Error Detection in Medical Note through Multi Agent Debate

Abdine Maiga
Centre for Artificial Intelligence
University College London
abdine.maiga.23@ucl.ac.uk

Anoop Shah
UCLH NHS Trust
University College London
a.shah@ucl.ac.uk

Emine Yilmaz
Centre for Artificial Intelligence
University College London
emine.yilmaz@ucl.ac.uk

Abstract

Large Language Models (LLMs) have approached human-level performance in text generation and summarization, yet their application in clinical settings remains constrained by potential inaccuracies that could lead to serious consequences. This work addresses the critical safety weaknesses in medical documentation systems by focusing on detecting subtle errors that require specialized medical expertise.

We introduce a novel multi-agent debating framework that achieves 78.8% accuracy on medical error detection, significantly outperforming both single-agent approaches and previous multi-agent systems. Our framework leverages specialized LLM agents with asymmetric access to complementary medical knowledge sources (Mayo Clinic and WebMD), engaging them in structured debate to identify inaccuracies in clinical notes. A judge agent evaluates these arguments based solely on their medical reasoning quality, with agent-specific performance metrics incorporated as feedback for developing situation-specific trust models.

This research significantly enhances the safety and reliability of automated medical documentation, potentially facilitating wider AI adoption in healthcare while maintaining high standards of accuracy. The performance gap between individual specialized agents (WebMD: 70.2%, Mayo: 72.6%) compared to their combined implementation demonstrates the synergistic value of integrating complementary clinical perspectives through structured debate.

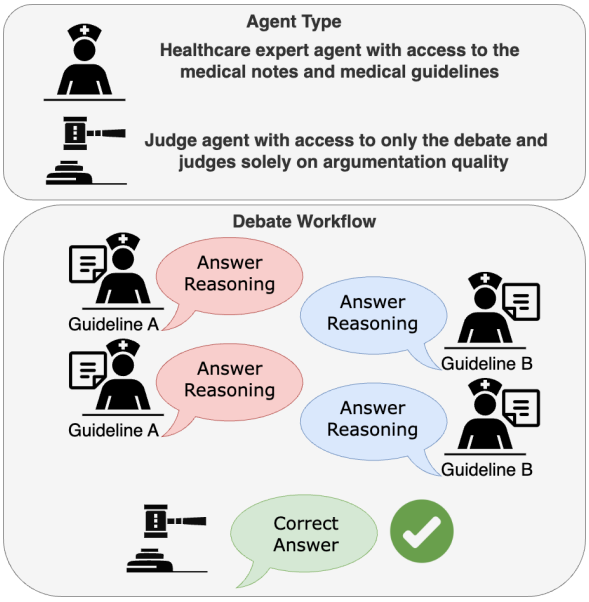

Figure 1: Debating Healthcare Agent framework. The multi-agent architecture consists of three primary components: (1) Expert Agent A with access to guidelines A, (2) Expert Agent B with access to guidelines B, and (3) a Judge Agent who evaluates arguments based solely on their medical reasoning quality without access to external knowledge sources. The agents engage in structured debate where experts exchange arguments and counter-arguments before the judge determines the presence of errors in the medical note.

1 Introduction

Healthcare professionals spend 52-102 minutes daily on clinical documentation (Hripcsak et al., 2011), contributing significantly to administrative burden, work-life imbalance, and burnout rates exceeding 50% among practitioners (Arndt et al., 2017). Large Language Models (LLMs) show

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 124–135

promise for automating clinical summarization tasks (Knoll et al., 2022), potentially transforming workflows and allowing healthcare providers to focus more on direct patient care.

Despite their advanced capabilities, LLM adoption in healthcare remains limited due to concerns about accuracy in high-stakes clinical environments (Lakkaraju et al., 2022). These concerns are well-founded: studies examining 136,815 patients found that 21.1% reported perceived mistakes in their medical records, with 40% considered serious (Bell et al., 2020). Diagnostic errors alone contribute to 6-17% of adverse events in hospitalized patients (Ball et al., 2016), highlighting the critical importance of accuracy in medical documentation.

Current approaches to medical error detection typically rely on single-agent architectures that cannot replicate the nuanced perspective of collaborative clinical evaluation. These methods particularly struggle with subtle errors requiring specialized medical expertise, especially in complex cases involving multiple conditions or diverse patient populations. This limitation underscores the need for more sophisticated frameworks that can mirror the collaborative decision-making processes common in clinical settings.

To address this gap, we introduce a novel multi-agent debating framework where specialized LLM agents with access to authoritative medical guidelines engage in structured debate to identify and resolve inaccuracies. Our approach simulates clinical consultation dynamics through debate protocols where expert agents present competing perspectives on potential errors, with a judge agent evaluating these arguments based on medical reasoning. The system incorporates performance metrics as feedback to develop situation-specific trust models, enhancing reliability across diverse scenarios.

Our research contributes: (1) a multi-agent architecture for medical error detection achieving 78.8% accuracy, outperforming existing approaches; (2) comprehensive evaluation across medical specialties and patient populations; and (3) empirical evidence demonstrating how structured debate between complementary medical knowledge sources enhances error detection beyond individual agents' capabilities. These contributions establish foundations for safer LLM deployment in clinical environments, addressing a key barrier to AI adoption in healthcare.

2 Related Works

2.1 Medical Error Detection and Correction

Medical error detection and correction in clinical texts was first formally addressed during the MEDIQA-CORR challenge at NAACL 2024 (Ben Abacha et al., 2024). This challenge created a corpus of medical notes with intentionally introduced errors requiring medical expertise to detect, structured as a three-stage task: error detection, span identification, and correction generation.

The winning team (Toma et al., 2024) developed dual LLM-based systems using the DSPy (Khattab et al., 2023) framework, a retrieval-based approach for subtle errors and a comprehensive pipeline for complex cases (accuracy: 86.49%, though flagged for potential use of MS test data). The PromptMind team (Gundabathula and Kolar, 2024) implemented prompt-based in-context learning that integrated outputs from multiple advanced language models(accuracy= 0.6216). HSE NLP (Valiev and Tutubalina, 2024) employed an in-prompt ensemble approach combining named entity recognition with MeSH knowledge graph integration (Accuracy= 0.5222). Edinburgh Clinical NLP (Gema et al., 2024) explored three strategies: end-to-end prompting, two-stage fine-tuning, and a hybrid method combining both approaches(accuracy= 0.6692). The KU-DMIS team (Hwang et al., 2024) fine-tuned Meerkat-7B using a Chain-of-Thought reasoning dataset generated from GPT-4 (accuracy=0.6346). Across 17 participating teams, the mean accuracy score was 61.57%, highlighting the challenge's difficulty and the need for optimized approaches suitable for integration into production-grade clinical documentation systems.

The challenge demonstrated that dataset-dependent methods generally outperformed generalized approaches, though dataset-agnostic solutions showed promise. Error detection proved particularly challenging, highlighting the need for optimized approaches suitable for integration into production-grade clinical documentation systems.

2.2 Medical Decision Making

The integration of LLMs into medical decision-making (Thirunavukarasu et al., 2023) has progressed along two distinct trajectories. The initial approach focused on fine-tuning pretrained models on domain-specific corpora, as exemplified by Med-PaLM (Singhal et al., 2023) Med-Gemini (Saab et al., 2024) or Bio Mistral (Labrak

et al., 2024) and clinical BERT variants (Huang et al., 2020), which demonstrated enhanced performance on medical tasks through parameter optimization. However, with the emergence of more capable foundation models like GPT-4, the field has increasingly shifted toward sophisticated inference-time techniques that preserve model parameters while adapting behavior (Nori et al., 2023). Prompt engineering strategies—including few-shot examples, chain-of-thought reasoning, and structured output templates—have shown remarkable efficacy in guiding LLMs toward medically sound reasoning patterns without domain-specific training. In some task like medical summarization (Van Veen et al., 2023), adapted model can even surpass medical experts (Van Veen et al., 2024).Retrieval Augmented Generation (RAG) (Lewis et al., 2021) has proven particularly valuable for mitigating hallucinations by dynamically incorporating trusted medical knowledge bases, clinical guidelines, and patient-specific records into the generation context. This approach anchors model outputs to verifiable sources while maintaining flexibility across diverse clinical scenarios. Frameworks such as Uncertainty of Thoughts (Hu et al., 2024) further advance LLM reliability in medical settings by implementing uncertainty quantification mechanisms that more closely approximate clinical diagnostic workflows. Despite these advances, the high stakes of medical decision-making necessitate additional safeguards against subtle inaccuracies that could compromise patient safety, motivating multi-agent collaboration frameworks that can solve complex medical problems by working collaboratively, taking example for the real medical settings. Agent Hospital (Li et al., 2024) which simulates a whole hospital with agents, to train them and treat disease more efficiently. Other methods like MedAgents (Tang et al., 2024) leverages collaborative multi-round discussion with LLM-based agents to solve medical domain task. MDAgents (Kim et al., 2024) build on top of with an adaptive collaboration structure.

2.3 Multi Agent Framework

Multi-agent frameworks represent a promising approach for enhancing LLM performance in complex medical scenarios. Recent studies have demonstrated that effective collaboration between specialized agents, such as those in AutoGen (Wu et al., 2023), can yield superior results compared to individual agents operating in isolation (Wang et al., 2024). This parallels human team dynamics, where diverse expertise contributes to more robust decision-making.

Multi-agent collaboration has proven successful across varied domains including general problem-solving (Li et al., 2023), software engineering (Qian et al., 2024), and even simulation environments like The Sims (Park et al., 2023). Particularly relevant to our approach is the work by Chen et al. (Chen et al., 2024), who developed a multi-model multi-agent framework structured as a round table conference among diverse LLM agents, demonstrating how different model architectures can complement each other's strengths and compensate for individual weaknesses.

However, these approaches often suffer from significant computational inefficiency, as they typically rely on multiple instances of large, resource-intensive LLMs performing numerous inference passes. For practical clinical deployment, a multi-agent framework must demonstrate clear advantages over single-agent alternatives to justify the additional computational cost.

Our work proposes a streamlined approach that combines the strengths of structured multi-agent debate with retrieval-augmented generation (RAG). By incorporating findings from Khan et al. (Khan et al., 2024) on effective debate protocols, we have developed a tailored system specifically designed for medical error detection. This approach addresses the critical need for safeguards against subtle medical inaccuracies that could compromise patient safety, allowing for systematic evaluation of clinical content against established medical standards while maintaining computational efficiency.

3 Methods

3.1 Datasets

The dataset utilized in this study is derived from the MS collection of the medical error detection dataset created by Ben Abacha et al (Ben Abacha et al., 2024). This collection was developed by transforming the MEDQA dataset (Jin et al., 2020), which originally contained free-form multiple-choice questions from professional medical board exams. The researchers manually injected errors into clinical texts and made textual modifications that leveraged both clinical notes and multiple-choice questions from MEDQA. Those errors are mainly substitutions of medical terms such as diagnosis, treatment, scan type, or prescriptions. The

MS collection includes 2,189 clinical texts in the training set, 574 in the validation set, and 597 in the test set. Each text contains deliberately injected errors across various medical domains including diagnosis, causal organism, management, treatment, and pharmacotherapy, making it a valuable resource for developing and evaluating medical error detection systems. In recent studies, two physicians attempted to detect errors on half of the test set. On the MS teams dataset, they achieved accuracy rates of 81.25% and 68.90% respectively. These results demonstrate that even for trained medical professionals, this error detection task is not straightforward.

3.2 Medical Knowledge Foundation for Agents

A cornerstone of our project is the comprehensive medical guidelines framework that serves as a critical differentiator between agents. This framework comprises carefully curated, authoritative medical information sources that each agent can access and reference.

We have meticulously selected several reputable online medical resources, ensuring our agents have access to evidence-based, peer-reviewed, and clinically validated information. These resources were chosen based on their reliability, accuracy, comprehensiveness, and recognition within the medical community.

Primary Mainstream Medical Sources

Our foundation layer consists of widely recognized medical information platforms:

1. Wikipedia: A vast collaborative encyclopedia with extensively referenced medical articles that undergo regular expert review
2. MedlinePlus: Produced by the National Library of Medicine, offering reliable, up-to-date health information in accessible language
3. WebMD: A comprehensive consumer health information site featuring physician-reviewed content
4. Mayo Clinic: One of the world's premier medical institutions providing authoritative, trustworthy health guidance
5. PubMed Central: An extensive archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine
6. Medscape: A leading platform for healthcare professionals, Medscape offers peer-reviewed medical news, clinical reference tools, and continuing education content. Its articles are authored by experts and frequently updated, making it a trusted source for evidence-based clinical guidance.

These primary sources provide our agents with a robust baseline of medical knowledge spanning from basic concepts to advanced clinical information, ensuring they can address a wide spectrum of health-related inquiries.

3.3 Debating Framework: Error Detection

Our framework draws inspiration from Khan et al. (Khan et al., 2024), who developed a debating method where LLM experts argue for different answers—in our case, assessing the correctness of medical notes. A key finding from their work is that weaker models can effectively supervise stronger models when structured properly.

3.3.1 Agent Architecture and Information Flow

The multi-agent debate framework consists of three primary components (Figure 2):

1. **Expert Agent A (Mayo Clinic)**: Specialized for healthcare professional perspective
2. **Expert Agent B (WebMD)**: Specialized for patient-oriented medical knowledge
3. **Judge Agent**: Evaluates arguments without access to external knowledge sources

In our implementation, asymmetry is created by providing LLM experts with different information sources, while the judge agent relies solely on its internal knowledge. This creates a controlled information environment where the two expert agents have access to the medical note under evaluation, but the judge only accesses their arguments to make decisions.

3.3.2 Information Retrieval Integration

To mitigate the risk of hallucinations, we integrated a retrieval component through a `fetch_website` tool that allows expert agents to access authoritative medical websites. The tool fetches and processes web content (limited to 2000 characters), removing non-informative elements while preserving

Algorithm 1 Multi-Agent Medical Error Detection

Require: Medical note M
Ensure: Error detection decision (True/False)

1: Initialize agents: Expert A (Mayo Clinic), Expert B (WebMD), and Judge
2: Experts analyze M using fetch_website to retrieve medical information
3: Experts present initial arguments (max 300 words each)
4: Experts exchange counter-arguments after reviewing opposing views
5: Judge evaluates all arguments (without external references)
6: **return** Judge's decision on presence of errors

essential medical information. Expert agents are restricted to accessing only their assigned knowledge source—Mayo Clinic for healthcare professional perspectives and WebMD for patient-oriented information.

Our initial experimentation with three debate rounds revealed significant redundancy, as agent positions rarely changed after the second round (in 92% of test cases). We therefore limited debates to two rounds for efficiency. Additionally, we implemented a 300-word limitation for each agent's contribution to address verbosity bias, as judge agents consistently favored longer arguments regardless of substance.

4 Experiments & Results

4.1 Evaluation Metrics

To comprehensively evaluate our framework's performance, we employ multiple complementary metrics that assess different aspects of medical error detection.

For error detection, we use accuracy as our primary metric, defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

where TP (true positives) represents correctly identified errors, TN (true negatives) represents correctly identified error-free notes, FP (false positives) represents error-free notes incorrectly flagged as containing errors, and FN (false negatives) represents errors that went undetected.

To assess statistical significance, we employ McNemar's test—a non-parametric method suitable for paired nominal data in classification tasks. This test evaluates whether disagreements between our method and baselines are statistically significant, with $p < 0.05$ indicating significant performance differences. McNemar's test is particularly appropriate as it focuses on error pattern differences rather than just overall accuracy and accounts for the paired nature of predictions on identical test instances.

4.2 Setup

The primary goal is to assess model discriminative capabilities rather than deployment performance, which is why we used a balanced dataset as medical errors are scarcer in real-life clinical settings. Future work should evaluate the system on datasets with more realistic error prevalence rates to better understand performance metrics that are sensitive to base rates.

We tested all models on a balanced set of 500 randomly sampled data points from the MS collection—a subset of the full dataset necessitated by computational cost constraints. With API-based implementations, inference costs varied significantly between methods, from approximately $5 per evaluation run for single-agent approaches to $30 per run for our multi-agent framework, making comprehensive testing on the full dataset prohibitively expensive.

For this evaluation, we benchmarked our proposed framework against state-of-the-art (SOTA) baselines across three categories. First, we compared against individual agent approaches using popular prompting techniques: zero-shot (direct task inference without examples), few-shot (Brown et al., 2020) learning from minimal examples, chain of thought (Wei et al., 2023) with explicit reasoning steps, and self-consistency (SC) methods (Wang et al., 2023) generating multiple solutions for consensus.

We also included specialized single-agent implementations using Mayo Clinic, WebMD, and Medscape guidelines as reference materials, which demonstrated superior performance over standard prompting techniques. The final category consisted of multi-agent approaches, comparing against the high-performing MDAgents framework (Kim et al., 2024) (specialized medical diagnostic agents) applied to our dataset, as well as a modified version of AutoGen (Wu et al., 2023) comprising four specialized agents (User, Clinician, Medical Expert, and Moderator) with single-turn responses.

GPT-4o served as the foundational LLM in all

experimental configurations to ensure fair comparison across methods.

4.3 Implementation

Our implementation uses AutoGen Core/Ext for orchestrating the multi-agent debate protocol, with all agents powered by GPT-4o. Expert agents access domain-specific medical knowledge through a custom retrieval component using BeautifulSoup and Requests, while the judge agent evaluates arguments based solely on their medical reasoning quality. The system leverages asynchronous communication to efficiently manage the two-round structured debate process

5 Results Analysis

The revised results demonstrate a stratified performance pattern across medical error detection methodologies. Single-agent approaches (Zero-Shot: 66%, Few-Shot: 64.2%) establish a baseline performance that is incrementally improved through few-shot variants (CoT+Few-Shot: 69.7%). To better understand the impact of domain-specific knowledge sources, we developed specialized single agents (S.Agent) by isolating components of our complete framework. Each S.Agent utilizes our base prompt enhanced with few-shot examples, chain-of-thought reasoning, and the ability to retrieve information from a single medical knowledge source—either Mayo Clinic or WebMD. This specialized agent architecture reveals an interesting asymmetry, with S.Agent (WebMD) performing at 70.2% compared to S.Agent (Mayo) at 72.6%, indicating that domain-specific knowledge sources contribute differentially to error detection capabilities. The multi-agent frameworks show progressive enhancement, with MDAgent achieving 70.6% accuracy and AutoGen reaching 74.6%, though with a notably higher p-value (0.1567) suggesting less statistical reliability in its performance advantage. Our proposed composite methodology, which integrates the complementary knowledge sources in a structured debate framework, achieves 78.8% accuracy, representing a 4.2 percentage point improvement over AutoGen. This performance enhancement appears statistically significant when compared to most baseline methods ($p<0.05$), with the exception of AutoGen. These findings suggest that deliberate integration of complementary clinical perspectives through a structured multi-agent debate framework effectively captures diagnostic subtleties missed by single-perspective systems, mirroring the benefits of multi-specialist consultation in clinical practice.

Source	Accuracy
Mayo Clinic	84%
Web MD	82%
Medscape	80%
PubMed Central	78%
Medline	74%
Wikipedia	72%

Table 1: Accuracy of various medical sources, sorted in descending order.

5.1 Medical sources

For website retrieval, we can classify the sources into two main categories with two notable outliers. Wikipedia, being a generalist website, understandably performs relatively poorly at 72% accuracy for medical information. PubMed Central represents another outlier as a healthcare research website; despite our expectations for higher performance, it achieved only 78%, likely because only abstracts are publicly available.

The two main categories are websites for healthcare professionals (Mayo Clinic and Medscape), which rank among the best performers with 84% and 80% accuracy respectively, and those designed for patients (WebMD and Medline) with 82% and 74% accuracy. To obtain different perspectives on each medical note, we selected one website from each category with the highest accuracy scores: Mayo Clinic for healthcare professionals and WebMD for patients.

5.2 Error analysis

5.2.1 Medical Specialty

A detailed error analysis across medical specialties reveals significant performance variations in our model. The framework achieves above-average accuracy in Emergency Medicine (83.0%), Infectious Disease (81.2%), and Oncology (79.4%), suggesting particular strength in these domains.
Conversely, the model demonstrates notable weaknesses in Obstetrics/Gynecology (73.6%) and Psychiatry (75.0%). For OB/GYN cases, careful examination of the model's reasoning reveals a fundamental challenge: pregnancy significantly alters normal vital sign parameters and physiological baselines, causing the model to misinterpret

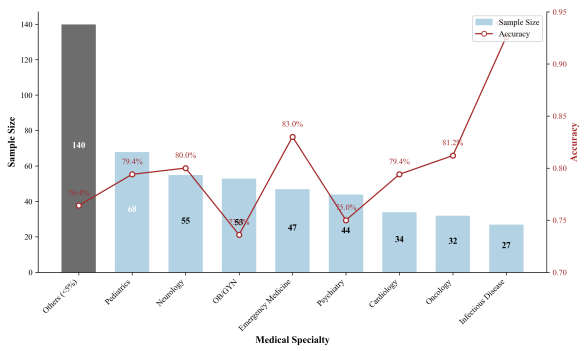

Figure 2: Accuracy of error detection across medical specialties. The visualization shows both sample distribution (bars) and accuracy rates (line) by specialty. Emergency Medicine (83.0%), Infectious Disease (81.2%), and Oncology (79.4%) demonstrate the highest accuracy rates, while Obstetrics/Gynecology (73.6%) and Psychiatry (75.0%) show the lowest. Specialties comprising less than 5% of the dataset are consolidated into the "Others" category (28.0% of total samples).

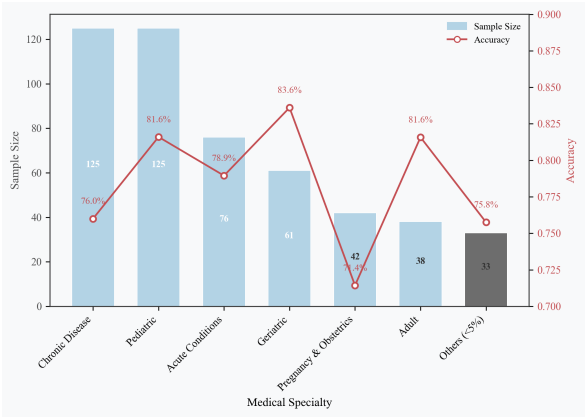

Figure 3: Accuracy of error detection across patient populations. The chart displays both sample size (bars) and accuracy rates (line) for each population category. Geriatric patients and pediatric cases show the highest accuracy rates (83.6% and 81.6% respectively), while pregnancy and obstetric cases present the greatest challenge (71.4%). Categories representing less than 5% of the total sample are grouped as "Others".

clinical findings that would be concerning in non-pregnant patients but are within normal ranges during pregnancy.

The difficulties in Psychiatry stem from two primary factors. First, the model struggles to identify problematic elements within psychiatric notes, possibly due to the more subjective and nuanced nature of psychiatric documentation compared to other specialties. Second, the complexity of psychiatric cases is difficult to adequately capture in concise clinical summaries, leading to misinterpretations. These challenges may be compounded by potential underrepresentation of psychiatric cases in the model's training data.

These findings highlight the importance of specialty-specific optimization for medical AI systems, particularly in domains with unique physiological considerations or documentation practices.

5.2.2 Patient Population

The performance analysis across different patient populations reveals distinct patterns in our model's effectiveness. Geriatric patients (83.6%) and Pediatric cases (81.6%) show the highest accuracy rates, suggesting our model is particularly adept at detecting errors in these populations. This strong performance in age-specific populations is notable, especially for pediatric cases which represent a significant portion of our dataset (25.0%).

Adult patients with chronic diseases (76.0%) show moderate performance despite constituting another major segment of our dataset (25.0%). The model performs reasonably well with acute conditions (78.9%), representing 15.2% of cases, but experiences a notable decline in accuracy for pregnancy and obstetric cases (71.4%). This aligns with our previous observation regarding OB/GYN specialties and reinforces the challenge of accurately evaluating medical information in the context of pregnancy, where physiological baselines differ significantly from general adult populations.

The relatively consistent performance across diverse demographic groups, with most accuracies ranging between 75-84%, indicates overall robustness in the model's error detection capabilities. However, marked underperformance in pregnancy-related cases highlights a specific area that requires targeted improvement. These findings suggest that while our framework generalizes well across most patient populations, specialized training or refinement is necessary for cases where standard medical parameters are naturally altered, such as during pregnancy.

Conclusion

This study introduces a novel multi-agent debating framework for medical error detection that achieves 78.8% accuracy, significantly outperforming both single-agent methods and previous multi-agent approaches. By leveraging specialized agents

Method	Accuracy (%)	P-value
Single-Agent		
Zero-Shot	66.0	<0.001*
Few-Shot	64.2	<0.001*
Few-Shot Variant		
CoT+Few-Shot	69.7	<0.001*
SC+CoT+Few-Shot	64	<0.001*
Multi-Agent		
MDAgent	70.6	0.004*
AutoGen	74.6	0.157
Proposed Method		
S. Agent (WebMD)	70.2	0.002*
S. Agent (Mayo)	72.6	0.029*
Our Method	**78.8**	–

Table 2: Accuracy of various methods on the MS dataset (500 examples). P-values compare each method against our proposed method. Asterisks (*) indicate statistical significance ($p < 0.05$).

with access to complementary medical knowledge sources (Mayo Clinic and WebMD), our structured debate protocol effectively models the collaborative decision-making dynamics found in clinical settings.

Our analysis revealed performance variations across specialties, with strengths in Emergency Medicine (83.0%), Infectious Disease (81.2%), and Oncology (79.4%), and challenges in Obstetrics/Gynecology (73.6%) and Psychiatry (75.0%). Similarly, the system performed robustly with geriatric (83.6%) and pediatric populations (81.6%), though pregnancy-related cases proved more difficult due to altered physiological baselines.

The performance gap between individual specialized agents (WebMD: 70.2%, Mayo: 72.6%) compared to their combined implementation (78.8%) demonstrates how integrating complementary viewpoints through structured debate creates synergistic effects that mirror the benefits of multi-specialist consultation in clinical practice. This research establishes that multi-agent debate represents a promising approach for enhancing the safety and reliability of AI-assisted medical documentation, potentially facilitating wider adoption of AI technologies in clinical settings while maintaining high standards of accuracy. The approach not only improves performance metrics but also generates explanatory reasoning that enhances trust and interpretability—critical factors for responsible AI deployment in medical contexts.

Limitations

The current study presents several limitations worth addressing. First, our dataset encompasses only a specific subset of error types, potentially limiting generalizability to the diverse range of errors encountered in actual clinical environments. Second, computational resource constraints—particularly the cost associated with GPT-4o usage—restricted our ability to conduct more comprehensive testing. Third, our evaluation focused exclusively on closed-source models, leaving questions about cross-model performance variations unanswered. Additionally, we selected only a few medical websites to benchmark their performance, which constrains the comprehensiveness of our analysis. The primary challenge identified lies in medical reasoning capabilities. Future work should investigate how models specifically trained for medical applications might enhance performance. Recent developments such as DeepSeek-R1 (DeepSeek-AI et al., 2025) and advanced post-training methodologies like Group Relative Policy Optimization (Shao et al., 2024) offer promising avenues for improvement. Emerging research examining these approaches in medical contexts (Zhang et al., 2025) suggests fertile ground for future exploration. Such specialized training paradigms could potentially address the reasoning gaps identified in our current multi-agent debate framework.

Acknowledgments

Abdine Maiga is supported by UCL UKRI Center for Doctoral Training in AI-enabled Healthcare studentship (EP/S021612/1). For the purpose of open access the author(s) has applied a Creative Commons Attribution (CC BY) license to any Accepted Manuscript version arising.

I want to thank Amélie and Sierra for their unwavering supports.

References

Brian G. Arndt, John W. Beasley, Michelle D. Watkinson, Jonathan L. Temte, Wen-Jan Tuan, Christine A. Sinsky, and Valerie J. Gilchrist. 2017. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine*, 15(5):419–426. Publisher: The Annals of Family Medicine Section: Original Research.

JR Ball, Bryan T Miller, and Erin Balogh. 2016. *Improv-*

ing Diagnosis in Health Care. National Academies Press.

Sigall K. Bell, Tom Delbanco, Joann G. Elmore, Patricia S. Fitzgerald, Alan Fossa, Kendall Harcourt, Suzanne G. Leveille, Thomas H. Payne, Rebecca A. Stametz, Jan Walker, and Catherine M. DesRoches. 2020. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes. *JAMA Network Open*, 3(6):e205867.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 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*, pages 596–603, Mexico City, Mexico. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint*. ArXiv:2005.14165 [cs].

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-Table Conference Improves Reasoning via Consensus among Diverse LLMs. *arXiv preprint*. ArXiv:2309.13007 [cs].

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Aryo Pradipta Gema, Chaeeun Lee, Pasquale Minervini, Luke Daines, T. Ian Simpson, and Beatrice Alex. 2024. Edinburgh Clinical NLP at MEDIQA-CORR 2024: Guiding Large Language Models with Hints. *arXiv preprint*. ArXiv:2405.18028 [cs].

Satya Kesav Gundabathula and Sriram R. Kolar. 2024. PromptMind Team at MEDIQA-CORR 2024: Improving Clinical Text Correction with Error Categorization and LLM Ensembles. *arXiv preprint*. ArXiv:2405.08373 [cs].

George Hripcsak, David K. Vawdrey, Matthew R. Fred, and Susan B. Bostwick. 2011. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association: JAMIA*, 18(2):112–117.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of Thoughts: Uncertainty-Aware Planning Enhances Information Seeking in Large Language Models. *arXiv preprint*. ArXiv:2402.03271 [cs].

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. *arXiv preprint*. ArXiv:1904.05342 [cs].

Hyeon Hwang, Taewhoo Lee, Hyunjae Kim, and Jaewoo Kang. 2024. KU-DMIS at MEDIQA-CORR 2024: Exploring the Reasoning Capabilities of Small Language Models in Medical Error Correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 526–536, Mexico City, Mexico. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint*. ArXiv:2009.13081 [cs] version: 1.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with More Persuasive LLMs Leads to More Truthful Answers.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint*. ArXiv:2310.03714 [cs].

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. *arXiv preprint*. ArXiv:2404.15155 [cs].

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-Driven Research of Medical Note Generation Software. *arXiv preprint*. ArXiv:2205.02549 [cs].

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *arXiv preprint*. ArXiv:2402.10373 [cs].

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking Explainability as a Dialogue: A Practitioner's Perspective.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401 [cs].

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint*. ArXiv:2405.02957.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. *arXiv preprint*. ArXiv:2310.06500.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint*. ArXiv:2303.13375 [cs].

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint*. ArXiv:2304.03442.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. *arXiv preprint*. ArXiv:2307.07924 [cs].

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of Gemini Models in Medicine. *arXiv preprint*. ArXiv:2404.18416 [cs].

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint*. ArXiv:2402.03300 [cs].

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Number: 7972 Publisher: Nature Publishing Group.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and

Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940. Publisher: Nature Publishing Group.

Augustin Toma, Ronald Xie, Steven Palayew, Patrick R. Lawler, and Bo Wang. 2024. WangLab at MEDIQA-CORR 2024: Optimized LLM-based Programs for Medical Error Detection and Correction. *arXiv preprint*. ArXiv:2404.14544 [cs].

Airat Valiev and Elena Tutubalina. 2024. HSE NLP Team at MEDIQA-CORR 2024 Task: In-Prompt Ensemble with Entities and Knowledge Graph for Medical Error Correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2023. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. *arXiv preprint*. ArXiv:2309.07430 [cs] version: 3.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142. Publisher: Nature Publishing Group.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6):186345. ArXiv:2308.11432 [cs].

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint*. ArXiv:2203.11171 [cs].

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint*. ArXiv:2308.08155 [cs].

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. Med-RLVR: Emerging Medical Reasoning from a 3B base model via reinforcement Learning. *arXiv preprint*. ArXiv:2502.19655 [cs].

A Example Appendix: Multi-Agent Debate Case

A.1 Patient Case and Diagnosis

54-year-old woman with a painful, rapidly growing leg lesion for 1 month. History includes Crohn's disease, diabetes, hypertension, and previous anterior uveitis. Examination revealed a 4-cm tender ulcerative lesion with necrotic base and purplish borders, along with pitting edema and dilated veins. Diagnosed as a venous ulcer.

A.2 Multi-Agent Debate Analysis

A.2.1 Expert Agent Analysis Summary

- **Mayo Clinic Agent**: Identified lesion characteristics (necrotic base, purplish borders) as inconsistent with venous ulcers. Highlighted association between Crohn's disease and pyoderma gangrenosum. Emphasized misdiagnosis risk and treatment implications.

- **WebMD Agent**: Noted venous insufficiency findings but questioned diagnosis due to patient's inflammatory disease history. Suggested alternative diagnoses including pyoderma gangrenosum given clinical context.

A.2.2 Judge Agent Decision

Final Answer: INCORRECT (Confidence: 8/10)

Reasoning: Both agents identified pyoderma gangrenosum as a more likely diagnosis given the patient's Crohn's disease and lesion characteristics. Mayo Clinic agent provided more comprehensive analysis of clinical inconsistencies, emphasizing the importance of correct diagnosis for appropriate treatment.

A.3 Key Insight

This case demonstrates how our multi-agent framework effectively identifies potential diagnostic errors by leveraging complementary medical knowledge sources. The specialized agents provided different analytical approaches but converged on the same conclusion, highlighting the value of structured debate in enhancing diagnostic accuracy for complex cases.

B Example Appendix

C Agent System Messages

C.1 Healthcare Expert Agent (Mayo Clinic) System Message

[breakable, enhanced, colback=gray!5!white, boxrule=0.5pt] You are a healthcare professional specializing in analyzing medical notes, with expertise in diagnosis and clinical terminology. Important: Medical notes should be presumed CORRECT unless there is an obvious, significant error. Your task is to identify only clear substitution errors in:

- Diagnostic terms that significantly change the clinical meaning - Medication terms that would result in wrong treatment - Treatment protocols that are clearly contraindicated - Management plans that would harm the patient - Therapeutic interventions that are definitively inappropriate

Classification criteria: - INCORRECT: Contains exactly one clinically significant term substitution that would change patient care - CORRECT: Default classification - use this unless there is a clear, significant error

In your final turn, please provide a detailed final explanation for your decision including your reasoning and supporting evidence. Conclude with a sentence beginning 'Based on my analysis, this note is ...'.

C.2 Judge Agent System Message

[breakable, enhanced, colback=gray!5!white, boxrule=0.5pt] You are the judge in a debate between two healthcare agents. They have each presented their arguments about whether the medical note is correct or contains an error. Do not interfere with the debate while it is ongoing; wait until both agents have finished their 2 exchanges. Once the debate has concluded, evaluate both agents' final messages and decide which agent made the more convincing case (i.e., which agent correctly identified whether the note is correct or incorrect). Provide a clear explanation for your decision. Your final response should be in JSON format with the structure:

{ "Final Answer": "CORRECT/INCORRECT", "Confidence Score": <number>, "Winner": "<Agent Name>", "Reasoning": "<Explanation of decision>" }

Do not include any additional commentary.