Original Paper

Cross-Site Stability of Embedding Model Rankings for Known-Item Retrieval From Clinical Notes: Two-Corpus Comparative Evaluation Study

Yngve Mikkelsen, MD, MSc, DBA

Saïd Business School, University of Oxford, Oxford, England, United Kingdom

Corresponding Author:
Yngve Mikkelsen, MD, MSc, DBA
Saïd Business School
University of Oxford
Park End Street
Oxford, England, OX1 1HP
United Kingdom
Phone: 44 1865 422800
Email: yngve.mikkelsen@sbs.ox.ac.uk

Abstract

Background: Retrieval-augmented generation over clinical text depends on the embedding model chosen—a choice that interacts with documentation genre. Cross-site stability of embedding-model rankings has not been evaluated on real electronic health record (EHR) narratives with genre held constant; public benchmarks are predominantly single-source, and multi-institution evaluations compare models and query formulations within tasks rather than ranking stability across institutions.

Objective: The aim of the study is to test whether embedding model rankings transfer across 2 institutional EHR corpora, with genre held constant; to examine the contributions of model, site/corpus, and genre and their dependence on scoring; and to test whether the dominant contributor depends on model-panel composition.

Methods: We compared 13 embedding models on a known-item passage-to-source-document task, each query a deterministic 2-sentence excerpt from its target, in a 2×2 design crossing site/corpus (Beth Israel Deaconess Medical Center via MIMIC-IV-Note [Medical Information Mart for Intensive Care IV Clinical Notes Module] and University of California, San Francisco via ER-Reason) with genre (discharge summary and imaging report). Query sentences were removed from their targets before indexing to prevent verbatim overlap, notes were deduplicated after normalization, models used their recommended pooling, long documents were scored as overlapping chunks by maximum similarity, and cell sizes were matched (N=1235 per cell). Effectiveness was measured by mean reciprocal rank at 10, rank transfer by Kendall τ and selection regret, and variance by factorial decomposition (η^2). Transfer, regret, and the contrastive-panel decomposition carry 95% patient-clustered bootstrap intervals.

Results: Among the 8 contrastive models, cross-site Kendall τ was 0.86 (95% CI 0.64-1.00) for discharge and 0.64 (95% CI 0.36-0.86) for imaging. The model×site interaction was small under every scoring choice (η^2=0.005-0.029); the site/corpus main effect was negligible under best-chunk scoring (η^2=0.000) but scoring-dependent (0.172 under mean-pooling). Cross-site selection regret was 0 for imaging and up to 0.033 for discharge (~10% of the destination's best mean reciprocal rank; intervals 0-0.071). Adding 5 nonretrieval masked-language-model comparators moved the largest term to model (η^2=0.874): the dominant-factor question is panel-sensitive, while the qualitative transfer conclusion was unchanged.

Conclusions: For known-item retrieval from clinical notes at 2 US academic medical centers, embedding-model rankings showed positive cross-site concordance, moderate for imaging and strong for discharge summaries—with intervals wide enough that direction, not magnitude, is the finding. Which factor dominates absolute effectiveness depends on the panel evaluated: primary contrastive-panel scoring favors genre, whereas adding nonretrieval comparators makes model choice dominant; so, claims about major determinants are panel-specific. Cross-site model selection incurred zero to modest performance loss. When documentation genre and retrieval configuration are matched, model-selection results from one institutional EHR corpus may be a useful starting point for another, although local confirmation remains warranted.

(JMIR AI 2026;5:e109305) doi: 10.2196/109305

XSL•FO
RenderX

KEYWORDS

clinical natural language processing; information retrieval; embedding models; retrieval-augmented generation; external validation; electronic health records

Introduction

Retrieval-augmented generation (RAG) has become a common architecture for applying large language models to clinical free text [1], and the quality of the retrieval stage is largely determined by the embedding model used to represent documents and queries. A growing body of evaluation has established that no single embedding model is uniformly best for clinical retrieval, and that performance depends not only on the model but also on characteristics of the task and corpus [2-4].

Prior work [2] established, in a factorial evaluation across 3 document collections, that clinical context variables—documentation genre and query formulation—collectively rival the choice of embedding model in explaining variance in retrieval effectiveness, and that model rankings are not fully portable across documentation genres or query strategies. That study was explicit about a boundary it could not cross: its 3 collections represented different documentation contexts (dictated transcriptions, published case reports, and synthetic notes) rather than different institutions, and it could not test whether embedding model effectiveness varies across institutions. It named that question directly, proposing multisite validation analogous to the Dartmouth Atlas approach to unwarranted variation [5].

Evaluation of clinical embedding models has broadened since, but not along the institutional axis. A large-scale benchmark compared 39 embedding models across 7 health care semantic tasks on Mount Sinai and MIMIC-IV (Medical Information Mart for Intensive Care IV) data, reframing each task as retrieval scored by mean reciprocal rank (MRR) and recall@k [4]. A public electronic health record (EHR) retrieval benchmark with graded semantic-match assessment provides a reusable testbed, though a single-institution and single-genre one [6]. A multi-institution evaluation compared 7 models with alternative pooling strategies and query formulations across 2 EHR sources and found that performance on public benchmarks does not necessarily transfer to clinical corpora [3]. A mechanistic companion study examined layer-level representation geometry across 13 transformer configurations and 3 clinical corpora, finding model rankings stable when evaluation was expanded within public sources and identifying retrieval-objective training and embedding geometry as determinants of effectiveness [7]. Together these establish that model choice matters and that effectiveness is task- and corpus-dependent. None reports whether the relative ordering of models is preserved when the institution changes and the documentation genre is held fixed, which is the question this study addresses.

The question has practical weight. If model rankings are unstable across institutions, a benchmark result from one hospital or a public corpus cannot guide model selection at another, and each deployment site would need to repeat evaluation locally. If rankings are stable, evaluation effort is transferable, and public benchmarks retain value. The question bears directly on how clinical informatics teams should budget for the nontrivial work of embedding model selection.

To our knowledge, cross-site stability of embedding-model rankings has not been evaluated on real institutional EHR narratives while holding the documentation genre constant. Public clinical retrieval benchmarks are predominantly single-source, and a recent multi-institution embedding evaluation [3] evaluated embedding models, pooling strategies, and query formulations across 2 EHR sources but did not quantify the stability of model rankings across institutions. We address this question using 2 real institutional EHR corpora that share 2 documentation genres. The second question carries a specific comparison. A factor disturbs a deployment decision only insofar as it changes the ordering of models, so the quantity of interest is the model-by-site interaction, and its magnitude is interpretable only against a benchmark. We use the model-by-genre interaction, since genre is already known to perturb rankings [2]. We therefore ask the following research questions (RQs):

- RQ1: Do embedding model rankings for clinical text retrieval transfer across institutional EHR corpora when documentation genre is held constant?
- RQ2: How do model, site/corpus, and genre, and their interactions, contribute to variation in retrieval effectiveness, and does the model-by-site interaction differ from the model-by-genre interaction?

Methods

Study Design

We used a 2×2 factorial design crossing site/corpus with documentation genre, evaluating 13 embedding models across all 4 cells. The site/corpus factor contrasted Beth Israel Deaconess Medical Center (BIDMC, Boston), represented by MIMIC-IV-Note (Medical Information Mart for Intensive Care IV, Clinical Notes Module), with the University of California, San Francisco (UCSF), represented by ER-Reason. The genre factor contrasted discharge summaries with diagnostic imaging reports, both present in both corpora. Holding genre constant across sites isolates the site comparison from document-type variation. We refer to the experimental factor as “site/corpus” rather than treating it as a pure institutional effect, because the 2 corpora differ not only in institution but also in note-selection pipeline, care setting, and deidentification, which the design cannot separate.

Data Sources and Governance

MIMIC-IV-Note [8-10] is a deidentified corpus of BIDMC notes; ER-Reason (version 1.0.0) [11] is a deidentified corpus of UCSF emergency-department and inpatient notes, structured at the encounter level with one column per note genre. Both require credentialed access under their data use agreements. All processing was performed locally; no note text was transmitted

XSL•FO
RenderX

to any third-party service. Because ER-Reason attaches historical notes to selected index encounters and its deidentification shifts dates, we did not treat the corpora as temporally disjoint (see Limitations section).

Retrieval Task and Overlap Control

We used a known-item, passage-to-source-document retrieval formulation. Throughout, query denotes the text submitted to the retriever, not a clinician-authored information need: for each document, a query was derived deterministically from the document's narrative body as a 2-sentence excerpt of narrative prose, and the target was that source document within a pool of same-genre, same-site documents. The task is therefore passage-to-source-document matching rather than question answering, and the implications for generalizability are discussed in the Limitations section. This approach requires no relevance labels. That matters here because the 2 corpora record diagnoses under incompatible schemes—MIMIC-IV-Note uses *International Classification of Diseases* (ICD) codes and ER-Reason uses free-text diagnosis names with no ICD field—so any shared category label would have to be derived differently at each site, introducing labeling error that is correlated with site.

Because a query drawn from its target shares vocabulary with it, we removed each query's sentences from its target document before indexing, so that a model must match the remainder of the note rather than the query's verbatim presence. Documents whose target still contained the query after removal (notes that repeat the same passage internally) were excluded. This overlap control materially lowers absolute MRR relative to a formulation that leaves the query in the target and yields a task that better approximates content-based retrieval.

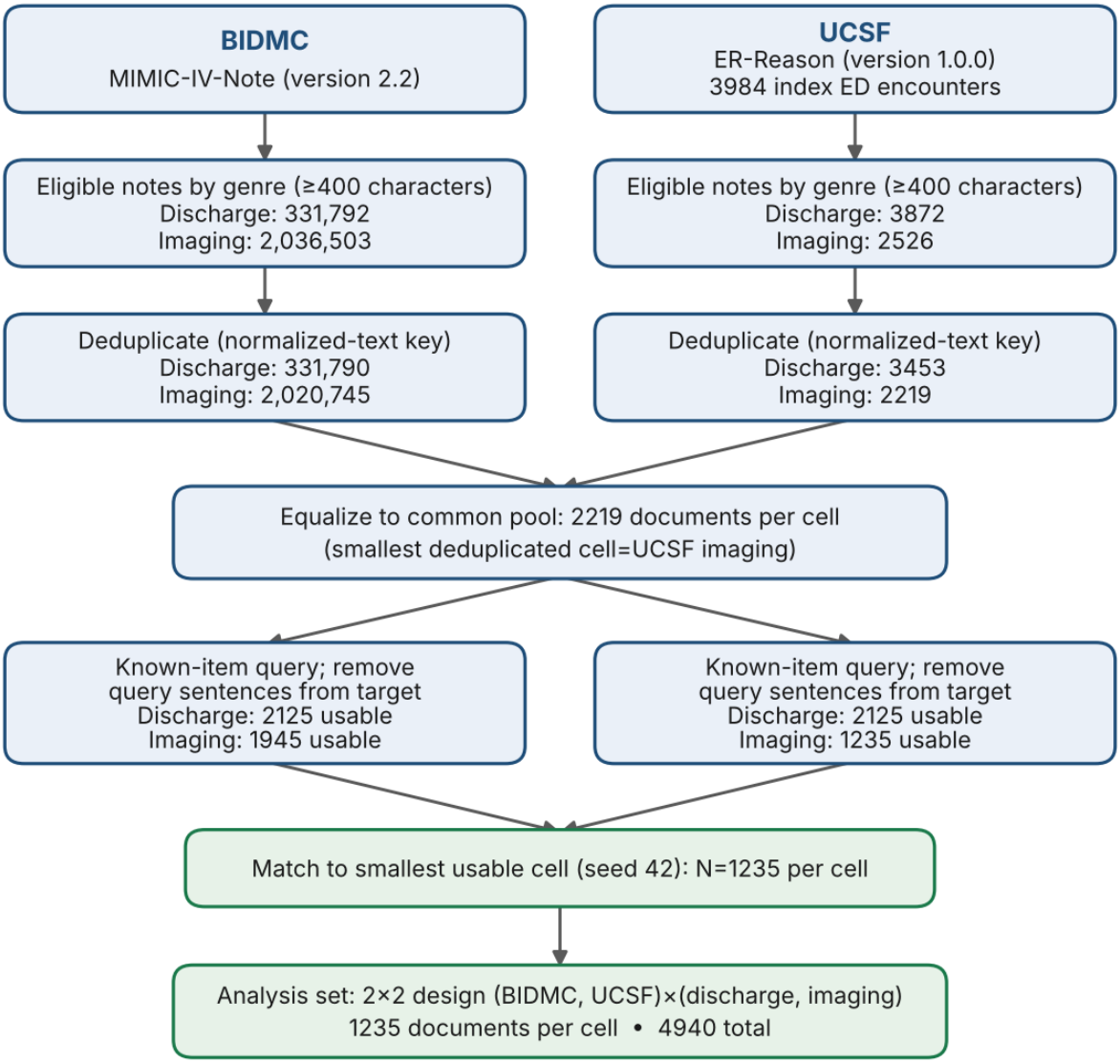
The query extractor is deterministic and identical at both sites. Three alternatives were considered and rejected. A randomly chosen sentence would frequently select header, field-value, or templated text, which carries little document-identifying content, so retrieval difficulty would depend on where the draw landed rather than on the model under test. A model-generated question would place a second model inside the measurement, so that a cross-site difference could originate in the generator rather than the retriever, and the queries would not be reproducible from the corpora alone. A clinician-authored question set would be the most realistic option, but it cannot be produced identically at 2 sites under separate data use agreements, and it would reintroduce the site-specific human variation this design holds constant. The deterministic rule therefore gives up realism in order to gain comparability, which is what a cross-site comparison requires. It selects the longest contiguous run of narrative prose in a document and returns its first 2 sentences (at least 10 words), excluding header and field-value lines and sentences that recur in more than 3 documents of a cell (corpus-specific boilerplate). A sentence qualifies as narrative prose only if it also matches a fixed, case-insensitive token list used as a heuristic marker of running prose (is, was, were, are, has, had, have, presents, presented, reports, denies, shows, showed, demonstrates, reveals, underwent, admitted, noted, seen, found, developed, complains, there, without, and status). Most entries are finite verbs, but the list also includes 3 function words (there, without, and status) that in these corpora reliably mark prose rather than tabulated fields. No parser or part-of-speech tagger is used, so the rule is deterministic and reproducible from the deposited code. In practice, it excludes the enumerated findings and measurement lines that dominate structured reports. Two data-release differences were normalized at both sites before extraction so they could not masquerade as site signal: differing deidentification (Medical Information Mart for Intensive Care [MIMIC] substitutes deidentification markers for dates and identifiers; ER-Reason retains shifted date-like strings; date and time expressions were scrubbed consistently at both sites before extraction) and differing line-wrapping (MIMIC hard-wraps at ~80 characters; ER-Reason encodes structure with runs of whitespace, which we reflowed at both sites).

Deduplication and Sampling

An audit found that 10.8% (419/3872) of ER-Reason discharge notes and 12.2% (307/2526) of imaging notes were duplicates after deterministic text normalization—the same historical note attached to a patient's multiple index encounters. Before deduplication, notes shorter than 400 characters were excluded from each site×genre cell at both sites, so that only substantive notes entered the pool. Because a duplicate target has more than one correct answer, we removed duplicate texts identified by a normalized-text key (exact equality after lowercasing, removing digits and other nonletter characters, and collapsing whitespace) at both sites. This operational rule may merge otherwise identical templated notes that differ only in numerical content and is therefore treated as a duplicate-removal procedure. In the ER-Reason corpus, among notes of at least 400 characters and before deduplication, equalization, and query extraction, approximately 12.6% (421/3335) of patients contributed more than one document of the same genre; therefore, all uncertainty intervals resample patients rather than queries. Within each final analyzed cell of 1235 documents, the corresponding figure is about 1%, because those cells are drawn from a much larger pool (see Limitations section). Because a patient may appear in both genres of a site, each bootstrap draws patients per site and applies the same draw jointly to that site's 2 genre cells. After duplicate removal, each site×genre cell was sampled without replacement to 2219 documents—the size of the smallest deduplicated cell (UCSF imaging)—using a fixed random seed (42), so that query extraction operated on an equal-sized pool per cell. A known-item query was then extracted from each document, and its sentences were removed from the target. Because the extracted query had already been normalized and scrubbed, the corresponding span was located in the raw target by canonical-token alignment before removal. Documents with fewer than 100 residual characters after removal, or in which the normalized query text still reappeared in the target, were excluded. The 4 usable cells were finally matched to the smallest usable cell (N=1235; random seed 42; Figure 1).

XSL•FO
RenderX

Figure 1. Document extraction and matched-sampling flow. Notes were extracted by genre from each corpus (BIDMC via MIMIC-IV-Note version 2.2 and UCSF via ER-Reason version 1.0.0), filtered to those of at least 400 characters, and deduplicated using a deterministic normalized-text key (lowercasing, removal of digits and other nonletter characters, and whitespace collapse). Because the corpora differ greatly in size, all cells were equalized to a common pool of 2219 documents (the smallest deduplicated cell, UCSF imaging) before query extraction. For each document, a known-item query was extracted, and its sentences were removed from the target to prevent verbatim overlap; documents with insufficient residual target text after query-sentence removal (fewer than 100 characters) or with residual query text were excluded, leaving the usable counts shown. Cells were then matched to the smallest usable cell (random seed 42), yielding 1235 documents per cell (4940 total) in a 2×2 design crossing site/corpus with documentation genre. N was set by the design rather than by a power calculation: it is the largest value common to all 4 cells, and equalizing it holds the candidate pool size constant so that retrieval difficulty is not confounded with corpus size. BIDMC: Beth Israel Deaconess Medical Center; ED: emergency department; MIMIC-IV-Note: Medical Information Mart for Intensive Care IV, Clinical Notes Module; UCSF: University of California, San Francisco.



Models and Configuration

We evaluated 8 contrastively trained retrieval or sentence-embedding models (BGE-base [12], GTE-base [13], E5-base [14], Nomic-embed [15], all-mpnet-base and all-MiniLM-L6 [16], MedCPT [17], and BioLORD-2023 [18]) and 5 masked-language-model (MLM) encoders (BERT-base [19], BioBERT [20], ClinicalBERT [21-23], PubMedBERT [24], and SciBERT [25]). The short names used throughout correspond to the following Hugging Face repositories: BGE-base, BAAI/bge-base-en-v1.5; GTE-base, thenlper/gte-base; E5-base, intfloat/e5-base-v2; Nomic-embed, nomic-ai/nomic-embed-text-v1.5; all-mpnet-base, sentence-transformers/all-mpnet-base-v2; all-MiniLM-L6, sentence-transformers/all-MiniLM-L6-v2; MedCPT, ncbi/MedCPT-Query-Encoder with ncbi/MedCPT-Article-Encoder; BioLORD-2023, FremyCompany/BioLORD-2023; BERT-base, bert-base-uncased; BioBERT, dmis-lab/biobert-v1.1; ClinicalBERT, medicalai/ClinicalBERT; PubMedBERT, microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext; and SciBERT, allenai/scibert_scivocab_uncased. Exact commit revisions are recorded in the repository's model manifest [26]. Each model used its recommended configuration: MedCPT with its separate query and article encoders and classification token (CLS) pooling; BGE-base with CLS pooling and its retrieval query instruction; E5-base and Nomic-embed with

XSL•FO
RenderX

their documented query or document prefixes; and GTE-base, all-mpnet-base, all-MiniLM-L6, and BioLORD-2023 with mean pooling. The 5 MLM encoders, which lack a canonical sentence-pooling method and are not designed as retrieval models, were standardized to CLS pooling and are reported as standardized-backbone comparators rather than as deployment-configured retrievers; our primary analyses are therefore reported both for the full 13-model panel and, as the deployment-relevant subset, for the 8 contrastive models. Cosine similarity ranked the pool for each query. Proprietary API models were not evaluated because both corpora are credentialed. Large decoder-based embedding models were excluded as a resource-aware design choice: the panel is restricted to encoder-scale models a hospital team can host on commodity hardware inside its own security perimeter, the setting this question addresses. Whether ranking stability extends to LLM-scale embedders is untested here.

Chunking

Discharge summaries average roughly 1500-1800 words; under single-window encoding at a 512-token limit, almost all discharge targets were truncated (BIDMC: 1234/1235, 99.9%; UCSF: 1206/1235, 97.7%), compared with a minority of imaging targets (BIDMC 188/1235, 15.2%; UCSF 333/1235, 27%); counts are for the contrastive models, and across all 13 models, the imaging proportions range from 11.3% to 37.8%. To prevent truncation from confounding the genre comparison, documents were embedded as overlapping 510-token windows with a 384-token stride (126-token overlap). For models that require a document prefix, the prefix was tokenized separately and prepended to every window rather than to the document before windowing, so each encoded chunk carries the instruction the model expects; window length was reduced by the prefix length so that no encoded unit exceeded the 512-token limit. The stride is unchanged by this adjustment, so for a model with an *n*-token document prefix, the content carried by each window is (510–*n*) tokens, and the content overlap between consecutive windows is (126–*n*) tokens; for the 2 prefixed models in the panel, this is 124 tokens (E5-base, 2-token prefix) and 122 tokens (Nomic-embed, 4-token prefix). Each candidate document was then assigned the maximum cosine similarity between the query and any of that document's chunks. This is a max-over-chunk multivector scoring scheme, conceptually related to late-interaction retrieval [27]. Under single-window (truncated) encoding, the discharge cross-site correlation among contrastive models was substantially lower ($\tau \approx 0.50$); chunking raised it to 0.86, indicating that fixed-window truncation materially depressed the discharge cross-site correlation in the uncorrected analysis; because truncation was near-universal for discharge documents but substantially less frequent for imaging reports, this also accounts for part of the apparent genre difference in that analysis. Because documents with more chunks have more opportunities to match, we report a chunk-count sensitivity analysis below.

Analysis

Analyses comprise primary end points, secondary descriptive retrieval metrics, and sensitivity analyses. The primary end points are mean reciprocal rank at 10 (MRR@10) under best-chunk scoring, the cross-site rank-transfer measures for RQ1, and the contrastive-panel variance decomposition for RQ2. Best-chunk scoring is the aggregation rule implemented in the primary retrieval pipeline, chosen because it preserves relevant local evidence anywhere in a long document without averaging it away or randomly omitting it. No scorer was prespecified: each alternative is a post hoc rescoring of the chunk embeddings cached by that primary run, developed in response to what the preceding analysis showed. Multimedia Appendix 1 sets out the chronology of these decisions and reports every scoring variant and run relevant to it, including the superseded single-seed realization and the scorers not reported there. Hit@1, Hit@10, and top-3 or top-5 rank overlap are secondary descriptive measures reported alongside them. The remaining quantities are sensitivity analyses testing whether the primary conclusions depend on scoring geometry; the mean-pooled and single-random-chunk scorers, the chunk-count regression, and the mean-centering check all followed the primary analysis and were in place at first submission, whereas the multiseed repetition of the single-chunk scorer was added during peer review; each is labeled as such where it is reported. Four quantities carry the inferential load. MRR@10 and Hit@k measure retrieval within a cell. Kendall τ, between the 2 sites' rankings within a genre, answers RQ1: whether a ranking learned at one institution holds at the other. Cross-site selection regret states the same in deployment terms, as the MRR forgone by choosing the other site's best model. η^2 answers RQ2: how much variation each factor and interaction explains, with model×site the quantity of interest. For RQ1, we computed Kendall τ between the 2 sites' model rankings within each genre (each ranking orders models by their cell MRR@10, so τ measures agreement across models, not across queries), stratified into full-panel and contrastive-only comparisons, and computed top-3 or top-5 overlap and cross-site selection regret (the MRR lost by deploying one site's best-ranked model at the other). For RQ2, we decomposed MRR@10 variance across the 4 cells×models into main effects, and all interactions (η^2) reported separately for the full panel and the contrastive subset. Bootstrap intervals (2000 patient-clustered resamples) were reported for Hit@1 and Hit@10, all rank-transfer and regret estimates, and the primary contrastive-panel variance decomposition; the secondary full-panel decomposition is presented descriptively as point estimates, except for the direct paired model×genre versus model×site contrast, for which a bootstrap interval is also reported. To directly compare the 2 interactions in RQ2, within each patient-clustered bootstrap replicate, we calculated $\Delta\ \eta^2 = \eta^2$ (model×genre) $- \eta^2$ (model×site) and reported the percentile 95% bootstrap interval for this paired difference. The point η^2 is computed on one MRR estimate per model×site×genre cell, so it partitions variance across cell-level means and does not describe the variance of individual query outcomes. Because the full factorial model partitions all model×site×genre cell means, the point-estimate decomposition is saturated at the cell-mean level and has no residual error term at that level; uncertainty is therefore derived separately from per-query outcomes within cells. Within each patient-clustered bootstrap replicate, queries are resampled by patient, the cell-level MRR values are recomputed, and the full factorial

XSL•FO
RenderX

decomposition is recalculated from those replicate means, so the reported intervals reflect regeneration of the entire partition rather than resampling of a fixed set of η^2 values. Sensitivity analyses addressed scoring geometry in 3 ways. First, the chunked documents were rescored under 2 alternative geometries: mean-pooling each document's chunk embeddings into a single document vector, and scoring each document by one randomly selected chunk, the latter repeated across 20 independent seeds and summarized as the mean and SD across seeds. Second, because the embedding spaces are anisotropic, each cell's query and document embeddings were mean-centered by subtracting the cell mean vector before rescoring, as an exploratory check on whether the genre component survives removal of the similarity floor rather than as a correction to the primary analysis. Third, to quantify the order-statistic advantage conferred by higher chunk counts, the maximum cosine similarity of mismatched query-document pairs was regressed on document chunk count within each model and cell, and the slope reported as similarity units per additional chunk. All are reported in the Robustness of the Genre Effect section; the alternative scoring and mean-centering results are summarized in the Robustness of the Genre Effect section.

Ethical Considerations

This study is a secondary analysis of 2 deidentified clinical note corpora accessed under credentialed data use agreements. MIMIC-IV-Note [8] is distributed by PhysioNet under the PhysioNet Credentialed Health Data Use Agreement, which requires completion of human-subjects research training and prohibits any attempt at reidentification. The collection of patient information and creation of that resource were reviewed by the institutional review board at BIDMC, which granted a waiver of informed consent and approved the data-sharing initiative. For ER-Reason [11], the collection and deidentification of patient information were conducted by the Information Commons team at UCSF, and approval to share the dataset was granted by the institution's compliance team. Access to both corpora was obtained only after completing the required credentialing and training. All processing was performed locally by the credentialed author, and no note text was transmitted to any third-party service, in compliance with the applicable data use agreements. This retrospective secondary analysis of previously deidentified records involved no new data collection or patient contact and reports only aggregate results. Because the work used only existing, deidentified data released under established data use agreements and did not involve human participants or identifiable private information, no additional institutional ethics review was sought at the author's institution; the governance basis is the credentialed access and deidentification provided by the data custodians described earlier.

Results

Retrieval performance

Table 1 reports MRR@10 with patient-clustered CIs for all 13 models across all 4 cells under the query-excluded, chunked configuration. With the query removed from its target, absolute retrieval is genuinely difficult: the contrastively trained models cluster around 0.15-0.33, while the standard backbone MLM encoders sit near 0.01-0.07, far below the contrastive models in every cell. As interpretable secondary metrics, mean performance across the 8 contrastive models placed the known-item target within the top 10 for 34%-46% of queries (Hit@10) and first for 11%-18% (Hit@1), with discharge summaries exceeding imaging reports at both sites (Table 2).

Table 1. Known-item mean reciprocal rank at 10 (MRR@10; 95% patient-clustered CI), query-excluded and chunked (N=1235 per cell).

Model	Type[a]	BIDMC[b] discharge, MRR@10 (95% CI)	BIDMC imaging, MRR@10 (95% CI)	UCSF[c] discharge, MRR@10 (95% CI)	UCSF imaging, MRR@10 (95% CI)
BGE-base	c	0.280 (0.260-0.302)	0.181 (0.164-0.199)	0.282 (0.261-0.303)	0.184 (0.166-0.202)
GTE-base	c	0.297 (0.276-0.319)	0.195 (0.177-0.214)	0.301 (0.280-0.323)	0.199 (0.180-0.219)
E5-base	c	0.233 (0.214-0.253)	0.182 (0.164-0.201)	0.252 (0.232-0.274)	0.158 (0.141-0.174)
Nomic-embed	c	0.329 (0.306-0.353)	0.189 (0.171-0.207)	0.298 (0.277-0.322)	0.177 (0.159-0.196)
all-mpnet-base	c	0.266 (0.245-0.287)	0.185 (0.167-0.204)	0.291 (0.270-0.313)	0.165 (0.148-0.182)
all-MiniLM-L6	c	0.233 (0.214-0.253)	0.171 (0.154-0.189)	0.264 (0.244-0.286)	0.156 (0.140-0.172)
MedCPT	c	0.152 (0.136-0.169)	0.146 (0.131-0.162)	0.174 (0.157-0.193)	0.146 (0.130-0.164)
BioLORD-2023	c	0.194 (0.176-0.214)	0.163 (0.145-0.181)	0.217 (0.198-0.238)	0.158 (0.141-0.175)
BERT-base	m	0.015 (0.011-0.020)	0.018 (0.012-0.025)	0.016 (0.011-0.022)	0.012 (0.008-0.017)
BioBERT	m	0.019 (0.013-0.025)	0.026 (0.019-0.034)	0.024 (0.018-0.031)	0.011 (0.007-0.015)
ClinicalBERT	m	0.025 (0.018-0.033)	0.069 (0.058-0.080)	0.033 (0.026-0.041)	0.043 (0.034-0.053)
PubMedBERT	m	0.029 (0.022-0.036)	0.049 (0.040-0.059)	0.036 (0.027-0.045)	0.031 (0.024-0.040)
SciBERT	m	0.018 (0.012-0.025)	0.021 (0.015-0.028)	0.012 (0.008-0.017)	0.011 (0.007-0.015)

[a]Type: c=contrastively trained; m=standardized-backbone masked-language-model comparator.

[b]BIDMC: Beth Israel Deaconess Medical Center.

[c]UCSF: University of California, San Francisco.

XSL•FO
RenderX

Table 2. Secondary retrieval metrics: contrastive-panel (8-model) Hit@1 and Hit@10 (95% patient-clustered CI), query-excluded and best-chunk scoring (N=1235 per cell)[a].

Cell	Hit@1 (95% CI)	Hit@10 (95% CI)
BIDMC[b] discharge	0.179 (0.164-0.194)	0.424 (0.405-0.445)
BIDMC imaging	0.112 (0.099-0.125)	0.341 (0.319-0.362)
UCSF[c] discharge	0.183 (0.168-0.198)	0.456 (0.436-0.478)
UCSF imaging	0.105 (0.093-0.118)	0.341 (0.320-0.362)

[a]Hit@k is the proportion of queries whose known-item target appears within the top k retrieved documents (equivalently Recall@k, one relevant target per query); Hit@1 equals the proportion ranked first. Values are means over the 8 contrastive models. Per-model Hit@1 or Hit@10 is provided in Multimedia Appendix 1.

[b]BIDMC: Beth Israel Deaconess Medical Center.

[c]UCSF: University of California, San Francisco.

RQ1: Rankings Transfer Across Sites

Model rankings were positively correlated across sites within each genre, though precision is limited by the small model panel (Table 3). Among the 8 contrastive models, cross-site Kendall τ was 0.86 (95% patient-clustered CI 0.64-1.00) for discharge summaries and 0.64 (95% patient-clustered CI 0.36-0.86) for imaging; across the full 13-model panel, cross-site Kendall τ was 0.92 (95% patient-clustered CI 0.80-0.95) and 0.82 (95% patient-clustered CI 0.69-0.92), respectively. Top-5 model overlap across sites was complete for discharge and 0.80 for imaging (Table 3). GTE-base was ranked first or second at both sites in both genres. Two adjacent pairs are separated by less than 0.0005 MRR: E5-base (0.2330) and all-MiniLM-L6 (0.2333) at BIDMC discharge, and E5-base (0.1577) and BioLORD-2023 (0.1578) at UCSF imaging. These effective ties nonetheless determine the rank order entering the τ and top-k overlap calculations, so both statistics are more sensitive to arbitrarily small differences than their point values suggest. Reported τ are point estimates; their bootstrap distributions are left-skewed toward the τ=1 ceiling, so the intervals are asymmetric.

Table 3. Cross-site rank transfer by model family (query-excluded, chunked; 95% patient-clustered CI).

Genre	Panel	Kendall τ (95% CI)	Top-3	Top-5
Discharge	Full panel (13-model panel)	0.923 (0.795-0.949)	0.67	1.00
Discharge	Contrastive (8 contrastive model)	0.857 (0.643-1.000)	0.67	1.00
Imaging	Full panel (13-model panel)	0.821 (0.692-0.923)	0.67	0.80
Imaging	Contrastive (8 contrastive model)	0.643 (0.357-0.857)	0.67	0.80

Observed cross-site selection regret was 0 for imaging—both sites selected GTE-base—and, for discharge, 0.003 (BIDMC→UCSF) and 0.033 (UCSF→BIDMC, ~10% of the destination's best MRR; Table 4). Patient-clustered bootstrap intervals ranged from 0 to 0.071, indicating a transfer cost that is zero to modest but imprecisely estimated at this panel size. The best model was reselected within each bootstrap replicate, so a zero observed regret can yield a positive bootstrap distribution when a different model occasionally ranks highest in a resample. Because regret is nonnegative, the interval beginning at zero is not a test against a null of zero regret; we report it as an interval, not a significance statement.

Table 4. Observed cross-site selection regret with patient-clustered bootstrap 95% intervals (contrastive models)[a].

Genre	Source→destination	Source best	Destination best	Observed	Bootstrap 95% CI
Discharge	BIDMC[b]→UCSF[c]	Nomic-embed	GTE-base	0.003	0.000-0.021
Discharge	UCSF→BIDMC	GTE-base	Nomic-embed	0.033	0.000-0.071
Imaging	BIDMC→UCSF	GTE-base	GTE-base	0.000	0.000-0.039
Imaging	UCSF→BIDMC	GTE-base	GTE-base	0.000	0.000-0.008

[a]Regret was calculated from unrounded mean reciprocal rank values; the mean reciprocal rank values displayed in Table 1 are rounded to 3 decimals, so differences taken from that table may not reproduce these figures exactly.

[b]BIDMC: Beth Israel Deaconess Medical Center.

[c]UCSF: University of California, San Francisco.

XSL•FO
RenderX

RQ2: Model and Genre Dominate; the Site Interaction Is Small

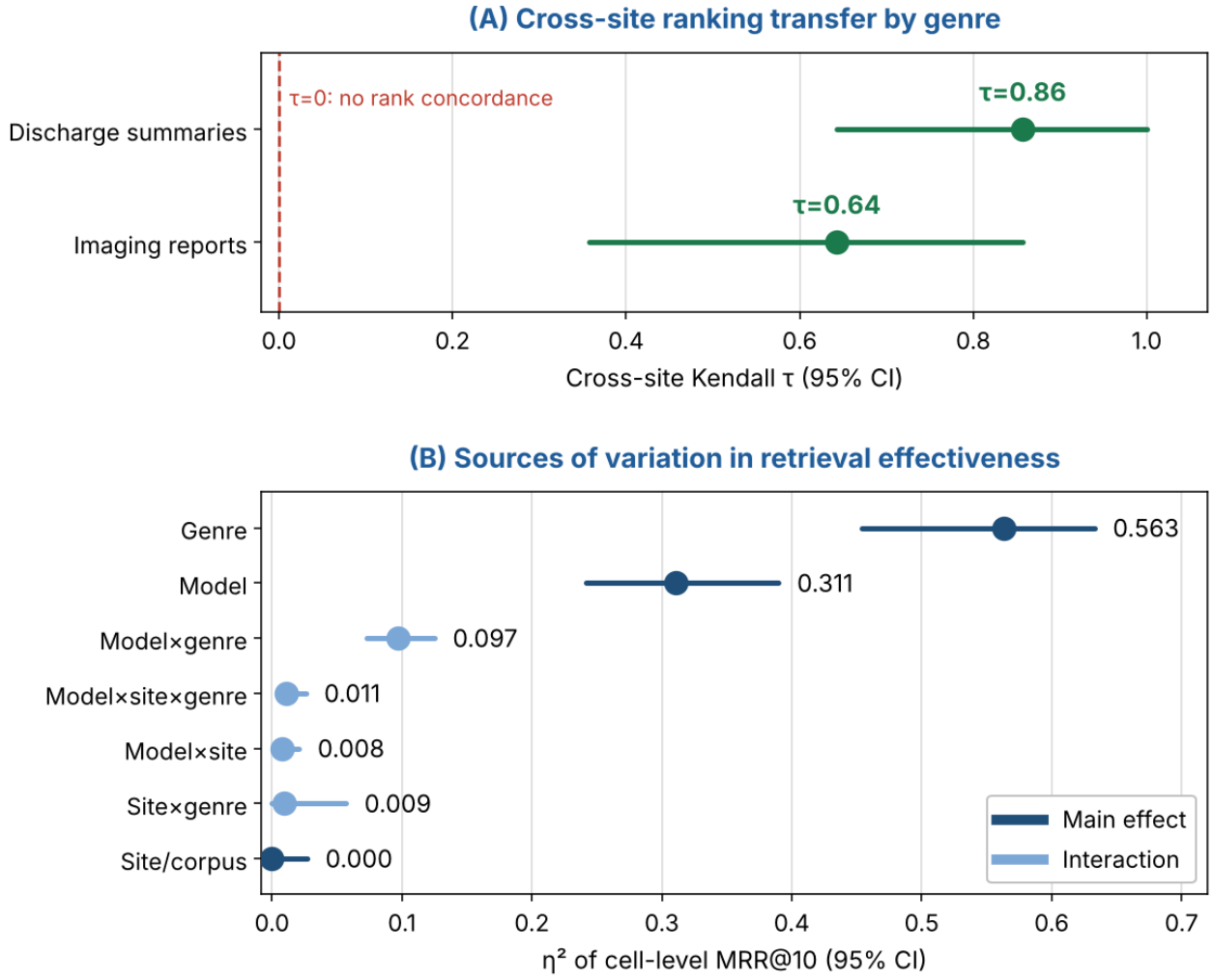
Among the 8 deployment-relevant contrastive models, model choice and documentation genre were the 2 primary sources of variation in retrieval effectiveness under primary best-chunk scoring (Table 5; Figure 2B). Genre accounted for η^2=0.563 (95% CI 0.454-0.633), and the model for η^2=0.311 (95% CI 0.242-0.390). These proportions partition variance among the cell-level mean MRR values that define the design, not among individual query outcomes, and should be read as describing where retrieval effectiveness differs across models, sites, and genres rather than how much of any one query's success is explained. However, this ordering is sensitive to scoring geometry (see Robustness of the Genre Effect section), so we report the 2 as comparably large rather than claiming a strict ranking between them. The site/corpus main effect was η^2=0.000 (upper bound 0.027), and the model×site interaction was η^2=0.008 (95% CI 0.004-0.021), both small. The model×genre interaction (η^2=0.097, 95% CI 0.073-0.125) was substantially larger than the model×site interaction (η^2=0.008, 95% CI 0.004-0.021). A paired patient-clustered bootstrap of the difference confirmed this directly: η^2(model×genre)–η^2(model×site)=+0.089 (95% CI +0.060 to +0.114), excluding zero, indicating reliably greater model-specific variation by documentation genre than by site/corpus. The same contrast in the full 13-model panel was also positive and excluded zero (+0.062, 95% CI +0.046 to +0.078). Figure 2 shows the cross-site transfer estimates and the contrastive-panel decomposition together. Panel composition matters for the apparent dominant factor. Adding the 5 nonretrieval MLM comparators to form a 13-model panel shifts the largest term to the model main effect (η^2=0.874) because those comparators sit far below the contrastive models at both sites, inflating between-model variance. We report both panels (Table 5) and treat the contrastive subset as the deployment-relevant one; the cross-site conclusion—small model×site interaction—is consistent across both panels.

Table 5. Variance decomposition of mean reciprocal rank at 10 under best-chunk scoring: point estimates for the full 13-model panel and point estimates with 95% patient-clustered bootstrap intervals for the contrastive subset (see Robustness of the Genre Effect section for chunk-count and mean-centering sensitivities).

Term	Full panel (13-model panel)	Contrastive (8 contrastive model) (95% CI)
Model	0.874	0.311 (0.242-0.390)
Genre	0.056	0.563 (0.454-0.633)
Model×genre	0.064	0.097 (0.073-0.125)
Model×site×genre	0.002	0.011 (0.006-0.026)
Model×site	0.002	0.008 (0.004-0.021)
Site×genre	0.002	0.009 (0.000-0.057)
Site	0.000	0.000 (0.000-0.027)

XSL•FO
RenderX

Figure 2. Sources of variation in retrieval effectiveness and cross-site ranking transfer, contrastive-model panel (8 models). Throughout, circles indicate point estimates, and horizontal lines indicate 95% patient-clustered bootstrap intervals (2000 resamples). (A) Cross-site Kendall τ between the 2 sites' model rankings within each genre; green denotes the observed τ, and the red dashed line denotes τ=0 (no rank concordance). Rankings transferred positively for both discharge summaries (τ=0.86) and imaging reports (τ=0.64). (B) Variance decomposition for the model and genre main effects and their interactions, computed as η^2 of the cell-level mean MRR@10 values rather than of individual query outcomes; dark blue denotes main effects, and light blue denotes interactions. Documentation genre (η^2=0.563) and model choice (η^2=0.311) were the 2 largest components, whereas the model×site interaction (η^2=0.008) and the site/corpus main effect (η^2=0.000) were small. Values are from the primary best-chunk scoring configuration. η^2: eta-squared (proportion of variance explained); Kendall τ: Kendall rank correlation coefficient; MRR@10: mean reciprocal rank at 10; site/corpus denotes the bundled site and corpus factor.



Robustness of the Genre Effect

Because the genre main effect is large and the retrieval task scores documents by their best-matching chunk, we checked whether it could be an artifact of 2 features of the setup: differential chunk counts across genres and embedding anisotropy.

Chunk Count

Discharge summaries yield more chunks (BIDMC: median 7, IQR 5-9; UCSF: median 10, IQR 7-13) than imaging reports (BIDMC: median 1, IQR 1-1; UCSF: median 1, IQR 1-2), and documents with more chunks offer more opportunities to achieve a high best-chunk similarity. For each model, we estimated the order-statistic advantage by regressing the maximum cosine similarity of mismatched query-document pairs on document chunk count. Estimated slopes ranged from 0.01 to 0.08 similarity units per additional chunk, indicating a limited increase in null maximum similarity with higher chunk counts. However, a shift in null similarity does not translate directly into a shift in MRR. The regression bounds how far the null baseline moves with chunk count, but it does not establish how much of the observed genre effect that movement explains. We therefore rely on convergent evidence from scorers that remove or randomize chunk count, rather than on the regression alone. Alternative scorers that remove or vary the number of chunks (mean-pooling the whole document; a single randomly chosen chunk, averaged over 20 seeds) left genre a major factor (η^2=0.21 and 0.72 [SD 0.03 across seeds], respectively), bracketing the best-chunk value. A fourth scorer, length-normalized maximum similarity, subtracted an empirical null curve of expected maximum similarity by chunk count from each document's observed score. It is not reported in Table 6 because the subtracted quantity is dominated by the anisotropy

XSL•FO
RenderX

floor rather than by chunk count: the null at k=1 was 0.53-0.80, while the increment from k=1 to k≥5 was only 0.013-0.075. Under this scorer, genre η^2 was 0.165 and model η^2 0.537, with model×site 0.019 and cross-site τ 0.71 in both genres (Multimedia Appendix 1). Mean-centering, below, removes the same floor by a per-dimension shift and is the geometric check reported instead.

Table 6. Genre and model η^2 for the contrastive panel across scoring choices and after mean-centering; site and model×site are shown for the transfer conclusion[a].

Scoring	Genre η^2	Model η^2	Site η^2	Model×site η^2	Discharge, Kendall τ	Imaging, Kendall τ
Best-chunk (primary, Table 5)	*0.563*	*0.311*	*0.000*	*0.008*	*0.86*	*0.64*
Mean-pooled document	0.213	0.390	0.172	0.029	0.71	0.71
Single random chunk, mean of 20 seeds (SD)	0.717 (0.027)	0.214 (0.025)	0.023 (0.008)	0.013 (0.005)	0.65	0.69
Mean-centered, best-chunk	0.356	0.404	0.002	0.005	0.86	0.57

[a]Row 1 presents the primary decomposition based on full cell-level mean reciprocal rank estimates (Table 5), with patient-clustered bootstrap uncertainty reported there. Rows 2-4 are descriptive sensitivity analyses, rescored from cached chunk embeddings using unweighted cell means, and do not carry the bootstrap uncertainty of the primary analysis; comparisons with row 1 should therefore be interpreted qualitatively rather than inferentially. SD is reported only for the single-random-chunk row, which is a mean across 20 independent seeds; the other rescorings are deterministic and have no associated spread. Values in italics are the primary analysis.

Anisotropy

The strongest retrievers embed clinical text into a markedly anisotropic space [28]: the mean cosine between unrelated document chunks, averaged over the 4 cells, was 0.67-0.84 for the higher-anisotropy models (E5-base 0.84, GTE-base 0.79, MedCPT 0.77, Nomic-embed 0.69, and BGE-base 0.67) and 0.30-0.46 for the lower-anisotropy models (all-mpnet-base 0.46, all-MiniLM-L6 0.43, and BioLORD-2023 0.30). These groupings describe embedding geometry, not retrieval performance; MedCPT, for example, is highly anisotropic but the weakest contrastive model in Table 1. A perfectly uniform cosine offset would not alter within-query rankings. Because empirical anisotropy can vary across models and documents, we mean-centered each cell's embeddings as an exploratory sensitivity analysis, not as a correction to the primary results. Centering removed the similarity floor (mean off-diagonal cosine for GTE-base 0.79→0.00) and changed the relative model and genre components (genre η^2 0.56→0.36, model 0.31→0.40, and model×genre 0.10→0.20) but left the model×site interaction small (0.005), the site/corpus main effect negligible (0.002), and cross-site τ positive (discharge: 0.86; imaging: 0.57). Genre remained a major component after mean-centering, indicating that its contribution was not eliminated by the anisotropy sensitivity analysis; the exact ordering of model and genre, however, depended on scoring geometry.

Two findings were stable across scoring choices: the model×site interaction remained small (η^2=0.005-0.029), and cross-site τ remained positive, supporting transfer of relative model rankings. Model and genre were consistently the 2 large factors, with only their relative ordering changing. The site/corpus main effect, by contrast, was scoring-dependent: negligible under best-chunk and mean-centered scoring but η^2=0.172 under mean-pooling, reflecting a corpus-level shift in absolute retrieval level rather than a change in model ordering. We therefore distinguish robust transfer of relative model performance from scoring-dependent differences in absolute performance between corpora.

Discussion

Principal Findings

Across 2 real institutional EHR corpora, embedding model rankings for known-item clinical retrieval showed positive cross-site concordance—strong for discharge summaries and moderate for imaging reports, with intervals wide enough that the direction rather than the magnitude is the finding—once verbatim query-target overlap, duplicate notes after text normalization, and document truncation were controlled, and models were run in their recommended configurations. The model×site interaction, which captures site-specific differences in model performance, remained small under every scoring choice tested (η^2=0.005-0.029); this was consistent with the positive cross-site Kendall correlations. The site/corpus main effect was negligible under best-chunk and mean-centered scoring (η^2≤0.002) but rose to η^2=0.172 under mean-pooled document scoring, indicating that corpus-level differences in absolute retrieval level are scoring-dependent even when relative model ordering transfers. Cross-site selection regret was zero for imaging and small for discharge (point estimate up to ~10% of the destination's best MRR, with a bootstrap interval including zero). Model choice and documentation genre were both major sources of variation in absolute effectiveness; their relative ordering depended on scoring geometry, and genre remained a major component after accounting for chunk-count and anisotropy sensitivities.

The practical implication is that, in this 2-corpus evaluation, a model-selection decision made at one site largely transferred to the other when documentation genre and retrieval configuration were matched: the models that ranked well at BIDMC also ranked well at UCSF, and adopting one site's choice at the other cost little, though not nothing for discharge summaries. Whether this extends to model selection performed

XSL•FO
RenderX

on public, predominantly single-source benchmarks is a reasonable hypothesis that follows from these results but was not directly tested here; such benchmarks may retain value as selection guides, provided the genre, query formulation, indexing strategy, and retrieval objective are matched.

Consistency With Prior Single-Source Work

That model choice and documentation genre are both major determinants of retrieval effectiveness, while model-specific site variation is small, is consistent with, and extends prior single-source evidence that clinical context variables rival model choice in embedding-based retrieval [2]. The earlier study could vary genre but not institution; here, with site/corpus added as a factor, genre remains a large source of variation. These findings narrow, rather than overturn, the earlier recommendation for mandatory local validation: re-evaluation appears particularly important when documentation genre or retrieval formulation changes, whereas relative model selection may transfer more readily across sites when these factors are held constant. We separate 2 claims. First, genre explains much of the variation in absolute effectiveness: it shifts the overall MRR level of a cell (discharge exceeds imaging at both sites) more than it reorders the models. Second, model-specific variation by genre is captured by the model×genre interaction (η^2=0.097), which is substantially larger than model×site (η^2=0.008) but small relative to the main effects. The large genre main effect is therefore primarily a level effect and should not be read as strong evidence that model rankings differ between genres.

Implications for Clinical RAG Model Selection

Cross-site ranking stability bears directly on how informatics teams allocate scarce evaluation effort when deploying a retrieval layer for clinical RAG. If relative model rankings did not transfer, every institution would need to repeat a full local embedding evaluation—assembling a labeled or known-item retrieval set, running the candidate models, and rescoring—before selecting a model, a costly prerequisite to deployment. Our results suggest that, when documentation genre and retrieval configuration are matched, a model-selection decision made at one site is a reasonable starting point at another, so local work can be reprioritized from reranking the full model panel toward lighter-weight confirmation. Concretely, a team could confirm the top 2 or 3 externally ranked models on a local validation set of the target genre rather than benchmarking the full panel and prioritize broader re-evaluation when documentation genre or retrieval formulation changes; when only the institution changes under otherwise matched conditions, begin with local confirmation of the externally ranked leaders. We frame this as deployment validation and resource allocation rather than a claim of patient benefit: the evidence here concerns the stability of model selection across sites, not downstream clinical outcomes, which remain to be established in prospective evaluation.

The Role of Truncation and Panel Composition

Two analytic choices proved consequential and are worth flagging for future evaluations. First, truncation: under single-window encoding, essentially all discharge summaries were truncated at 512 tokens, and the discharge cross-site correlation among contrastive models fell to about 0.50; chunked max-similarity retrieval increased it to 0.86, indicating that fixed-window truncation materially contributed to the apparent genre difference in an uncorrected analysis. Second, panel composition: including nonretrieval MLM comparators shifts the apparent dominant factor from genre to model because such comparators inflate between-model variance. Both observations caution that conclusions about which factor most destabilizes rankings depend on task and panel choices and should be drawn from a controlled task with a deployment-relevant model set. Within the contrastive panel, MedCPT and BioLORD-2023 were the weakest models in most cells despite their biomedical specialization. One possible explanation is a mismatch between their training objectives—PubMed query-article retrieval and biomedical concept representation—and the task used here; because model family, training corpus, and objective are not independently varied, this comparison cannot establish which is responsible.

Comparison With Other Work

Compared with the single-source factorial study that motivated this analysis [2], we add the site/corpus axis it identified as missing. This study established, across 3 document collections, that documentation genre and query formulation together rival model choice in explaining variance in retrieval effectiveness, but it could not separate institution from collection, because its collections differed in provenance rather than in the institution that produced them. The present study holds genre fixed, varies the institution, and reports 3 quantities that the previous study could not: the cross-site rank correlation, the model-by-site interaction measured against the model-by-genre interaction, and the effectiveness cost of transferring a model choice between sites. We find less variation by site than by genre. A recent mechanistic companion study found that embedding-model rankings were stable when evaluation was expanded within 2 public clinical-text sources, while also identifying representation geometry and retrieval-objective training as important determinants of retrieval performance [7]. This study did not evaluate independent institutional EHR corpora or cross-site ranking transfer with documentation genre held constant; the present study addresses that transportability question directly. Compared with a recent public EHR retrieval benchmark [6], which is single-institution and single-genre, we add a second site and a second genre and shift the question from within-task effectiveness to cross-site ranking stability. Compared with a recent multi-institution embedding evaluation [3], which spans 2 health systems but compares models, pooling strategies, and query formulations within tasks, we provide the cross-site ranking comparison that its design does not report. A recent large-scale benchmark evaluated 39 embedding models across 7 health care semantic tasks on Mount Sinai and MIMIC-IV data, reframing each task as retrieval scored by MRR and recall@k [4]; that study spans many models and tasks but does not address whether relative model rankings transfer across institutions with genre held constant, which is the specific question examined here.

XSL•FO
RenderX

Limitations

First, this is known-item retrieval: the query is derived from its target. Although we removed the query sentences from the target and excluded internally repeated notes, the query and the target remainder still originate from the same clinical document; the task is therefore best described as within-document semantic retrieval, and it preserves a semantic dependence between query and target that need not hold when a clinician poses a prospective question. We regard it as a controlled assay of embedding portability—holding the query-target relationship fixed while varying site and genre—rather than a simulation of prospective clinical querying, and absolute MRR values are specific to this task. The geometry of this task also differs from prospective querying in a way that may favor particular model families. Because the extractor returns the first 2 sentences of the longest run of discriminative prose, the resulting queries resemble narrative summaries (discharge: at BIDMC: median 36, IQR 29-44 words and at UCSF: median 39, IQR 29-52 words; imaging: at BIDMC: median 26, IQR 16-34 words and at UCSF: median 20, IQR 13-29 words) rather than the short, targeted questions typical of clinical RAG. Matching a passage of narrative prose to the remainder of its own note is a symmetric document-to-document problem, whereas a prospective clinical question is asymmetric and shares less vocabulary with its answer. Models trained for symmetric passage similarity may therefore be advantaged here relative to models tuned for asymmetric question answering, and the ranking we report should be read as specific to this query construction.

Second, the site/corpus factor is not cleanly isolated. The 2 corpora differ in the note-selection pipeline, care setting, patient mix, and deidentification, and their eras cannot be exactly aligned. MIMIC-IV documents an overall collection period of 2008-2019 and provides approximate anchor-year groups, but exact original calendar dates are not directly recoverable from the deidentified timestamps. ER-Reason retains shifted date-like strings spanning 1983-2024, with 601/3872 (15.5%) discharge notes plausibly contemporaneous with MIMIC-IV. We therefore interpret site/corpus as a bundled contrast rather than an isolated institutional effect. Because the era difference cannot be quantified, neither can its contribution to the measured site/corpus variance; documentation conventions change over time, so a residual era mismatch could contribute to that component, but its magnitude and direction are not identified by this design. The 2 sampling frames also differ in scale by roughly 2 orders of magnitude: 331,792 eligible BIDMC discharge notes against 3872 at UCSF, because ER-Reason samples historical notes attached to a fixed set of index emergency encounters, while MIMIC-IV-Note draws from a broad inpatient population. Equalizing the analyzed pool does not equalize what it represents, so the 2 discharge cells may be differently representative of routine documentation at their institutions. This is distinct from the bundled-contrast caveat above and is an alternative explanation for the size of the site/corpus component in either direction.

Third, the rank correlations and interaction terms are based on a 13-model panel, including an 8-model contrastive subset, which limits the precision of τ and the η^2 estimates; a larger model panel would sharpen both. The resulting intervals are wide enough that the direction of transfer is far better supported than any particular value: the imaging contrastive-panel interval spans 0.36 to 0.86, and the discharge interval reaches 1.00. Readers should take the sign, not the point estimate, as the finding. The relative sizes and ordering of the model, genre, and interaction terms should be interpreted as panel- and scoring-specific rather than as general constants. Because 2 sites do not constitute a random sample of institutions, the decomposition is a descriptive partition of variance within this fixed experimental panel; the small site component (η^2=0.000, upper bound 0.027) should not be read as a population-level estimate that institutions generally explain almost none of the retrieval variation.

Fourth, the variance decomposition uses one MRR observation per model×site×genre cell for the point η^2, with uncertainty derived from within-cell per-query outcomes.

Fifth, chunk count varies by genre and can favor a max-similarity scorer; the sensitivity analyses reduce but do not eliminate uncertainty about scoring geometry.

Sixth, deterministic duplicate removal after digit deletion can merge distinct templated notes that differ only numerically. We audited this exhaustively rather than by sampling. Of 10,110 merged groups, 9707 (96%) were correct merges: byte-identical, differing only in whitespace, punctuation, or case while carrying identical digits, or identical after the pipeline's own date, time, and identifier scrubbing. The remaining 403 groups (1007 documents) differed in content after scrubbing and are false-positive merges, concentrated in templated procedure and measurement reports—vascular surveillance summaries, catheter placement and paracentesis notes, and dosimetry fields—whose narrative is fixed and whose discriminative content is numeric. They represent 0.049% (992/2,036,503) of the MIMIC-IV radiology corpus and 0.387% (15/3872) of ER-Reason discharge notes, with none in the MIMIC-IV discharge corpus and none in the ER-Reason imaging cell that determined N. Because deduplication precedes sampling of the equalized pool, these merges removed candidate targets rather than altering the analyzed set. Removing numerically distinct but otherwise templated documents could change the difficulty of the retrieval pool in either direction, so we do not claim a guaranteed direction of bias; the relevant argument is the very low frequency of such merges.

Seventh, the scope of the design bounds every claim above. BIDMC and UCSF are both large US academic medical centers under similar documentation paradigms and record conventions, so ranking stability between them is a weaker test than transfer to community hospitals, safety-net settings, or health systems outside the United States. Our findings are evidence that institutional differences of this kind did not substantially destabilize model rankings, not evidence of institutional invariance. Proprietary API models were omitted because their terms of service are incompatible with the governing data use agreements; since many enterprise deployments rely on such

XSL•FO
RenderX

end points, the stability reported here applies strictly to the tested open-weight models.

Eighth, extraction yield differed sharply across cells, and the surviving documents are not a random subset. From an equalized pool of 2219 documents per cell, 2125 discharge notes at each site, and 1945 BIDMC imaging reports yielded a usable query, against 1235 at UCSF imaging—the cell that consequently determined N for the whole design. Of the 984 UCSF imaging documents excluded, 943 yielded no query—930 because no contiguous run of discriminative prose met the extractor's requirements and 13 because no sentence survived segmentation and filtering—41 were removed by the residual-leak guard and none by the minimum residual-length rule. Applying the same 3 criteria to the other cells, BIDMC discharge lost 1 document to the absence of an extractable query, 93 to the residual-leak guard, and none to the minimum residual-length rule; UCSF discharge lost 74, 20, and none; and BIDMC imaging lost 238, 32, and 4. Each set sums to the difference between the equalized pool of 2219 documents and the usable count shown in Figure 1. The cause is genre-by-corpus rather than genre alone: ER-Reason imaging sentences failed the prose-marker requirement at roughly twice the BIDMC rate (4961/24,248, 20.5% vs 3040/26,650, 11.4% of all segmented sentences), consistent with more heavily templated radiology reporting. The analyzed UCSF imaging cell—which determines N for the whole design—is therefore enriched for the more narrative reports in that corpus, and its absolute values should not be read as representative of UCSF imaging documentation.

Ninth, query length varies systematically by genre: discharge queries average 40.4 words against 25.1 for imaging, a standardized mean difference of about 1.0. Query length is therefore associated with genre and is a potential genre-correlated source of variation that this design cannot separate from genre itself. Because the design compares sites within genre, the informative comparison is within-genre rather than site-averaged. Discharge queries average 37.8 words at BIDMC against 43.0 at UCSF, and imaging queries 27.2 against 23.0; the 2 differences run in opposite directions and cancel in the site averages (32.5 vs 33.0 words), which is why the site-averaged figures are uninformative here. The estimates these within-genre differences could affect are the site/corpus main effect, the model×site interaction, and the model×site×genre interaction, each of which is identified by within-genre contrasts between the 2 sites. All 3 were small (η^2=0.000, 0.008, and 0.011), so any length-driven contribution to them is either correspondingly small or offset by an opposing site effect; the design cannot separate a residual within-genre length difference from a site difference, and we do not claim that it can. A separate and distinct length effect operates on the document side: the mean-pooled sensitivity analysis, which removes the best-chunk advantage of documents with more chunks, reduces the genre component from η^2=0.563 to 0.213 (Table 6). This analysis addresses scoring geometry and document chunk count, not query length, and is not direct evidence about the query-length confounder. Full distributions are given in Multimedia Appendix 2. Relatedly, the boilerplate filter excludes any sentence appearing in more than 3 documents within a cell; measured against otherwise eligible sentences, this removed 5.2% (5744/109,835) at BIDMC discharge, 6.8% (844/12,369) at BIDMC imaging, 10.6% (13,327/126,007) at UCSF discharge, and 12% (1144/9573) at UCSF imaging, so the threshold prunes most aggressively in exactly the templated genre where near-duplicate phrasing is expected, but does not approach the level at which distinctive clinical narrative would be removed wholesale. The concern is directional as well as quantitative: because the rule preferentially removes recurrent wording, it may bias extracted queries toward more distinctive narrative phrasing, particularly in templated imaging reports, and the magnitude of any resulting effect on retrieval cannot be identified from this analysis.

Tenth, the 5 MLM encoders were standardized to CLS pooling because they lack a canonical sentence-pooling method and were never trained with a pooling head for retrieval. Mean pooling over token embeddings can outperform CLS for such backbones, so their reported values are a floor rather than a fair estimate of their achievable performance, and the gap between them and the contrastive models should not be read as a measure of how much contrastive training contributes.

Eleventh, approximately 12.6% (421/3335) of ER-Reason patients contribute more than one document of the same genre before deduplication, and clinical narrative is frequently copied forward between a patient's notes. Where a nontarget note contains the same copied passage, retrieving it counts as an error even though the model has located semantically correct text, which deflates the reported scores. Within the analyzed cells, this affects a small share: the 4 cells contain 1219, 1226, 1210, and 1221 distinct patients across 1235 queries, so only about 1% of patients contribute more than one document to a candidate pool.

Twelfth, we measure retrieval only. High MRR or Hit@k on a known-item task indicates that a model places the correct document near the top of a ranked list; it does not establish that a downstream generation system built on that retriever will produce factually accurate or clinically sound output. Retrieval quality is necessary but not sufficient for RAG, and model selection for a deployed system should be validated on generation outcomes as well as on ranking metrics. We nonetheless regard retrieval-ranking stability as the appropriate target here: the model-selection decision is made on retrieval evidence before a generator is chosen, and end-to-end quality confounds the retriever with the generator, prompt, and rubric, so a cross-site difference observed end-to-end could not be attributed to the retriever. Isolating retrieval is what makes the site contrast interpretable.

Conclusions

For known-item retrieval from clinical notes across 2 US academic medical centers, embedding-model rankings showed positive cross-site concordance after controlling for verbatim overlap, duplication, and truncation, moderate for imaging, and strong for discharge summaries, with intervals wide enough that the direction rather than the magnitude is the finding. The site/corpus main effect was negligible under primary best-chunk scoring, and the model×site interaction remained small across scoring sensitivities; cross-site reuse of a model-selection decision incurred zero to modest retrieval-performance loss.

XSL•FO
RenderX

Model choice and documentation genre were the primary determinants of absolute effectiveness among deployment-relevant models, but which of the 2 appeared dominant depended on panel composition: genre was largest among the contrastive models, whereas model became dominant once the nonretrieval comparators were included. For relative model selection, documentation genre produced more model-specific variation than the site/corpus contrast, whereas absolute site-level differences depended on scoring geometry. Whether these findings extend to realistic query workloads, larger model panels, additional institutions, and proprietary models requires contemporaneous, pipeline-matched multi-institutional evaluation.

Data Availability

The datasets analyzed during this study are not publicly available because they are governed by data use agreements requiring credentialed access: MIMIC-IV-Note (Medical Information Mart for Intensive Care IV, Clinical Notes Module; version 2.2) and ER-Reason (version 1.0.0) are available to credentialed users via PhysioNet. No clinical note text is reproduced. The full analysis pipeline—query extraction and overlap removal, deterministic duplicate removal, chunked embedding, retrieval, and statistical analysis including the patient-clustered bootstrap—together with fixed random seeds, sampling manifests, exact model identifiers and revisions, tokenizer and pooling configurations, and aggregate outputs, is available in the Zenodo repository [26] and maintained in a public source repository [29].

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Generative AI Use

Claude (Anthropic) models, including Claude Opus 4.8, served as a coding assistant in developing the Python analysis scripts. All code was independently reviewed, validated, and executed by the author, who takes full responsibility for the analysis. During the peer-review revision, the assistant was also used to help with editing and grammar in response to reviewer comments and to check numerical consistency between the analysis outputs, the tables, and the text. Details of this assistance and the analysis scripts are provided in Multimedia Appendix 3.

Authors' Contributions

YM was involved in conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing the original draft, review and editing, and visualization. The author read and approved the final manuscript.

Conflicts of Interest

The author serves as the Chief Medical Officer of AlgiPharma AS. AlgiPharma has no products or commercial interests related to embedding models, retrieval-augmented generation, or clinical informatics systems. No other competing interests are disclosed.

Multimedia Appendix 1

Per-model secondary retrieval metrics. Hit@1 and Hit@10 for each of the 13 models in each site×genre cell, with 95% patient-clustered bootstrap CIs; Table 2 reports the mean of the 8 contrastive models in each cell. The appendix also documents the chronology of the scoring-geometry decisions and reports every scorer that was run, including those not shown in Table 6.
[DOCX File , 25 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Query length distributions across the 4 cells. Mean, SD, variance, median, IQR, and extremes of extracted query length in words and characters for each site×genre cell, reported to assess whether query length confounds the genre main effect.
[DOCX File , 17 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Declaration of generative AI assistance and analysis scripts. Describes the scope of generative AI use during analysis and manuscript preparation and lists the analysis scripts used to produce the reported results.
[DOCX File , 80 KB-Multimedia Appendix 3]

References

1. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv Neural Inf Process Syst. 2020;33:9459-9474. [doi: 10.48550/arXiv.2005.11401]
2. Mikkelsen Y. Clinical context variables collectively rival model choice in embedding-based retrieval: multi-corpus benchmark study. JMIR Med Inform. 2026;14:e94241. [FREE Full text] [doi: 10.2196/94241] [Medline: 42097608]

XSL•FO
RenderX

3. Myers S, Miller TA, Gao Y, Churpek MM, Mayampurath A, Dligach D, et al. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. J Am Med Inform Assoc. 2025;32(2):357-364. [doi: 10.1093/jamia/ocae308] [Medline: 39703187]
4. Soffer S, Omar M, Gendler M, Glicksberg BS, Kovatch P, Efros O, et al. A scalable framework for benchmark embedding models in semantic health-care tasks. J Am Med Inform Assoc. 2025;32(12):1877-1887. [doi: 10.1093/jamia/ocaf149] [Medline: 40977370]
5. Wennberg JE, Fisher ES, Skinner JS. Geography and the debate over medicare reform. Health Aff (Millwood). 2002;Suppl Web Exclusives:W96-114. [doi: 10.1377/hlthaff.w2.96] [Medline: 12703563]
6. Zhao Z, Yuan H, Liu J, Chen H, Ying H, Zhou S, et al. CliniQ: a multi-faceted benchmark for electronic health record retrieval with semantic match assessment. ArXiv. Preprint posted online on February 10, 2025. [doi: 10.48550/arXiv.2502.06252]
7. Mikkelsen Y. Effects of model choice, corpus context, and post hoc correction on layer-level embedding degradation in clinical document retrieval: experimental study. JMIR Med Inform. 2026;14:e99639. [FREE Full text] [doi: 10.2196/99639] [Medline: 42455615]
8. Johnson A, Pollard T, Horng S, Celi L, Mark R. MIMIC-IV-Note: deidentified free-text clinical notes (version 2.2). PhysioNet. 2023. [doi: 10.13026/1n74-ne17]
9. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. [FREE Full text] [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
10. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
11. Molina M, Mehandru N, Golchini N, Alaa A. ER-REASON: a benchmark dataset for LLM-based clinical reasoning in the emergency room (version 1.0.0). PhysioNet. 2025. [FREE Full text] [doi: 10.13026/55s7-3c27]
12. Xiao S, Liu Z, Zhang P, Muennighoff N, Lian D, Nie J. C-Pack: packed resources for general Chinese embeddings. 2024. Presented at: SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; July 14-18, 2024:641-649; Washington, DC, United States. [doi: 10.1145/3626772.3657878]
13. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. ArXiv. Preprint posted online on August 7, 2023. [doi: 10.48550/arXiv.2308.03281]
14. Wang L, Yang N, Huang X, Jiao B, Yang L, Jiang D, et al. Text embeddings by weakly-supervised contrastive pre-training. ArXiv. Preprint posted online on December 7, 2022. [doi: 10.48550/arXiv.2212.03533]
15. Nussbaum Z, Morris J, Duderstadt B, Mulyar A. Nomic Embed: training a reproducible long context text embedder. ArXiv. Preprint posted online on February 2, 2024. [doi: 10.48550/arXiv.2402.01613]
16. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019:3982-3992; Hong Kong, China. [doi: 10.18653/v1/D19-1410]
17. Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. Bioinformatics. 2023;39(11):btad651. [FREE Full text] [doi: 10.1093/bioinformatics/btad651] [Medline: 37930897]
18. Remy F, Demuynck K, Demeester T. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. J Am Med Inform Assoc. 2024;31(9):1844-1855. [FREE Full text] [doi: 10.1093/jamia/ocae029] [Medline: 38412333]
19. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. Presented at: Proceedings of NAACL-HLT 2019; June 2-7, 2019:4171-4186; Minneapolis, MN, United States. [doi: 10.18653/v1/N19-1423]
20. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-1240. [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
21. Wang G, Liu X, Ying Z, Yang G, Chen Z, Liu Z, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. Nat Med. 2023;29(10):2633-2642. [FREE Full text] [doi: 10.1038/s41591-023-02552-9] [Medline: 37710000]
22. medicalai/ClinicalBERT. Hugging Face. URL: https://huggingface.co/medicalai/ClinicalBERT [accessed 2026-08-10]
23. Liu X, Liu H, Yang G, Jiang Z, Cui S, Zhang Z, et al. A generalist medical language model for disease diagnosis assistance. Nat Med. Mar 2025;31(3):932-942. [doi: 10.1038/s41591-024-03416-6] [Medline: 39779927]
24. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2021;3(1):1-23. [doi: 10.1145/3458754]
25. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019:3615-3620; Hong Kong, China. [doi: 10.18653/v1/D19-1371]

XSL•FO
RenderX

26. Mikkelsen Y. Cross-site stability of embedding model rankings for known-item retrieval from clinical notes. Zenodo. 2026. [doi: 10.5281/zenodo.21434088]
27. Khattab O, Zaharia M. ColBERT: efficient and effective passage search via contextualized late interaction over BERT. 2020. Presented at: SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; July 25-30, 2020:39-48; Virtual Event. [doi: 10.1145/3397271.3401075]
28. Ethayarajh K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019:55-65; Hong Kong, China. [doi: 10.18653/v1/D19-1006]
29. Mikkelsen Y. cross-site-embedding-rankings: source code repository. GitHub. URL: https://github.com/yngvemikkelsen/cross-site-embedding-rankings [accessed 2026-10-20]

Abbreviations

BIDMC: Beth Israel Deaconess Medical Center
CLS: classification token
EHR: electronic health record
ICD: International Classification of Diseases
MIMIC: Medical Information Mart for Intensive Care
MIMIC-IV: Medical Information Mart for Intensive Care IV
MIMIC-IV-Note: Medical Information Mart for Intensive Care IV, Clinical Notes Module
MLM: masked language model
MRR: mean reciprocal rank
MRR@10: mean reciprocal rank at 10
RAG: retrieval-augmented generation
RQ: research question
UCSF: University of California, San Francisco

Edited by A Coristine; submitted 11.Aug.2026; peer-reviewed by M Reiter-Haas, Y Xin; comments to author 09.Sep.2026; revised version received 09.Oct.2026; accepted 13.Oct.2026; published 11.Nov.2026

Please cite as:
Mikkelsen Y
Cross-Site Stability of Embedding Model Rankings for Known-Item Retrieval From Clinical Notes: Two-Corpus Comparative Evaluation Study
JMIR AI 2026;5:e109305
URL: https://ai.jmir.org/2026/1/e109305
doi: 10.2196/109305
PMID:

©Yngve Mikkelsen. Originally published in JMIR AI (https://ai.jmir.org), 11.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX