Mastering ACR Medical Appropriateness Criteria: Aligning Language Agents with Clinical Logic via Group Relative Policy Optimization

Anni Tziakouri

Master of Science

Data Science

School of Informatics

University of Edinburgh

2025

Abstract

The number of unnecessary imaging procedures is increasing, harming patients and straining healthcare systems. Although the ACR Criteria offer evidence-based guidance on selecting appropriate imaging, they remain underutilized in clinical workflows. With the growing capabilities of LLM-based reasoning, there is now an opportunity to bridge this gap by enabling more trustworthy and transparent imaging referrals. This study introduces an LLM-based Reasoning Agent trained via Reinforcement Learning (RL), specifically Group Relative Policy Optimization (GRPO), to replicate expert clinical reasoning and recommend appropriate imaging; marking the first application of RL leveraging structured reasoning from the ACR Criteria. It is also the first to systematically compare reasoning-focused reward functions and evidence integration strategies in medicine, placing reasoning quality at the core, to build clinician trust and enable real-world deployment. Our best lightweight RL model, MedReason-Embed, outperforms the baseline by 18% in macro F1, achieves significantly higher reasoning capabilities, and surpasses both larger models and those trained with alternative strategies, showing that reasoning-aligned supervision enables efficient, trustworthy clinical AI. To that end, we also develop a modular end-to-end agentic system that replicates the full imaging referral process, incorporating PubMed-based evidence retrieval to generate well-justified recommendations. The system aims to generalize beyond static guidelines and operates fully autonomously, with potential for continuous updates. This work highlights the promise of reasoning-focused RL within full-system architectures to enable autonomous, trustworthy, and explainable clinical decision-making in radiology.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Anni Tziakouri*)

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my family for their unwavering support, encouragement, and belief in me. Your love and constant presence throughout this journey have meant everything. The sacrifices you have made and your faith in my potential have shaped not only this Dissertation, but the person I am today. Your support helped me discover and pursue what I truly love, and inspired me to strive for excellence in everything I do. I truly owe everything to you.

I am also sincerely grateful to everyone who contributed to this Dissertation, especially my supervisor, Professor Filippo Menolascina, for his truly generous time, continuous guidance, and the many long conversations that helped shape the direction of this work. His insightful feedback, thoughtful suggestions, and support were invaluable throughout every stage of this project. I would also like to thank the professors at the University of Edinburgh who supported me along the way, as well as our collaborator from Oxford University, Junde Wu, for his contribution and insights.

Table of Contents

Chapter 1

Introduction

1.1   Background and Problem Statement

Despite recent advances in medicine, there has been a substantial increase in low-value medical imaging; namely imaging procedures where potential risks outweigh expected diagnostic benefits for patients. Unnecessary use of imaging has risen sharply, with studies estimating that 20–50% of CT scans in the U.S. may not be clinically justified [34, 114, 118, 132]. Other studies similarly report 35-80% of imaging procedures falling outside established clinical standards, varying considerably across settings [7, 31, 49, 70, 83, 99, 115, 118]. Unnecessary imaging harms patients, clinicians, and healthcare systems. These procedures expose patients to harmful ionizing radiation linked to increased cancer risk [88, 118] and can cause overdiagnosis, patient anxiety, delayed critical diagnoses with life-threatening consequences, and overall reduced quality of care [7, 48, 83]. At the healthcare system level, inappropriate imaging drives up costs and strains resources, with estimates indicating that reducing unnecessary procedures could save thousands to millions annually [63, 83, 115].

To address inappropriate imaging, the American College of Radiology (ACR) developed the ACR Appropriateness Criteria [94], a set of evidence-based guidelines designed to assist clinicians in selecting suitable imaging for specific medical scenarios [21, 65]. For instance, for women under 30 experiencing breast pain, the ACR recommends ultrasound to avoid radiation, while mammography is preferred for women over 40 due to higher cancer risk. The ACR Criteria are the product of a rigorous, expert-driven process. Firstly, panels of human clinicians review medical literature to identify high-quality evidence for specific medical conditions. Then, they synthesize the evidence to determine whether an imaging procedure is appropriate while also

explaining the reasoning behind each recommendation [65]. The guidelines are based on the RAND/UCLA method, which defines appropriate care as that in which "benefits outweigh harms" [38], balancing diagnostic value, patient risk, and resource use. As such, they serve as expert proxies for imaging justification [65].

1.2   AI in Medical Imaging Referrals

Although guidelines are essential for reducing unnecessary imaging [7, 100, 118], their adoption remains limited with fewer than 1% of clinicians using them as their primary reference for referrals [14] and as also reflected by the ongoing prevalence of inappropriate imaging [100]. Many studies suggest AI-based models could improve imaging referrals and support guideline adherence [100, 122, 139], but adoption remains low due to mistrust from poor transparency, hallucinations, and regulatory issues. Reliable medical AI requires clear, stepwise explanations to build trust, ensure accuracy, and comply with regulations like the EU AI Act [130, 137]. However, most models still lack sufficient explainability, limiting their adoption in high-stakes care [96, 104, 134]. One form of explainability is the model's reasoning output, allowing clinicians to assess rationales for imaging referrals and build trust by revealing step-by-step logic [107].

Recently, there has been a growing shift in LLM development toward reasoning models that move beyond fast, heuristic "System 1" thinking to more analytical, step-by-step "System 2" reasoning [75], a crucial trend for medicine that requires robust reasoning. However, recent findings have revealed shortcomings of Supervised Fine-Tuning (SFT), the main method for adapting LLMs to medical tasks [92], including weak reasoning, overfitting, and shortcut learning [92, 76, 26, 96]. This has driven interest in Reinforcement Learning (RL), which instead uses reward signals to train models and has been shown to improve reasoning, scalability, and generalization in LLMs [96, 26]. Notably, Group Relative Policy Optimization (GRPO) [112], the RL method used in DeepSeek-R1 [45], replaces computationally expensive neural reward models with rule-based comparisons across a group of responses; making RL practical for large-scale, reasoning-focused LLM training under limited resources. RL has shown strong results in math and coding tasks [56, 124], with growing promise in medical domains [68, 96, 119]; however, its use and exploration in medicine remain limited.

While RL has improved reasoning tasks, its success depends heavily on the reward design [77]. Most RL-based approaches for adapting models to medical tasks use a dual-reward setup: one reward to ensure the output follows the required format, and

another to reward correctness of the final answer [68, 96], but these often fail to capture the full complexity of clinical reasoning [96, 148]. For example, a model might correctly reject a CT scan for mild headaches, not by knowing it offers no clinical benefit, but by spotting superficial cues like the word "mild". Such shortcut strategies, reinforced by answer-only rewards, can yield high answer accuracy but shallow reasoning that lacks clinical validity [32, 76, 96]. Indeed, models trained with rewards that evaluate the intermediate reasoning steps, have been shown to outperform those trained solely on answer rewards [27, 57, 77, 133].

A key research gap in clinical Natural Language Processing is the lack of reward models that encourage valid reasoning, since most current models prioritize final accuracy over aligning intermediate reasoning with clinical standards [32]. In this context, the ACR criteria offer *tremendous value* as a fully supervised, evidence-based resource, providing high-quality reasoning and recommendations that form a solid foundation for developing agentic LLM systems to improve imaging referrals and clinical decision-making.

1.3   Objectives and Research Questions

With this work, we aim to build an innovative and deployment-ready agentic architecture for imaging referrals. At its core is a GRPO-trained LLM agent that determines, given a specific medical scenario (i.e. patient's conditions/symptoms), whether a given imaging procedure is appropriate while generating high-quality reasoning, which is crucial for making this system trustworthy once deployed. We develop specialized *reasoning* reward functions and use techniques like context integration to explicitly align the model's reasoning with expert clinical workflows. We then compare models based on performance, efficiency, and reasoning alignment, highlighting key trade-offs. Notably, this study is the first to apply RL while leveraging expert-validated reasoning from the ACR Criteria. It also is the first study to systematically compare custom clinical reasoning-focused rewards and evidence integration strategies in medicine, to improve justification strength beyond simple answer matching. RL research in medicine is still in its early stages [101, 96], with most studies emerging only recently in the last year or so, highlighting clinical reasoning as a novel and rapidly evolving application area for LLMs.

A crucial part of this research is developing a comprehensive end-to-end system that replicates the entire ACR process, not just the reasoning component. It integrates

agents for evidence retrieval and quality assessment from medical databases, which the reasoning agent synthesizes into informed imaging recommendations. Prior work shows grounding outputs in evidence improves accuracy and clinician trust [134]. Therefore another key innovation is the automatic identification, evaluation, and incorporation of high-quality medical literature as external knowledge, surpassing conventional LLM pretraining and enabling robust generalization beyond fixed ACR guidelines. This empowers physicians to confidently recommend imaging even in scenarios lacking established guidelines, a major practical advance. Additionally, the system supports continual learning and autonomous updates as new literature emerges, positioning it at the forefront of adaptive, trustworthy AI for clinical decision support.

Overall, the agentic system takes as input a clinical note describing a diagnosis and a proposed imaging procedure, ultimately producing a recommendation (e.g., "procedure X is appropriate") supported by detailed reasoning. To accurately interpret the clinical note and facilitate deployment, the system uses the widely adopted ICD coding system [39] employed by most hospital systems, enabling better alignment with ACR criteria. When the diagnosis falls outside predefined ACR criteria, the system retrieves and filters high-quality evidence from medical databases, which the reasoning agent synthesizes to generate a well-justified imaging recommendation. The system aims to serve as an expert "second opinion" to support clinicians, reduce low-value imaging, and improve equitable healthcare by ensuring only patients with true need receive procedures, promoting better care and fairness.

As outlined, the main research questions are:

- To what extent can GRPO-adapted reasoning language models accurately recommend clinically appropriate imaging procedures, while providing transparent justifications in alignment with established diagnostic guidelines, compared to existing approaches?
- How do (a) the integration of high-quality literature evidence and (b) reasoning-centric reward functions influence model performance, reasoning quality, and real-world deployability?
- To what extent can we replicate the ACR process and how well can it be extended and generalized to other conditions not covered by the ACR while maintaining robustness and clinical reliability?

The report is structured as follows: Chapter 2 reviews relevant literature and background, Chapter 3 details the methodology, and Chapters 4 and 5 present the results and conclusions respectively. The GitHub repository is available here.

Chapter 2

Background

2.1 ACR Criteria and their role in Imaging Referrals

The American College of Radiology (ACR) developed the ACR Criteria as evidence-based guidelines to help clinicians select appropriate imaging for medical conditions, with the key goals of promoting transparency and standardization [21]. The development of the ACR Criteria involves several key steps: selecting broad clinical conditions (e.g., Breast Pain), and for each condition defining multiple specific variants incorporating details such as patient demographics (e.g., age groups); conducting systematic literature searches in medical databases; assessing the quality of retrieved evidence using the widely-used GRADE scale [110]; and synthesizing the filtered high-quality evidence into recommendations that classify imaging procedures for each variant as "Usually Appropriate", "May Be Appropriate", or "Usually Not Appropriate", while also providing detailed reasoning. These criteria are regularly updated based on new evidence [65]. However, they are not exhaustive, with notable gaps such as pediatric coverage [21, 83], limiting universal applicability. As of July 2025, the ACR Criteria cover 257 different conditions [94], which can be accessed via the ACR Criteria website.

| Select clinical condition and specific variants | Conduct systematic literature searches | Assess evidence quality using GRADE scale. | Synthesize evidence into justified recommendations |
| (e.g. Breast Pain and women under 30) | Retrieve papers by searching medical database | Assign strength of evidence to papers | Procedure X "May be Appropriate" for this variant because.. |

Figure 2.1: The ACR Criteria creation process

Multiple studies have explored automating ACR-based imaging referrals with LLMs, but these remain limited in scope, show variable performance, and rarely evaluate the

transparency of the model's reasoning, a critical requirement for deployment. General-purpose LLMs can reach up to 88% accuracy in well-defined, guideline-covered scenarios such as breast screening, but performance may drop by up to 30% in more ambiguous domains [102], limiting generalizability [13, 91, 143]. Also, nearly all prior work focuses on end-point agreement with ACR recommendations, without assessing the clinical validity of the underlying reasoning. A separate line of research uses retriever augmented systems that directly retrieve answers from the ACR Criteria, achieving referral accuracies up to 83% and in some cases surpassing radiologists [103, 122, 139]. However, these systems primarily optimize for retrieving the correct ACR document, reducing the task to locating an existing recommendation, and thus they work only for guideline-covered cases. By contrast, the approach proposed in this work is designed to handle both guideline-covered and non-covered scenarios. It explicitly trains an LLM on the ACR Appropriateness Criteria using RL with a dedicated focus on reasoning. This enables the model not only to produce clinically aligned, evidence-based recommendations when guidelines exist, but also to generalize reasoning patterns to cases where the ACR provides no direct answer.

2.2 ICD coding

To align with best practices in clinical workflows, we use the International Classification of Diseases (ICD) [39] to unequivocally identify patient conditions. Maintained by the World Health Organization (WHO) [135], ICD is the global standard for clinical documentation, billing, research, and healthcare communication [9, 33]. Versions like ICD-9 organize thousands of conditions hierarchically with standardized codes and descriptions (e.g., code 611.71 for "Mastodynia" under the broader 611 category for "Disorders of breast"). Our system maps clinical diagnostic notes to ICD codes as an *intermediate layer* before linking them to ACR guidelines. This not only simplifies integration into hospital systems but also offers the added benefit of standardizing patient conditions using a single, internationally validated controlled vocabulary.

ICD coding, the process of assigning standardized ICD codes to clinical documentation, is a well-studied but challenging task; firstly due to the imbalanced label space; for example, in the benchmark dataset of patient records MIMIC-III [60], 10% of codes account for 85% of assignments, while 22% of codes appear twice or less [146]. Another major challenge is the wide variation in clinical documentation length, format, and style, which often includes ambiguous language, synonyms, abbreviations, and

clinicians' personal writing styles, making automated ICD coding difficult for both humans and AI systems [33]. Manual ICD coding is labor-intensive and error-prone, with human accuracy averaging 83% in the UK but varying widely (50–98%) [19], further highlighting the need for automated, consistent and standardized solutions [33]. ICD coding initially relied on rule-based methods [36] but has evolved to use embedding models [40, 87], neural networks like LSTMs [11] and pretrained language models [52], often incorporating code hierarchies or synonyms to improve accuracy [25, 142].

LLMs have recently gained interest for automating ICD coding due to their advanced language understanding [82, 141]. While general LLMs like GPT-4 and LLaMA2 show improved understanding, they often hallucinate details and struggle with rare or complex cases, achieving around 46% ICD match rates, comparable to human coders on difficult cases [89, 116]. Some studies report better results, but their generalizability remains limited [2]. To address these issues, Retrieval Augmented Generation (RAG) uses ICD code descriptions to reduce hallucinations and improve matching accuracy [64]. Combining RAG with LLM-based reranking [12, 66] can further boost performance; for example, MedCoder [12] reached an F1 of 0.60 on a synthetic dataset.

However, performance varies across datasets due to clinical settings and annotation differences, highlighting the need for validation and human oversight in LLM-based ICD coding [2, 117, 141]. Reported accuracies are modest and inconsistent, highlighting the need to tailor models to specific use cases rather than rely on benchmarks [33].

2.3 Reasoning

2.3.1 Reasoning Models

According to [75], the LLM field is shifting from fast heuristic "System 1" thinking to analytical "System 2" reasoning. While older models excel at quick decisions, newer ones like OpenAI's o1 series [56], DeepSeek's R1 [45], and Gemini 2.5 [41] employ step-by-step reasoning for expert-level performance on complex tasks. This shift is crucial in medicine, where transparent, robust reasoning underpins safe clinical use.

Reinforcement learning (RL) has recently emerged as a powerful method for improving LLM reasoning. Supervised Fine-Tuning (SFT), previously the dominant approach for injecting domain knowledge in the post-training phase [92, 137], only encourages reasoning implicitly, often requires large amounts of data, and suffers from critical limitations such as overfitting and shortcut learning, where models rely on superficial

patterns rather than genuine reasoning [26, 45, 96, 112, 120]. These consequences are especially problematic in medicine, where adaptability to new data is crucial [68, 119]. Unlike SFT, which learns from fixed input-output pairs, RL updates model behavior using reward signals that reflect task-specific performance, offering greater flexibility for optimizing complex objectives like multi-step reasoning quality.

This has led to growing interest in RL-based post-training paradigms that further enhance reasoning performance. Recent examples include GRPO and its variants, such as Dr. GRPO [79] and Group Sequence Policy Optimization (GSPO), just introduced in July 2025 [144], along with other methods focused on improving reasoning and sample efficiency. In this work, we adopt GRPO due to its demonstrated effectiveness on structured reasoning tasks and easier implementation.

RL methods, such as GRPO, improve transparent reasoning, showing clear benefits in medical tasks, even with limited data. MedVLM-R1, a lightweight 2B parameter model trained with GRPO on just 600 radiology Visual Question Answering (VQA) samples from public medical VQA benchmarks such as PathVQA [47], achieved a 20% accuracy boost compared to the baseline, reaching 78.22% on the evaluation set and also demonstrated strong out-of-distribution generalization [96]. Similarly, Med-R1, trained with RL and evaluated on the OmniMedVQA benchmark [51], achieved nearly a 30% accuracy improvement over its base model (Qwen2-VL-2B) and even outperformed the much larger Qwen2-VL-72B model [68]. Also, multiple studies report that RL-adapted models consistently outperform their SFT counterparts, particularly in generalization to unfamiliar tasks, some of them noting performance gains of 16–35% [68, 96, 113, 119], with a study summarizing this as "SFT memorizes, RL generalizes" [26].

Most RL-adapted models in medicine currently use two main reward functions: one to ensure the final answer is correct, and another to encourage proper formatting, helping the model produce clear and well-organized reasoning. However, several studies have argued that these generic reward functions fail to produce models with genuine clinical reasoning capabilities, as optimizing for the correct answer alone does not guarantee valid or robust reasoning in medical tasks [24, 27, 37, 69, 93]. Indeed, in the MedVLM-R1 study that used the standard dual reward function [96], the authors found that while the final answer was sometimes correct, the LLM's underlying reasoning and approach to reaching that answer were often flawed. To address this, the authors of [148] introduced a binary semantic reward using an LLM to evaluate whether generated reasoning was supported by clinical evidence. This approach not only improved final accuracy by 2%, but also enhanced overall reasoning quality. Therefore, given the

importance of reward design in guiding RL behavior and inspired by the authors of [148], our study not only evaluates standard reward functions, but also proposes novel ones that leverage ACR-annotated justifications to promote genuine clinical reasoning.

2.3.2 Reward design

Well-designed rewards are essential in RL training since they guide learning. Poorly specified ones can lead to reward hacking, where models exploit shortcuts to maximize rewards without achieving true reasoning abilities [10, 77]. Thus, robust reward functions aligned with expert goals are essential for effective RL applications in medicine. Many studies argue that clinical reasoning should be the primary goal for LLMs in medicine [18, 98, 140] and yet, most works still prioritize simple answer accuracy as a reward metric [24, 27, 37, 69, 96, 149].

This contrast can also be framed in terms of the supervision strategy of the reward model, that guides the LLM reasoning. A key design choice is between Outcome-supervised Reward Models (ORMs), which use answer accuracy, and Process-supervised Reward Models (PRMs), which evaluate intermediate steps with harder-to-obtain annotations but offer greater interpretability [27, 57, 77, 133]. For example, PRMs may reward a model for correctly identifying relevant clinical risk factors before reaching a diagnostic conclusion, which an ORM would reward directly. The reward signal in PRMs allows the model to understand the reasoning *process* rather than *just the final answer*. The foundational study by [77] showed that PRMs achieved higher accuracy and better reasoning quality than ORMs on complex tasks. PRMs are especially valuable in multi-step reasoning, detecting intermediate errors, providing richer feedback, promoting interpretable, aligned behavior, and reducing issues like reward hacking [29, 128]. Despite their benefits, PRMs remain underutilized.

We hypothesize that, given PRMs' advantages over ORMs in evaluating intermediate reasoning steps [81, 148], incorporating reasoning rewards will enhance performance and support real-world deployment. Although progress has been made, aligning reasoning remains a challenge, as there is no consensus among clinicians on what constitutes "good reasoning" or which aspects are most meaningful to measure [71, 129]. Additionally, there is still no unified framework on the most *scalable* automatic method for evaluating reasoning [46], and most studies actually rely on human annotators, which is not scalable [24, 77, 128, 133].

2.3.3 Evaluating clinical reasoning

Recent literature seeks automated, scalable metrics for evaluating reasoning chains, meaning whether the step-by-step thought process is sound and aligns with gold-standard reasoning. Common surface-level metrics based on word or n-gram overlap such as BLEU [97], often fail to capture deeper semantic nuances and yield inconsistent results across tasks [24, 108]. Semantic approaches like verifier training offer deeper evaluation [74] but are costly due to fine-grained annotation requirements [77].

An increasingly common approach is to use *LLMs-as-critics*, prompted or fine-tuned to directly assess reasoning quality, with studies showing that effective prompt design is beneficial [44]. Two notable models, LlamaV-o1 [126] and PathVLM-R1 [136], use this approach by evaluating the generated reasoning across multiple criteria, such as medical knowledge rationality, faithfulness, and semantic coverage, to provide fine-grained, reproducible assessments, leading to improved reasoning abilities. Defining exact evaluation criteria for LLM-based assessment is helpful, as it enables detailed evaluation and shows strong alignment with human judgments [72, 120, 126, 147]; however, using LLMs-as-critics remains computationally intensive.

To conclude this section on reasoning, we highlight the need to distill clinical reasoning for agents handling medical referrals. Building on the advantages of PRMs over ORMs [81, 148], the goal of better-aligned reasoning [18, 98, 140], and the clinician trust it fosters [129, 137], this study explores multiple reasoning reward functions comparing generated reasoning with gold-standard reasoning to "push" the model to "think" like a clinical expert. Inspired by Deepseek-R1 [45], we stress that reward and evaluator design must fit the dataset, domain, and goals, balancing computational cost and performance [71]. Admittedly, evaluating reasoning alignment is complex, but our focus is on comparing several reward functions, both existing and custom, built around key reasoning elements in our dataset. While not exhaustive, this represents a principled, dataset-driven approach.

2.4 Medical retrieval of high quality evidence

Clinical decision support requires both expert-aligned reasoning and grounding in high-quality evidence [134]. To this end, our approach combines reasoning rewards with medical literature retrieval, aiming to improve performance and trust. By supplying

LLMs with domain-specific context, we can reduce hallucinations and enhance factual accuracy, even without prior in-domain training [122]. For example, RetCare [134] uses RAG to ground outputs in clinical evidence from PubMed's 38 million biomedical citations [1], enhancing accuracy and reliability. Clinicians report that literature-based explanations greatly increase trust in model recommendations [134], which is crucial for deployment and therefore central to our research. While RetCare's keyword-based retrieval could be improved with advanced query expansion and filtering, it effectively boosts trust through evidence-based explanations.

Developing effective search strategies for evidence retrieval is essential yet challenging due to rapid growth in medical publications [80]. Researchers have enhanced PubMed's default search with methods like automatic query expansion, Boolean logic, publication filters and semantic relevance checks [80, 85, 105].

A recent promising approach, DeepRetrieval [58], trained an LLM with RL to rewrite queries using Boolean logic, without supervised labels. For example, it rewrites the query "best diagnostic imaging for breast pain" into a Boolean string like "((Breast Pain OR Mastalgia) AND (Imaging OR Mammography OR MRI OR CT))", expanding with synonyms and related terms to improve search. By exploring query variations and optimizing retrieval metrics like NDCG, their LLM outperforms SFT methods that just mimic reference queries, achieving 65% recall versus the previous SOTA of 25%. In a follow-up, the authors introduced S3 [59], an efficient framework that separates search from generation and optimizes retrieval based on downstream utility, using a novel reward function, and is the current SOTA for evidence retrieval.

An effective search strategy alone is not enough; efficiently identifying high-quality evidence from the retrieved literature remains challenging, expert-dependent, and time-consuming, highlighting the need for automation in Evidence-Based Medicine [3, 22]. Machine learning (ML) approaches have been proposed to predict evidence quality using features such as study design, sample size, and article metadata, with study design often proving the strongest predictor and classification accuracy reaching approximately 60–70% [3, 78, 106]. However, differences in algorithms, feature sets, and evaluation settings make direct comparisons difficult, and challenges in modeling factors such as publication bias further limit their effectiveness [121]. Nevertheless, automation remains a promising avenue for predicting strength of evidence.

This component of our end-to-end system searches medical databases for relevant evidence, identifies and filters high-quality studies, and provides them as context to the reasoning component.

Chapter 3

Methodology

3.1 Architecture and Design

The final system uses a modular, agent-based architecture for transparent, extensible clinical decision-making, as can be seen in Figure 3.1. Each agent handles a distinct pipeline stage, passing outputs forward in a structured, pseudo-sequential flow mirroring clinical reasoning: from interpreting clinical input to retrieving evidence and generating recommendations. The workflow is coordinated via LangGraph [23].
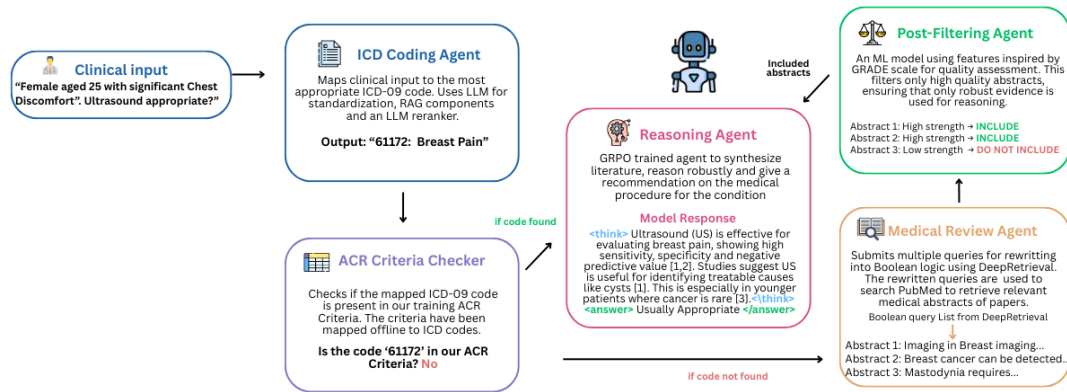

Figure 3.1: Overview of the multi-agent system architecture.

- **ICD Coding Agent**: Given a clinical note maps it to the most appropriate ICD-9 code, serving as the initial standardization step.
- **ACR Criteria Checker**: Verifies if the mapped ICD-9 code is in our predefined the ACR Criteria list, that have also been mapped to ICD codes. If it is, the corresponding ACR medical evidence is passed to the Reasoning Agent; if not,

the pipeline proceeds to the Medical Review and Post-Filtering Agents to gather evidence.

- **Medical Review Agent**: Uses DeepRetrieval [58] with a set of queries to retrieve relevant medical evidence from PubMed for the specific condition.
- **Post-Filtering Agent**: Filters retrieved literature for high-quality studies by using various features that aim to replicate the GRADE [110] scale.
- **Reasoning Agent**: The core of the architecture is the reasoning agent, trained with GRPO to replicate the reasoning from the ACR Criteria. Different variants include incorporating medical evidence and using different reward functions that emphasize reasoning alignment.

Therefore, our architecture is built to address the focused real-world question a physician often faces: "For my patient with condition Y, is ordering procedure X the right thing to do?" by progressing through each component. Details on the implementation of each agent are provided in the sections below.

3.2 ICD Coding Agent

The first component of our system maps clinical diagnostic notes to ICD codes, which is challenging because medical notations vary; for example the clinical notes, "female with significant chest discomfort" and "severe focal breast pain" are used interchangeably but refer to the same ICD code (see Section 2.2). To address this, we use a sample of 6,500 synthetic records of medical imaging prescriptions, with about 500 unique codes. Each record contains a clinical note, given as a natural language description (in Italian) of the patient's diagnosis along with its ICD-09 code, as well as a description of the prescribed imaging procedure; notably only the first two fields are used here for mapping.

Notably, this dataset has shorter, less ambiguous clinical notes (average word count: 7.29) than prior studies [55, 73], with a single assigned code per record, highlighting ICD coding's data-specific nature and suggesting no single method performs best across all datasets [33]. We also observe a long-tail distribution (Appendix D), as highlighted in other research [60, 146]. Importantly, over 35% of records show ICD code mismatches, as clinical notes do not match the assigned code's description, which was confirmed by large embedding distances across several medical embedding systems. This is expected, as the synthetic data is based on real doctor annotations and matches known clinician-

level ICD coding error rates; a quality issue also noted in prior studies [73], making initial cleaning essential. To address this, we used embedding similarity thresholds to identify mislabeled or ambiguous records and corrected them, improving the dataset's accuracy for modeling.

After data cleaning, we use our *ICD Coding Agent*, which combines LLM and RAG components, proven to enhance performance in prior work [28, 64, 66]. The agent first uses LLaMA 3.1 8B [42] to convert the unstructured Italian clinical notes into *standardized* English medical text, chosen for its efficiency. We hypothesize that this LLM standardization step converts the noisy clinical notes to ICD-compatible vocabulary, supporting the RAG component, consistent with studies showing LLMs' effectiveness in clinical language standardization [4, 139]. The standardized medical text is then matched to the top 5 most similar ICD code descriptions encoded and stored in a FAISS index [61] for fast retrieval. For efficiency, only ICD codes observed in the dataset were indexed, as in prior work. Another helpful addition is that if retrieval is ambiguous, as shown by a low top similarity score (below 0.7) or low variance among the retrieved codes (below 0.005), with thresholds set intentionally low, the same LLM is used for reranking to select the best match. To our knowledge, this hybrid pipeline which combines standardization, fast retrieval and accurate reranking is novel and can be easily extended to other versions like ICD-10. Evaluation of this agent uses standard ICD coding metrics: top-1 and top-5 accuracy, top-1 hierarchical accuracy (matching the first three digits of the ICD code which represent the main condition), and Mean Reciprocal Rank (MRR), which measures how highly the correct code is ranked.

We also used the same method (standardization, retrieval, reranking) to map ACR conditions to ICD codes, leveraging the static ACR dataset for offline processing and human validation. We note here that one ACR condition can map to multiple ICD codes and these mappings will be loaded in the *ACR Criteria Checker* Agent. Because variants contain additional demographic and clinical details not captured by ICD codes, we built a custom function to extract this information from clinical notes and assign the most accurate variant, defaulting to variant 1 when details are unavailable.

While not our main focus, mapping clinical notes and ACR criteria to the *intermediate ICD code layer* is very valuable for deployment, as most clinicians use ICD codes. This approach demonstrates deployment potential and provides a novel method for aligning these coding systems.

3.3 Reasoning Agent

3.3.1 Preprocessing the ACR Criteria and extracting reasoning

The core of our architecture, the *Reasoning Agent*, was developed by extracting structured data from the ACR. We selected a diverse subset of 30 condition documents from the ACR corpus, spanning different categories, body regions, and clinical presentations to ensure varied reasoning patterns. From these, we parsed the "Narrative and Rating" sections, which provide, for each condition–variant–procedure triplet, an appropriateness rating ("Usually Appropriate", "May Be Appropriate", or "Usually Not Appropriate") along with expert-authored reasoning supporting the rating. Parsing was performed using a custom script with manual corrections for formatting issues, yielding approximately 1,800 structured triplets after excluding cases where the ACR panel did not reach consensus on the recommendation.

Initial analysis shows that reasoning texts are lengthy and complex, with implicit logic that makes them hard to replicate. To streamline usage, we distill them into concise "reasoning traces": clear, self-contained sentences capturing key rationales as atomic, verifiable claims, following best practices [53]. We use LLaMA 4 Scout 17B 16E Instruct, chosen for its strong summarization and reasoning abilities [6]. This standardized format improves interpretability, supports downstream tasks as the traces are notably shorter, and eases human review. Distillation is feasible here since ACR reasoning mainly consists of *unordered factual points* rather than complex stepwise arguments. Still, rigorous human oversight was required to address potential hallucinations and omissions, with thorough double-checking to ensure accuracy and reliability. More details in AppendixA.4.
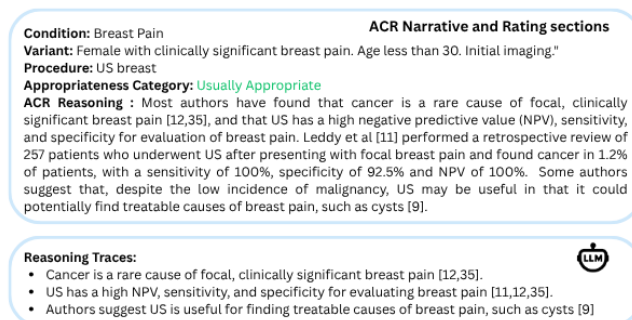

ACR Narrative and Rating sections

Condition: Breast Pain
Variant: Female with clinically significant breast pain. Age less than 30. Initial imaging."
Procedure: US breast
Appropriateness Category: Usually Appropriate
ACR Reasoning : Most authors have found that cancer is a rare cause of focal, clinically significant breast pain [12,35], and that US has a high negative predictive value (NPV), sensitivity, and specificity for evaluation of breast pain. Leddy et al [11] performed a retrospective review of 257 patients who underwent US after presenting with focal breast pain and found cancer in 1.2% of patients, with a sensitivity of 100%, specificity of 92.5% and NPV of 100%. Some authors suggest that, despite the low incidence of malignancy, US may be useful in that it could potentially find treatable causes of breast pain, such as cysts [9].

Reasoning Traces:
- Cancer is a rare cause of focal, clinically significant breast pain [12,35].
- US has a high NPV, sensitivity, and specificity for evaluating breast pain [11,12,35].
- Authors suggest US is useful for finding treatable causes of breast pain, such as cysts [9]

Figure 3.2: Condensed ACR preprocessing and reasoning trace example for "Breast Pain"

3.3.2 Group Relative Policy Optimization (GRPO)

RL is a popular post-training technique used to fine-tune LLMs by optimizing for desired behaviors through reward signals. Group Relative Policy Optimization (GRPO), introduced in DeepSeekMath [112], improves upon Proximal Policy Optimization (PPO) [109] by removing the need for a separate value network to estimate rewards. Instead, for each prompt q sampled from the dataset $\mathcal{D}$, GRPO generates a group of G outputs $\{o_i\}_{i=1}^{G}$ using the old policy $\pi_{\theta_{old}}$ and each output o_i is assigned a reward r_i. The advantage is then computed relative to the group's average reward, which both simplifies training and reduces computational cost by eliminating the value network.

To maintain stable updates, GRPO applies a clipped surrogate objective similar to PPO, which uses the importance sampling ratio $\frac{\pi_{\theta_{new}}}{\pi_{\theta_{old}}}$ between the new and old policies. This ratio is clipped within the interval $[1-\varepsilon, 1+\varepsilon]$ to prevent large policy updates. Additionally, GRPO includes a Kullback–Leibler (KL) divergence penalty D_{KL} scaled by a coefficient β, which discourages the updated policy $\pi_{\theta_{new}}$ from drifting too far from a reference policy π_{ref}, i.e., the frozen pretrained model. This regularization helps maintain fluency and coherence in the generated outputs. Formally, the GRPO objective maximizes the expected clipped advantage across the group of outputs while minimizing the KL divergence penalty:

$$
\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{o_i\} \sim \pi_{\theta_{old}}(O|q)} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(r_i \cdot \frac{\pi_{\theta_{new}}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, \right. \right.
$$
$$
\left. \left. \text{clip}\left(\frac{\pi_{\theta_{new}}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon \right) r_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta_{new}} \| \pi_{ref}) \right]
$$

$$(3.1)$$

To efficiently adapt our LLM, we combine GRPO with Low-Rank Adaptation (LoRA) [50], a parameter-efficient method that updates only a small subset of weights via low-rank decomposition, reducing memory and computation costs, using the Unsloth framework [30].

3.3.3 Model Variants and Reward Designs

We evaluate multiple aspects of the *Reasoning Agent*, which is built on the LLaMA 3.1 8B backbone, chosen for its open-source availability, size, and Unsloth compatibility. Model variants and rewards are described below, with a summary in Table 3.1 and an example in Figure 3.3.

Baseline model: Our first model was trained using standard rewards common in prior work: one binary "answer" reward for the correct appropriateness label and another binary "format" reward for properly enclosing reasoning within `<think></think>` followed by the answer in `<answer></answer>` tags.

Citations model: Secondly, we hypothesized that prior RL-based medical models overly depend on the LLM's existing knowledge [43, 54, 96], limiting integration of up-to-date specific evidence and increasing the risk of factual errors. The *Citations* model tests whether adding specific, relevant, high-quality evidence to the context improves performance by grounding recommendations in factual literature rather than pre-trained knowledge. This marks a novel approach for RL-trained LLMs in medicine which can be crucial in building trust with clinicians [134]. We incorporated the medical evidence cited by the ACR to support their recommendations, retrieving references via PubMed IDs (PMIDs) via the PubMed API. Unfortunately, due to journal access restrictions only the abstracts of the papers were freely accessible. For efficiency, we further condensed these abstracts to just their results/conclusions subsections, as these capture the key points behind appropriateness. Although this reduces content depth, it still enables us to assess the value of incorporating external evidence to the agent, while preserving computational efficiency with a smaller context window. The *Citations* model therefore uses the same rewards as the *Baseline*, but incorporates condensed medical abstracts as additional context.

Next, we recognize that the valuable reasoning in the ACR Criteria can be leveraged to "teach" the model to reach correct conclusions logically. To this end, the following two models, *LLM Eval* and *MedReason-Embed*, incorporate reward functions that align the model's generated reasoning with expert "gold" ACR reasoning.

LLM Eval model: Inspired by approaches in [126, 136], where a separate LLM assesses the factuality of outputs from the trained LLM, we extend this idea by adding an explicit judgment component that evaluates the quality of the generated reasoning. Building upon the *Citations* model, which incorporates medical evidence as additional context, the *LLM Eval* model extends this by integrating a separate LLM-based reward that scores the alignment between the generated and gold reasoning. This evaluator assesses three key criteria: overlap of medical concepts, logical consistency of reasoning, and presence of supporting evidence, producing a score between 0 and 1. By explicitly rewarding reasoning quality during training, we aim to go beyond simply incorporating evidence, and instead teach the model what constitutes "good" reasoning, encouraging outputs that better align with expert gold reasoning. Thus, the *LLM Eval* model

combines the same medical evidence and reward functions as the *Citations* model (answer and format rewards) with the new LLM-based reward score. The factuality judge uses the smaller, efficient Qwen1.5 1.8B model [125], selected after initial testing confirmed its suitability for the task.

MedReason-Embed: Lastly, we also propose another model with a *novel* custom reward function, motivated by multiple studies showing that models may give correct answers with flawed reasoning [24, 27, 96]. This reward jointly evaluates answer correctness and reasoning sentence alignment: for each gold reasoning sentence, we compute its embedding and compare it to all generated sentence embeddings, taking the highest cosine similarity score. We then average these maximum scores across all gold sentences and multiply the result by the binary answer reward. Thus, the reward is proportional to the degree of reasoning alignment but is granted only when the final answer is correct, linking this reward directly to both reasoning quality and outcome correctness, an innovation not previously explored. We hypothesize that this reward will improve both the model's reasoning and its final answers. By rewarding reasoning alignment only when the referral decision is correct, the model can perhaps learn to weigh risks and benefits from medical evidence, fostering a deeper, more faithful understanding of clinical reasoning and its link to the final decision. Although multiplying by a binary outcome can seem brittle, since "good" reasoning is ignored if the final answer is wrong, this design enforces the principle that valid reasoning must lead to the correct answer. Therefore, the *MedReason-Embed* also includes all the medical evidence and has two reward functions: the binary format reward and this innovative reward outlined above. Mathematically, this reasoning reward is defined as:

$$R_{joint} = \mathbb{I}_{\text{gen\_answer=gold\_answer}} \times \frac{1}{N} \sum_{i=1}^{N} \max_{j} \left(\cos \left(\mathbf{e}_i^{\text{gold}}, \mathbf{e}_j^{\text{gen}} \right) \right)$$

where $\mathbb{I}$ is the indicator function for answer correctness, cos denotes cosine similarity, and $\mathbf{e}_i^{\text{gold}}, \mathbf{e}_j^{\text{gen}}$ are reasoning embeddings of gold and generated sentences respectively.

Model	Evidence	r_{answer}	r_{format}	$r_{reasoning}$
Baseline	None	✓	✓	None
Citations	✓	✓	✓	None
LLM Eval	✓	✓	✓	LLM-based
MedReason-Embed	✓	✓ (via reasoning)	✓	Joint reward

Table 3.1: Comparison of Reasoning Agent Variants

Finally, as in other studies [68, 96], we also wanted to compare our RL-based models with other standard approaches. We trained a standard SFT model on the same LLaMA 3.1 8B backbone, using cross-entropy loss to predict the appropriateness label for each triplet, with matching training parameters to ensure fair comparison. We also evaluated the larger LLaMA 3.1 405B model from the same series [5] on the test set for comparison. Notably, these models are designed solely to output an appropriateness label, without producing explicit reasoning, as in [96]. Full implementation details for all models are provided in Appendix A.
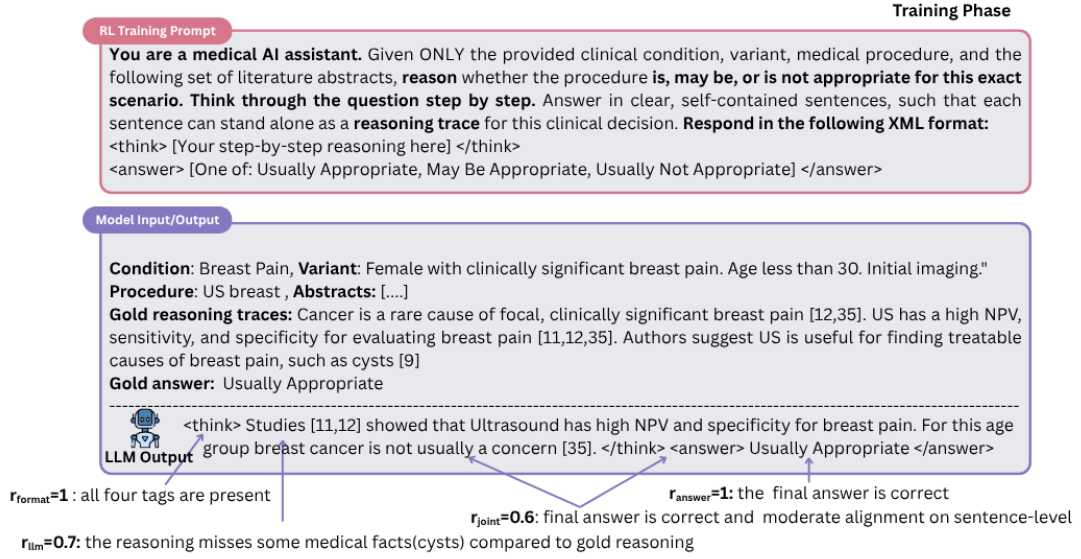
Figure 3.3: Example of training and respective rewards.

3.3.4 Evaluation Metrics

Our reasoning model evaluation approach focuses on three main aspects: *predictive performance*, *reasoning alignment*, and *training efficiency*.

- **Predictive performance:** We evaluate predictive performance using the F1 score, which combines Precision (Pr) $(\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}))$ and Recall (Re) $(\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}))$ into a single measure $F1 = \frac{2 \cdot \mathrm{Pr} \cdot \mathrm{Re}}{\mathrm{Pr} + \mathrm{Re}}$, making F1 suited for imbalanced datasets because it balances the trade-off between identifying positive cases and avoiding false positives. Given our own dataset's imbalance (about 64% "Usually Not Appropriate"), we report both **Macro F1**, which averages the F1 score for each class equally and highlights minority class performance, and **Weighted F1**, which averages F1 scores according to class prevalence to reflect overall

performance. In our setting, we calculated that achieving a macro F1 above 50% and a corresponding weighted F1 above 60% represents competitive performance, indicating a solid balance between common and rare appropriateness categories and reflecting strong overall precision and recall. All scores are averaged over three runs to reduce the effect of stochastic variation, enabling fair comparison across RL-adapted models, SFT, and the larger LLaMA 3.1 405B model on the test set.

- **Reasoning alignment:** We assess the quality and alignment of generated reasoning against gold traces using two metrics:

    1. *LLM-align-score:* The LLaMA 3.1 8B LLM, scores the similarity between generated and gold reasoning on the test set based on explicit criteria such as clinical relevance and medical knowledge, giving an alignment score out of 10, consistent with other research [148].

    2. *NER Embedding F1:* We use OpenMed's NER pipeline [95] to extract clinical entities from both the generated and gold reasoning, and supplement it with custom rule-based phrases (e.g., "no relevant clinical literature") to capture key specific signals. These entities are embedded and F1 score is computed based on maximum cosine similarity between the sets. This metric assesses how well the model captures expert-identified concepts (recall) while avoiding unsupported or hallucinated content (precision). By focusing on semantic overlap rather than exact wording, it offers a robust measure of reasoning quality, while using medical entities ensures the evaluation centers on clinically meaningful concepts rather than entire sentences.

- **Relative Training Time:** Training times are reported as multiples of the *Baseline* to simplify comparison, ensure fairness, and normalize for hardware differences. Inference times are similar across models.

3.4   Medical Review and Post-Filtering Agent

We now turn to the *Medical Review* and *Post-Filtering* Agents. To investigate whether our system generalizes beyond ACR-covered conditions, we replicate their evidence-gathering approach by performing expert-defined query searches in structured databases [21, 65] and perform filtering of the retrieved results by modeling the GRADE scale to retain only high-quality evidence.

Firstly, for the *Medical Review Agent*, we chose the Deep-Retrieval-PubMed-3B model [123] because it provides a user-friendly API and direct access to PubMed. Analysis of the ACR evidence-gathering process showed that a single query can not capture all relevant aspects; for example, the ACR's search for "Breast Pain" includes related conditions (e.g., "Breast Neoplasm"), synonyms (e.g., "Mastodynia"), and procedure-specific terms. To replicate this, we generated and submitted multiple queries to DeepRetrieval for rewriting, ensuring comprehensive coverage of all relevant clinical concepts, with two examples of such queries shown in Figure 3.4. We also applied filters for publication year, language, and removed duplicates. However, we found that it is not possible to retrieve exactly the same medical papers as the ACR. This is because the ACR's search is conducted in the Medline database (which lacks a standalone API) and involves *extensive* manual review of a large set of results.

Instead, we focused on retrieving papers that capture the *same clinical concepts and findings* as those used by ACR, aiming to replicate the reasoning behind the recommendations, not to match the exact documents. This reflects clinical practice, where different studies can support the same conclusions about imaging appropriateness. Accordingly, our evaluation examines how using alternative but relevant evidence influences *final recommendations*, rather than whether the original documents are retrieved.

To ensure the retrieved evidence closely aligned with the content of the gold-standard ACR papers, we focused on a small subset of conditions and tested multiple search strategies, varying the query wording, the number of documents retrieved, and other parameters. Each query was rewritten using DeepRetrieval, with the phrase "Diagnostic Imaging" added to improve clarity and relevance. We then compared the retrieved evidence with the ACR references by embedding and clustering their results sections, measuring distances between cluster centers to assess conceptual similarity. Additionally, we applied Latent Dirichlet Allocation (LDA) [16] to both sets to extract topic distributions, enabling comparison of their thematic coverage. While this approach simplifies the comparison and may not capture every nuance, it balances clinical relevance with our computational constraints. An example can be seen in Appendix B.1.

After retrieving documents using the *Medical Review Agent*, our *Post-Filtering Agent* enhances precision by applying the ACR panel's GRADE scale [110] to select the most relevant evidence. This scale assesses strength of evidence (SOE) based on factors such as study design, risk of bias and precision, categorizing papers into four

levels: high, moderate, low, and very low. For this, we use structured PubMed features (e.g., publication type) and data extracted from their abstracts (e.g., cohort size) to train an ML model that predicts each paper's SOE. We use the provided combinations of PMIDs and their SOE given by the ACR "Appendix" sections of our set and experiment with various features inspired by previous studies and multiple ML models. The main evaluation metric is the recall of papers with "high" strength of evidence (SOE), as our aim is to reliably identify and include only these highest-quality studies in the model's context. This prioritizes capturing all top-tier studies over maximizing overall accuracy. This approach aims to automate and scale evidence review, closely mimic ACR expert evaluation, and test whether these structured, easily accessible features can approximate a process that typically requires specialized clinical expertise.

Following the ACR's approach, we search for literature at the *condition* level instead of the more granular *variant* level to reduce irrelevant results and noise. After filtering, papers are assigned to condition-variant-procedure triplets using the following simple regular expressions and heuristic rules: papers are included in the triplet if their abstracts mention both the procedure and condition (or synonyms), and either contain relevant evaluation keywords or are a recognized publication type such as review or clinical trial.
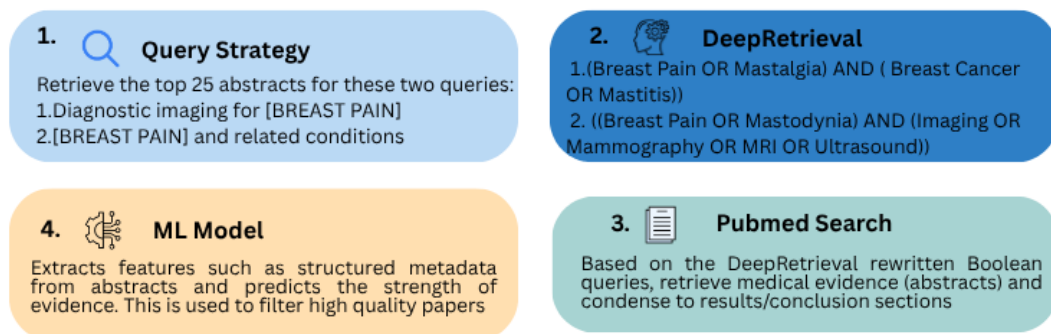

Figure 3.4: An example of evidence-seeking strategy for "Breast Pain"

To conclude this section, our methodology centers on optimizing query formulations for DeepRetrieval to retrieve medical evidence with similar content to the ones in ACR. DeepRetrieval rewrites input queries to maximize relevance, and we append "Diagnostic Imaging" before querying PubMed via API for the top-k results. The retrieved evidence is then filtered using a quality prediction ML model, and papers with a predicted SOE of "high" are assigned to specific condition-variant-procedure triplets using the heuristic rules described above.

3.5   Generalization

Finally, our last goal is to evaluate the generalization of our approach. We therefore created a "generalization set" of five previously unseen conditions, selected for their clinical relevance and minimal similarity to the training set (Appendix A.5). These conditions deliberately span different anatomical regions (e.g., wrist), clinical domains (e.g., neurology, gynecology), and diagnostic purposes (e.g., chronic vs. acute scenarios) from the initial set. Care was taken to match the distribution of labels to the original set, with 67% of the generalization set labeled as "Usually Not Appropriate". For these cases, we used the full pipeline described above to automatically retrieve evidence and our reasoning component, simulating how the system would function on conditions outside the ACR, since no experts were available to construct new cases.

Notably, this analysis differs from our earlier experiments in two ways: (a) condition novelty, as the generalization set contains conditions not present in the train/test set, and (b) evidence source, using evidence retrieved by our pipeline vs with the ACR citations we used so far. To separate these effects, we also ran inference on the generalization set using the ACR's "gold" citations. We then compare two sets of results:

- Generalization set vs. test set performance using ACR citations: captures the effect of condition novelty while keeping the evidence source constant (ACR).
- ACR citations vs. our own retrieved evidence on the generalization set: captures the effect of evidence source while keeping the conditions constant.

This dual analysis quantifies the true generalization gap for each model, providing a realistic measure of robustness in unfamiliar scenarios and different sets of evidence; advancing toward a fully autonomous, end-to-end decision-support system.

Chapter 4

Results

4.1 ICD Coding Agent

The performance of the *ICD Coding Agent* is shown in Table 4.1. Here, we note that the LLM standardization step proved highly effective, frequently mapping Italian clinical notes to the correct ICD code description with near-exact wording. This demonstrates the LLM's strength in normalizing clinical language despite real-world variability in notation. In rare instances, some contextual details (such as patient history) were overlooked, but no hallucinated or extra information was introduced.

Metric	Value
Top-1 Accuracy	80.45%
Top-5 Accuracy	91.97%
Mean Reciprocal Rank (MRR)	85.46%
Top-1 Hierarchical Accuracy	91.47%

Table 4.1: Evaluation Results ICD Agent

The ICD Coding Agent achieved a strong top-1 accuracy of 80.45% and a hierarchical accuracy (3-digit ICD match) of 91.47%, indicating strong identification of the key clinical condition, but sometimes missing some specificity. The MRR of 85.46% further demonstrates that the correct ICD code is typically ranked near the top, confirming the reliability of our pipeline, with the exact correct code presented in the top-5 codes 91.97% of the time. We also note that the LLM reranking component contributed a small improvement, increasing top-1 accuracy by approximately 0.5%. Finally, these results further underscore the value of integrating LLMs for nuanced clinical understanding

and semantic matching, consistent with findings from recent literature.

Overall, these results are promising and indicate strong potential for broader application of this methodology. Nevertheless, it is important to acknowledge that the favorable outcomes may be partly attributable to the nature of the dataset, which included relatively short and less ambiguous clinical notes.

4.2 Reasoning

4.2.1 Model performances

The Reasoning Agent is trained on 1,800 condition-variant-procedure triplets from 30 conditions, with stratified 70/30 train-test split. Table 4.2 compares the four RL models, the SFT, and the larger LLaMA 3.1 405B, while Table 4.3 reports reasoning quality and training metrics for the RL models.

Model / Config	Macro Avg F1	Weighted F1
Baseline	33.5%	37.1%
Citations	45.6%	56.6%
LLM Eval	52.7%	65.6%
MedReason-Embed	51.6%	65.6%
SFT model	36.7%	56.2%
LLaMA 405B	47.0%	53.7%

Table 4.2: Model Predictive Performance Results

Model/Config	LLM-align-score (/10)	NER Embedding F1	Relative Training Time
Baseline	5.64	37.9%	1.0
Citations	7.28	61.2%	1.2
LLM Eval	7.57	65.4%	1.8
MedReason-Embed	7.67	65.5%	1.3

Table 4.3: Model Reasoning Alignment and Training efficiency

The *Baseline*, trained only with format and answer rewards, achieves macro and weighted F1 scores of 33.5% and 37.1%. In comparison, a naive majority-class classifier predicting only "Usually Not Appropriate", would achieve a lower macro F1 of 26%

but a higher weighted F1 of 50% on our dataset, indicating that while our *Baseline* model captures class diversity better and its more deployable, it still struggles to optimize performance across imbalanced labels and does not learn many meaningful distinctions. Reasoning alignment scores are also relatively low (5.64/10 and 37.9%), as we would have expected a well-aligned model to score much higher. This indicates poor alignment with ACR reasoning and makes the model unsuitable for deployment. Interestingly, adding medical evidence as context in the *Citations* model significantly improves performance, with macro and weighted F1 scores rising to 45.6% and 56.6%. The 12% macro F1 and 20% weighted F1 increase compared to the *Baseline* show that contextualized LLMs improve performance across all classes, including minority classes, reflecting better synthesis of information. This is key because because it shows that common foundation models can be adapted for domain-specific tasks by adding high-quality, relevant context to boost performance. Reasoning alignment also improves considerably, as reflected in the LLM-align score and NER Embedding F1, which increase to 7.28/10 (+1.64 points) and 61.2% (+23.3%) respectively, compared to the *Baseline*. Relative training time is comparable with the *Baseline* despite the larger context window. Overall, this shows that high-quality, relevant context can boost predictive and reasoning performance while staying computationally efficient.

The next two models use ACR reasoning traces: the *LLM Eval* and *MedReason-Embed* models. The *LLM Eval* model further increases F1 scores to 52.7% (macro) and 65.6% (weighted), marking significant gains over the *Citations* model in both overall and minority-class performance. The LLM-align-score and NER Embedding F1 also show better alignment with ACR reasoning than the *Citations* model with scores of 7.57/10 and 65.4% respectively, although smaller gains than expected. This suggests that rewarding reasoning quality improves both reasoning alignment and final answer performance. However, this model is the most computationally demanding, as its training process requires an additional LLM call for every generated answer to evaluate reasoning alignment. Finally, our innovative *MedReason-Embed* model achieves 51.6% macro F1 and 65.6% weighted F1, improving significantly over the *Citations* model by 6% and 9% respectively and performing similarly to the *LLM Eval* model. It also roughly matches the *LLM Eval* model in reasoning alignment while requiring *significantly* fewer computational resources. This shows that a well-designed, task-specific reward can match more resource-intensive approaches, making *MedReason-Embed* the most efficient and effective model for our needs. We also performed McNemar's test [86], which compares paired classification errors (i.e., cases

where one model is correct and the other is wrong) between models; showing that all RL model results differ significantly with each other (using $\alpha = 0.01$) except the *LLM Eval* and *MedReason-Embed* models, where no significant difference was found, confirming that they perform similarly (see Appendix A.7).

Lastly, the *SFT* model trained under the same conditions, achieved a macro F1 of 36.7% and a weighted F1 of 56.2%. While this outperforms the *Baseline*, its predictions are mostly limited to "Usually Not Appropriate", clearly showing overfitting as also reflected in the big disparity between the F1 scores, and does not outperform the other three RL models. Inference with the much larger *LLaMA 3.1 405B* model yielded macro F1 of 47.0% and weighted F1 of 53.7%, performing similarly to our *Citations* model but still not reaching the performance of the two RL models using reasoning rewards. Although this larger model benefits from broader medical knowledge and may have been indirectly exposed to ACR-like data during training, its high resource demands limit its practicality in real-world deployments [62]. Overall, our RL models using both reasoning-based rewards and medical evidence, *LLM Eval* and *MedReason-Embed* significantly outperformed their SFT counterpart and even much larger models.

Across our experiments, two consistent patterns emerged. Firstly, incorporating high-quality medical context led to substantial gains in both performance and reasoning alignment compared to models without such context, demonstrating the added value of evidence beyond the model's prior training. Secondly, models trained with task-specific reasoning rewards achieved even higher reasoning alignment and predictive performance, surpassing both the SFT and much larger models. Notably, these improvements can be achieved using resource-efficient methods, as demonstrated by our *MedReason-Embed* model, showing that knowledge integration and reasoning-focused reward design can deliver strong results without increasing model size and while being computationally efficient, underscoring the value of thoughtful reward strategies.

We include an example illustrating reasoning alignment between generated and gold reasoning in Figure 4.1, along with a brief qualitative error analysis of the *MedReason-Embed* model. Notably, human evaluation found no signs of overfitting or reward hacking (e.g., verbose but "clinically empty" reasoning). While deeper review by medical experts is needed, our initial review shows that misclassifications mostly stemmed from missing key details necessary for conclusions, often due to limited evidence provided as context (as we provided condensed abstract sections rather than full papers) in the *Reasoning Agent*. We observed that the model generally avoided hallucinations and interpreted the literature accurately, but deeper evaluation is needed.
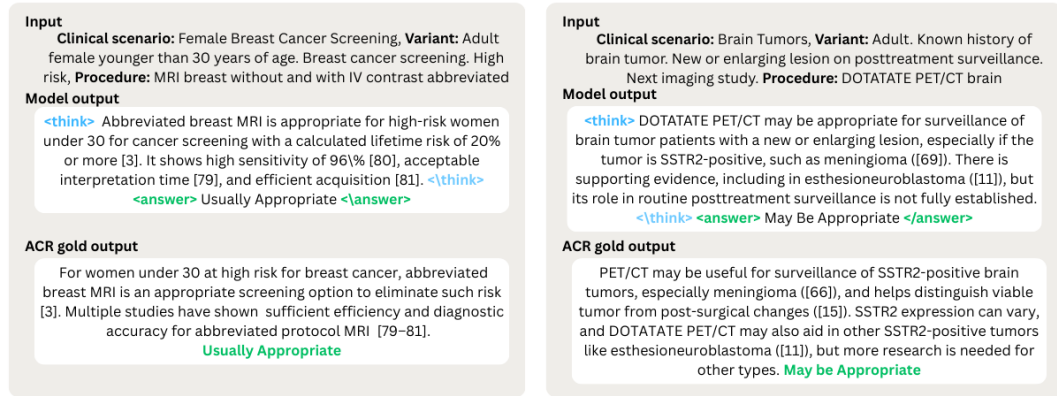
Figure 4.1: Example of model answers from MedReason-Embed Model

4.2.2 Training Reward Trajectories

We also analyzed the *smoothed* training reward curves, averaging rewards over a sliding window of 25 training steps to reduce noise, for all RL-adapted models to examine how different reward functions influence learning. Although only two epochs were used for the final training runs to balance performance gains with computational cost, we plot curves up to four epochs to better visualize trends; therefore, performance is reported at epoch 2. As a reminder for the *Baseline*, *Citations*, and *MedReason-Embed* models, total reward $r_i \in [0,2]$, while the *LLM Eval* model's $r_i \in [0,3]$. From Figure 4.2 we observe that all four models quickly learn to produce correctly formatted outputs, stabilizing format rewards after about 1 epoch. The *Baseline* model shows slow, unstable answer reward improvements without a clear turning point, likely due to weak reward signals, the variety of conditions and its limited medical knowledge. Adding medical evidence in the *Citations* model leads to a sharp increase in answer reward near the end of the first epoch, resulting in healthier training curves. The overall reward reaches 1.6 out of 2 by the second epoch, though some instability appears afterward. The *LLM Eval* model, with its three reward signals, shows only modest gains in answer reward and notably a decline in the LLM-based reasoning alignment reward. Despite the model's competitive performance, the use of multiple rewards may have confused the learning process, with the model not fully understanding how to maximize them effectively. Further research is needed to better understand this effect. By epoch 2 its reward achieves approximately an 1.8 out of 3. Lastly, the *MedReason-Embed* model shows the healthiest training dynamics, with a sharp reward increase after epoch 1 followed by stable performance, suggesting an "aha moment". This phenomenon, observed in

several studies, is described as the point where the model "autonomously develops advanced problem-solving strategies, including reflection and self-correction" [45, 145], which is quite impressive with such little training. It is plausible that, at this point, the model moves beyond simple pattern matching and begins to develop more abstract problem-solving strategies, such as weighing specific clinical concepts like "radiation risk" against "diagnostic sensitivity", much like an expert would. It later plateaus around the second epoch at about 1.4 out of 2, where the training was concluded. We note that this increase is more significant than in the *Citations* model, as it stabilizes considerably afterward. This further highlights the potential of our innovative reward function, demonstrating its effectiveness as a strong training signal. Overall, examining training reward trajectories is essential, since models with high apparent performance (like *LLM Eval*) may not exhibit stable or consistent learning.

(a) Baseline

(b) Citations

(c) LLM Eval

(d) MedReason-Embed
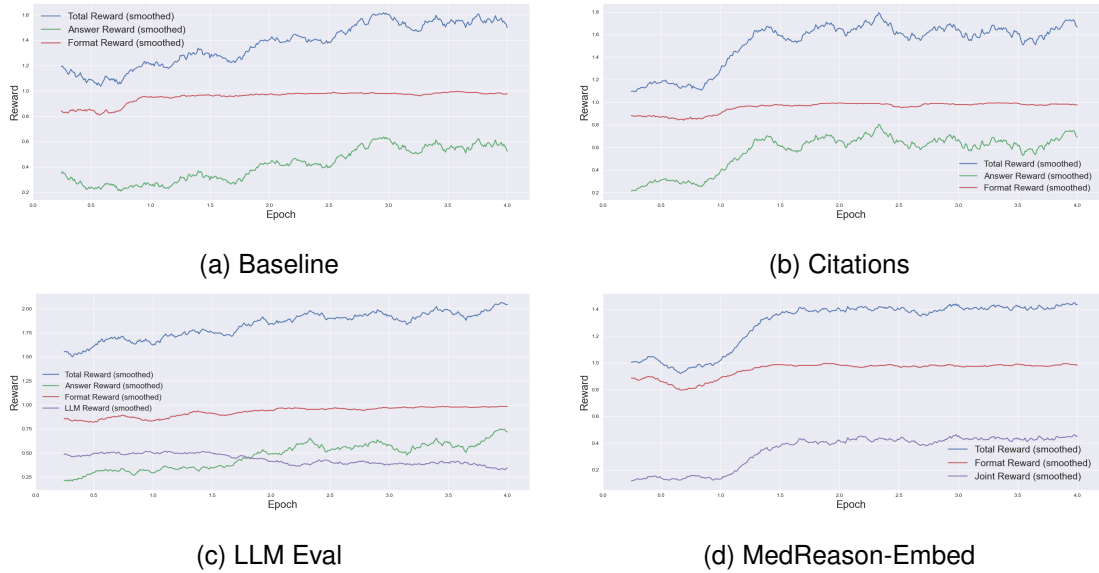
Figure 4.2: **Smoothed training reward trajectories**: plotted different components of the reward for each RL model, with window size 25, over 4 epochs for better visualization.

4.3 Evidence Retrieval and Post-Filtering

For the *Medical Review Agent*, we implemented the Deep-Retrieval-PubMed-3B model [123] and tested it, aiming to retrieve conceptually similar, rather than identical, evidence to the ACR citations, that will lead to the same predictions. We explored various search strategies by analyzing conceptual similarity, using clustering distances and LDA for concept alignment on a small subsample of conditions. After some experimentation

we found that our most efficient strategy was to use a combination of distinct queries, covering related conditions, synonyms, and imaging-specific clinical evidence (Appendix B.2), as input to DeepRetrieval to generate rewritten Boolean queries, retrieving the top 25 PubMed papers for each rewrite.

For the *Post-Filtering Agent*, we developed an ML algorithm to assess the quality of retrieved papers. After testing different features and models used in other studies, a Random Forest approach [15] using features such as study design and year of publication (full list in Appendix C) achieved the best results. Despite using simple features, the model achieved a recall of *0.74* for identifying studies with "high" strength of evidence, with no signs of overfitting based on the stratified 80/20 train-test split. This strong performance is critical, as only "high" strength studies are included in the LLM's context. Beyond its practical role, the model's ability to automatically identify strong evidence is valuable, as this task is typically time-consuming, expert-driven, and could benefit from further automation, given our promising results. Feature importance analysis confirmed study design, journal quality, and publication year as key predictors (Appendix C).

4.4   Generalization results

This section evaluates the model's ability to generalize using a "generalization set" of five new clinical conditions with minimal similarity to the training set, spanning diverse anatomical regions, clinical domains, and diagnostic purposes (approximately half the size of the original test set). We assess performance of the generalization set in two scenarios: (1) using the gold-standard ACR provided citations, and (2) using citations retrieved by our medical retrieval pipeline, for the reasons stated in Section 3.5. Results for all RL-adapted models, SFT, and the larger LLaMA 3.1 405B are presented. Table 4.4 displays performance using *the ACR "gold" citations*, while Table 4.5 shows results with our *own retrieved citations*, both on the generalization set.

Firstly, we compare generalization performance on the test and generalization test sets using ACR citations to capture the effect of condition novelty (Tables 4.2 and 4.4). For the *Baseline* model, there is only a modest drop in macro and weighted F1 scores (about 2.5%), which is expected and acceptable given the differences between the two sets. In contrast, the *Citations*, *LLM Eval*, and *MedReason-Embed* models each show a more notable 5–7% decrease in macro F1 and a 2–3% decrease in weighted F1 compared to their counterparts on the test set, indicating a bigger, but still manageable generalization gap, consistent with gaps in other studies [96]. This suggests that while

Model / Config	Macro Avg F1	Weighted F1
Baseline	31.0%	34.5%
Citations	40.5%	53.0%
LLM Eval	46.6%	63.6%
MedReason-Embed	44.5%	63.3%
SFT	43.3%	65.1%
LLaMA 405B	51.7%	60.0%

Table 4.4: Generalization dataset performance with **ACR gold citations**

medical context and richer supervision improve performance, these models remain somewhat sensitive to distribution shifts. The modest gap likely reflects condition heterogeneity in the generalization set, which introduces new anatomical regions and clinical domains requiring perhaps unfamiliar reasoning strategies. With no signs of overfitting or reward hacking, the small gap seems expected rather than exploitative, indicating the model generalizes adequately. Importantly, the RL models' performance ranking remained consistent, with the same best and worst performers.

The SFT model, while outperforming the *Citations* model in both F1 scores, it does not outperform our two reasoning-based RL models in macro F1 and continues to overpredict "Usually Not Appropriate", reflecting earlier issues and underscoring its limited generalization; an expected outcome also noted in prior studies [26]. The much larger LLaMA 405B model actually showed the best results, even outperforming its test set performance, but has other deployment issues due to its size.

The key takeaway is that reasoning-based RL models outperformed the SFT approach, maintaining acceptable and expected generalization gaps, whereas the SFT model again showed overfitting, overpredicting "Usually Not Appropriate". Future work should build a larger generalization set to assess this more thoroughly and aim to further reduce the gap by moving beyond random condition sampling toward optimal experimental design, which better reflects real-world clinical distributions and has improved RL performance in prior studies [101].

Next, we evaluate the model using our own citations retrieved via DeepRetrieval (Table 4.5) aiming to to capture the effect of *evidence source* in performance. Unlike earlier experiments with ACR "gold" citations, this uses automatically retrieved evidence from our pipeline to assess the system's ability to operate fully autonomously; retrieving, filtering, and reasoning without human curation.

Model / Config	Macro Avg F1	Weighted F1
Baseline	31.0%	34.5%
Citations	40.4%	46.6%
LLM Eval	43.8%	54.9%
MedReason-Embed	45.9%	55.0%

Table 4.5: Generalization dataset performance with **our own citations**

Here, we compare the performance of models using ACR-provided citations against those using our own retrieval strategy, to assess the impact of different medical evidence sources on prediction performance in the same generalization set (Tables 4.4 and 4.5). Firstly, the *Baseline* model's performance was unchanged, as it does not use any citations. The *Citations* model's macro F1 remained stable at around 40%, while its weighted F1 dropped by 6%, suggesting performance declined slightly on the majority class while minority-class performance was unaffected. Notably, as seen in previous analysis, adding citations yielded a clear gain (9% macro and 12% weighted F1) over the *Baseline*, when using our own citations. The *LLM Eval* model showed a small 3% drop in macro F1 and comparable 8% drop in weighted F1 compared to its ACR-citation counterpart, in line with expectations given the dataset imbalance and it still outperformed the *Citations* model in both metrics. Interestingly, the *MedReason-Embed* model achieved a 1% increase in macro F1 and an 8% drop in weighted F1, indicating more robustness to diverse evidence sources and good generalization to new citations.

The key takeaway is that using entirely different evidence sets yields *comparable* results, as shown by the small differences in macro F1 scores. This suggests that in medicine, different high-quality papers often lead to similar clinical conclusions. This confirms that the model's recommendations are robust to the choice of supporting literature, reflecting both the consistency of the medical evidence base and the strengths of our *Reasoning Agent*. We also note that minor performance drops may stem from evolving research, as changes in evidence can shift what is considered appropriate. While the generalization set is limited in size and diversity, these findings are promising and indicate the system can operate effectively in a fully autonomous setting without heavy reliance on human-curated sources. Broader evaluations across more conditions, and especially ones that are not currently covered by the ACR, are needed to confirm robustness under varying citation contexts, as the strong influence of evidence quality and outcomes may lead to greater variance in model predictions as this space expands.

Chapter 5

Conclusions and Discussion

5.1 Summary of Key Results

In this work, we developed a comprehensive end-to-end agentic system that replicates the full clinical referral pipeline; from ICD coding to literature retrieval, filtering, and structured reasoning, achieving strong performance despite being trained on only 30 conditions compared to the 257 in the full ACR corpus. Our *Reasoning Agent*, trained with GRPO, and in particular our best model *MedReason-Embed*, is a lightweight 8B-parameter model with fast inference suitable for practical deployment. By incorporating novel reasoning-aligned reward functions and medical knowledge, it demonstrates promising capabilities in generating clinically relevant justifications and appropriate imaging recommendations that align closely with expert-authored ACR criteria. These findings reinforce the potential of adapted LLMs not just as answer generators, but as structured clinical reasoning tools. Uniquely, the system adapts seamlessly to both ACR-covered cases and scenarios where no existing guidelines; a capability unmatched by other systems, offering exceptional value through robustness in real-world settings and the capacity for continual learning. We also note that assembling domain experts to develop the ACR criteria is complex, time-consuming, and costly. While our system is not yet a replacement, it offers a promising and economically efficient complement to expert-driven processes; particularly relevant amid ongoing debates about AI's role in medicine, such as predictions of radiologists being displaced by AI [127]. The main conclusions, linked to the research questions (Section 1.3) and additional insights, are presented below following the order of our end-to-end system architecture for clarity.

Strong Performance of Our ICD Coding Agent: Our novel approach, combining LLM-based standardization, RAG, and LLM reranking, achieved strong results on

real-world clinical data. The LLM reliably mapped diverse and noisy inputs to ICD-compatible formats, reinforcing the emerging role of *LLMs as standardizers* in clinical NLP [4, 139]. While direct comparison is limited due to differences in datasets and metrics, our pipeline which achieved a top-1 accuracy of 80.45% and top-5 accuracy of 91.97%, is strongly competitive with current methods such as MedCodER (F1: 0.60) [12], and approaching average human accuracy in the UK (83%) [19], notably with less variance. While below commercial systems like Ambience AI (95% top-1 ICD-10 accuracy [8]), their lack of public methodology prevents direct comparison. Overall our results encourage broader adoption of this pipeline, though further research is needed to assess performance in more complex and ambiguous cases.

The power of RL trained models: Our results show that RL provides a lightweight yet effective way to adapt general-purpose LLMs, outperforming both SFT and much larger models. The SFT model, trained under identical conditions, consistently overfit, mostly predicting "Usually Not Appropriate", making it unsuitable for deployment. While the 405B-parameter LLaMA 3.1 performed comparably to the *Citations* model on the test set, it failed to outperform our 8B parameter models, *LLM Eval* and *MedReason-Embed*, that used reasoning reward functions, despite being 50× larger. As noted in recent work [62], smaller LLMs are preferable in medicine for privacy and efficiency, yet often struggle with complex reasoning. Our two reasoning RL-trained models overcome this, delivering stronger performance while remaining resource-efficient, in line with recent findings [68, 96, 113, 119]. We argue that RL is key to enabling more structured, analytical "System 2" reasoning in clinical AI [75]. We also believe that our work provides a blueprint for developing highly capable reasoning agents in low-resource domains.

Contextualized RL-adapted LLMs: Providing medical evidence as context to our RL-trained *Citations* model, boosted macro F1 by 12% over the *Baseline*, and improved alignment with ACR reasoning, consistent with prior work [134]. This is significant, since this simple yet effective strategy adapts general-purpose LLMs to specialized clinical domains without additional fine-tuning. It enhances both performance and explainability, making it a promising direction for real-world deployment in medicine and in any other field where transparency is essential.

The Power of Reasoning Rewards: Our results show that models trained with reasoning-specific rewards, *LLM Eval* and *MedReason-Embed*, outperformed the *Citations* model (that also has access to medical evidence) by 6–7% in macro F1 and about 4% in NER embedding F1, improving both predictive performance and reasoning

quality. This supports prior findings PRMs outperform ORMs. [29, 77, 128]. These findings reinforce growing concerns that generic answer-based rewards are insufficient for medical LLMs [24, 37, 71, 96, 149]. Despite increasing recognition that clinical reasoning should serve as the primary training signal [71], most models still prioritize final answer accuracy [27, 37]. These results further underscore the value of reasoning-based rewards for aligning LLMs with expert thinking. Admittedly, while our rewards improved performance, they may still miss subtle clinical nuances. Therefore, future work should prioritize developing grounded, robust reasoning rewards and shared frameworks for clinical reasoning alignment with the goal of further improving performance. Bridging this gap will require close collaboration between clinicians and AI researchers and is essential for building trustworthy, transparent, and deployable medical AI systems.

Our MedReason-Embed Model Performed the Best: The *MedReason-Embed* model achieved the best overall tradeoff between performance, alignment, and efficiency. It demonstrated stable training and an "aha moment" of self-correction, suggesting the potential development of meaningful reasoning capabilities [45, 145]. Performing similarly to the *LLM Eval* model while using fewer resources, it stands out as our most effective and scalable solution; which is particularly significant since this is a novel contribution of our work. These results further highlight the value of reasoning-based rewards and that future work should focus on designing rewards that align closely with goals of the task and developing shared frameworks to guide this effort.

Robust and Effective Literature Evidence Gathering: This component addressed the challenge of handling cases lacking formal guidelines [139]. Using the rewritten queries obtained using DeepRetrieval that include synonyms and related conditions, significantly outperformed standard PubMed searches while remaining simple to implement. Heuristics such as adding "Diagnostic Imaging" and procedure specific searches further improved relevance. Importantly, our *Post-Filtering Agent* consistently identified high-quality studies, achieving a recall of 0.74 which is competitive and in line with prior work, which reports classification accuracies of approximately 60–70% on all categories [3, 78, 106]. This highlights the strong potential of our feature selection and overall model to automate a process that traditionally requires intensive manual review by clinical experts. Notably, our generalization study found that performance remained *comparable* even with an entirely different set of retrieved literature. This resilience to source variation indicates strong robustness and real-world applicability, ensuring reliable performance across diverse and evolving clinical settings.

Generalization Abilities: Generalization is essential for real-world deployment, so we evaluated our models on a separate *generalization set* of five previously unseen conditions, covering different anatomical regions and diagnostic purposes. The *Baseline* model showed only a modest performance drop, whereas the *Citations*, *LLM Eval*, and *MedReason-Embed* models experienced a more noticeable - though still acceptable and expected - decline. This pattern suggests mild sensitivity to out-of-distribution cases and is consistent with prior work [96], which reported similar drops under distribution shifts, and greater robustness in RL-trained models over SFT. Similarly, our SFT model overfitted and failed to generalize, while both reasoning RL-trained models maintained stronger performance on unseen conditions. Overall, these results suggest that our approach can replicate the ACR reasoning process and shows encouraging potential for extension to conditions not explicitly covered by the ACR, while maintaining robustness and clinical reliability. Expanding the generalization set with additional ACR-like resources and conducting deeper error analyses will be key to further improving robustness and understanding the limits of reasoning transfer.

Our Architecture Enables Scalable Deployment: The system was designed with deployment, efficiency, and further extensions in mind. By integrating ICD coding, an industry standard in clinical workflows, it aligns naturally with hospital systems, easing real-world adoption. The evidence-gathering pipeline is fast and adaptable, allowing for the inclusion of emerging research, which is especially important for conditions not covered by the ACR but also support the continuous refinement of recommendation when new (or contradicting) research comes forward. Our post-filtering approach performed well even with simple features and can be further enhanced. The architecture also supports extensions, such as incorporating multiple reasoning agents with distinct roles and expertise (e.g., radiologists, physicians etc.) to better reflect real-world referral dynamics. Its modular, graph-based design ensures scalability, flexibility, and practical integration into clinical settings.

5.2 Limitations and Future Work

During this study, we identified several and challenges and opportunities for further research.

While reasoning-based reward functions improved both performance and alignment, our current methods, relying on embeddings or LLMs, may miss deeper logical errors and domain-specific nuances. A promising direction for future work is to integrate our

approach with Neurosymbolic AI by mapping each reasoning step to structured medical knowledge graphs like UMLS [17], allowing explicit checks that the relationships between medical concepts are accurate and clinically valid and in line with the ACR reasoning. This could provide a more interpretable and rigorous basis for assessing clinical reasoning. Moreover, understanding where and why a model fails requires close collaboration with medical professionals, who can identify errors, gaps in logic, or clinically irrelevant justifications that automated metrics may overlook. Therefore, developing better benchmarks and metrics for clinical reasoning remains an open challenge, and collaboration with clinical experts will be essential to fully understand and evaluate model reasoning.

A key limitation is that some aspects of clinical reasoning require expert knowledge not always captured in the LLM's knowledge and context given, especially since we provide only condensed abstracts rather than full-text papers. Licensing restrictions limit access to full texts, which often contain critical details missing from abstracts and therefore constrained our models' performance. Moreover, our retrieval strategy also relies on a small number of papers and PubMed's default relevance rankings, which may overlook important evidence. To address this, future work could integrate tools like MedGraphRAG [138] for injecting structured domain knowledge through their knowledge graph and also build a large vector database of medical literature (can also be potentially summarized by medical LLMs for efficiency) to provide richer, consistent and more reliable context to the reasoning agent.

Another key limitation is the small size and content of the generalization set. Expanding it is essential to draw more reliable conclusions about the models' generalization abilities. The performance drops we observed were expected and relatively modest, highlighting the promise of our approach. Moving forward, it is important to generate additional generalization cases, ideally in collaboration with clinical experts and extend beyond the ACR to assess reasoning in a broader range of medical scenarios.

An important improvement would be to leverage advanced medical LLMs, such as the recent MedGEMMA [111] as the bakbone model for the *Reasoning Agent*. Many studies also suggest first fine-tuning models on medical data before RL adaptation [136, 148]. These strategies can better align language models with clinical logic and enhance step-wise reasoning before RL.

There are many opportunities to incorporate more data, broadening the applications. While this study is a proof of concept, similar training could be applied to other clinical guidelines, such as the UK NICE guidelines [90], ESR iGuide [35], and Canadian

Association of Radiologists' recommendations [20]. This approach can also extend to other model-derived evidence-based clinical protocols. Recent work has challenged the belief that curating reasoning datasets is prohibitively expensive [68, 84]; verified synthetic datasets created using LLMs, like those developed in MedReason [137] or ReasonMed [120] have already shown scalable, high-quality and effective reasoning chain curation. Both real and synthetic data sources offer significant opportunities for expanding and refining our system.

To conclude, while our primary focus on this study was on medical imaging appropriateness and clinical reasoning, the methodology presented here is not limited to radiology or even to healthcare. The combination of expert-derived reasoning traces, process-supervised RL, and reasoning-centric reward functions offers a generalizable framework for any domain where decisions depend on structured, multi-step reasoning and particularly those with limited data but deep, high-value expertise. By demonstrating that adapted LLMs can remain lightweight and can learn to reason in a verifiable and faithful way, this work lays the groundwork for extending such systems to other fields as well.

5.3   Closing Remarks

Overall, this study demonstrates significant potential and offers novel insights by leveraging RL with ACR guidelines for clinical recommendations, an approach that, to our knowledge, has not been explored in prior work. By uniquely combining RL and reasoning alignment methods, we have unlocked new potential for automated, evidence-based decision support. As a proof of concept with 30 conditions, our research highlights both the promise and practical challenges of this methodology, answering key questions and providing valuable direction for future development. While small-scale, this experiment lays the groundwork for scaling, full deployment, and even commercialization. We are committed to advancing this system, write a publishable paper and fully utilize the breadth of all 257 ACR guidelines (a process already in motion!). Ultimately, our architecture opens new directions for creating scalable, trustworthy, and efficient AI-driven clinical decision support aligned with evidence-based standards.

Bibliography

[1] About pubmed. `https://pubmed.ncbi.nlm.nih.gov/about/`. Accessed: 2025-07-01.

[2] Yasir Abdelgadir, Charat Thongprayoon, Jing Miao, Supawadee Suppadungsuk, Justin H Pham, Michael A Mao, Iasmina M Craici, and Wisit Cheungpasitporn. Ai integration in nephrology: evaluating chatgpt for accurate icd-10 documentation and coding. *Frontiers in Artificial Intelligence*, 7:1457586, 2024.

[3] Wael Abdelkader, Tamara Navarro, Rick Parrish, Chris Cotoi, Federico Germini, Alfonso Iorio, R Brian Haynes, Cynthia Lokker, et al. Machine learning approaches to retrieve high-quality, clinically relevant evidence from the biomedical literature: systematic review. *JMIR medical informatics*, 9(9):e30401, 2021.

[4] M Agrawal, S Hegselmann, H Lang, Y Kim, and D Sontag. Large language models are zero-shot clinical information extractors. arxiv, 2022. *arXiv preprint arXiv:2205.12689*, 2023.

[5] Meta AI. Llama 3.1 405b. `https://huggingface.co/meta-llama/Llama-3.1-405B`, 2025. Accessed: July 2025.

[6] Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`, 2025. Accessed: 2025-07-23.

[7] Haitham Alahmad, Ahmed Hobani, Mohammed Alasmi, Abdulrhman M Alshahrani, Ahmad Abanomy, Mohammad Alarifi, Abdulmajeed Alotaibi, Khaled Alenazi, and Mansour Almanaa. Investigating the potential overuse of pan-computed tomography (panct) examinations in trauma cases in emergency departments. *Medicina*, 60(11):1742, 2024.

[8] Ambience Healthcare. Ambience healthcare, 2025. Accessed: 2025-08-07.

[9] American Academy of Professional Coders (AAPC). What is medical coding? Online, 2022. Accessed: 2025-06-09.

[10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[11] Sheikh Shams Azam, Manoj Raju, Venkatesh Pagidimarri, and Vamsi Chandra Kasivajjala. Cascadenet: An lstm based deep learning model for automated icd-10 coding. In *Future of Information and Communication Conference*, pages 55–74. Springer, 2019.

[12] Krishanu Das Baksi, Elijah Soba, John J Higgins, Ravi Saini, Jaden Wood, Jane Cook, Jack Scott, Nirmala Pudota, Tim Weninger, Edward Bowen, et al. Medcoder: A generative ai assistant for medical coding. *arXiv preprint arXiv:2409.15368*, 2024.

[13] Yiftach Barash, Eyal Klang, Eli Konen, and Vera Sorin. Chatgpt-4 assistance in optimizing emergency department radiology referrals and imaging selection. *Journal of the American College of Radiology*, 20(10):998–1003, 2023.

[14] Andre B Bautista, Anthony Burgos, Barbara J Nickel, John J Yoon, Amish A Tilara, and Judith K Amorosa. Do clinicians use the american college of radiology appropriateness criteria in the management of their patients? *American journal of roentgenology*, 192(6):1581–1585, 2009.

[15] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

[16] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

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

[18] Peter G Brodeur, Thomas A Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdulnour, Adrian Haimovich, Jason A Freed, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv preprint arXiv:2412.10849*, 2024.

[19] Elaine M Burns, E Rigby, R Mamidanna, A Bottle, P Aylin, P Ziprin, and OD Faiz. Systematic review of discharge coding accuracy. *Journal of public health*, 34(1):138–148, 2012.

[20] Canadian Association of Radiologists. CAR Appropriateness Criteria. `https://car.ca/clinical-resources/appropriateness-criteria/`. Accessed: 2025-07-01.

[21] Sherwin S Chan, Michael L Francavilla, Ramesh S Iyer, Cynthia K Rigsby, David Kurth, and Boaz K Karmazyn. Clinical decision support: the role of acr appropriateness criteria. *Pediatric radiology*, 49:479–485, 2019.

[22] Cathy Charles, Amiram Gafni, and Emily Freeman. The evidence-based medicine model of clinical practice: scientific teaching or belief-based preaching? *Journal of evaluation in clinical practice*, 17(4):597–605, 2011.

[23] Harrison Chase and LangChain AI. Langgraph: Library for building stateful, multi-agent workflows with llms. `https://github.com/langchain-ai/langgraph`, 2024. Accessed: 2025-06-18.

[24] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599, 2025.

[25] Yuwen Chen and Jiangtao Ren. Automatic icd code assignment utilizing textual descriptions and hierarchical structure of icd code. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 348–353. IEEE, 2019.

[26] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[27] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro

Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[28] Marco Cremaschi, Davide Ditolve, Cesare Curcio, Anna Panzeri, Andrea Spoto, and Andrea Maurino. Decoding the mind: A rag-llm on icd-11 for decision support in psychology. *Expert Systems with Applications*, 279:127191, 2025.

[29] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

[30] Michael Han Daniel Han and Unsloth team. Unsloth, 2023.

[31] Theony Deshommes, Gabrielle Freire, Natalie Yanchar, Roger Zemek, Marianne Beaudin, Antonia Stang, Matthew John Weiss, Sasha Carsen, Isabelle J Gagnon, Belinda J Gabbe, et al. Low-value clinical practices in pediatric trauma care. *JAMA Network Open*, 7(10):e2440983–e2440983, 2024.

[32] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

[33] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. Automated clinical coding: what, why, and where we are? *NPJ digital medicine*, 5(1):159, 2022.

[34] European Commission. Radiation protection no. 180: Medical radiation exposure of the european population, 2015. Accessed: 2025-03-26.

[35] European Society of Radiology. ESR iGuide: Clinical Decision Support. `https://www.myesr.org/esriguide`. Accessed: 2025-07-01.

[36] Richárd Farkas and György Szarvas. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, pages 1–9. Springer, 2008.

[37] Mohsen Fayyaz, Fan Yin, Jiao Sun, and Nanyun Peng. Evaluating human alignment and model faithfulness of llm rationale. *arXiv preprint arXiv:2407.00219*, 2024.

[38] Kathryn Fitch, Steven J Bernstein, Mary Dolores Aguilar, Bernard Burnand, Juan Ramon LaCalle, Pablo Lazaro, Mirjam van het Loo, Joseph McDonnell, Janneke Vader, and James P Kahan. *RAND/UCLA appropriateness method user's manual*. RAND corporation Santa Monica, CA, 2000.

[39] National Center for Health Statistics (US). *The International Classification of Diseases, 9th Revision, Clinical Modification: Procedures: tabular list and alphabetic index*, volume 3. US Department of Health and Human Services, Public Health Service, Health . . . , 1980.

[40] Gonçalo Gomes, Isabel Coutinho, and Bruno Martins. Accurate and well-calibrated icd code assignment through attention over diverse label embeddings. *arXiv preprint arXiv:2402.03172*, 2024.

[41] Google AI. Gemini thinking models api documentation. `https://ai.google.dev/gemini-api/docs/thinking`, 2024. Accessed: 2025-06-28.

[42] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[43] Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642, 2025.

[44] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[45] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[46] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024.

[47] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[48] R Edward Hendrick and Mark A Helvie. United states preventive services task force screening mammography recommendations: science ignored. *American Journal of Roentgenology*, 196(2):W112–W116, 2011.

[49] Marion Houot, Souraya Arnaud, Marie Mongin, Gabriel Pop, Michaël Soussan, Annie Lannuzel, and Bertrand Degos. Relevance of 123i-fp-cit spect prescriptions for the diagnosis of parkinsonian syndromes. *Scientific Reports*, 14(1):25088, 2024.

[50] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[51] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[52] CW Huang, SC Tsai, and YN Chen. Plm-icd: automatic icd coding with pretrained language models, arxiv. *arXiv preprint arXiv:2207.05289*, 2022.

[53] Heyuan Huang, Alexandra DeLucia, Vijay Murari Tiyyala, and Mark Dredze. Medscore: Factuality evaluation of free-form medical answers. *arXiv preprint arXiv:2505.18452*, 2025.

[54] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[55] Jinmiao Huang, Cesar Osorio, and Luke Wicent Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177:141–153, 2019.

[56] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[57] J Jiang, Z Chen, Y Min, J Chen, X Cheng, J Wang, Y Tang, H Sun, J Deng, WX Zhao, et al. Enhancing llm reasoning with reward-guided tree search. *arXiv*, 2411, 2024.

[58] Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*, 2025.

[59] Pengcheng Jiang, Xueqiang Xu, Jiacheng Lin, Jinfeng Xiao, Zifeng Wang, Jimeng Sun, and Jiawei Han. s3: You don't need that much data to train a search agent via rl. *arXiv preprint arXiv:2505.14146*, 2025.

[60] 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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[61] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Faiss: A library for efficient similarity search and clustering of dense vectors. *arXiv preprint arXiv:1702.08734*, 2017.

[62] Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, et al. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240, 2025.

[63] Joanne SM Kim, Joyce Z Dong, Stacey Brener, Peter C Coyte, and Y Raja Rampersaud. Cost-effectiveness analysis of a reduction in diagnostic imaging in degenerative spinal disorders. *Healthcare policy*, 7(2):e105, 2011.

[64] Eyal Klang, Idit Tessler, Donald U Apakama, Ethan Abbott, Benjamin S Glicksberg, Monique Arnold, Akini Moses, Ankit Sakhuja, Ali Soroush, Alexander W Charney, et al. Assessing retrieval-augmented large language model performance in emergency department icd-10-cm coding compared to human coders. *medRxiv*, 2024.

[65] David A Kurth, Boaz K Karmazyn, Christine A Waldrip, Mythreyi Chatfield, and Mark E Lockhart. Acr appropriateness criteria® methodology. *Journal of the American College of Radiology*, 18(11):S240–S250, 2021.

[66] Keith Kwan. Large language models are good medical coders, if provided with tools. *arXiv preprint arXiv:2407.12849*, 2024.

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

[68] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[69] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[70] Matthew D Lavery, Rylen A Williamson, Jason Curran, July Wilkey, and Kirk McCarroll. Canadian ct head rule adherence in a rural hospital without in-house computed tomography. *Canadian Journal of Rural Medicine*, 29(4):167–172, 2024.

[71] Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. *arXiv preprint arXiv:2502.12289*, 2025.

[72] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*, 2023.

[73] Rumeng Li, Xun Wang, and Hong Yu. Exploring llm multi-agents for icd coding. *arXiv preprint arXiv:2406.15363*, 2024.

[74] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. Making large language models better reasoners with step-aware verifier. *arXiv preprint arXiv:2206.02336*, 2022.

[75] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. From

system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

[76] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3), 2024.

[77] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[78] Jou-Wei Lin, Chia-Hsuin Chang, Ming-Wei Lin, Mark H Ebell, and Jung-Hsien Chiang. Automating the process of critical appraisal and assessing the strength of evidence with information extraction technology. *Journal of evaluation in clinical practice*, 17(4):832–838, 2011.

[79] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[80] Zhiyong Lu. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036, 2011.

[81] Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *arXiv preprint arXiv:2407.00782*, 2024.

[82] Thomas M Maddox, Peter Embí, Jackie Gerhart, Jennifer Goldsack, Ravi B Parikh, and Troy C Sarich. Generative ai in medicine—evaluating progress and challenges. *New England Journal of Medicine*, 2025.

[83] Maura Marin, Flora Maria Murru, Francesco Baldo, Gianluca Tamaro, Elena Faleschini, Egidio Barbi, and Gianluca Tornese. Minimizing unnecessary brain magnetic resonance imaging in pediatric endocrinology: a retrospective cohort analysis. *Frontiers in Endocrinology*, 15:1456541, 2024.

[84] Tyler McDonnell, Mucahid Kutlu, Tamer Elsayed, and Matthew Lease. The many benefits of annotator rationales for relevance judgments. *International Joint Conferences on Artificial Intelligence*, 2017.

[85] Liam McKeever, Van Nguyen, Sarah J Peterson, Sandra Gomez-Perez, and Carol Braunschweig. Demystifying the search button: a comprehensive pubmed search strategy for performing an exhaustive literature review. *Journal of parenteral and enteral nutrition*, 39(6):622–635, 2015.

[86] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[87] George Michalopoulos, Michal Malyska, Nicola Sahar, Alexander Wong, and Helen Chen. Icdbigbird: a contextual embedding model for icd code classification. *arXiv preprint arXiv:2204.10408*, 2022.

[88] Diana L Miglioretti, Eric Johnson, Andrew Williams, Robert T Greenlee, Sheila Weinmann, Leif I Solberg, Heather Spencer Feigelson, Douglas Roblin, Michael J Flynn, Nicholas Vanneman, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA pediatrics*, 167(8), 2013.

[89] Akram Mustafa, Usman Naseem, and Mostafa Rahimi Azghadi. Large language models vs human for classifying clinical documents. *International Journal of Medical Informatics*, page 105800, 2025.

[90] National Institute for Health and Care Excellence. About NICE. `https://www.nice.org.uk/about`. Accessed: 2025-07-01.

[91] Lleayem Nazario-Johnson, Hossam A Zaki, and Glenn A Tung. Use of large language models to predict neuroimaging. *Journal of the American College of Radiology*, 20(10):1004–1009, 2023.

[92] Fatemeh Nazary, Yashar Deldjoo, and Tommaso Di Noia. Chatgpt-healthprompt. harnessing the power of xai in prompt-based healthcare decision support using chatgpt. In *European Conference on Artificial Intelligence*, pages 382–397. Springer, 2023.

[93] Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought

in multi-hop reasoning with knowledge graphs. *arXiv preprint arXiv:2402.11199*, 2024.

[94] American College of Radiology. Acr appropriateness criteria®. `https://acsearch.acr.org/`, 2025. Accessed: 2025-07-02.

[95] OpenMed. Openmed/openmed-ner-pathologydetect-pubmed-v2-109m. `https://huggingface.co/OpenMed/OpenMed-NER-PathologyDetect-PubMed-v2-109M`, 2024.

[96] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

[97] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[98] Badar Patel, Galina Gheihman, Joel T Katz, Arabella Simpkin Begin, and Sonja R Solomon. Navigating uncertainty in clinical practice: a structured approach. *Journal of General Internal Medicine*, 39(5):829–836, 2024.

[99] Juana María Plasencia-Martínez, Marta Sánchez-Canales, Elena Otón-González, Nuria Isabel Casado-Alarcón, Belén Molina-Lozano, Estefanía Cotillo-Ramos, Herminia Ortiz-Mayoral, and José María García-Santos. Inappropriate requests for cranial ct scans in emergency departments increase overuse and reduce test performance. *Emergency Radiology*, 30(6):733–741, 2023.

[100] Jaka Potočnik, Edel Thomas, Aonghus Lawlor, Dearbhla Kearney, Eric J Heffernan, Ronan P Killeen, and Shane J Foley. Machine learning and deep learning for classifying the justification of brain ct referrals. *European Radiology*, 34(12):7944–7952, 2024.

[101] Zhongxi Qiu, Zhang Zhang, Yan Hu, Heng Li, and Jiang Liu. Open-medical-r1: How to choose data for rlvr training at medicine domain. *arXiv preprint arXiv:2504.13950*, 2025.

[102] Arya Rao, John Kim, Meghana Kamineni, Michael Pang, Winston Lie, Keith J Dreyer, and Marc D Succi. Evaluating gpt as an adjunct for radiologic decision making: Gpt-4 versus gpt-3.5 in a breast imaging pilot. *Journal of the American College of Radiology*, 20(10):990–997, 2023.

[103] Alexander Rau, Stephan Rau, Daniela Zoeller, Anna Fink, Hien Tran, Caroline Wilpert, Johanna Nattenmüller, Jakob Neubauer, Fabian Bamberg, Marco Reisert, et al. A context-based chatbot surpasses radiologists and generic chatgpt in following the acr appropriateness guidelines. *Radiology*, 308(1):e230970, 2023.

[104] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[105] José Antonio Salvador-Oliván, Gonzalo Marco-Cuenca, and Rosario Arquero-Avilés. Errors in search strategies used in systematic reviews and their effects on information retrieval. *Journal of the Medical Library Association: JMLA*, 107(2):210, 2019.

[106] Abeed Sarker, Diego Mollá, and Cécile Paris. Automatic evidence quality prediction to support evidence-based decision making. *Artificial intelligence in medicine*, 64(2):89–103, 2015.

[107] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

[108] Natalie Schluter. The limits of automatic summarisation according to rouge. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 41–45. Association for Computational Linguistics, 2017.

[109] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[110] Holger J Schünemann, Andrew D Oxman, Jan Brozek, Paul Glasziou, Roman Jaeschke, Gunn E Vist, John W Williams, Regina Kunz, Jonathan Craig, Victor M

Montori, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Bmj*, 336(7653):1106–1110, 2008.

[111] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[112] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[113] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025. *URL https://arxiv. org/abs/2504.07615*, 2025.

[114] Rebecca Smith-Bindman, Marilyn L Kwan, Emily C Marlow, Mary Kay Theis, Wesley Bolch, Stephanie Y Cheng, Erin JA Bowles, James R Duncan, Robert T Greenlee, Lawrence H Kushi, et al. Trends in use of medical imaging in us health care systems and in ontario, canada, 2000-2016. *Jama*, 322(9):843–856, 2019.

[115] Olawale A Sogbein, Aaron G Chen, J Andrew McClure, Jennifer Reid, Blayne Welk, Brent A Lanting, and Ryan M Degen. Unnecessary interventions for the management of hip osteoarthritis: a population-based cohort study. *Canadian Journal of Surgery*, 67(4):E300, 2024.

[116] Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. Assessing gpt-3.5 and gpt-4 in generating international classification of diseases billing codes. *medRxiv*, pages 2023–07, 2023.

[117] Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040, 2024.

[118] Henriettæ Ståhlbrandt, Ida Björnfot, Torsten Cederlund, and Anja Almén. Ct and mri imaging in sweden: retrospective appropriateness analysis of large referral samples. *Insights into Imaging*, 14(1):134, 2023.

[119] Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025.

[120] Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *arXiv preprint arXiv:2506.09513*, 2025.

[121] Simon Šuster, Timothy Baldwin, Jey Han Lau, Antonio Jimeno Yepes, David Martinez Iraola, Yulia Otmakhova, and Karin Verspoor. Automating quality assessment of medical evidence in systematic reviews: model development and validation study. *Journal of medical Internet research*, 25:e35568, 2023.

[122] Jin Rong Tan, Daniel YZ Lim, Quan Le, Gita Y Karande, Lai Peng Chan, Yeong Huei Ng, Daniel SW Ting, Sudharsan Madhavan, Hiok Yang Chan, Anh NT Tran, et al. Chatgpt performance in assessing musculoskeletal mri scan appropriateness based on acr appropriateness criteria. *Scientific Reports*, 15(1):7140, 2025.

[123] DeepRetrieval Team. Deepretrieval-pubmed-3b. `https://huggingface.co/DeepRetrieval/DeepRetrieval-PubMed-3B`, 2024. Accessed: 2025-06-26.

[124] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[125] Qwen Team. Qwen1.5-1.8b-chat. `https://huggingface.co/Qwen/Qwen1.5-1.8B-Chat`, 2024. Accessed: 2025-07-24.

[126] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

[127] New York Times. Ai and radiologist jobs at mayo clinic. *The New York Times*. Accessed via web; URL restricted to subscribers.

[128] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[129] Dmitriy Umerenkov, Galina Zubkova, and Aleksandr Nesterov. Deciphering diagnoses: how large language models explanations influence clinical decision making. *arXiv preprint arXiv:2310.01708*, 2023.

[130] European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 12 july 2024 on artificial intelligence (ai act), 2024. Official Journal of the European Union, L 168, 12 July 2024.

[131] Unsloth. R1 reasoning: World's first grpo reasoning model, 2024. Accessed: 2025-06-10.

[132] U.S. Food and Drug Administration. Appropriate use, 2022. Accessed: 2025-03-26.

[133] Hanyin Wang, Qiping Xu, Bolun Liu, Guleid Hussein, Hariprasad Korsapati, Mohamad El Labban, Kingsley Iheasirim, Mohamed Hassan, Gokhan Anil, Brian Bartlett, et al. Process-supervised reward models for clinical note generation: A scalable approach guided by domain expertise. *arXiv preprint arXiv:2412.12583*, 2024.

[134] Zixiang Wang, Yinghao Zhu, Junyi Gao, Xiaochen Zheng, Yuhui Zeng, Yifan He, Bowen Jiang, Wen Tang, Ewen M Harrison, Chengwei Pan, et al. Retcare: Towards interpretable clinical decision making through llm-driven medical knowledge retrieval. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.

[135] World Health Organization. World health organization (who) official website, 2025. Accessed: 2025-08-05.

[136] Jianyu Wu, Hao Yang, Xinhua Zeng, Guibing He, Zhiyu Chen, Zihui Li, Xiaochuan Zhang, Yangyang Ma, Run Fang, and Yang Liu. Pathvlm-r1: A reinforcement learning-driven reasoning model for pathology visual-language tasks. *arXiv preprint arXiv:2504.09258*, 2025.

[137] Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

[138] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024.

[139] Michael S Yao, Allison Chae, Charles E Kahn Jr, Walter R Witschey, James C Gee, Hersh Sagreiya, and Osbert Bastani. Evidence is all you need: Ordering imaging studies via language model alignment with the acr appropriateness criteria. *arXiv preprint arXiv:2409.19177*, 2024.

[140] Shahram Yazdani and Maryam Hoseini Abardeh. Five decades of research and theorization on clinical reasoning: a critical review. *Advances in medical education and practice*, pages 703–716, 2019.

[141] Youngju Yoo and Sewon Kim. How to leverage large language models for automatic icd coding. *Computers in Biology and Medicine*, 189:109971, 2025.

[142] Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic icd coding. *arXiv preprint arXiv:2203.01515*, 2022.

[143] Hossam A Zaki, Andrew Aoun, Saminah Munshi, Hazem Abdel-Megid, Lleayem Nazario-Johnson, and Sun Ho Ahn. The application of large language models for radiologic decision making. *Journal of the American College of Radiology*, 21(7):1072–1078, 2024.

[144] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[145] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

[146] Tong Zhou, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Kun Niu, Weifeng Chong, and Shengping Liu. Automatic icd coding via interactive shared representation networks with self-distillation mechanism. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5948–5957, 2021.

[147] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

[148] Wenhui Zhu, Xuanzhao Dong, Xin Li, Peijie Qiu, Xiwen Chen, Abolfazl Razi, Aris Sotiras, Yi Su, and Yalin Wang. Toward effective reinforcement learning fine-tuning for medical vqa in vision-language models. *arXiv preprint arXiv:2505.13973*, 2025.

[149] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

Appendix A

Reasoning Agent

A.1 Reasoning Agent Implementation

All experiments were run on an NVIDIA A100 80GB GPU. We used GRPO with Unsloth [30], which leverages vLLM [67] and LoRA [50] for efficient training [131], with rank set to 32. For LoRA adaptation, we injected low-rank trainable matrices into key transformer components (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) while keeping the rest of the model frozen, enabling memory-efficient fine-tuning with gradient checkpointing for long-context reasoning. Base hyperparameters such as the learning rate and optimizer were unchanged, but we set `per_device_train_batch_size` to 6 and `gradient_accumulation_steps` to 2 to optimize GPU usage. We tested different values for `num_generations`, which controls how many candidate outputs are generated per input to promote diversity and enhance learning, and found that 6 offered the best balance between performance and efficiency. Models were trained for 2 epochs, with the Baseline model completing in approximately 2 hours.

A.2 Reward functions in more detail

- **Answer Reward** ($r_{\mathbf{ans}}$): Binary reward for correctly predicting the appropriateness label. *Purpose:* Ensures correct clinical recommendations.

- **Format Reward** ($r_{\mathbf{fmt}}$): Binary reward for using proper `<think>` and `<answer>` tags. *Purpose:* Enforces consistent output formatting to support structured thinking and improve performance.

- **LLM Evaluator Reward** ($r_{\mathbf{LLM}}$): LLM scores reasoning alignment with gold

examples (scaled to 0–1 scale; Fig. A.2), rewarding medically relevant, expert-like reasoning (*LLM-Eval* model).

- **MedReason-Embed Reward (r_{joint}):** Combines answer correctness with reasoning trace alignment (avg. max cosine similarity × binary correctness), promoting both correct answers and well-aligned reasoning (*MedReason-Embed* model).

Figure A.1: The LLM-eval based reward given by prompting Qwen1.5-1.8B to assign a score based on alignment of ACR "gold" and generated reasonings.

A.3 Prompt used for RL training

Figure A.2: The prompt given to train the RL models *Citations*, *LLM Eval* and *MedReason-Embed*

A.4   Extracting reasoning traces

Figure A.3: Prompt for LLaMA 4 Scout 17B 16E to extract reasoning traces from the complex and lengthy ACR reasoning

A.5   ACR Processing

The following conditions were selected from the ACR for training, test and generalization sets.

Figure A.4: All the ACR conditions in our datasets (train/test and generalization)

The end train/test dataset had respective value counts for each condition-variant-procedure triplet: 'Usually Not Appropriate': 1115, 'Usually Appropriate': 321, 'May Be Appropriate': 300, total size 1736. The train/test split was done on a 70/30 ratio. The generalization dataset had a similar distribution with respective value counts for each condition-variant-procedure triplet: 'Usually Not Appropriate': 175, 'Usually Appropriate': 52, 'May Be Appropriate': 32, total size 259.

A.6 SFT training details

We trained our SFT model using the trl library's 'SFTTrainer', aligning all key hyperparameters with those in the GRPO training phase to ensure a fair and consistent comparison. Specifically, we used the same learning rate (5e-6), cosine learning rate scheduler, weight decay (0.1), and memory-efficient 8-bit optimizer (adamw_8bit) across both training regimes. The maximum sequence length, batch size configuration (via gradient accumulation), and tokenization strategies were also matched. We set max_steps=300 to approximately equal 2 full epochs over the training data, to match the RL model.

A.7 Pairwise McNemar's test p-values

	Base	Citations	MedReason-Embed	LLM Eval
Base	—	0.0000	0.0000	0.0000
Citations		—	0.0000	0.0000
MedReason-Embed			—	0.4222
LLM Eval				—

Table A.1: Pairwise McNemar's test p-values between RL models. Values below 0.01 indicate statistically significant differences

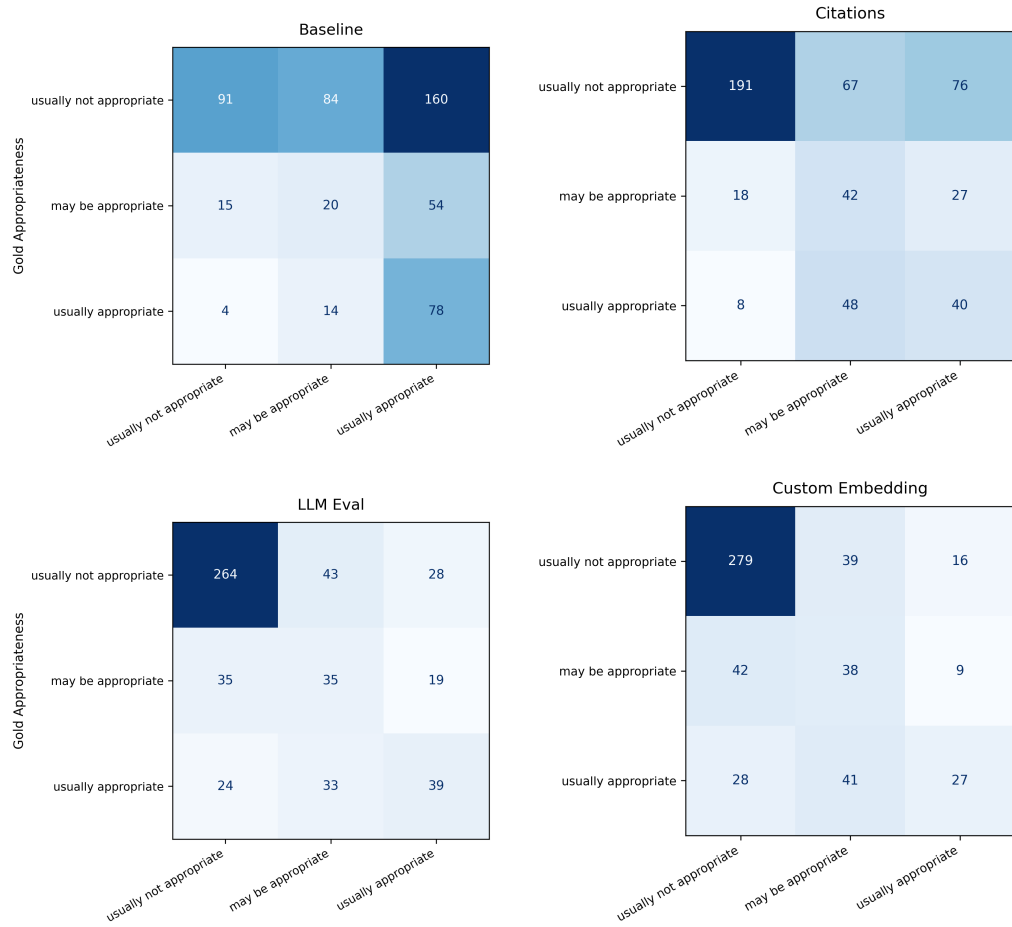## A.8   Confusion matrices and other metrics

Figure A.5: Confusion matrices for the four RL adapted models

The *Baseline* model performs poorly, with strong confusion between "Usually Appropriate" and "Usually Not Appropriate", which is the most critical misclassification in a clinical context, highlighting its lack of clinical reasoning. In contrast, the *Citations* model shows better separation, particularly by anchoring predictions more toward the "May Be Appropriate" category, suggesting that grounding in evidence helps avoid extreme misclassifications. The *LLM Eval* model further reduces false positives in the "Usually Appropriate" class, indicating improved caution in uncertain cases. Finally, the *MedReason-Embed* model achieves the best overall balance, with clearer boundaries across all three classes and particularly for "Usually Not Appropriate".
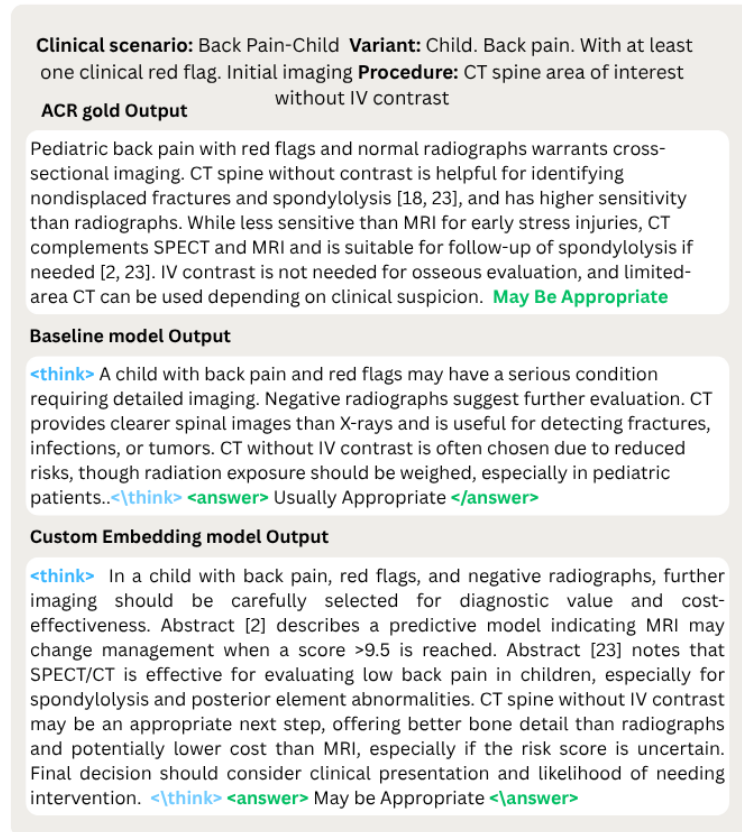
A.9   More reasoning examples

Clinical scenario: Back Pain-Child **Variant:** Child. Back pain. With at least one clinical red flag. Initial imaging **Procedure:** CT spine area of interest without IV contrast

ACR gold Output

Pediatric back pain with red flags and normal radiographs warrants cross-sectional imaging. CT spine without contrast is helpful for identifying nondisplaced fractures and spondylolysis [18, 23], and has higher sensitivity than radiographs. While less sensitive than MRI for early stress injuries, CT complements SPECT and MRI and is suitable for follow-up of spondylolysis if needed [2, 23]. IV contrast is not needed for osseous evaluation, and limited-area CT can be used depending on clinical suspicion. **May Be Appropriate**

Baseline model Output

<think> A child with back pain and red flags may have a serious condition requiring detailed imaging. Negative radiographs suggest further evaluation. CT provides clearer spinal images than X-rays and is useful for detecting fractures, infections, or tumors. CT without IV contrast is often chosen due to reduced risks, though radiation exposure should be weighed, especially in pediatric patients..<\think> <answer> Usually Appropriate </answer>

Custom Embedding model Output

<think> In a child with back pain, red flags, and negative radiographs, further imaging should be carefully selected for diagnostic value and cost-effectiveness. Abstract [2] describes a predictive model indicating MRI may change management when a score >9.5 is reached. Abstract [23] notes that SPECT/CT is effective for evaluating low back pain in children, especially for spondylolysis and posterior element abnormalities. CT spine without IV contrast may be an appropriate next step, offering better bone detail than radiographs and potentially lower cost than MRI, especially if the risk score is uncertain. Final decision should consider clinical presentation and likelihood of needing intervention. <\think> <answer> May be Appropriate <\answer>

Figure A.6: Reasoning Alignment and Appropriateness Accuracy: MedReason-Embed vs. Baseline Model

For illustration, we show a condensed example of the gold reasoning and outputs from two models. The *MedReason-Embed* model gives the correct recommendation ("May Be Appropriate") and closely mirrors ACR reasoning by using contextual evidence and citing relevant studies. In contrast, the *Baseline* model offers a generic explanation and the wrong label ("Usually Appropriate"). This highlights how evidence context and reasoning rewards improve both reasoning quality and final decisions.

Appendix B

DeepRetrieval implementation

B.1   Retrieval Strategy: Clustering Analysis example
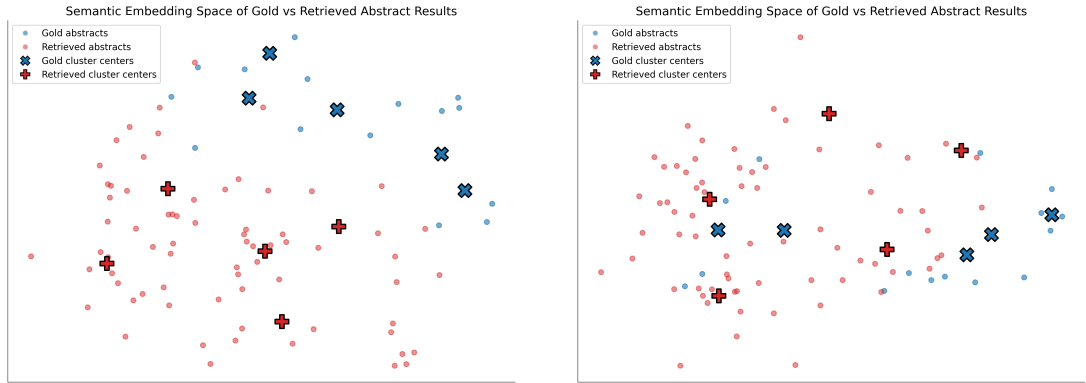
A small example illustrating our process for designing a search strategy using the semantic embedding space of the conclusion sections of retrieved and gold evidence projected into a 2D space for the condition *Breast Pain*. On the left, we apply a basic retrieval strategy using a single generic clinical query rewritten by DeepRetrieval. On the right, we enhance the query strategy using multiple queries for DeepRetrieval rewriting and also incorporating relevant terms such as "Diagnostic Imaging" to refine the search, which produces retrieved references more closely aligned with the gold citations.

B.2   Recommended query strategies

Input Queries to DeepRetrieval

[Breast Pain]
Diagnostic imaging for [Breast Pain]
Clinical evidence for the use of diagnostic imaging in the evaluation of [Breast Pain]
[Breast Pain] and related conditions
Affected conditions [Breast Pain] and synonyms
Other way to say [Breast Pain]
Clinical evidence for the use of [Mammography] in the evaluation of [Breast Pain]
P: [Breast Pain] ; I: [Mammography]; C: Alternative imaging procedures; O:Diagnostic accuracy, risk and benefits"

Rewritten Queries by DeepRetrieval

((Breast Pain OR Mastalgia) AND (Breast Cancer OR Breast Cancer Risk OR Cancer Risk) AND (Diagnostic Imaging))
((Breast Pain OR Mastalgia) AND (Imaging OR Mammography OR Magnetic Resonance Imaging OR Ultrasound OR Breast Imaging) AND (Diagnostic Imaging))
((Mammography OR mammography) AND (Breast Pain OR Mastalgia OR Mastodynia) AND (Diagnostic Imaging))

Figure B.1: The suggested input queries to DeepRetrieval with an example for the **Breast Pain** condition and **Mammography** procedure. Then a *sample* of the rewritten queries from DeepRetrieval with our addition of the term "Diagnostic Imaging"

Appendix C

SOE predictor

Figure C.1: Features used for modeling the SOE predictor
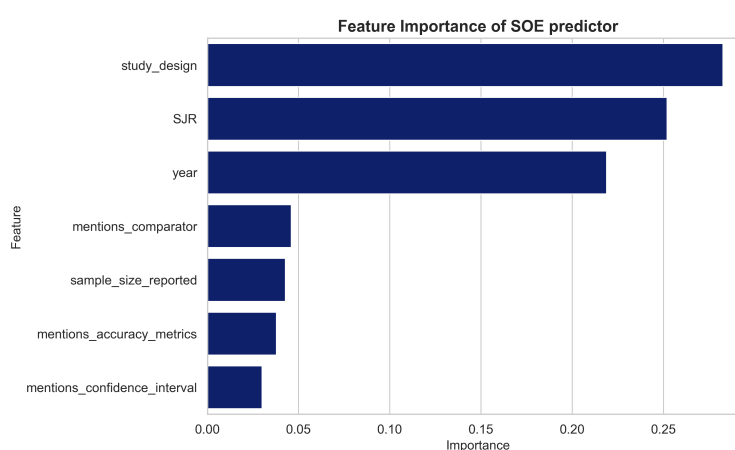
Figure C.2: Feature importances observed using a Random Forest predictor

Appendix D

ICD Coding

Figure D.1: LLM prompt for LLaMA 3.1 8B, used to convert noisy Italian clinical notes into ICD-compatible language for the RAG component.
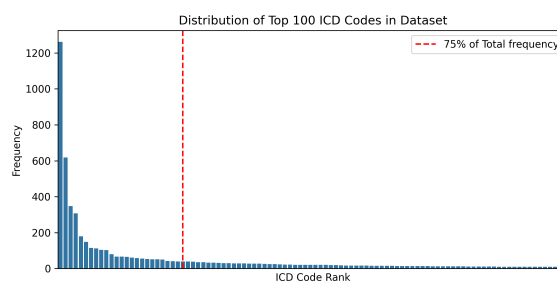

Figure D.2: **Distribution of the top 100 ICD codes:** The red dashed line shows where cumulative frequency reaches 75% of occurrences (at code rank 25), highlighting the long-tail pattern seen in ICD distributions.