Error Correction in Radiology Reports: A Knowledge Distillation-Based Multi-Stage Framework

Jinge Wu[1*], Zhaolong Wu[2*], Ruizhe Li[3], Tong Chen[4], Abul Hasan[5], Yunsoo Kim[1], Jason Pui-Yin Cheung[2 †], Teng Zhang[2 †], Honghan Wu[1,6 †]

[1]University College London,
[2]The University of Hong Kong,
[3]University of Aberdeen,
[4]The University of Sydney,
[5]University of Oxford,
[6]University of Glasgow
honghan.wu@ucl.ac.uk, {cheungjp, tgzhang}@hku.hk

Abstract

The increasing complexity and workload of clinical radiology leads to inevitable oversights and mistakes in their use as diagnostic tools, causing delayed treatments and sometimes life-threatening harm to patients. While large language models (LLMs) have shown remarkable progress in many tasks, their utilities in detecting and correcting errors in radiology reporting are limited. This paper proposes a novel dual-knowledge infusion framework that enhances LLMs' capability for radiology report proofreading through systematic integration of medical expertise. Specifically, the knowledge infusion combines medical knowledge graph distillation (MKGD) with external knowledge retrieval (EXKR), enabling an effective automated approach in tackling mistakes in radiology reporting. By decomposing the complex proofreading task into three specialized stages of detection, localization, and correction, our method mirrors the systematic review process employed by expert radiologists, ensuring both precision and clinical interpretability. To perform a robust, clinically relevant evaluation, a comprehensive benchmark is also proposed using real-world radiology reports with real-world error patterns, including speech recognition confusions, terminology ambiguities, and template-related inconsistencies. Extensive evaluations across multiple LLM architectures demonstrate substantial improvements of our approach: up to 31.56% increase in error detection accuracy and 37.4% reduction in processing time. Human evaluation by radiologists confirms superior clinical relevance and factual consistency compared to existing approaches.

code — https://github.com/knowlab/MedKIC-Radiology-Proofreading

Introduction

The landscape of medical documentation is undergoing rapid transformation, with radiology reports becoming increasingly complex to support precision medicine and comprehensive patient care (Brady 2017; Wu et al. 2022; Collins and Varmus 2015; Topol 2019). However, this evolution has introduced significant challenges that threaten documentation quality and reliability, with critical implications for patient safety and healthcare efficiency.

Retrospective studies indicate that approximately 30% of radiological examinations contain documentation errors, ranging from terminology inconsistencies and negation mistakes to contextual contradictions (Kasalak et al. 2023; Berlin 2007; Wu, Kim, and Wu 2024; Wu et al. 2023). These errors manifest through speech recognition misinterpretations (e.g., "effusion" becoming "infusion") and template-based inconsistencies, leading to delayed diagnoses, inappropriate treatments, and compromised patient safety (Pinto dos Santos et al. 2018). The clinical consequences of such errors extend beyond immediate patient care, affecting treatment planning, follow-up decisions, and medicolegal documentation.

Manual review processes are becoming increasingly unsustainable as examination volumes increase by 3-5% annually while specialist availability grows at less than 2% per year (Sunshine and Burkhardt 2006; Kruskal et al. 2011; Wu, Hasan, and Wu 2024). This growing workload burden particularly affects busy clinical environments and resource-constrained healthcare facilities, where comprehensive manual proofreading becomes impractical, leading to potential quality compromises in medical documentation.

Current automated approaches face four critical limitations that hinder their clinical adoption and effectiveness. Existing methods lack systematic medical knowledge integration, operating with linguistic correctness that ignores clinical appropriateness—unable to distinguish between medically distinct terms like "consolidation" and "atelectasis" (Huang, Altosaar, and Ranganath 2019; Wang et al. 2018; Wu et al. 2024). They provide black-box decision-making incompatible with clinical environments where practitioners must validate recommendations (Cab-

*These authors contributed equally.
†These authors are corresponding authors.

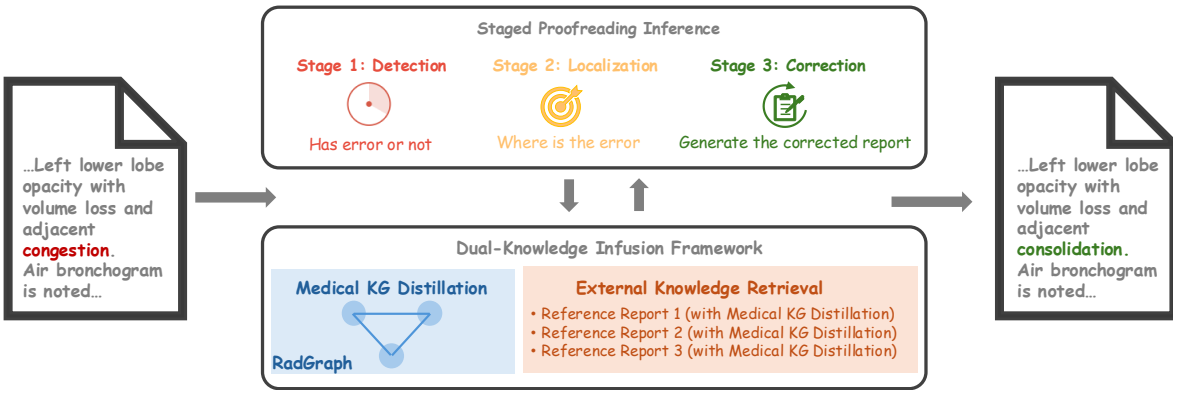

Figure 1: An overview of our medical report proofreading framework. The staged inference process (top) breaks down error correction into detection (identifying error presence), localization (pinpointing error terms like "congestion"), and correction (providing proper replacements like "consolidation"). The dual-knowledge infusion framework (bottom) supports this process through MKGD's structural analysis and EXKR's domain knowledge integration, enabling accurate and clinically sound corrections.

itza, Rasoini, and Gensini 2017; Sendak, Balu, and Schulman 2020). Additionally, these approaches treat error correction as monolithic tasks without modeling expert cognitive processes, and rely on static training paradigms that cannot adapt to evolving medical knowledge like COVID-19 terminology (Ng et al. 2020)—limiting their utility in dynamic clinical practice.

To address these challenges, we present a novel framework combining staged proofreading inference with dual-knowledge infusion, as illustrated in Figure 1. Our staged approach decomposes error correction into detection, localization, and correction phases, mirroring expert review processes while enabling transparent decision-making that can be validated by clinical practitioners. Our dual-knowledge mechanism combines medical knowledge graph distillation (MKGD) with external knowledge retrieval (EXKR), transforming clinical reports into structured representations and leveraging clinically relevant reference cases without expensive fine-tuning—providing scalable quality assurance that enhances rather than replaces clinical expertise.

Our work makes four key contributions with demonstrated benefits for clinical practice: 1) **Methodological Innovation**—staged inference framework enabling transparent AI-assisted decision support that integrates seamlessly with clinical workflows, 2) **Knowledge Integration Architecture**—dual-knowledge infusion for scalable domain adaptation that leverages existing medical knowledge without requiring extensive retraining, 3) **Clinical Validation Framework**—comprehensive benchmark with authentic error patterns validated by practicing radiologists, ensuring real-world applicability and clinical relevance, and 4) **Demonstrated Clinical Impact**—significant improvements (up to 31.56% in error detection, 37.4% processing time reduction) that directly enhance report quality and reduce radiologist workload, particularly benefiting high-volume clinical environments where efficiency gains translate to improved patient care.

Related Work

Research in automated medical text error detection reveals critical limitations that highlight the need for more sophisticated clinical solutions.

General Text Correction Approaches achieve success in linguistic domains (Fang et al. 2024; Kamoi et al. 2024) but fail in medical contexts where clinical appropriateness supersedes linguistic correctness. While correcting "recieve" to "receive," they cannot distinguish between "consolidation" and "atelectasis"—legitimate medical terms representing different pathophysiological processes requiring distinct treatments.

Medical Domain-Specific Approaches like MEDIQA-CORR 2024 (Ben Abacha et al. 2024) focus on general clinical notes using synthetic data rather than specialized radiology reports. Recent medical language models show promise through fine-tuning (Zhou et al. 2023; Abacha et al. 2023a) but operate as black boxes incompatible with clinical environments where radiologists must validate automated recommendations.

Knowledge-Enhanced Methods incorporate structured medical knowledge like UMLS concepts (Rajpurkar et al. 2022) but treat error correction as monolithic tasks without modeling expert cognitive processes. A radiologist would systematically: (1) assess error presence, (2) identify "congestion" as inappropriate in pulmonary context, and (3) correct to "consolidation." Current methods cannot decompose this reasoning chain.

Static Training Limitations present the most fundamental challenge. Current approaches freeze medical knowledge at training time, creating vulnerabilities in rapidly evolving practice (Lazaridou et al. 2021; Sun et al. 2025; Singhal et al. 2023). Models struggle with novel terminology like "COVID pneumonia" and cannot incorporate evolving guidelines without expensive retraining.

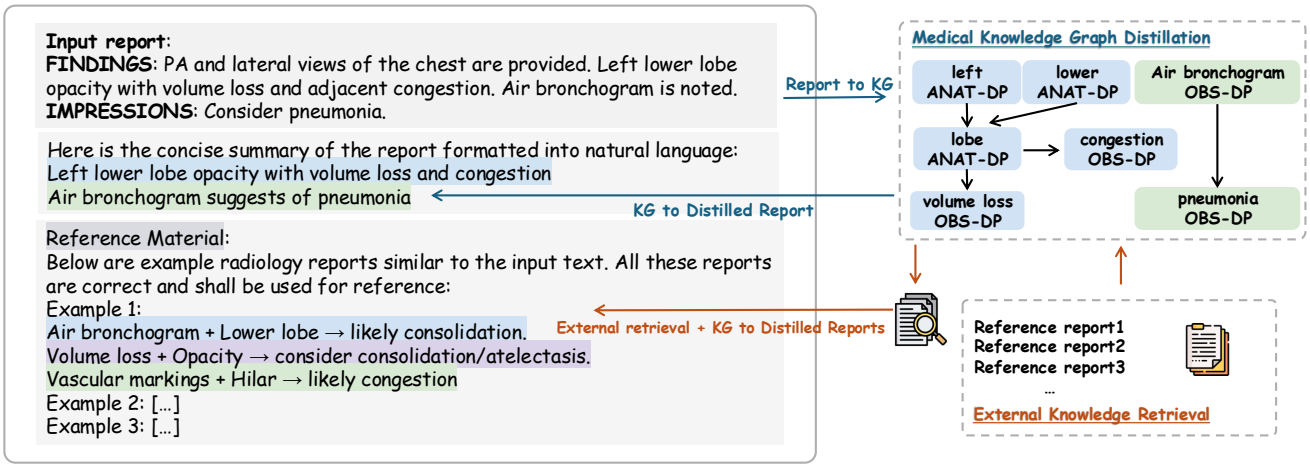

Figure 2: Illustration of our dual-knowledge infusion framework. Left: Input medical report with task description and reference examples. Right: MKGD transforms the report into a structured graph representation capturing anatomical entities (ANAT-DP) and observations (OBS-DP/DA) with their relationships (modify, located_at, suggestive_of), while EXKR provides relevant domain knowledge from reference reports to guide the correction process.

Methodology

Staged Proofreading Inference

We decompose error correction task into three specialized stages that mirror expert radiologist review processes, enabling transparent and clinically-grounded reasoning.

- **Stage 1: Error Detection** performs binary classification to identify whether a report contains errors. The model analyzes global patterns and medical consistency indicators to distinguish between correct and problematic reports, focusing on medical-specific characteristics that indicate potential documentation issues
- **Stage 2: Error Localization** conducts fine-grained analysis to pinpoint specific error locations within reports identified as problematic. By examining individual medical entities, relationships, and broader clinical context to identify the precise terms or phrases requiring correction.
- **Stage 3: Error Correction** generates clinically appropriate corrections based on detected error types and locations. The correction process is guided by both local context and broader medical knowledge to ensure medical accuracy and report consistency while maintaining clinical validity.

Each stage is designed with medical-specific reasoning mechanisms that enable precise and reliable error handling while maintaining broader clinical contexts. This structured decomposition allows the model to focus on specific aspects of medical error analysis at each stage, leading to more accurate and interpretable results compared to generic language model applications. Moreover, this step-by-step approach helps reduce hallucination and encourages more targeted corrections that align with medical knowledge and practices.

Dual-Knowledge Infusion Framework

Detecting domain-specific errors requires bridging local textual coherence with medical knowledge that extends beyond the immediate document. To that end, we propose a *dual-retrieval* mechanism comprised of medical knowledge graph distillation (MKGD) with external knowledge retrieval (EXKR), as shown in Figure 2.

Medical Knowledge Graph Distillation (MKGD) We transform clinical reports into structured medical entity graphs using RadGraph (Jain et al. 2021), a specialized framework designed for clinical information extraction from radiology reports. RadGraph has been extensively validated and demonstrates high performance in extracting clinically relevant entities and relationships from chest X-ray reports.

Entity Extraction identifies two main categories that capture the fundamental components of radiological descriptions. Anatomical entities (ANAT) represent specific body structures mentioned in reports, encompassing organs, tissues, and anatomical landmarks such as lungs, ribs, cardiac structures, and vascular components. Observational entities (OBS) capture clinical findings and are systematically classified into three distinct certainty levels that preserve diagnostic confidence: Definitely Present (DP) indicates confirmed findings with high clinical certainty, Uncertain (U) represents possible or equivocal findings requiring further evaluation, and Definitely Absent (DA) explicitly documents ruled-out conditions or normal findings. This classification system ensures that the extracted knowledge graph preserves the nuanced diagnostic reasoning inherent in radiological interpretations.

Relationship Modeling captures three fundamental relation types that form the semantic structure of medical knowledge. The *suggestive of* relation encodes diagnostic reasoning chains by linking observational findings to their potential clinical implications, such as when specific imaging patterns suggest particular pathological conditions. The *located at* relation establishes precise anatomical localization by connecting observational findings to their corresponding anatomical sites, enabling accurate spatial mapping of

clinical findings within body structures. The *modify* relation captures hierarchical and descriptive relationships between entities, including anatomical modifiers that specify spatial characteristics (e.g., "left", "lower", "bilateral") and observational qualifiers that describe finding attributes such as severity ("mild", "severe") or morphological characteristics ("nodular", "linear").

Graph-to-Text Conversion transforms the structured representation back into human-readable sentences through systematic rule-based integration that preserves both clinical accuracy and linguistic naturalness. The conversion process operates through several coordinated phases: entity classification categorizes nodes by type and certainty level, semantic integration combines related entities into meaningful phrases through hierarchical sorting of modifiers according to clinical conventions (e.g., combining ⟨lower, modify, lobe⟩ and ⟨opacity, located at, lobe⟩ into "lower lobe opacity"), logical reasoning applies domain-specific transformations with attention to negations and uncertainty indicators (converting ⟨opacity, located at, lobe⟩ with DA certainty to "no lobe opacity"), and sentence construction groups phrases by anatomical regions while applying grammatical rules to generate well-formed sentences. This systematic approach ensures that the generated natural language maintains both clinical accuracy and readability while preserving the semantic richness of the original knowledge graph representation.

External Knowledge Retrieval (EXKR) In addition to knowledge graph distillation, our method integrates knowledge from external clinical sources by leveraging a curated database of error-free reference reports to provide real-world expertise and domain-specific knowledge patterns. This external knowledge component addresses the limitation of analyzing reports in isolation by incorporating broader clinical experience and established medical knowledge patterns.

Reference Selection employs semantic similarity matching using e5-large-unsupervised embeddings to identify the top-k most relevant reference reports based on cosine similarity scores with the input report. We empirically determined that k=4 provides optimal performance across different model architectures while maintaining computational efficiency. This similarity-based approach ensures that retrieved references are topically relevant and contextually appropriate for the specific clinical scenario under analysis.

Knowledge Standardization processes selected reference reports through the same entity extraction and graph-to-text conversion pipeline used for input reports, creating structured natural language statements that provide domain-specific knowledge rules and clinical patterns. This standardized processing ensures consistent knowledge representation while enabling direct comparison between input reports and reference examples. The resulting knowledge statements capture both explicit medical relationships (e.g., "air bronchogram + lower lobe $\rightarrow$ likely consolidation") and implicit clinical patterns derived from validated practice.

Contextual Integration combines the standardized reference knowledge with input report analysis to provide comprehensive medical context. This integration process considers both the structural similarities captured through MKGD representations and the semantic relationships identified through EXKR patterns, enabling detection of subtle inconsistencies that might be missed by either approach alone. The dual-level framework leverages both the internal logic of individual reports and the accumulated wisdom of clinical experience encoded in reference examples.

Integration with LLM

We carefully design prompt engineering strategies that integrate both knowledge sources into natural language instructions, enabling effective communication of complex medical context to language models while maintaining clarity and focus. The prompt template architecture follows a consistent structure across all three stages: (1) professional role definition establishing the LLM as a radiologist with specific expertise in chest radiology and diagnostic report writing, (2) task-specific instructions tailored to the current stage's objectives, clearly specifying the expected analysis focus and decision criteria, (3) structured knowledge integration that presents both MKGD-generated summaries and EXKR reference examples in organized, digestible formats, and (4) explicit output format specifications that ensure consistent responses suitable for downstream processing and clinical validation.

Benchmark Data Construction

As there was no prior study for this task when we started this work, we constructed a comprehensive benchmark that reflected real-world clinical scenarios by introducing real-world scenario derived error patterns into validated radiology reports. This approach ensured that our evaluation captured clinically relevant documentation challenges, providing meaningful assessment of system performance in clinical contexts.

The dataset foundation was the MIMIC-CXR dataset (Johnson et al. 2019), a large-scale collection of real-world radiology reports. We constructed two distinct sets: (1) a reference set of 112,251 error-free radiology reports that serve as the knowledge base for our framework's EXKR component, and (2) an evaluation set of 1,622 radiology reports comprising 512 error-free examples for baseline comparison and 1,110 reports containing systematically introduced errors for comprehensive testing. The following three strategies were adopted to introduce errors. Although the original reports are assumed to be error-free, minor real-world inaccuracies may still exist.

General Error Introduction Strategy focused on clinically realistic scenarios that reflect real-world documentation challenges encountered in clinical practice. We introduced one error per report in either Findings or Impression section. This was to ensure that errors could be detected using cross-sectional information available within the report. This constraint ensured that our evaluation focused on text-based error detection capabilities while maintaining clinical plausibility. The single-error constraint also enabled precise attribution of detection and correction performance to specific types and locations of errors.

Model	Error Detection (Acc%)		Error Localization (Acc%)		Error Correction (AggNLG)		
	Baseline	Our Method (MKGD+EXKR)	Baseline	Our Method (MKGD+EXKR)	Baseline (End-to-End)	Baseline (Staged)	Our Method (MKGD+EXKR)
Medical Domain Models							
MMedLM2	41.49	**73.05** (+31.56)	30.94	**46.05** (+15.11)	47.80	58.27	**53.50** (-5.70)
Llama3-Aloe	45.31	**67.26** (+21.95)	43.34	**51.35** (+8.01)	63.33	90.17	**74.77** (+11.44)
General Purpose Models							
Phi3-mini	67.26	**73.06** (+5.80)	47.71	**52.65** (+4.94)	74.36	74.08	**78.85** (+4.49)
Phi3-small	79.03	**80.21** (+1.18)	63.44	**65.04** (+1.60)	80.03	86.57	**86.67** (+6.64)
Phi3-medium	73.67	**79.04** (+5.37)	69.73	63.44 (-6.29)	84.47	90.25	**92.25** (+7.78)
Llama3-8B	37.79	**62.27** (+24.48)	37.29	**53.14** (+15.85)	84.34	94.29	**94.43** (+10.09)
Average	57.43	**72.48 (+15.05)**	48.74	**55.28 (+6.54)**	72.39	82.27	**80.08 (7.69)**

Table 1: Overall Performance Comparison: Results of the staged inference framework with dual-knowledge infusion (MKGD + EXKR). For Error Detection and Error Localization, values represent accuracy (%). For Error Correction, results are reported using the aggregate NLG score (AggNLG), computed as the average of ROUGE-1, BERTScore, and BLEU. Two baselines are compared: (1) End-to-End, where the model directly generates corrected reports, and (2) Staged Baseline, which applies staged inference without knowledge infusion. Our method adds dual-knowledge infusion (MKGD + EXKR) on top of the staged framework.

Negation Error Implementation was implemented by converting positive findings to negative assertions or vice versa, creating errors that could fundamentally alter clinical interpretations. We utilized RadGraph's negation detection capabilities to identify sentences containing explicit negation markers such as "no," "without," "absence of," or "ruled out." The error generation process removed these negation indicators to convert negative findings into positive assertions (e.g., "no pleural effusion" becomes "pleural effusion"), or conversely added negation markers to positive findings to create false negative statements. This approach ensured that negation errors were introduced in syntactically appropriate locations while preserving overall sentence structure and medical terminology, creating realistic documentation errors that reflect common transcription and template-related mistakes.

Entity-Based Clinical Inconsistency Generation was a strategy to introduce terminology confusions that reflect real-world documentation challenges. Working with practicing radiologists, we identified 12 critical radiological findings that commonly appear in chest X-ray reports: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, and Pneumothorax. For each finding, we established categorized replacement alternatives: speech recognition/spelling confusions reflected common transcription errors (e.g., "effusion" $\rightarrow$ "infusion"), terminology ambiguity represented similar medical terms that may be confused in clinical documentation (e.g., "congestion" $\rightarrow$ "consolidation"), template-related terms captured standardized alternatives used in reporting systems (e.g., "cardiac enlargement" $\rightarrow$ "cardiomegaly"), and other clinical conditions represented related but distinct diagnoses that might be inappropriately substituted.

All introduced errors were validated by practicing radiologists to ensure clinical plausibility and realistic documentation challenges, providing a comprehensive evaluation framework that reflects real-world clinical scenarios.

Experiments

For evaluation, we select diverse state-of-the-art LLMs spanning general and medical domains. From the general domain, we use LLaMA3-8B and Phi3 variants (mini-3.8B, small-7B, medium-14B) (Abdin et al. 2024), allowing us to study the impact of model scale. We also evaluate two medical-specific models, Llama3-Aloe-8B-Alpha and MMedLM2 (Qiu et al. 2024; Gururajan et al. 2024). These models built on LLaMA3-8B, and have been further fine-tuned on a wide range of medical instructional datasets, synthetic medical data, medical textbooks, medical websites, etc.

Our evaluation follows the three stages of our framework. For the first two stages, we calculate accuracy for error detection and localization. Regarding the error correction stage, which is only triggered when errors are detected, we employ an aggregate Natural Language Generation (NLG) score (AggNLG) that combines ROUGE-1 (Lin 2004), BERTScore (Zhang et al. 2019), and BLEU (Sellam, Das, and Parikh 2020):

$$\text{AggNLG} = \frac{\text{ROUGE-1} + \text{BERTScore} + \text{BLEU}}{3} \quad (1)$$

This composite metric has been used in the related work with Abacha et al. (2023b) and demonstrated strong correlation with human judgment.

Regaring our staged approach, correction is only performed when errors are detected in the first stage. In this

Model	Error Detection (Acc%)			Error Localization (Acc%)			Error Correction (AggNLG)			Processing Time (s)		
	RAG	Ours	Δ	RAG	Ours	Δ	RAG	Ours	Δ	RAG	Ours	Gain%
Medical Domain Models												
MMedLM2	52.65	**73.05**	**+20.40**	35.88	**46.05**	**+10.17**	48.13	**53.50**	**+5.37**	30.51	**22.00**	**27.90**
Llama3-Aloe	58.81	**67.26**	**+8.45**	44.51	**51.35**	**+6.84**	74.28	**74.77**	**+0.49**	25.60	**16.10**	**37.30**
General Purpose Models												
Phi3-mini	65.59	**73.06**	**+7.47**	42.97	**52.65**	**+9.68**	75.76	**78.85**	**+3.09**	24.10	**15.50**	**35.60**
Phi3-small	72.51	**80.21**	**+7.70**	58.93	**65.04**	**+6.11**	84.56	**86.67**	**+2.11**	25.00	**15.90**	**36.40**
Phi3-medium	73.13	**79.04**	**+5.91**	60.91	**63.44**	**+2.53**	88.87	**92.25**	**+3.38**	31.40	**21.90**	**30.10**
Llama3-8B	51.84	**62.27**	**+10.43**	46.92	**53.14**	**+6.22**	88.34	**94.43**	**+6.09**	23.20	**14.50**	**37.40**
Average	**62.42**	**72.48**	**+10.06**	**48.35**	**55.28**	**+6.93**	**76.66**	**80.08**	**+3.42**	**26.64**	**17.65**	**34.10**

Table 2: Performance and Efficiency Comparison between Simple RAG and Our Method. Our Method refers to the dual-knowledge infusion framework (MKGD + EXKR). Δ represents absolute improvement in accuracy/score, while Gain% shows processing time reduction.

case, we compute AggNLG scores only when both the model output and ground truth indicate the presence of errors. The generally high NLG scores observed in our results can be attributed to the nature of radiological report errors, which typically involve modifying only a few key terms while preserving the majority of the report content.

Experiment Setups. All the experiments are conducted with 4 NVIDIA GeForce RTX 3090 24576MiB. For our LLM experiments, we utilize the AutoModelForCausalLM from the Hugging Face Transformers library. The model is loaded from the specified model_id, with the torch_dtype parameter set to torch.bfloat16. We set the max_new_tokens=300, do_sample=True, temperature=0.001 and top_p=0.8, all other hyperparameters remains unchanged as default values.

RAG Pipeline. We use the e5-large-unsupervised model (Wang et al. 2022) to transform reports into a vectorized database. This process involves applying cosine similarity to find the most similar texts based on the input radiology reports. The experiment uses parameters set to chunk_size=1000 and chunk_overlap=100. We choose top k = 4 reports for the RAG evaluation as it in genral achieves the best performance.

Overall Performance

As shown in Table 1, our framework consistently outperforms baselines across all three tasks. The largest gains are observed in error detection (+15.05%) and error localization (+6.54%), demonstrating that structured, stage-wise reasoning helps models better identify and pinpoint clinically inconsistent expressions.

In the error correction stage, we evaluate three paradigms—an end-to-end baseline, a staged reasoning baseline, and our staged framework with dual-knowledge infusion. Staged inference already improves correction quality over the end-to-end approach by decomposing the task into interpretable reasoning steps. Building upon this, our dual-knowledge design further enhances factual accuracy and contextual coherence, achieving an average improvement of +7.69% over the staged baseline, with notable gains for Llama3-Aloe (+11.44%) and Llama3-8B (+10.09%). Although small declines appear in MMedLM2, qualitative review indicates fewer hallucinations and higher factual precision. Overall, these results confirm that integrating domain knowledge with structured reasoning yields more accurate, interpretable, and clinically trustworthy corrections across diverse model architectures.

Comparison with Simple Retrieval

Table 2 compares our dual-knowledge infusion framework against conventional Retrieval-Augmented Generation (RAG) approaches. While simple RAG retrieves entire documents based on surface-level text similarity, our method employs structured knowledge extraction to guide targeted retrieval of clinically relevant information. Our approach consistently outperforms simple RAG across all metrics, achieving average improvements of 10.06%, 6.93%, and 3.42% for detection, localization, and correction tasks respectively. Medical domain models show the largest gains, with MMedLM2 achieving a 20.40% improvement in error detection. These results validate our hypothesis that structured, relationship-aware knowledge retrieval is more effective than generic document-level retrieval for medical applications, where understanding anatomical relationships and clinical patterns is crucial for accurate error detection.

Computational Efficiency Analysis

In addition, our more sophisticated framework achieves significant efficiency gains, reducing processing time by 27.9%-37.4% across all models while maintaining superior performance. This efficiency improvement stems from our structured approach to information processing: MKGD enables focused entity extraction and relationship analysis, while EXKR provides targeted knowledge integration that

avoids unnecessary context processing. The average processing time reduction of 8.99 seconds per instance (from 26.64s to 17.65s) demonstrates that structured reasoning can be both more accurate and computationally efficient than brute-force approaches. These efficiency gains are particularly valuable for clinical deployment, where timely analysis is crucial for workflow integration.

Human Evaluation

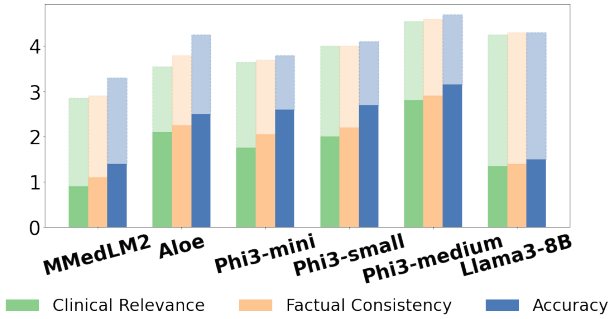

Figure 3: Human evaluation results comparing our proposed staged proofreading inference framework ("Ours") with the baseline model across three metrics: Accuracy, Factual Consistency, and Clinical Relevance. The lower stacked bars represent the baseline performance, while the upper stacked bars represent the improvements achieved by our method.

To provide a more comprehensive validation of our method's clinical relevance, we conducted a human evaluation with two practicing radiologists on 50 representative cases across six model architectures, comparing baseline inference with our staged proofreading framework using 3 five-point Likert scale (1–5). Figure 3 shows that our method consistently and significantly outperforms all baselines across three key dimensions—accuracy, factual consistency, and clinical relevance. The improvements are most pronounced for models such as Phi3-medium and Llama3-8B, where both factual correctness and clinical interpretability were notably enhanced, while MMedLM2 demonstrates particularly large gains in factual consistency and relevance to real-world diagnostic practice. These results highlight the essential role of expert evaluation in assessing medical text generation systems and confirm that our framework produces corrections that are not only linguistically precise but also clinically meaningful, trustworthy, and aligned with radiologists' reasoning processes.

Ablation Study

Table 3 presents a focused ablation study on Llama3-8B to understand the individual contributions of our framework components. The results reveal distinct patterns across different tasks and highlight the synergistic effects of our dual-knowledge infusion approach. MKGD alone provides minimal improvements across all tasks, with changes ranging from -0.30% to +0.20%, indicating that structured knowledge representation alone is insufficient. However, EXKR demonstrates substantial improvements in detection (+20.35%) and localization (+12.92%) but minimal impact

MKGD	EXKR	Det. (%)	Loc. (%)	Corr. (NLG)	Δ Avg. (%)
✗	✗	37.79	37.29	94.29	-
✓	✗	37.92	36.99	94.49	+0.01
✗	✓	58.14	50.21	94.38	+11.12
✓	✓	**62.27**	**53.14**	**94.43**	**+13.49**

Table 3: Ablation Study on Llama3-8B for the three tasks. Δ Avg. represents average improvement across three tasks.

on correction (+0.09%), suggesting that external knowledge retrieval is particularly valuable for error identification but less critical for generating corrections. The combination of MKGD + EXKR yields the best performance across all tasks (+24.48%, +15.85%, +0.14%), with the synergistic effect being most pronounced in detection and localization tasks. This pattern demonstrates that while EXKR drives the primary improvements, MKGD provides essential structural guidance that enables EXKR to work more effectively, particularly for tasks requiring precise understanding of medical entity relationships and clinical context.

Conclusion

This work addressed the challenge of automated error detection and correction in radiology reports through a framework combining staged proofreading inference with dual-knowledge infusion. Our approach decomposes error correction into three phases—detection, localization, and correction—while integrating structured medical knowledge. Evaluations across six language models demonstrate improvements in all three subtasks. Our structured approach also achieves efficiency improvements, reducing processing time by up to 37.4% while maintaining performance.

The framework transforms general-purpose language models into medical quality assurance tools without requiring expensive domain-specific fine-tuning. This addresses critical challenges in clinical practice: high error rates in radiological examinations, the growing gap between examination volumes and specialist availability, and the need for transparent, validatable AI systems that integrate with clinical workflows. Our approach provides automated proofreading that enhances clinical expertise, benefiting busy clinical environments and resource-constrained healthcare facilities where comprehensive manual review can be challenging. The principles established—staged inference modeling, dual-knowledge integration, and transparent reasoning—offer a foundation for developing AI systems that address healthcare documentation challenges while maintaining the interpretability and safety standards needed for clinical practice.

Despite these promising results, our work still has several limitations that point to future research directions: (1) extending the framework to multilingual and cross-specialty medical reports; (2) designing adaptive knowledge retrieval that evolves with clinical feedback; and (3) incorporating multimodal inputs—including images, previous reports, and patient history—to achieve more holistic error detection.

Acknowledgments

This research was supported by the Health and Medical Research Fund [Grant Nos. 19200911 and 21223141] and the National Natural Science Foundation of China Young Scientists Fund [Grant No. 82303957]. It also received support from the UK's Engineering and Physical Sciences Research Council (EPSRC; UKRI2701: PAIR: Building a Cloneable Pipeline for Utilizing Foundation AI on EHRs), the Medical Research Council (MR/S004149/1, MR/X030075/1), and the British Council (Facilitating Better Urology Care With Effective and Fair Use of Artificial Intelligence—A Partnership Between UCL and Shanghai Jiao Tong University School of Medicine). HW's role in this research was partially funded by the Legal & General Group through a research grant to establish the independent Advanced Care Research Centre at the University of Edinburgh. The funders had no role in the conduct of the study, data interpretation, or the decision to submit this work for publication. We sincerely thank all funding agencies for their support.

References

Abacha, A. B.; Yim, W.-w.; Fan, Y.; and Lin, T. 2023a. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2291–2302.

Abacha, A. B.; Yim, W.-w.; Michalopoulos, G.; and Lin, T. 2023b. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2575–2588.

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Ben Abacha, A.; Yim, W.-w.; Fu, V.; Sun, Z.; Xia, F.; and Yetisgen, M. 2024. Overview of the MEDIQA-CORR 2024 Shared Task on Medical Error Detection and Correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, 1–10. Mexico City, Mexico: Association for Computational Linguistics.

Berlin, L. 2007. Malpractice issues in radiology: perceptual and cognitive errors. *Seminars in Ultrasound, CT and MRI*, 28(3): 170–177.

Brady, A. P. 2017. Error and discrepancy in radiology: inevitable or avoidable? *Insights into Imaging*, 8(1): 171–182.

Cabitza, F.; Rasoini, R.; and Gensini, G. F. 2017. Unintended consequences of machine learning in medicine. *JAMA*, 318(6): 517–518.

Collins, F. S.; and Varmus, H. 2015. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9): 793–795.

Fang, T.; Wong, D. F.; Zhang, L.; Jin, K.; Zhang, Q.; Li, T.; Hou, J.; and Chao, L. S. 2024. LLMCL-GEC: Advancing Grammatical Error Correction with LLM-Driven Curriculum Learning. *arXiv preprint arXiv:2412.12541*.

Gururajan, A. K.; Lopez-Cuena, E.; Bayarri-Planas, J.; Tormos, A.; Hinjos, D.; Bernabeu-Perez, P.; Arias-Duart, A.; Martin-Torres, P. A.; Urcelay-Ganzabal, L.; Gonzalez-Mallo, M.; et al. 2024. Aloe: A Family of Fine-tuned Open Healthcare LLMs. *arXiv preprint arXiv:2405.01886*.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Kamoi, R.; Das, S. S. S.; Lou, R.; Ahn, J. J.; Zhao, Y.; Lu, X.; Zhang, N.; Zhang, Y.; Zhang, R. H.; Vummanthala, S. R.; et al. 2024. Evaluating LLMs at Detecting Errors in LLM Responses. *arXiv preprint arXiv:2404.03602*.

Kasalak, Ö.; Alnahwi, H.; Toxopeus, R.; Pennings, J. P.; Yakar, D.; and Kwee, T. C. 2023. Work overload and diagnostic errors in radiology. *European Journal of Radiology*, 167: 111032.

Kruskal, J. B.; Anderson, S.; Yam, C.-S.; and Sosna, J. 2011. Changes to the reporting radiologist workforce: implications for radiology residents and practicing radiologists. *Radiology*, 259(3): 616–621.

Lazaridou, A.; Kuncoro, A.; Gribovskaya, E.; Agrawal, D.; Liska, A.; Mairesse, F.; and Blunsom, P. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34: 29348–29363.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Ng, M.-Y.; Lee, E. Y.; Yang, J.; Yang, F.; Li, X.; Wang, H.; Lui, M. M.-s.; Lo, C. S.-Y.; Leung, B.; Khong, P.-L.; et al. 2020. Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiology: Cardiothoracic Imaging*, 2(1): e200034.

Pinto dos Santos, D.; Hempel, J.-M.; Mildenberger, P.; Klöckner, R.; and Persigehl, T. 2018. Speech recognition in radiology: how to implement a successful solution. *Insights into Imaging*, 9(6): 891–895.

Qiu, P.; Wu, C.; Zhang, X.; Lin, W.; Wang, H.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Towards Building Multilingual Language Model for Medicine. *arXiv preprint arXiv:2402.13963*.

Rajpurkar, P.; Chen, E.; Banerjee, O.; and Topol, E. J. 2022. AI in health and medicine. *Nature medicine*, 28(1): 31–38.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Sendak, M. P.; Balu, S.; and Schulman, K. A. 2020. Human factors that influence the adoption of a sepsis clinical decision support system. *Applied Clinical Informatics*, 11(3): 424–431.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Sun, C.; Teichman, K.; Zhou, Y.; Critelli, B.; Nauheim, D.; Keir, G.; Wang, X.; et al. 2025. Generative large language models trained for detecting errors in radiology reports. *Radiology*, 315(2): e242575.

Sunshine, J. H.; and Burkhardt, J. H. 2006. Radiology groups' workload in relative value units and factors affecting productivity. *Radiology*, 240(3): 504–515.

Topol, E. J. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1): 44–56.

Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; and Wei, F. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S.; et al. 2018. Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics*, 77: 34–49.

Wu, H.; Wang, M.; Wu, J.; Francis, F.; Chang, Y.-H.; Shavick, A.; Dong, H.; Poon, M. T.; Fitzpatrick, N.; Levine, A. P.; et al. 2022. A survey on clinical natural language processing in the United Kingdom from 2007 to 2022. *NPJ digital medicine*, 5(1): 186.

Wu, J.; Dong, H.; Li, Z.; Wang, H.; Li, R.; Patra, A.; Dai, C.; Ali, W.; Scordis, P.; and Wu, H. 2024. A hybrid framework with large language models for rare disease phenotyping. *BMC Medical Informatics and Decision Making*, 24(1): 289.

Wu, J.; Hasan, A.; and Wu, H. 2024. RadBARTsum: Domain Specific Adaption of Denoising Sequence-to-Sequence Models for Abstractive Radiology Report Summarization. *arXiv preprint arXiv:2406.03062*.

Wu, J.; Kim, Y.; Keller, E. C.; Chow, J.; Levine, A. P.; Pontikos, N.; Ibrahim, Z.; Taylor, P.; Williams, M. C.; and Wu, H. 2023. Exploring multimodal large language models for radiology report error-checking. *arXiv preprint arXiv:2312.13103*.

Wu, J.; Kim, Y.; and Wu, H. 2024. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhou, H.; Gu, B.; Zou, X.; Li, Y.; Chen, S. S.; Zhou, P.; Liu, J.; Hua, Y.; Mao, C.; Wu, X.; et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.