs. variants): Kruskal–Wallis non-parametric tests were used to assess overall group differences, followed by Dunn's post-hoc tests (adjusted p-values) comparing each condition against the base model. Shapiro–Wilk tests were performed separately for each metric and condition (n = 500 per condition). All distributions violated the assumption of normality (α = 0.05), motivating the use of non-parametric statistical tests throughout.

Metric	Base	Base vs 10% R-LM	Base vs 25% R-LM	Base vs 50% R-LM	Base vs 75% R-LM	Base vs 90% R-LM	Base vs Finetuning
ROUGE-L Kruskal-Wallis **χ^2 = 271.34**	0.24 ± 0.15 (Control)	0.25 ± 0.16 (ns)	0.27 ± 0.18 (**)	0.30 ± 0.22 (***)	030 ± 0.22 (**)	0.30 ± 0.22 (**)	0.43 ± 0.25 (****)
Perplexity Kruskal-Wallis **χ^2 = 572.04**	8.81 ± 6.90 (Control)	5.88 ± 4.08 (****)	5.57 ± 4.00 (****)	5.78 ± 4.50 (****)	6.93 ± 6.10 (****)	9.42 ± 9.55 (ns)	0.43 ± 0.25 (****)

Significance codes: ns = not significant; ** p< 0.01; *** p < 0.001; **** p < 0.0001.

Table S8 - Sample Generated Outputs and Hallucination Examples with R-LMs at 50% Retrieval Strength. The full set of results reported in this article is uploaded to Physionet.

Reference	Generated	ROUGE-L	Perplexity	Hallucination (Y/N)
impression: no acute cardiopulmonary process.	impression: no acute intrathoracic process. no evidence of free air.	0.53	1.74	Y
impression: technically successful fine needle aspiration of the right thyroid nodule. no periprocedural complications. cytology is pending.	impression: uncomplicated fine needle aspiration of the right thyroid nodule. no periprocedural complications. cytology is pending.	0.90	1.3	N
impression: stable post-treatment changes on the left. no evidence for new or recurrent malignancy. findings were discussed with the patient. bi-rads 2 - benign findings. routine annual followup suggested.	impression: no significant interval changes with some subtle residual fibroglandular densities. bilateral vascular calcifications on multiple prior exams without specific mr evidence.	0.15	5.38	Y
impression: no pleural effusion.	impression: no evidence of pleural effusion.	0.8	1.97	N
impression: ap chest compared to through : left picc line follows a non-standard course in the chest descending to the left of the midline. it is either in left internal mammary vein or persistent left-sided superior vena cava, distinguishable by the course of the line on a lateral view which should be obtained. there is no pneumothorax or pleural effusion. lungs are clear. heart size top normal. transvenous right atrial and right ventricular pacer leads follow their expected courses. right jugular line tip projects over the mid svc. findings were discussed by dr. with the clinical care team.	impression: technical artifact obscures the detail in the left mid and lower lung fields.	0.09	11.28	Y