CRAF: A Clinical Reasoning-Adaptive Framework via Reinforcement Learning for Similar Case Retrieval

Jie Lin[1*], **Lei Jiang**[1*], **Zongyi Chen**[2], **Liansheng Wang**[1,2†]

[1]Department of Computer Science at School of Informatics, Xiamen University
[2]National Institute for Data Science in Health and Medicine, Xiamen University
{jaylin, jianglei, zongyichen}@stu.xmu.edu.cn, lswang@xmu.edu.cn

Abstract

With the advancement of information retrieval (IR) technologies toward deep semantic understanding, reasoning-based methods—featuring explicit chain-of-thought generation—have demonstrated significant advantages in multi-hop and causal reasoning tasks. However, in complex clinical case retrieval scenarios, implicit reasoning cues within clinical data often hinder current models from effectively capturing deep semantic associations between queries and cases. Query rewriting and expansion techniques based on reasoning offer a promising solution to this challenge by uncovering and completing the latent clinical intent behind user queries, thereby enhancing semantic coverage and reasoning sensitivity. In this paper, we propose CRAF, a clinically adaptive reasoning framework tailored for similar case retrieval. Our method generates clinical reasoning paths and incorporates a fine-grained semantic reward mechanism, enabling efficient query rewriting through reinforcement learning. Experimental results on the PMC-Patients benchmark demonstrate that CRAF consistently delivers robust improvements across multiple retrieval tasks, achieving reasoning performance comparable to that of commercial models.

Code — https://github.com/jaylinio/CRAF
Datasets — https://arxiv.org/pdf/2202.13876
Extended version — https://github.com/jaylinio/CRAF

Introduction

For complex multi-hop retrieval queries where relevant documents exhibit complete lexical and semantic divergence from the query, traditional search systems (Zhu et al. 2023) relying on lexical overlap (Robertson and Zaragoza 2009) or shallow semantic similarity (Ma et al. 2024; Chen et al. 2024b; Devlin 2018; Liu 2019) prove inadequate. Accurate retrieval requires deep reasoning (Su et al. 2024) to both locate pertinent documents and bridge the gap between user intent and factual answers. The development of reasoning retrieval can be broadly categorized into two complementary approaches. The first involves training novel retrievers or rerankers using task-specific reasoning data, aiming to better accommodate complex reasoning requirements (Weller et al. 2025; Shao et al. 2025). The second leverages query reasoning and rewriting techniques (Jiang et al. 2025; Niu et al. 2024; Jagerman et al. 2023), harnessing the advanced reasoning capabilities of large language models (LLMs) (Guo et al. 2025; Face 2025). Through chain-of-thought prompting (Wei et al. 2022), these methods generate intermediate reasoning steps that serve as "post-reasoning queries" to replace the original queries during document retrieval. These two approaches are functionally complementary. Query reasoning techniques enhance the informativeness and expressiveness of the original queries via LLMs and chain-of-thought prompts, while retrievers designed specifically for reasoning tasks can further exploit these post-reasoning queries to better model reasoning semantics, thereby significantly improving retrieval performance.

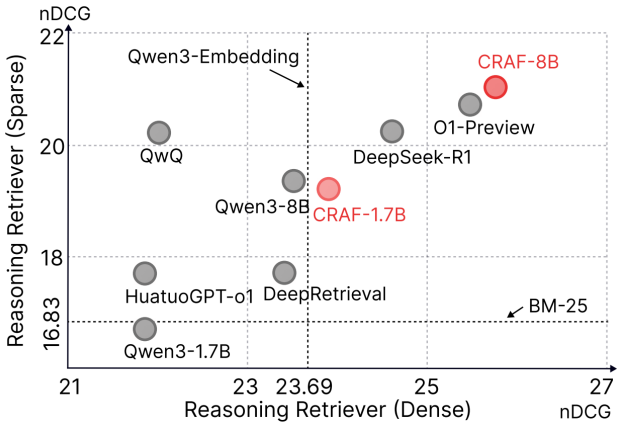

Figure 1: Clinical case retrieval via query reasoning performance: sparse vs. dense retrievers. See Table. 1 for details.

However, in the context of clinical text data, training retrievers (Wang et al. 2025) or rerankers (Shao et al. 2025) using task-specific reasoning data poses greater challenges compared to query reasoning and rewriting techniques, primarily in terms of data annotation and model training. On the one hand, clinical reasoning is often implicit, involving complex causal relationships and multi-hop logic that are difficult to capture through rule-based or weakly supervised

*These authors contributed equally.
†Corresponding author.

approaches. Constructing large-scale, high-quality reasoning annotations heavily relies on domain experts, making the process costly and limited in coverage. On the other hand, it is difficult to construct effective supervision signals for training reasoning-oriented retrievers. Relevance-based supervision fails to adequately capture the semantics of reasoning, often requiring the introduction of more complex model architectures to encode deep semantic paths, which can lead to issues such as training instability and poor generalization. Moreover, such models are not easily compatible with existing pretrained retrieval frameworks and typically require redesigning both model architecture and training objectives. In contrast, query reasoning and rewriting techniques leverage large language models (LLMs) to generate "post-reasoning queries," which enhance the informativeness of original queries without relying on costly annotated reasoning data, offering greater scalability and practical feasibility. In clinical case retrieval tasks, the core challenge for reasoning retrieval lies in accurately bridging clinicians' diagnostic exploration intent with fragmented yet critical medical evidence scattered across case reports. This retrieval paradigm requires not only comprehending surface-level symptom descriptions (Huang, Lin, and Demner-Fushman 2006; Schardt et al. 2007; Riva et al. 2012; Eriksen and Frandsen 2018), but more critically inferring the underlying clinical reasoning logic, then pinpointing reference-worthy analogous cases from vast repositories.

We propose CRAF, a query reasoning and rewriting pipeline specifically designed for reasoning retrieval tasks in clinical settings. The CRAF employs a structured reward function tailored to the Group Relative Policy Optimization (GRPO) framework(Shao et al. 2024), enabling reinforcement learning for query reasoning tasks. Additionally, we construct an automated data curation pipeline based on the publicly available PMC-Patients dataset (Zhao et al. 2023), which efficiently selects high-quality reasoning examples to train clinical reasoning-oriented query rewriting models. In benchmark evaluations, CRAF performance is comparable to that of commercial reasoning models such as o1-preview[1] and DeepSeek-R1 (Guo et al. 2025), while offering significantly lower inference costs. More importantly, when used in conjunction with sparse or dense retrievers (Robertson and Zaragoza 2009; Zhang et al. 2025), CRAF further improves overall performance, demonstrating strong adaptability and practical utility in retrieval pipelines. The main contributions of this work are summarized as follows:

- We propose the CRAF model family (including 1.7B and 8B versions), specifically designed for reasoning retrieval tasks. These medium-sized language models achieve strong performance on clinical similar cases retrieval, comparable to leading commercial models. Through co-working with various existing retrievers, CRAF significantly enhances overall retrieval performance, demonstrating the practicality and generalizability of query reasoning in real-world clinical scenarios.
- We construct an automated data curation pipeline specifically designed for training query rewriting models. It enables efficient allocation of positive and negative samples, providing more informative supervision data for model training, thereby significantly improving the model's generalization capability and training stability.
- We design a novel reward function that combines global semantic similarity with localized clinical entity structural relevance. This formulation preserves semantic alignment capabilities while incorporating explicit modeling of clinical knowledge structures, leading to improved learning efficiency and reasoning accuracy for language models.

[1]https://openai.com/index/introducing-openai-o1-preview/

Methods

Clinical query reasoning for similar case retrieval. For clinical similar case retrieval, given a case query q and a candidate case galleries $G = \{g_1, \ldots, g_n\}$, the goal is to identify and return a relevant subset $G^+ = \{g_1^+, \ldots, g_m^+\}$ (where $m \ll n$) using a retriever RT. In reasoning clinical retrieval, the relevance between q and G^+ is not solely determined by surface-level symptom descriptions or diagnostic labels, but rather through alignment with specific clinical reasoning pathways associated with each case—such as differential diagnosis processes, pathophysiological mechanisms, or treatment decision-making logic. These clinical reasoning pathways typically involve key steps such as parsing patient complaints and history, constructing a differential diagnosis framework, analyzing laboratory and imaging results, and deriving diagnostic conclusions based on clinical evidence. Such reasoning chains are often implicitly embedded within physicians' clinical thinking and cannot be directly extracted from case reports. This renders traditional retrieval methods that rely solely on structured codes or keyword matching inadequate. Following Eq. 1, we refer to the process of constructing such diagnostic reasoning pathways for a given case as *clinical query reasoning*.

$$G^+ = RT\left(LLM(\mathcal{P}, q)\right) \tag{1}$$

Where $\mathcal{P}$ denotes the prompt instructions for guiding the LLM in reasoning and query rewriting, while G^+ represents the set of cases retrieved by the retriever (RT) based on the LLM-rewritten query (i.e., $LLM(\mathcal{P}, q)$).

Automated data curation pipeline. During preprocessing of the original PMC-Patients dataset, each article was matched with standardized MeSH (Medical Subject Headings) system terms, including main headings and potential qualifiers. We also recorded whether each MeSH term represented a core focus of the article. All extraction results were aggregated by PubMed Central ID, establishing a mapping between PMC articles and MeSH terms. To accurately identify disease-related MeSH terms, we parsed the descriptor metadata released by the U.S. National Library of Medicine (NLM)[2]. This metadata includes key information such as the official names, definitions, and hierarchical structures of MeSH terms. According to the MeSH classification system, disease-related entries are assigned specific categories of tree numbers. Therefore, we defined terms with such tree

[2]https://meshb.nlm.nih.gov/treeView

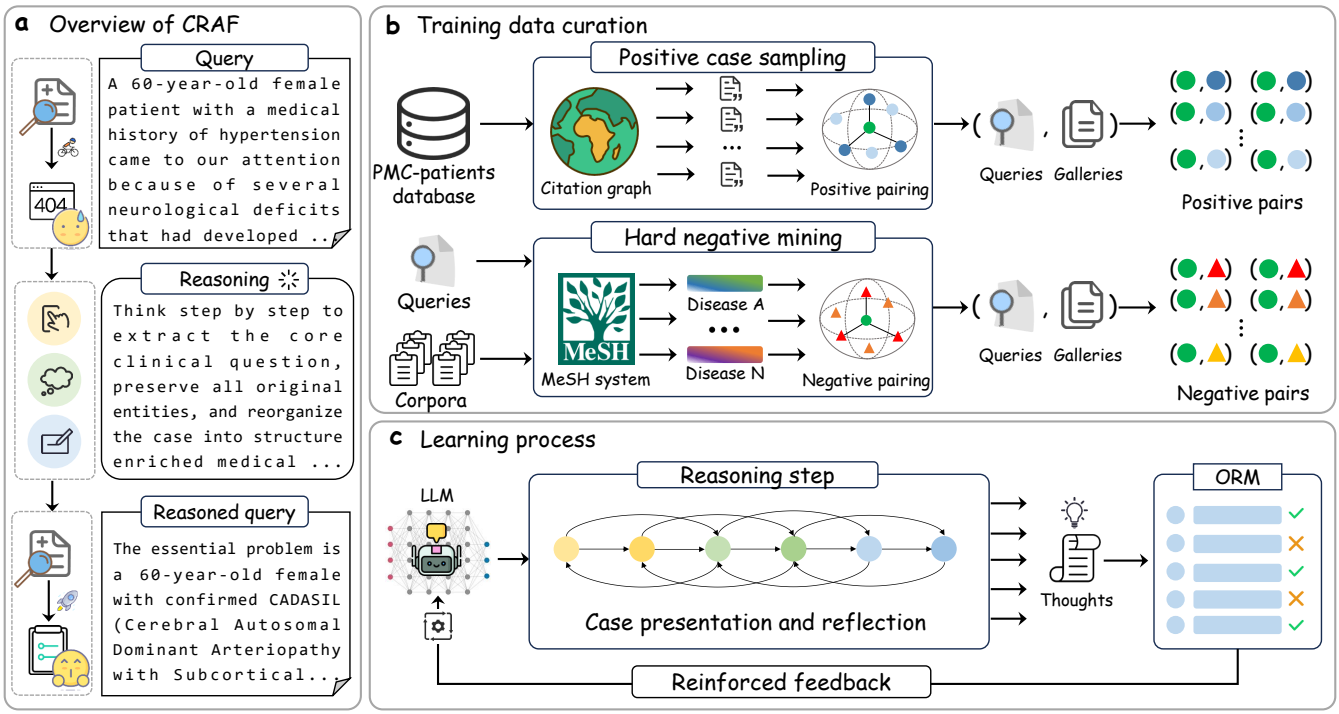

Figure 2: Overview of the proposed CRAF. During the data processing stage, we follow the PMC-Patients methodology, treating the original case report as the query and pairing it with its cited reference to form a positive sample. To construct negative samples, we mine disease-related cases with similar but distinct MeSH labels. In the training stage, we adopt GRPO-based reinforcement learning to align the model's reasoning process with clinical understanding.

numbers as disease terms and constructed a disease term vocabulary accordingly. Based on this vocabulary, we matched the MeSH terms associated with each publication mentioned earlier and extracted those that fall within the disease category. To further standardize the mapping between terms and medical classifications, we utilized the tree number in the MeSH metadata to extract the primary disease-related tree number corresponding to each disease term. This allowed us to construct a mapping from each term to a standardized medical disease classification code. This process effectively integrates the PMC-Patients literature with the official semantic structure of MeSH, establishing a bridge from natural language annotations to structured disease codes.

For query construction, the original PMC-Patients directly treats each patient summary as a "query patient" and automatically constructs positive samples based on the PubMed citation graph. Specifically, for a given query patient p, if there exists a citation relationship between another article a' and the case report $a(p)$ containing p ($a' \rightarrow a(p)$ or $a(p) \rightarrow a'$), or if a' is identical to $a(p)$, then a' is considered a positive sample for p. This citation-based strategy effectively captures literature that is semantically related to the patient case and provides a reliable foundation for downstream analysis. In the clinical retrieval domain, the complexity of medical knowledge and the wide variety of diseases pose significant challenges. Traditional training strategies relying solely on positive samples are insufficient for enabling models to learn subtle distinctions and deep relationships between diseases, resulting in poor generalization and inaccurate retrieval of relevant literature in complex cases. Diseases under the same MeSH tree node may share pathological mechanisms or treatment plans, yet exhibit critical differences. If these nuances are not well captured during training, models may suffer from false positives or negatives in real-world applications. To enhance the model's capability in handling hard cases, we introduce a dynamic allocation strategy for hard negative mining. Specifically, for any query patient p, we first locate the corresponding node $n(p)$ in the disease ontology tree. If p is associated with a common disease (occurrence count ≥ 50), we randomly sample N documents (with $N \leq 5$) as negative examples from different diseases under the direct child nodes $\{c_1, \ldots, c_k\}$ of $n(p)$. This operation leverages the hierarchical structure of the disease tree to select semantically related yet distinct diseases, aiming to construct more informative and challenging negative samples. In contrast, if p is associated with a rare disease (occurrence count < 50), we expand the sampling scope to include sibling diseases under the parent node $f(n(p))$, allowing negative examples to be sampled from a broader range. This cross-level sampling mitigates the issue of class imbalance caused by data sparsity and enhances the model's ability to distinguish rare disease cases.

Fine-grained semantic alignment reward function for reinforcement learning. Inspired by large-scale reasoning models (Guo et al. 2025), we adopt a GRPO reinforcement

Algorithm 1: Fine-grained Semantic Alignment Reward

Require:
q: Query, c: Completion, s: Solution $\{\text{obj}, G\}$
E_g, E_l: Embedders, α, β, γ: Weights
Ensure: R_{total}

$q' \leftarrow \text{RegexRemove}(c, \texttt{<think>...<\think>})$
$G \leftarrow s.\text{galleries}$
function SIM(q, G)
  $s_g, s_l \leftarrow 0$
  for $g \in G$ **do**
    $s_g + \cos\big(E_g(q), E_g(g)\big)$
    $P_q \leftarrow \text{ExtractParagraphs}(q)$
    $P_g \leftarrow \text{ExtractParagraphs}(g)$
    for $p \in \{1, \ldots, 5\}$ **do**
      if $P_q[p]$ and $P_g[p]$ exist **then**
        $s_l \mathrel{+}= \cos\big(E_l(P_q[p]), E_l(P_g[p])\big)$
  $\lambda \leftarrow \text{softmax}([\lambda_g, \lambda_l])$
  $s_{\text{all}} \leftarrow \lambda_0 s_g + \lambda_1 s_l$
  return $(s.\text{obj} = \text{pos})\ ?\ \alpha s_{\text{all}} : -\beta s_{\text{all}}$
$s^{before}_{contrast} \leftarrow \text{Sim}(q, G)$
$s^{after}_{contrast} \leftarrow \text{Sim}(q', G)$
$R_{\text{inc}} \leftarrow (s^{after}_{contrast} - s^{before}_{contrast})/|G|$
$\text{KL} \leftarrow D_{\text{KL}}\big(\pi_\theta \| \pi_g\big)$
$R_{\text{total}} \leftarrow R_{\text{inc}} - \gamma \cdot \text{KL}$
return R_{total}

learning algorithm to train large language models (LLMs) for clinical query reasoning. The core objective of query rewriting is to ensure that the rewritten query q' achieves a higher relevance score than the original query q. We propose a reward mechanism tailored for reasoning-based clinical retrieval, which guides the model to generate high-quality, structured, and semantically enriched rewrites. To this end, we design a reward function within the GRPO framework that leverages positive and negative sample guidance. Specifically, we train the model to maximize semantic alignment with relevant cases (positive samples) while minimizing similarity with irrelevant distractors (negative samples). Each training sample, as described in Section. , is represented as a triplet $\langle q, G, obj \rangle$, where G denotes either a set of positive cases G^+ or a set of negative cases G^-. The field $obj \in \{pos, neg\}$ indicates whether the training objective is to encourage the query to move closer to the positive cases or farther away from the negative cases.

The reward computation process begins by removing any internal deliberation traces embedded in the model's output, as such content may interfere with accurate similarity evaluation. We then compute the reward signal by integrating two complementary similarity metrics: global embedding similarity and local (section-wise) embedding similarity. For global similarity, we use the Qwen3-embedding model to encode the entire rewritten query and each gallery as a whole, followed by cosine similarity computation between the rewrite and each case. The summed similarities form the global similarity score, formally as Eq. 2.

$$s_{\text{global}}(q') = \sum_{g \in G} \cos\left(E_g(q'), E_g(g)\right) \tag{2}$$

For local similarity, inspired by (Riley et al. 2017), we aim to capture the fine-grained alignment of clinical reasoning structures. Specifically, we extract five standardized sections from each rewritten query and gallery: patient information, clinical presentation, diagnostic assessment, therapeutic intervention, and follow-up outcome. For each query, the model is prompted to summarize content under these five sections. For the candidate galleries G, these section summaries are pre-extracted using DeepSeek-R1. Each section is independently encoded using the domain-specific PubMedBERT model. Cosine similarity is then computed between corresponding sections of the rewritten query and galleries, and the section-wise similarities are summed to obtain the local similarity score, formally as Eq. 3.

$$s_{\text{local}}(q', G) = \sum_{g \in G} \sum_{s \in S} \cos\left(E_l^{(s)}(q'), E_l^{(s)}(g)\right) \tag{3}$$

The final semantic alignment score is the sum of the global and local similarity scores, formally as Eq. 4.

$$s(q', G) = \lambda_g s_{\text{global}}(q') + \lambda_l s_{\text{local}}(q', G) \tag{4}$$

where λ_g and λ_l denote the global and local weighting coefficients respectively, obtained by applying the softmax function to the trainable scalar parameters $\hat{\lambda}_g$ and $\hat{\lambda}_l$. Extracting and summarizing the five standardized sections serves to enhance the model's ability to structurally parse clinical narratives—isolating and recognizing core modules such as patient information and clinical findings while understanding the internal logic of clinical notes. Moreover, we employ differentiated handling strategies—real-time summarization for queries and offline extraction for candidates—to improve task adaptability and execution consistency. This ensures uniformity in both output format and content dimensions across dynamic queries and a static database. Additionally, this process deepens the model's domain-specific understanding and reasoning ability: as the five sections span the key stages of the clinical workflow, the model must grasp the domain-specific semantics of each section (e.g., distinguishing between diagnostic assessments and treatment interventions) in order to extract medically coherent content, thereby reinforcing its capacity to reproduce the clinical reasoning chain.

To reflect the reference supervision objective, we further introduce a contrastive reward term, which adapts to whether the model is guided to move closer to positive cases or away from negative ones. Given an objective, the contrastive reward is defined as Eq. 5.

$$s_{\text{contrast}}(q', G) = \begin{cases} \alpha\, s\big(q',\, G^+\big) & \text{if } obj = pos, \\ -\beta\, s\big(q',\, G^-\big) & \text{if } obj = neg, \end{cases} \tag{5}$$

where $s(q', G^+)$ denotes the alignment score between the rewritten query q' and the positive case set G^+, while

$s(q', G^-)$ measures the alignment with the negative case set G^-. The coefficients α and β are scalar hyperparameters that control the magnitude of reward under positive and negative objectives, respectively. We selected the optimal parameters $\alpha = 10$ and $\beta = 10$ in our experiments. This formulation provides directional guidance for learning—rewarding relevance when seeking positives, and penalizing similarity when avoiding negatives. This contrastive reward structure enables the model to simultaneously learn to approach relevant samples and avoid irrelevant ones. Empirically, it yields more semantically rich and retrieval-effective rewrites for clinical queries.

We define the per-sample incremental reward as the average increase in alignment to the galleries G when rewriting the original query q into q'. Formally as Eq. 6.

$$
\begin{aligned}
R_{\text{inc}}(q, q') &= \frac{s_{\text{contrast}}(q', G) - s_{\text{contrast}}(q, G)}{|G|} \\
R_{\text{total}}(q, q') &= R_{\text{inc}}(q, q') - \gamma \ \text{KL}\big(\pi_\theta(\cdot \mid q) \parallel \pi_{\text{g}}(\cdot \mid q)\big)
\end{aligned}
\tag{6}
$$

where $s(q, G)$ and $s(q', G)$ denote the similarity scores between the queries (q or q') and $|G|$ represents the number of cases. This reward quantifies the relevance improvement, encouraging the rewritten query q' to better match the positive set compared to the original query q. The KL penalty coefficient γ serves as a hyperparameter controlling the tolerance for deviation between the model and reference policy. The current policy's probability distribution $\pi_\theta(\cdot \mid q)$ represents the likelihood of generating rewrite q' given original query q, while the reference policy $\pi_{\text{g}}(\cdot \mid q)$ provides the baseline probability distribution on q. The Kullback-Leibler divergence $\text{KL}(\cdot \parallel \cdot)$ quantifies the difference between current and reference policies, computed as a token-level average. To stabilize training across heterogeneous query groups and mitigate issues arising from sparse or highly variable reward distributions, we normalize the reward signal via a group-wise advantage function. Specifically, the normalized advantage $\hat{A}(q, q')$ is computed as Eq. 7.

$$
\hat{A}(q, q') = \frac{R_{total}(q, q') - \mu_g}{\sigma_g + \delta} \tag{7}
$$

where $R(q, q')$ denotes the total reward received by the rewritten query q', which can incorporate both incremental and contrastive components. The terms μ_g and σ_g represent the mean and standard deviation of rewards within the same group g, respectively, and δ is a small constant added for numerical stability. Additionally, a group-specific weight coefficient w_g can be introduced to modulate the contribution of each group based on inter-group reward variance. This normalization strategy effectively enhances learning stability by aligning the reward scale across diverse query groups.

Experiments

Implementation details. We conducted training for both the 1.7B and 8B versions of CRAF using the SWIFT (Zhao et al. 2024) on a server equipped with 8 NVIDIA A800-80G GPUs, starting from the Qwen3-1.7B and Qwen3-8B pretrained models. To optimize GPU memory utilization, we implemented DeepSpeed's (Rasley et al. 2020) ZeRO-3 memory optimization strategy and enabled gradient checkpointing. We allocated 2 GPUs for vLLM inference services (Kwon et al. 2023) while dedicating the remaining 6 GPUs to model training. The complete training cycle required approximately 26 hours for the 1.7B model and 98 hours for the 8B model. For hyperparameter configuration, we set the per-GPU batch size to 16, base learning rate to $1e{-}6$, and KL divergence coefficient to 0.008. During inference, we generated 16 candidate responses per input prompt for advantage estimation in GRPO. To accommodate different embedding models' input length constraints, we established maximum generation lengths of 8192 tokens for Qwen3-Embedding and 512 tokens for PubMedBERT.

Main results. As shown in Table. 1, CRAF-8B, when paired with both a sparse retriever (BM25) and a dense retriever (Qwen3-Embedding), outperforms all baseline methods based on individual sparse or dense retrieval in terms of overall performance on two benchmark retrieval tasks. It also achieves a comparable level of performance to proprietary reasoning models such as o1-preview and DeepSeek-R1. Similarly, CRAF-1.7B demonstrates nearly equivalent performance, highlighting its efficiency in resource-constrained scenarios. By balancing reasoning capability and computational efficiency, CRAF emerges as a promising solution for query reasoning tasks. Given the superior performance of Qwen3-Embedding under reasoning queries, we further combine CRAF-generated reasoning queries with this retriever to explore their joint potential in clinical reasoning retrieval. Experimental results show that, when combined with Qwen3-Embedding, CRAF-1.7B surpasses Qwen3-8B, a model with five times more parameters, and CRAF-8B outperforms commercial models. Specifically, under Qwen3-Embedding, the average overall performance of CRAF-8B improves from 36.33 to 39.87, a significantly larger gain compared to the improvement under BM25 from 27.42 to 34.35. These findings indicate that our method exhibits strong adaptability and can effectively synergize with dense retrievers to further amplify performance advantages.

Qualitative Clinical Assessment. The t-SNE visualizations as show in Figure. 3, reveals that our rewritten clinical notes produce markedly more dispersed embeddings in the Qwen3-Embedding space than their raw counterparts. For both PAR and PPR retrieval, rewritten embeddings yield larger inter-cluster distances and reduced color overlap, whereas original embeddings form densely entangled clusters. At the macro level, our rewriting accentuates separation between disease systems. It preserves latent sub-phenotypes, as evidenced by discernible sub-cluster protrusions within individual color groups. This dual-scale fidelity indicates that the rewriting amplifies global discriminability without discarding clinically relevant nuances. Moreover, the resulting embedding manifold exhibits stronger topological stability—evidenced by lower variance across t-SNE hyperparameters—thus diminishing sensitivity to downstream retrieval settings and delivering more reliable representations for PAR and PPR tasks. This validates the efficacy of disease-guided training strategy.

Method	Patient-to-Article Retrieval (ReCDS-PAR)				Patient-to-Patient Retrieval (ReCDS-PPR)				AVG↑
	MRR↑	Prec↑	nDCG↑	Recall↑	MRR↑	Prec↑	nDCG↑	Recall↑	
Sparse Retriever									
BM25 (Robertson and Zaragoza 2009)	48.58	10.00	15.36	29.99	22.79	4.68	18.29	69.69	27.42
Dense Retriever									
ClinicalBERT (Alsentzer et al. 2019)	24.94	8.56	10.20	48.93	10.24	2.62	7.82	67.43	22.59
PubMedBERT (Gu et al. 2021)	42.96	16.08	19.51	**62.94**	19.37	5.05	16.30	79.35	32.70
BioLinkBERT (Yasunaga et al. 2022)	46.41	15.33	18.47	62.44	21.20	5.59	18.06	80.49	33.50
BGE (Chen et al. 2024c)	43.59	10.39	14.83	32.35	18.80	4.23	14.91	66.86	25.75
GTE (Li et al. 2023)	58.56	16.90	22.86	49.12	**28.02**	6.82	23.79	81.68	35.97
Qwen3-Embedding (Zhang et al. 2025)	**62.23**	**17.86**	**24.43**	48.24	27.05	6.51	22.95	81.35	**36.33**
ReasonIR (Shao et al. 2025)	59.19	15.59	21.80	46.46	28.01	**6.88**	**23.96**	**82.52**	35.55
Reasoning Rewritten with BM25 Retriever									
DeepSeek-R1 (Guo et al. 2025)	58.44↑9.86	14.40↑4.40	20.59↑5.23	45.26↑15.27	24.30↑1.51	5.34↑0.66	19.91↑1.62	77.14↑7.45	33.17↑5.75
o1-preview[1]	60.03↑11.45	15.32↑5.32	21.59↑6.23	**50.60**↑20.61	23.15↑0.36	5.75↑1.07	19.86↑1.57	77.56↑7.87	34.23↑6.81
HuatuoGPT-o1 (Chen et al. 2024a)	55.58↑7.00	13.40↑3.40	19.30↑3.94	42.47↑12.48	19.81↓-2.98	4.23↓-0.45	16.07↓-2.22	69.25↓-0.44	30.01↑2.59
QwQ (Team 2025b)	58.28↑9.70	14.22↑4.22	20.42↑5.06	45.13↑15.14	24.33↑1.54	5.39↑0.71	**20.03**↑1.74	77.02↑7.33	33.10↑5.68
DeepRetrieval (Jiang et al. 2025)	51.62↑3.04	10.90↑0.90	16.61↑1.25	33.05↑3.06	23.43↑0.64	4.82↑0.14	18.79↑0.50	71.53↑1.84	28.84↑1.42
Qwen3-1.7B (Team 2025)	52.61↑4.03	12.72↑2.72	18.22↑2.86	42.35↑12.36	18.65↓-4.14	4.12↓-0.56	15.14↓-3.15	70.90↑1.21	29.34↑1.92
Qwen3-8B (Team 2025)	56.69↑8.11	14.36↑4.36	20.28↑4.92	47.85↑17.86	22.37↓-0.42	5.02↑0.34	18.42↑0.13	74.55↑4.86	32.44↑5.02
CRAF-1.7B	55.04↑6.46	13.48↑3.48	19.28↑3.92	41.11↑11.12	23.37↑0.58	5.16↑0.48	19.14↑0.85	74.77↑5.08	31.42↑4.00
CRAF-8B	**60.92**↑12.34	**15.66**↑5.66	**22.09**↑6.73	47.86↑17.87	**24.94**↑2.15	**5.76**↑1.08	19.98↑1.69	**77.59**↑7.90	**34.35**↑6.93
Reasoning Rewritten with Qwen3-Embedding Retriever									
DeepSeek-R1 (Guo et al. 2025)	65.50↑3.27	19.28↑1.42	26.14↑1.71	51.44↑3.20	26.94↓-0.11	6.69↑0.18	23.12↑0.17	82.24↑0.89	37.67↑1.34
o1-preview[1]	66.84↑4.61	20.10↑2.24	27.14↑2.71	53.12↑4.88	23.42↑-3.63	6.90↑0.39	23.85↑0.90	83.34↑1.99	38.09↑1.76
HuatuoGPT-o1 (Chen et al. 2024a)	58.57↓-3.66	17.85↓-0.01	23.79↓-0.64	50.81↑2.57	23.40↓-3.65	5.80↓-0.71	19.96↓-2.99	80.69↓-0.66	35.11↓-1.22
QwQ (Team 2025b)	60.04↓-2.19	17.67↓-0.19	23.82↓-0.61	50.17↑1.93	23.48↓-3.57	6.04↓-0.47	20.25↓-2.70	80.16↓-1.19	35.20↓-1.13
DeepRetrieval (Jiang et al. 2025)	62.11↓-0.12	17.89↑0.03	24.46↑0.03	48.77↑0.53	26.22↓-0.83	6.43↓-0.08	22.39↓-0.56	81.24↓-0.11	36.19↓-0.14
Qwen3-1.7B (Team 2025)	59.32↓-2.91	16.64↓-1.22	22.83↓-1.60	47.26↓-0.98	24.74↓-2.31	6.06↓-0.45	20.93↓-2.02	80.45↓-0.90	34.78↓-1.55
Qwen3-8B (Team 2025)	63.60↑1.37	18.28↑0.42	24.97↑0.54	50.25↑2.01	25.88↓-1.17	6.39↓-0.12	22.09↓-0.86	81.63↑0.28	36.64↑0.31
CRAF-1.7B	64.50↑2.27	19.53↑1.67	26.22↑1.79	51.93↑3.69	27.45↑0.40	6.98↑0.47	23.63↑0.68	83.04↑1.69	37.91↑1.58
CRAF-8B	**68.34**↑6.11	**21.50**↑3.64	**28.65**↑4.22	**55.04**↑6.80	**28.33**↑1.28	**7.81**↑1.30	**24.91**↑1.96	**84.36**↑3.01	**39.87**↑3.54

Table 1: Performance comparison on PMC-Patients (Zhao et al. 2023). Best in bold, second-best underlined.

Ablation study of reward components. Ablation studies on reward components reveal distinct patterns in how PAR and PPR respond to component configurations. As shown in Figure 4-(a), PAR exhibits a three-stage progression—local optimization, global adjustment, and synergistic gain—across both sparse and dense retrievers. Its performance gains rely heavily on multi-component synergy and sample diversity, particularly benefiting from the higher capacity of dense retrievers. This indicates a need for holistic and system-level optimization strategies. In contrast, PPR demonstrates greater stability with low sensitivity to component variations, as its performance is more grounded in robust retrieval logic, with reward components serving only as auxiliary enhancements. From a system design perspective, dense retrievers provide a more favorable architecture for integrating complex reward mechanisms, while sparse retrievers require more careful tuning and compatibility alignment between components and retrieval logic.

Ablation study of sample components. To further analyze the impact of different sampling strategies on retrieval performance, we observe a significant pattern of adaptation between sampling mechanisms and metric characteristics. As shown in Figure 4-(b), PAR exhibits high sensitivity to variations in sampling strategies, whereas PPR maintains a relatively stable, gradual improvement trend. This phenomenon is consistent across different retriever architectures. From a mechanistic perspective, the performance gain of PAR is strongly correlated with sample diversity, underscoring the critical role of negative sampling in representation learning. Notably, in the sparse retriever setting, the synergy between positive and negative sampling components is particularly prominent, indicating that comprehensive coverage of the sample space is essential for architectures with lower parameter efficiency. In contrast, dense retrievers, with their superior semantic representation capabilities, are able to fully leverage the supervision signals provided by the sampling components, thereby demonstrating greater potential for performance improvement. Compared to PAR, PPR shows weaker dependence on sampling strategies, with its optimization process exhibiting a clear pattern of diminishing marginal returns. This suggests that PPR relies more heavily on the fundamental modeling capabilities of the retrieval system, with sampling strategies serving primarily as auxiliary optimization tools. This further validates above findings on metric characteristics: PAR metrics are more suitable as benchmarks for optimizing sampling strategies, while PPR metrics are better suited for evaluating the fundamental retrieval capabilities of the system.

Ablation study of training data scale. Through an analysis of training data scale, we observe distinct dependency patterns between PAR and PPR. PAR exhibits a clear three-phase performance trajectory—rapid gains (0–33%), gradual improvement (33–66%), and eventual saturation (66–100%)—highlighting its strong reliance on large-scale data for uncovering deep semantic associations. This nonlinear trend holds across both sparse and dense retrievers, with dense retrievers achieving higher baselines even in early stages. In contrast, PPR demonstrates stable performance

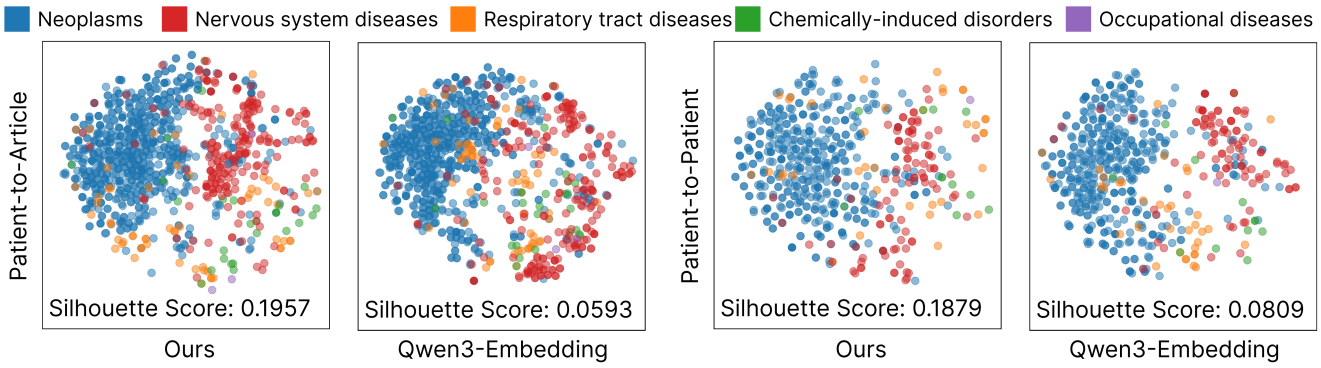

Figure 3: The t-SNE visualization comparing disease feature space of CRAF-reasoning queries and original queries.

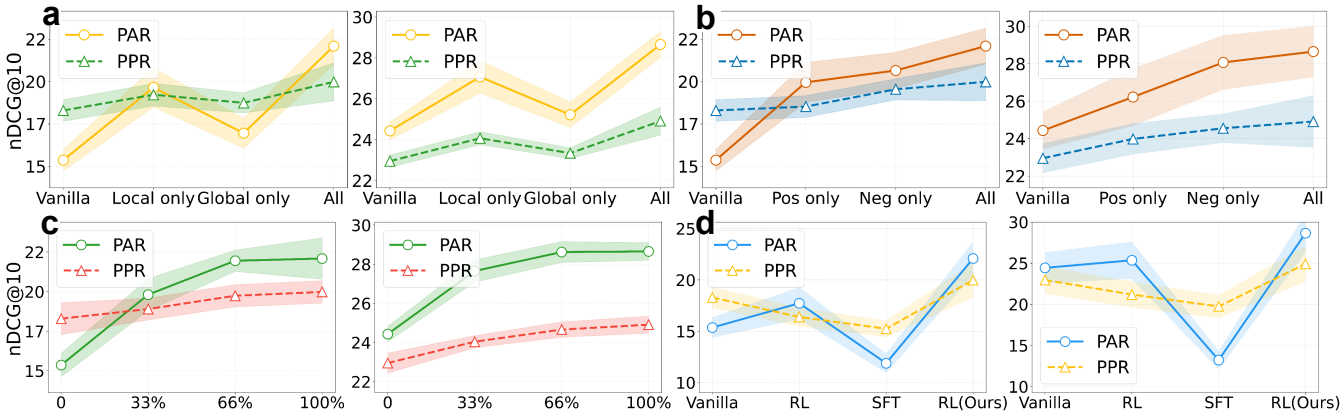

Figure 4: The ablation studies consist of four components: reward module ablation (top-left a), sample module ablation (top-right b), training data scale (bottom-left c), and training paradigm (bottom-right d). The analysis compares ReCDS-PAR and ReCDS-PPR metric variations under both sparse and dense retriever settings. All ablation studies were tested across multiple trained model checkpoints, with background data showing standard deviation (std).

even with minimal data and responds to increased data in a smooth, progressive manner. This robustness remains consistent across retriever types, indicating that PPR is structurally driven, relying more on model design than data volume. These differences underscore the fundamentally distinct optimization mechanisms behind the two metrics.

Ablation study of training paradigm. We investigate the performance differences between RL methods based on reward functions and traditional supervised SFT in the task of clinical case retrieval. As shown in Figure 4-(d), when trained with PAR, the conventional SFT approach suffers from a noticeable performance degradation. This degradation is primarily attributed to the higher semantic complexity of PAR—its reference cases tend to be more information-sparse, making token-level exact matching problematic and leading to catastrophic forgetting during training. In contrast, our proposed reinforcement learning approach alleviates this issue by incorporating a reward mechanism based on semantic embedding similarity. Rather than enforcing rigid token-level alignment, our method emphasizes maintaining global semantic consistency between the generated output and the reference text. This design proves particularly advantageous in clinical case retrieval, a scenario where semantic fidelity is critically important. Notably, compared to existing RL methods that typically rely on coarse-grained alignment strategies (Jiang et al. 2025), our approach introduces a more fine-grained reward formulation. This not only ensures semantic fidelity but also enhances the model's ability to generalize across diverse clinical expressions.

Conclusion

This paper presents CRAF, a clinical reasoning-adaptive framework designed to enhance similar case retrieval through query rewriting and reinforcement learning. By integrating fine-grained semantic alignment rewards and an automated data curation pipeline, CRAF effectively captures implicit clinical reasoning pathways and bridges the gap between diagnostic intent and relevant case evidence. Our findings highlight the critical role of reasoning-based query rewriting in clinical retrieval tasks and suggest that medium-sized models, when paired with RL optimization, can achieve strong performance with favorable efficiency.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005).

References

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024a. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. arXiv:2412.18925.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024b. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. ArXiv:2402.03216 [cs].

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024c. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eriksen, M. B.; and Frandsen, T. F. 2018. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *Journal of the Medical Library Association: JMLA*, 106(4): 420.

Face, H. 2025. Open R1: A fully open reproduction of DeepSeek-R1. https://github.com/huggingface/open-r1. Openr1.

Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; and Poon, H. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, X.; Lin, J.; and Demner-Fushman, D. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, 359.

Jagerman, R.; Zhuang, H.; Qin, Z.; Wang, X.; and Bendersky, M. 2023. Query Expansion by Prompting Large Language Models. ArXiv:2305.03653 [cs].

Jiang, P.; Lin, J.; Cao, L.; Tian, R.; Kang, S.; Wang, Z.; Sun, J.; and Han, J. 2025. DeepRetrieval: Hacking Real Search Engines and Retrievers with Large Language Models via Reinforcement Learning. arXiv:2503.00223.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180.

Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; and Zhang, M. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Ma, X.; Wang, L.; Yang, N.; Wei, F.; and Lin, J. 2024. Fine-Tuning LLaMA for Multi-Stage Text Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, 2421–2425. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0431-4.

Niu, T.; Joty, S.; Liu, Y.; Xiong, C.; Zhou, Y.; and Yavuz, S. 2024. JudgeRank: Leveraging Large Language Models for Reasoning-Intensive Reranking. arXiv:2411.00142.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, 3505–3506. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379984.

Riley, D. S.; Barber, M. S.; Kienle, G. S.; Aronson, J. K.; von Schoen-Angerer, T.; Tugwell, P.; Kiene, H.; Helfand, M.; Altman, D. G.; Sox, H.; et al. 2017. CARE guidelines for case reports: explanation and elaboration document. *Journal of clinical epidemiology*, 89: 218–235.

Riva, J. J.; Malik, K. M.; Burnie, S. J.; Endicott, A. R.; and Busse, J. W. 2012. What is your research question? An introduction to the PICOT format for clinicians. *The Journal of the Canadian Chiropractic Association*, 56(3): 167.

Robertson, S. E.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3: 333–389.

Schardt, C.; Adams, M. B.; Owens, T.; Keitz, S.; and Fontelo, P. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC medical informatics and decision making*, 7: 1–6.

Shao, R.; Qiao, R.; Kishore, V.; Muennighoff, N.; Lin, X. V.; Rus, D.; Low, B. K. H.; Min, S.; tau Yih, W.; Koh, P. W.; and Zettlemoyer, L. 2025. ReasonIR: Training Retrievers for Reasoning Tasks. arXiv:2504.20595.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Su, H.; Yen, H.; Xia, M.; Shi, W.; Muennighoff, N.; Wang, H.-y.; Liu, H.; Shi, Q.; Siegel, Z. S.; Tang, M.; et al. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Team, Q. 2025b. QwQ-32B: Embracing the Power of Reinforcement Learning. https://qwenlm.github.io/blog/qwq-32b/. QwQ-32B.

Wang, Z.; Cao, L.; Jin, Q.; Chan, J.; Wan, N.; Afzali, B.; Cho, H.-J.; Choi, C.-I.; Emamverdi, M.; Gill, M. K.; Kim, S.-H.; Li, Y.; Liu, Y.; Ong, H.; Rousseau, J.; Sheikh, I.; Wei, J. J.; Xu, Z.; Zallek, C. M.; Kim, K.; Peng, Y.; Lu, Z.; and Sun, J. 2025. A foundation model for human-AI collaboration in medical literature mining. arXiv:2501.16255.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Weller, O.; Ricci, K.; Yang, E.; Yates, A.; Lawrie, D.; and Durme, B. V. 2025. Rank1: Test-Time Compute for Reranking in Information Retrieval. arXiv:2502.18418.

Yasunaga et al. 2022. LinkBERT: Pretraining Language Models with Document Links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8003–8016.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; Huang, F.; and Zhou, J. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zhao, Z.; Jin, Q.; Chen, F.; Peng, T.; and Yu, S. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10(1): 909.

Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Chen, H.; Liu, Z.; Dou, Z.; and Wen, J.-R. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.