# CaseReportCollective: A Large-Scale LLM-Extracted Dataset for Structured Medical Case Reports

**Xiao Yu Cindy Zhang** **Wyeth Wasserman** **Melissa Fong** **Jian Zhu**
University of British Columbia
czhang@cmmt.ubc.ca

## Abstract

Case reports provide critical insights into rare and atypical diseases, but extracting structured knowledge remains challenging due to unstructured text and domain-specific terminology. We introduce **CaseReportCollective**, an LLM-extracted dataset of 85,961 open-access case reports spanning 37 years, validated through programmatic and human evaluation. Our dataset reveals key publication and demographic trends, including a significant increase in open-access case reports over the past decade, shifts in focus from oncology to COVID-19 and sex disparities in reporting across different medical conditions. Using **CaseReportCollective**, we further explore embedding-based retrieval for similar medical topics through accumulated similarity scores across extracted structured information. We also conducted detailed error analyses on the retrieval ranking, finding that highly reported topics dominate retrieval and the retrieval is driven by lexical overlap rather than underlying clinical relevance, often failing to distinguish between semantically similar yet mechanistically distinct conditions. Future work should focus on clinically aware embeddings adjusted for long-tailed case distributions to improve retrieval accuracy.

## 1 Introduction

Case reports, structured summaries outlining individual patient profiles and distinctive medical conditions (Venes, 2017), have historically played a critical role in rare disease discovery, novel treatment vigilance, and pandemic surveillance (Nissen and Wynn, 2014; Wu and Sung, 2003; Hymes et al., 1981). As of September 2023, over 2.41 million cases have been published (Parums, 2023), capturing a wealth of clinical details, including patient history, review of systems, laboratory findings, and imaging results. Leveraging this vast repository of medical knowledge has the potential to advance medical research and clinical education significantly. However, extracting structured knowledge from case reports remains a major challenge. Clinical narratives often contain domain-specific terminology, abbreviations, and colloquial descriptions, making automated extraction difficult without a foundational understanding of medical language. Additionally, key metadata such as patient sex and age are frequently implied rather than explicitly stated, requiring common-sense reasoning for accurate interpretation. The manual process of perusing case reports and distilling actionable insights is both labor-intensive and time-consuming, hindering large-scale systematic analysis. Furthermore, traditional rule-based natural language processing (NLP) approaches struggle with the semantic variability and unstructured nature of medical text, limiting their ability to aggregate and standardize case report data effectively.

In this study, we leveraged LLMs and rule-based algorithms to extract granular details from open-access medical case reports in Pubmed Central(PMC) into medical categories standard for patient assessments. Leveraging the metadata along with the fine-grain LLM extractions from this dataset, we analyzed the case report trends for publication years, sex, and patient age. With these fine-grained extractions from case reports, we demonstrate how this dataset can be used for information retrieval for similar cases. Our primary contribution is the construction of a large-scale, LLM-structured case report corpus. The demographic analyses are included to illustrate the dataset's clinical coverage and its potential for supporting diagnostic research across diverse patient populations and medical conditions. Specifically, we highlight:

- **CaseReportCollective:**[1] An LLM-extracted dataset of 85,961 open-access medical case

[1] *CaseReportCollective* is publicly available at https://huggingface.co/datasets/cxyzhang/CaseReportCollective_V1.0.



*Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025)*, pages 249–262

reports spanning 37 years, with structured extractions across 14 clinical categories and quality control via programmatic metrics and human evaluation.

- Uncovering **significant differences in sex distribution across age groups, publication years, and medical topics**. Balanced sex representation is observed only in the 42–65 age group, with more males in the 65+ and pediatric categories, and more females in the 18–41 age group. Over time, we observed the inclusion of intersex individuals in case reports. Additionally, certain conditions are disproportionately reported in one sex, with both biological factors and potential sex biases influencing the findings.

- Identifying **systematic biases in embedding-based disease retrieval**, including **prevalence bias, textual co-occurrence bias, and pathophysiological mismatches**. We demonstrate how high-frequency diseases (e.g., tuberculosis) dominate retrieval results, often suppressing rarer but clinically significant conditions. Additionally, semantic similarity alone proves insufficient for clinically accurate retrieval, as it frequently retrieves conditions based on surface-level word overlap rather than true clinical relevance. We suggest **context-aware embeddings and prevalence-adjusted ranking mechanisms** as future directions to improve retrieval accuracy.

## 2 Related Work

### 2.1 Medical Information Extraction

Rule-based systems and ontology-driven pipelines have been foundational in clinical NLP. Tools such as MetaMap (Aronson, 2001), Regextractor (Hinchcliff et al., 2012), MedLEE (Friedman et al., 1995), and cTAKES (Savova et al., 2010) extract clinical concepts using predefined grammars and the Unified Medical Language System (UMLS) (Bodenreider, 2004). While these systems offer transparency and have been trusted by clinicians, they require expert rule engineering, are costly to maintain, and struggle with terminological variation, leading to lower recall in open-domain scenarios.

To improve generalizability, hybrid models and deep learning have been proposed. Precursor-induced CRFs outperform traditional CRFs by propagating token context (Lee and Choi, 2019), while models like BioBERT and BiLSTM-CRF have shown strong results in biomedical NER tasks (Schulz et al., 2020). However, these approaches rely heavily on large-scale annotated corpora and may underperform on rare disease data. Notably, fine-tuned BioClinicalBERT has achieved high accuracy in extracting rare disease phenotypes from unstructured narratives(Shyr et al., 2024).

Recently, instruction-tuned large language models (LLMs) have emerged as general-purpose extractors capable of operating with minimal supervision. For example, InstructGPT extracted pediatric foreign body injury data across languages (Sciannameo et al., 2024), and ChatGPT outperformed BioClinicalBERT in rare disease phenotype extraction in one-shot settings (Shyr et al., 2024). These results suggest LLMs encode latent biomedical knowledge learned from large-scale corpora. While LLMs are not always superior to traditional NER architectures for structured or narrow-domain tasks, we leverage them in this work for their domain transferability and their ability to perform dense, multi-category extraction with minimal annotation effort.

Different from prior work in structuring clinical case reports (Zhao et al., 2022; Raza and Schwartz, 2023; Sciannameo et al., 2024), CaseReportCollective dataset is at a substantially larger scale, with LLMs applied across 14 categories and 85,961 case reports. This work complements existing clinical corpora such as MIMIC-III (Johnson et al., 2016), MedNLI (Romanov and Shivade, 2018), and N2C2 datasets (Stubbs et al., 2019), which focus on discharge summaries or specific annotation tasks. In contrast, our corpus standardizes narrative case reports into structured data that enables downstream demographic analysis and diagnostic benchmarking.

### 2.2 Sex Disparities in Clinical Narratives

Clinical narratives have historically reflected sex-based disparities in disease prognosis, presentation, diagnosis, and treatment (Bello and Mosca, 2004). These inequalities can introduce biases in clinical decision-making, ultimately affecting patient outcomes. For instance, one study found that males receive a diagnosis at a younger age than females, highlighting potential delays in recognition and intervention for female patients (Alcalde-Rubio et al., 2020). Additionally, an analysis of word embeddings applied to biomedical text revealed system-

atic biases, where substance use disorders were more frequently associated with males, while psychiatric disorders were more commonly linked to females, reinforcing harmful stereotypes in medical literature(Rios et al., 2020).

Our study leveraged knowledge of pretrained LLMs to perform dense extraction across multiple distinct medical categories. Unlike previous studies, we perform fine-grained dense extraction performance across multiple medical domains and demonstrate the utility of LLM-extracted data in biomedical research. In contrast to studies primarily focused on Named Entity Recognition (NER) for certain medical specialties (Abiha, 2024; Turchioe et al., 2022), **CaseReportCollective** provides a structured dataset spanning multiple medical specialties. Additionally, its metadata facilitates investigations into sex- and age-related differences in disease presentation, showcasing LLMs' ability to extract meaningful clinical trends from unstructured text. Furthermore, per-category extractions enable a fine-grained evaluation of embedding-based retrieval.

## 3 Methods

To construct **CaseReportCollective**, we leveraged publicly available clinical case reports and implemented a structured LLM-based extraction and evaluation pipeline.

### 3.1 Dataset construction

**CaseReportCollective** is developed using clinical case reports from the non-commercial PubMed Central (PMC) Open Access subset, sourcing full-text articles under CC BY-NC, CC BY-NC-SA, and CC BY-NC-ND licenses, accessed via the PMC FTP [2] on February 3, 2024. To extract structured clinical information, we instructed an LLM to identify 14 key clinical categories adapted from a specific standardized approach used inpatient Work-Up and monitoring for healthcare professionals [3]: **Vitals_Hema** (Vitals and Hematology Findings), **EENT** (Eyes, Ears, Nose, and Throat), **NEURO** (Neurology), **CVS** (Cardiovascular System), **RESP** (Respiratory System), **GI** (Gastrointestinal System), **GU** (Genitourinary System), **MSK** (Musculoskeletal System), **DERM** (Dermatology), **LYMPH** (Lymphatic System), **ENDO** (Endocrinology), **Pregnancy**, **Lab_Image** (Laboratory and Imaging), and **History**.

[2] https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/oa_noncomm/xml/

[3] https://blogs.ubc.ca/oeetoolbox/2019/02/patient-work-up-from-sample-template-inpatient/

### 3.2 Preprocessing LLM Extraction

We applied the few-shot and category-specific prompt templates used for structured information extraction from case report narratives, as described in (Zhang et al., 2025), where for each clinical category, prompts include a task-specific instruction followed by output formatting constraints. Due to the limitations encountered with earlier LLM frameworks for generating JSON-formatted output, a multi-step preprocessing approach was implemented. All nested categories were converted into a single list of strings. Subkeys within the JSON document were concatenated with their values to preserve context. For example, "blood pressure" is connected with the corresponding value "120/80 mmHg". There were observed instances when LLM extraction failed to retrieve relevant information due to either (a) a lack of detected relevant information or (b) formatting issues, such as the incorrect use of double quotes instead of single quotes, which led to JSON parsing errors. In these cases, we attempted to standardize the format by replacing single quotes with double quotes and then reattempted the LLM extraction.

### 3.3 LLM-Based Diagnostic Label Extraction and NER Supplementation

The LLM was instructed to extract the medical conditions from the title of each case report. For comparison, we performed NER using SciSpacy (Neumann et al., 2019). Since rare conditions are often under-represented in SciSpacy without additional context, we provided both keywords and labels to improve recognition. However, keywords often contain extraneous or broad information that is not the main focus of the case report (e.g., "pregnancy" in the context of "pregnancy luteoma"), which can dilute the core medical condition being described. To address this, we prioritized the LLM-extracted labels as the primary diagnostic labels. The NER output was only used to supplement these labels when the LLM failed to extract the relevant condition.

### 3.4 Demographic Attribute Extraction

Biological ages in case reports typically follow a standard format (e.g., "X-year-old"). To enable efficient and deterministic extraction, we applied

rule-based keyword extraction for age identification. Ages were categorized into predefined clinical groups: **Neonatal (0–1 month)**, **Infancy (1–18 months)**, **Childhood (1.5–11 years)**, and **Adolescence (11–16 years)** (Blau et al., 2014). Adulthood was further divided into **16–41 years**, **41–64 years**, and **>64 years**. Cases without age data were labeled as "Unspecified."

In contrast, biological sex is expressed more variably and often implicitly, requiring an LLM for context-dependent extraction. For instance, when a patient is described as "nulliparous," LLM may leverage its foundational knowledge to infer the patient as biologically female. Additionally, the LLM was instructed to recognize intersex category—characterized by physical, hormonal, or genetic traits—affecting approximately 1.7% of the population (Sax, 2002; Zeeman and Aranda, 2020).

The Chi-Square ($\chi^2$) test for independence is performed in investigating relationships between **age**, **sex**, **publication years** and **medical topics** in **CaseReportCollective**.

### 3.5 Implementation

We performed structured extraction of category-specific clinical information and diagnostic labels from case report texts and titles using few-shot prompting tailored for verbatim information capture. For each clinical category, we designed task-specific prompts that requested outputs in a standardized dictionary format. These prompts followed a consistent template with explicit formatting instructions to facilitate post-processing, as detailed in Zhang et al. (2025). For example, in the Neurological category, prompts instructed models to extract findings such as "neurological", "cognitive", "neurological tests and imaging" with outputs keyed by clinical feature types.

Initial large-scale extraction was conducted using LLaMA 3-8B-Instruct (Dubey et al., 2024), running under the Ollama framework[4] with 4-bit quantization on an NVIDIA Tesla V100 GPU, selected for its availability and computational efficiency. Benchmarking results from Zhang et al. (2025) showed that Qwen2.5-7B-Instruct (Hui et al., 2024) yielded better alignment with clinician judgments for dense clinical information extraction, supporting its use in subsequent inference tasks to extract biological sex from case report texts. This model was deployed using 16-bit floating point precision under the vLLM framework (Kwon et al., 2023). All models were set to a temperature of 0 to ensure deterministic outputs.

### 3.6 Evaluation of Extracted Texts with Automated Metrics and Human Assessment

Since the LLM was tasked with extracting verbatim text from case reports, we assessed extraction fidelity using dual string-based metrics: **Exact Match (EM)** and **Token Set Ratio (TSR%)**, implemented via the `fuzzywuzzy` library[5]. **EM** measures the proportion of extractions that exactly match the original text (ranging from 0 to 1), while **TSR (%)** quantifies partial similarity (ranging from 0 to 100) by allowing slight variations. To assess the fidelity of LLM-extracted text compared to the original case report, we compute the **Token Set Ratio (TSR)**. TSR is a partial similarity metric that captures approximate matches between texts by comparing token-level overlap and differences.

Let $T_1$ denote the set of tokens from the original case report text, and $T_2$ the set of tokens from the LLM-extracted output. We compute:

$$I = T_1 \cap T_2, \quad D_1 = T_1 \setminus T_2, \quad D_2 = T_2 \setminus T_1$$

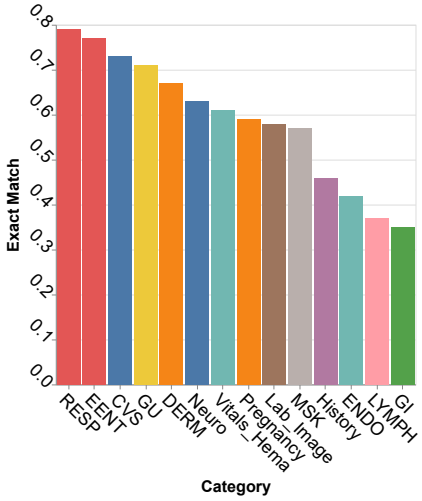
Here, $I$ denotes the shared tokens, $D_1$ represents tokens found only in the original text, and $D_2$ those found only in the LLM extraction. These token groups are each converted into strings, and string similarity is then assessed using the Levenshtein distance as implemented in the `fuzzywuzzy`.

Evaluation of a randomly selected subset of 400 LLM-extracted labels against their respective case report titles was performed by a student, guided by medical oversight. The evaluation focused on three criteria: **relevance**—alignment of the extracted entity with the title, **specificity**—correct identification of primary diseases or conditions, and **completeness**—thorough extraction of all relevant medical conditions. The detailed annotation guidelines are provided in Appendix A.

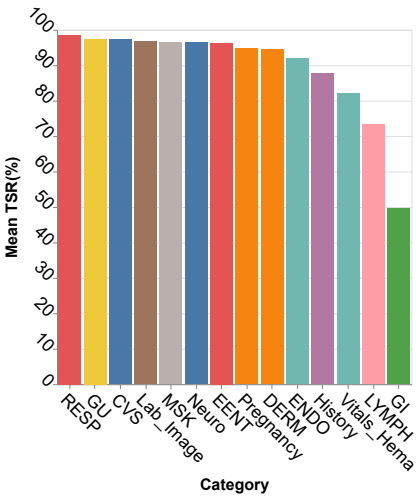
Additionally, a student, guided by medical oversight, evaluated a randomly selected subset of 400 LLM-extracted labels against the original article title, comparing them against their respective case report titles. The evaluation focused on three key criteria: **relevance**–alignment of the extracted entity with the title, **specificity**–correct identification

[4] https://github.com/ollama/ollama

[5] https://github.com/seatgeek/fuzzywuzzy




(A)




(B)

Figure 1: Programmatic Evaluation Results for LLM Per-Category Extraction. (A) Exact Match Score for Extracted Strings against Case Text. (B) Token Set Ratio of Extracted Strings against Case Text

of primary diseases or conditions, and **completeness**– thorough extraction of all relevant medical conditions. The detailed annotation guidelines are provided in Appendix A.

### 3.7 CaseReportCollective as Information Retrieval (IR) System

Medical conditions frequently involve multiple body systems, making it difficult to retrieve precise information from case reports. Analyzing entire case reports can obscure system-specific details and introduce confounding effects. We hypothesize that system-specific LLM extractions from CaseReportCollective can improve diagnosis retrieval by preserving relevant information within distinct medical categories.

For this IR task, we first converted the LLM-extracted category-specific texts into embeddings using MedEmbed (Balachandran, 2024). To evaluate retrieval across varying disease prevalences, we sampled 100 topics each from the top, middle, and bottom of the global frequency distribution—representing high, medium, and low-frequency groups—ensuring one unique case per topic. These queries were excluded from the retrieval corpus, which comprised the remaining 80K cases. Retrieval was performed based on L2-normalized embedding similarity via FAISS[6].

The accumulated similarity score for each test case is computed by first retrieving the top-K most similar disease topics from each clinical category. retrieved topics and their similarity scores were collected separately per category. If a topic appeared in multiple categories, its scores were averaged across categories to compute an accumulated similarity score. Final rankings per query were generated by sorting retrieved topics based on these averaged scores, reflecting cross-category semantic consistency.

Finally, Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain(NDCG@50), and Precision@50 were used for IR evaluation.

## 4 Results and Discussions

### 4.1 Dataset Composition

**CaseReportCollective** comprises 85,961 open-access case reports covering 53K unique combinations of medical topics published between 1986 and February 2024 (but notably with most of the full-text open-access case reports appearing in the past decade). On average, case reports contain 3,462 ± 1,920.66 words. The mean number of reports per condition is 2.88 ± 10.49, with COVID-19 (410 cases) being the most frequently reported topics, highlighting a skewed distribution where a small subset of topics dominates the dataset.

The amount of LLM-extracted information varies, with total extraction item counts 27.77 ± 81.57 across 14 categories. Example entries of **CaseReportCollective** can be found in Appendix C. **Lab_Image**, which includes all laboratory tests and imaging across body systems, along with **History**, have the highest extracted string

[6] https://github.com/facebookresearch/faiss/

counts due to their broad and inclusive nature. **CVS** has the third highest extracted string count, followed by **MSK** and **Vitals_Heme**. In contrast, the **GI** category has an extremely low extracted count in this dataset, which may reflect either the inherently limited description of gastrointestinal-related information in clinical case reports or the LLM's difficulty in recognizing such information. Appendix B shows the string count distribution per category.

### 4.2 LLM Extraction Quality

Although the mean EM score is at 0.59 ± 0.14, a high mean TSR(%) of 87.25 ± 10.79 is achieved, suggesting that LLM-extracted content effectively captures the original text but may introduce minor variations in wording or structure. As shown in Fig. 1, the RESP category exhibits the highest EM, indicating that respiratory-related extractions have the highest alignment. In contrast, the GI category has the lowest scores, suggesting that the LLM struggled to extract gastrointestinal-related information accurately, potentially due to variability in how such details are reported.

Out of 400 extracted medical topics for human evaluation, 19 cases (4.75%) were labeled as hallucinations by the human reviewer, where the LLM generated terms that were unrelated to the input text, overgeneralized, or misclassified (e.g., procedural terms instead of medical conditions). These errors likely stem from insufficient contextual information in the article title and biases toward frequently mentioned conditions in the LLM's training data, warranting further analysis. Despite these hallucinations, most extractions were clinically relevant, with mean scores of 2.94 ± 0.32 for relevance, 2.81 ± 0.39 for specificity, and 2.87 ± 0.36 for completeness. These results demonstrate strong performance, as detailed in Appendix A.

### 4.3 Temporal Trends

The publication of open-access case reports has increased significantly over the past decade. Figure 2 illustrates this trend, showing sporadic case report publications between 1986 and 2002, followed by a notable rise in recent years. This growth reflects the broader adoption of open access and a growing appreciation for case reports in clinical care.

The trend of case report topics has shifted over time, reflecting evolving clinical priorities. Before 2020, case reports predominantly focused on cancers (e.g., squamous cell carcinoma, hepatocarcinoma, renal cell carcinoma) and vascular conditions (e.g., aneurysms). During 2020-2021, COVID-19-related case reports surged, highlighting the role of case reports in rapid knowledge dissemination during global health crises. Post-2021, the focus changed to oncological and rare conditions (e.g., mucormycosis).

### 4.4 Age and Sex Stratification

Overall, **CaseReportCollective** consists of 31.61% Adulthood (42–65 yr), 28.12% Adulthood (18–41 yr), 18.50% Adulthood (>65 yr), 10.97% Childhood, 4.75% Infancy, 4.27% Adolescence, and 0.36% Neonatal cases, with 1.57% missing age extraction. Regarding sex distribution, the dataset comprises 55.60% Female, 44.10% Male, and 0.10% Intersex cases, with 0.20% missing sex assignment.

### 4.5 Sex Distribution Across Age Groups, Years, and Medical Topics

Sex composition varies significantly across age groups ($\chi^2 = 192.03$, $df = 12$, $p < 1.44 \times 10^{-34}$) (Fig. 3). Intersex cases are rare across all age groups, with the highest frequency observed in childhood (15 cases). These findings suggest a dependency between sex and age groups, potentially influenced by age-stratified biological factors, reporting practices, or selection biases.

We found significant variation in both age ($\chi^2 = 862.39$, $df = 252$, $p = 8.74 \times 10^{-68}$) and sex ($\chi^2 = 108.18$, $df = 72$, $p = 0.0037$) distributions across publication years. As shown in Figure 2, Female cases have generally been reported more frequently than male cases across all years, with the disparity widening over time. Intersex cases remain rare, appearing only after 2011.

The chi-square test ($\chi^2 = 401.70$, $df = 174$, $p < 4.73 \times 10^{-20}$) indicates a significant difference in sex distribution across high-frequency medical topics ($\geq 100$ occurrences). This suggests that certain medical conditions are disproportionately reported in one sex over the other (Figure 4). While some disparities may be attributed to sex-specific physiology and pathological differences, as reported in prior studies, others may result from systematic biases.

### 4.6 Can Embedding Models Reliably Retrieve Clinically Relevant Diseases?

We evaluated **CaseReportCollective** as a retrieval-based disease-ranking method that leverages em-

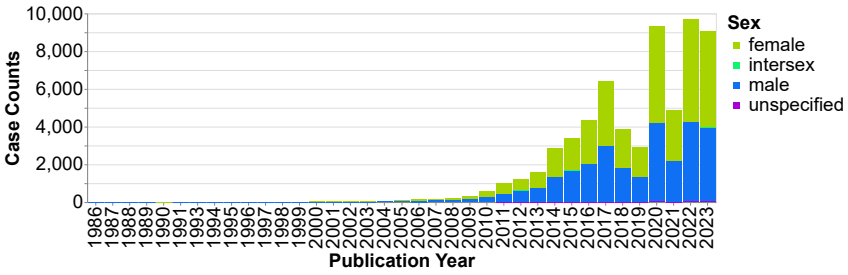



Figure 2: Biological Sex Distribution between 1986 and 2023

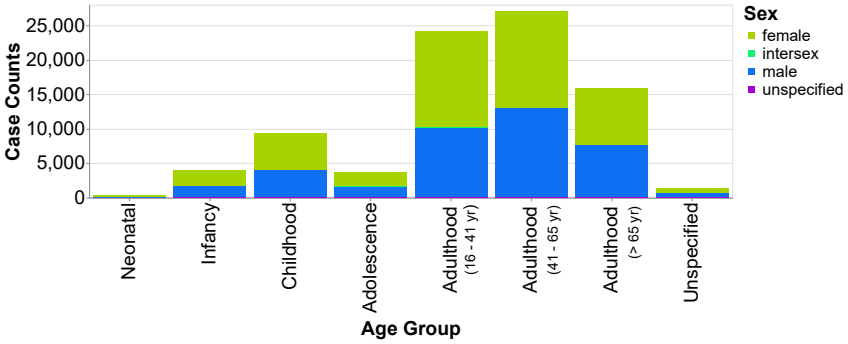



Figure 3: Biological Sex Distribution across Age Groups.

bedding similarity, retrieval frequency, and topic prevalence within the dataset. For each test case, we used category-specific embeddings to perform nearest-neighbor retrieval using FAISS, a fast vector similarity search library. Specifically, we retrieved the top 50 most similar topics for such an evaluation.

#### 4.6.1 Limitations of Traditional IR Metrics

While traditional IR metrics such as MRR, NDCG@50, and Precision@50 provide useful benchmarks for retrieval performance, they may underestimate the capabilities of embedding-based methods when applied to complex clinical narratives. This is particularly true in medical settings where semantically similar conditions may be expressed using diverse terminologies, synonyms, or compositional phrases that differ from canonical labels. Moreover, our case report dataset frequently presented multiple medical topics within single cases (e.g.,"adenomatous polyps, Lynch syndrome"), both of which were represented in the textual descriptions, making it challenging to distinguish. Although we initially considered standardizing medical topics using ontologies like UMLS, we found such mappings insufficient for less common medical conditions, leading to substantial information loss. Hence, we opted out of ontology-based standardization for this study.

In the evaluation, we permitted partial matching between retrieved and query topics, allowing matches such as "cystic fibrosis, multidrug-resistant pseudomonas infection" with "cystic fibrosis." The IR results (Fig. 5) show that our retrieval system has a suboptimal MRR of 0.026 for high-frequency topics, 0.01 for medium-frequency topics, and 0.0 for low-frequency topics, and struggles with ranking consistency as indicated by NDCG@50 scores of 0.19 for high-frequency topics, 0.05 for medium-frequency topics, and 0.07 for low-frequency topics. The system performs better for high-frequency topics in terms of NDCG, compared medium- and low-frequency topics. However, the overall low NDCG scores suggest that the system's ability to rank clinically significant diseases, including rare, low-frequency conditions, is limited. Furthermore, the extremely low Precision@50 for all topics indicates that many retrieved topics result from semantic linkage rather than true diagnostic relevance, highlighting a key limitation in the system's precision for clinical applications.

#### 4.6.2 Systematic Errors in Retrieval

We analyzed tuberculosis—a frequent topic—and highlighted some of the representative failure cases using our category-specific embedding-based system in Table 1. Despite its high frequency in the dataset, the retrieval system often over-prioritized

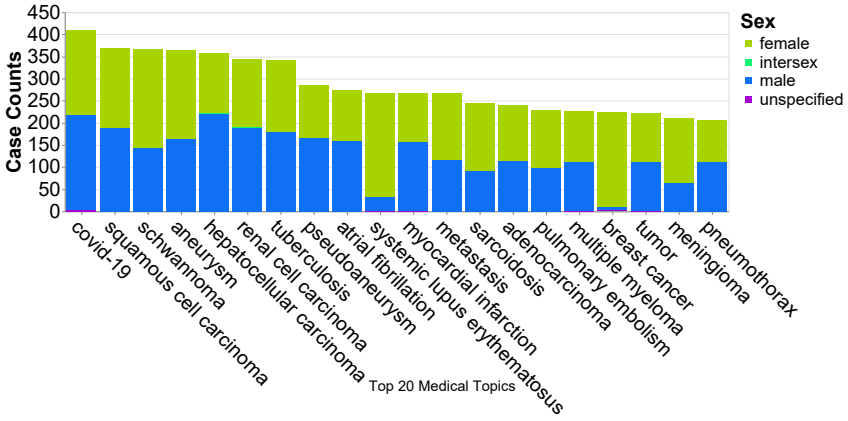



Figure 4: Biological Sex Distribution over Top 20 Medical Conditions.

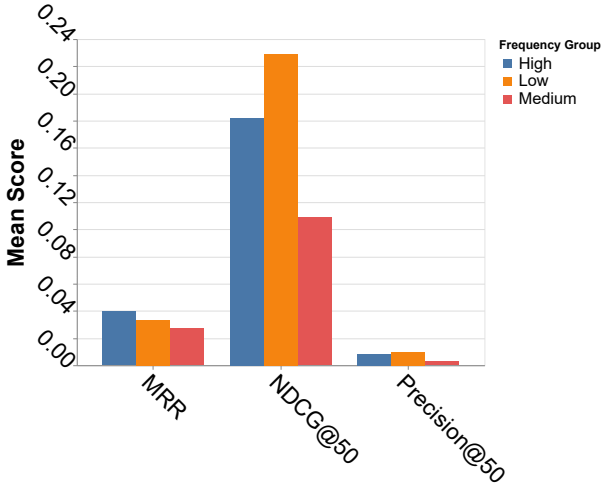



Figure 5: Mean Scores for Evaluation Metric across Three Frequency Groups

tuberculosis due to multiple failure modes. These include: **(1) Semantic Drift**, where chronic dermatologic conditions like nevus sebaceous were retrieved due to shared descriptors of persistent lesions; **(2) Anatomical Misalignment**, such as tracheal diverticula, arising from co-mentions in thoracic imaging contexts; **(3) Co-Treatment Artifact**, where conditions like steroid withdrawal syndrome appear due to shared treatment settings; **(4) Overgeneralized Infection Embedding**, where retrieval conflates unrelated infections like omphalitis or liver abscess; **(5) Anatomic Generalization**, where genitourinary tuberculosis cues led to retrievals like renal stone or UTI; **(6) Surface-Level Embedding Similarity**, as seen in matches like hemophilia B, driven by shared symptoms such as inflammation or bleeding; **(7) Rare Co-occurrence Confusion**, where diseases common in immunocompromised hosts (e.g., EBV/HLH) are incorrectly linked; and **(8) Entity Type Mismatch**, where congenital anomalies (e.g., anorectal malformation) are retrieved despite fundamentally differing etiology. Notably, many of these spurious matches yielded high similarity scores (>0.86), underscoring the embedding model's reliance on lexical and contextual overlap rather than clinically meaningful distinctions. Our findings indicate that the current embedding model is insufficient to fully capture the complexity of differential diagnosis.

## 5 Conclusion

In this study, we present CaseReportCollective, a large-scale structured dataset of medical case reports. Our analysis of the case reports suggest that the sex disparities in medical case reports have been decreasing temporally. Our findings demonstrate that, while leverage LLM-extracted category-wised information for embedding-based retrieval, there are still systematic failure modes that compromise clinical reliability, especially when unrelated conditions share surface-level linguistic features or co-occur in similar narrative contexts. Future work should explore the integration of structured clinical knowledge, prevalence-aware ranking mechanisms, and context-sensitive embedding models to improve medical retrieval systems.

| Index | Retrieved Topic | Query Topic | Norm Similarity | Issue | Failure Type | Possible Explanation for High Similarity |
|---|---|---|---|---|---|---|
| 4543 | nevus sebaceous, syringocystadenoma papilliferum | tuberculosis | 0.878 | Skin tumor unrelated to TB | Semantic Drift | Shared mention of chronic lesions or dermatological findings |
| 4544 | tracheal diverticula | tuberculosis | 0.872 | Airway abnormality unrelated to TB | Anatomical Misalignment | Co-occurrence in chest imaging discussions |
| 4545 | depression, steroid withdrawal syndrome | tuberculosis | 0.869 | Psychological syndromes unrelated to infection | Co-Treatment Artifact | TB and steroid use both appear in chronic illness contexts |
| 4546 | omphalitis, pyogenic liver abscess | tuberculosis | 0.868 | Different infection types | Overgeneralized Infection Embedding | Embedding captures general infection-related semantics |
| 4547 | renal stone, urinary tract infection | tuberculosis | 0.865 | Genitourinary disease not specific to TB | Anatomic Generalization | Overlap via genitourinary TB mentions |
| 4548 | hemophilia b, subgaleal hematoma | tuberculosis | 0.864 | Hematological condition | Embedding Surface Similarity | Shared features like bleeding or inflammation |
| 4549 | chronic active EBV infection, HLH, NK cell lymphoma | tuberculosis | 0.863 | Viral and hematologic malignancies | Rare Co-occurrence Confusion | TB sometimes mentioned in immunocompromised patients |
| 4550 | Churg-Strauss syndrome, neuroendocrine carcinoma | tuberculosis | 0.862 | Vasculitis and cancer unrelated to TB | Multisystem Similarity Confusion | Both may affect multiple organs, mentioned with granulomas |
| 4551 | anorectal malformation, ileal perforation | tuberculosis | 0.862 | Congenital/anatomical vs acquired infection | Entity Type Mismatch | Shared surgical or gastrointestinal mentions |
| 4552 | trichilemmal carcinoma | tuberculosis | 0.861 | Skin cancer unrelated to TB | Lexical Overlap | Chronic cutaneous conditions may trigger similarity |

Table 1: Issues in Retrieval of Tuberculosis: High Similarity but Incorrect Matches, Categorized by Failure Type

## References


Syeda Abiha. 2024. Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. *Journal of Artificial Intelligence and Health*, 1(2):57–65.

Lorena Alcalde-Rubio, Ildefonso Hernández-Aguado, Lucy Anne Parker, Eduardo Bueno-Vergara, and Elisa Chilet-Rosell. 2020. Gender disparities in clinical practice: are there any solutions? scoping review of interventions to overcome or reduce gender bias in clinical practice. *International journal for equity in health*, 19:1–8.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Natalie Bello and Lori Mosca. 2004. Epidemiology of coronary heart disease in women. *Progress in cardiovascular diseases*, 46(4):287–295.

Nenad Blau, Marinus Duran, K Michael Gibson, and Carlo Dionisi Vici. 2014. *Physician's guide to the diagnosis, treatment, and follow-up of inherited metabolic diseases*, volume 213. Springer.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. 1995. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108.

Monique Hinchcliff, Eric Just, Sofia Podlusky, John Varga, Rowland W Chang, and Warren A Kibbe. 2012. Text data extraction for a prospective, research-focused data mart: implementation and validation. *BMC medical informatics and decision making*, 12:1–7.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

KennethB Hymes, JeffreyB Greene, Aaron Marcus, DanielC William, Tony Cheung, NeilS Prose, Harold Ballard, and LindaJ Laubenstein. 1981. Kaposi's sarcoma in homosexual men—a report of eight cases. *The Lancet*, 318(8247):598–600.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Wangjin Lee and Jinwook Choi. 2019. Precursor-induced conditional random fields: connecting separate entities by induction for improved clinical named entity recognition. *BMC Medical Informatics and Decision Making*, 19:1–13.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Trygve Nissen and Rolf Wynn. 2014. The clinical case report: a review of its merits and limitations. *BMC research notes*, 7:1–7.

Dinah V Parums. 2023. the increasing relevance of case reports in medical education and clinical practice–and how to write them. *The American Journal of Case Reports*, 24:e942670–1.

Shaina Raza and Brian Schwartz. 2023. Entity and relation extraction from clinical case reports of covid-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, 23(1):20.

Anthony Rios, Reenam Joshi, and Hejin Shin. 2020. Quantifying 60 years of gender bias in biomedical research with word embeddings. In *Proceedings of the 19th SIGBioMed workshop on biomedical language processing*, pages 1–13.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Leonard Sax. 2002. How common is lntersex? a response to anne fausto-sterling. *Journal of sex research*, 39(3):174–178.

Sarah Schulz, Jurica Ševa, Samuel Rodriguez, Malte Ostendorff, and Georg Rehm. 2020. Named entities in medical case reports: corpus and experiments. *arXiv preprint arXiv:2003.13032*.

Veronica Sciannameo, Daniele Jahier Pagliari, Sara Urru, Piercesare Grimaldi, Honoria Ocagli, Sara Ahsani-Nasab, Rosanna Irene Comoretto, Dario Gregori, and Paola Berchialla. 2024. Information extraction from medical case reports using openai instructgpt. *Computer methods and programs in biomedicine*, 255:108326.

Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. 2024. Identifying and extracting rare diseases and their phenotypes with large language models. *Journal of Healthcare Informatics Research*, 8(2):438–461.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Ozlem Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J. Am. Medical Informatics Assoc.*, 26(11):1163–1171.

Meghan Reading Turchioe, Alexander Volodarskiy, Jyotishman Pathak, Drew N Wright, James Enlou Tcheng, and David Slotwiner. 2022. Systematic review of current natural language processing methods and applications in cardiology. *Heart*, 108(12):909–916.

Donald Venes. 2017. *Taber's cyclopedic medical dictionary*. FA Davis.

Eugene B Wu and Joseph JY Sung. 2003. Haemorrhagic-fever-like changes and normal chest radiograph in a doctor with sars. *The Lancet*, 361(9368):1520–1521.

Laetitia Zeeman and Kay Aranda. 2020. A systematic review of the health and healthcare inequalities for people with intersex variance. *International Journal of Environmental Research and Public Health*, 17(18):6533.

Xiao Yu Cindy Zhang, Carlos R. Ferreira, Francis Rossignol, Raymond T. Ng, Wyeth Wasserman, and Jian Zhu. 2025. Casereportbench: An llm benchmark dataset for dense information extraction in clinical case reports. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL), JMLR Workshop and Conference Proceedings*. To appear.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.

# A Human Evaluation Guidelines for LLM-Extracted Diagnostic Labels

**Objective:** Assess the accuracy, specificity, and clinical relevance of the LLM-generated labels in relation to the case report title. Use the Likert scale below for evaluation.

**Likert Scale for Evaluation**

| Score | Rating | Description |
|---|---|---|
| **3 - Excellent** | Perfect Match | Fully relevant, specific and complete. No improvement is needed. |
| **2 - Acceptable** | Partially Correct | The label is relevant but lacks key details (e.g. too broad or missing very few conditions). Minimal modification needed. |
| **1 - Unacceptable** | Incorrect or Misleading | Clinically wrong, misleading, or too vague to be useful. A major revision is needed. |

Table 2: Likert Scale for Evaluation

## A.1 Evaluation Criteria

**Evaluation Metrics:**

- **Relevance (1-3)**: Does the label relate to the case report title?
- **Specificity (1-3)**: Is the label precise and not too broad?
- **Completeness (1-3)**: Does the label capture the full diagnosis?

### A.1.1 1. Clinical Relevance

**Acceptable:**

- The label correctly identifies the primary disease, condition, or syndrome/symptom described in the title.
- The label is a well-recognized medical term or diagnosis.

**Not Acceptable:**

- The label is unrelated or related but too general (e.g., “disease” instead of “trigeminal schwannoma”).
- The label is misleading or incorrect.

### A.1.2 2. Specificity

**Acceptable:**

- The label captures the exact medical condition (e.g., "cardiac sarcoidosis" instead of just "sarcoidosis").
- The label includes relevant qualifiers when necessary (e.g., "trigeminal schwannoma" instead of just "schwannoma").

**Not Acceptable:**

- The label is too broad (e.g., for “brain abscess” extract as only “abscess”).
- The label adds unnecessary information that is not in the title.

### A.1.3 3. Completeness

**Acceptable:**

- The label correctly reflects all critical clinical elements in the title.
- If the title describes multiple conditions, the label should capture the main diagnosis.

**Not Acceptable:**

- The label only captures one part of a compound diagnosis when both are equally important (e.g., did not extract both “neuropathy” and “diabetes” in "neuropathy secondary to diabetes").

**Note:** Rather than evaluating individual entity completeness, the “completeness” metric is used to assess the full extraction of all entities, regardless of whether each concept is fully extracted. The specificity metric, however, will be used to evaluate the quality of each extracted entity.

# B Distribution of Extracted Strings Counts Across Clinical Categories

The bar plot shows the distribution of extracted string counts across different categories. Lab_Image and History contain the most string extractions with GI the least extractions.

# C Example Layout of CaseReportCollective

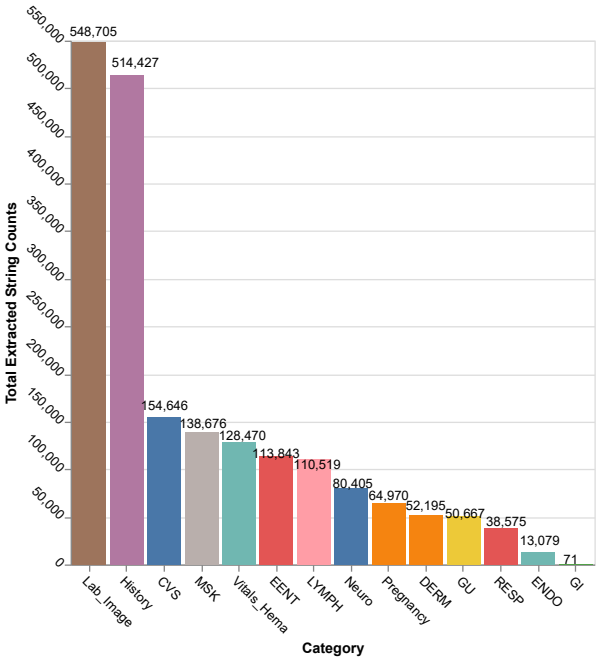



Figure 6: Distribution of Extracted String Counts Across Clinical Categories.

| pmcid | year | age | sex | topic | title | case | length | Vitals_Hema ... (Omitted 13 Clinical Category Columns) |
|---|---|---|---|---|---|---|---|---|
| **8116089** | 2021 | Adulthood (41-65 yr) | female | atrial septal defect | Transcatheter Device Closure of Secundum Atria... | We present a case of female Bosnian patient 50... | 209 | [pulse: 83/min, respiratory_rate: 15 breaths/m... |
| **8464474** | 2021 | Adulthood (41-65 yr) | female | hip revision | Total hip revision with custom-made spacer and... | A 61-year-old woman presented to our orthopaed... | 440 | [hematological_conditions: raised erythrocyte ... |
| **8433115** | Un-known | Adulthood (41-65 yr) | female | cardiac haemangioma | Totally endoscopic resection of epicardial car... | We report on a case of an incidentally found t... | 217 | [pulse: 72 bpm, blood_pressure: 125/70 mmHg] |

Figure 7: Example Layout of CaseReportCollective. Only **Vitals_Hema** (Vitals and Hematology Findings) is shown, other omitted categories are **EENT** (Eyes, Ears, Nose, and Throat), **NEURO** (Neurology), **CVS** (Cardiovascular System), **RESP** (Respiratory System), **GI** (Gastrointestinal System), **GU** (Genitourinary System), **MSK** (Musculoskeletal System), **DERM** (Dermatology), **LYMPH** (Lymphatic System), **ENDO** (Endocrinology), **Pregnancy**, **Lab_Image** (Laboratory and Imaging), and **History**