TRIAGEAGENT: Towards Better Multi-Agents Collaborations for Large Language Model-Based Clinical Triage

Meng Lu [1] **Brandon Ho** [2] **Dennis Ren** [2] **Xuan Wang** [1]

Abstract

The global rise in emergency department (ED) visits poses significant challenges to efficient patient management, especially in clinical triage. Traditionally managed by human professionals, triage is affected by substantial variability and high workload. Large language models (LLMs) offer promising reasoning and understanding capabilities, but using them directly in clinical triage is challenging due to the need for domain-specific accuracy and the complexity of the triage process. To address these issues, we introduce TRIAGEAGENT, a novel heterogeneous multi-agent framework enhancing collaborative decision-making for clinical triage. TRIAGEAGENT utilizes LLMs for role-playing, incorporating self-confidence and early-stopping mechanisms in multi-round discussions to improve reasoning and proficiency. It also employs the medical ESI handbook through a retrieval-augmented generation (RAG) approach for precise triage knowledge and integrates both coarse- and fine-grained ESI-level predictions. Extensive experiments show that TRIAGEAGENT outperforms state-of-the-art GPT-based methods on three professional clinical datasets. Additionally, we have released the first public benchmark dataset for clinical triage with ESI levels and provided benchmark human expert performance for comparison.

1. Introduction

Emergency Departments (EDs) are a crucial component of the healthcare system, operating around the clock to assess and prioritize patients based on the urgency and severity of their conditions. With a global increase in patient numbers, EDs face growing challenges in quickly and accurately classifying cases using the **Emergency Severity Index** (ESI). This process, known as **clinical triage**, is essential for prioritizing treatment and effectively allocating resources.

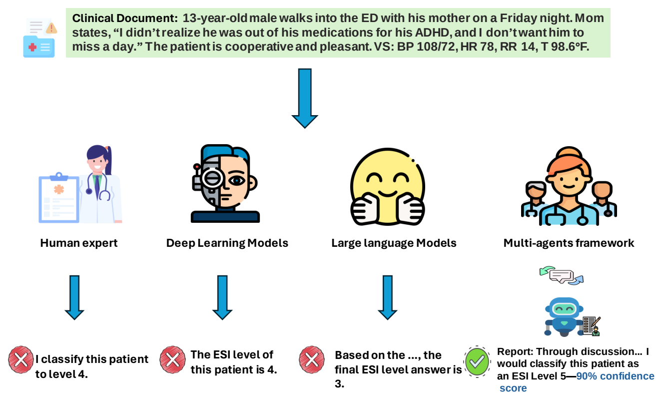


Figure 1. Workflow illustrations of clinical triage task.

Currently, the standard process of ESI classification in hospitals relies on human experts who read clinical notes to assess the urgency of cases (shown in Figure 1). While humans excel at this task, the increasing patient volume and the complexity of the clinical triage process highlight the limitations of relying solely on human performance. This approach is time-consuming, labor-intensive, error-prone, and significantly burdens clinical staff, often resulting in staff fatigue and reduced performance.

As a result, AI methods for automated ESI classification are in high demand. Traditional deep learning (DL) models (Yao et al., 2021; Sánchez-Salmerón et al., 2022) have attempted to assist in clinical triage but often fall short due to the nuanced and dynamic nature of the task. These models require extensive labeled data and struggle with generalization and real-time adaptation in emergency settings, leading to limited accuracy and reliability for widespread clinical adoption. The emergency of LLMs like GPT (Kojima et al., 2023), Med-PaLM (Chowdhery et al., 2022), and Llama (Touvron et al., 2023) offer a promising solution. With advanced text understanding capabilities, LLMs can quickly

[1]Department of Computer Science, Vriginia Tech, Blacksburg, USA [2]Children's National Hospital, Washington DC, USA. Correspondence to: Meng Lu <menglu@vt.edu>, Brandon Ho <bho2@childrensnational.org>, Ren Dennis <dmren2@childrensnational.org>, Xuan Wang <xuanw@vt.edu>.

ICML 2024 AI for Science workshop, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

interpret and categorize clinical notes by analyzing symptoms and diagnostic information, reducing time costs and minimizing human errors. They can also leverage external tools, such as knowledge base APIs (Qin et al., 2023; Zhuang et al., 2023), to enhance domain-specific knowledge, improving adaptability, speed, and accuracy in patient assessments, making LLMs more effective than traditional models. However, a single LLM struggles with the complexity of emergency scenarios, such as complex medical diagnostic processes. Therefore, a multi-agent approach is needed to handle these complexities and enhance LLM effectiveness in the process of clinical triage.

Recent advancements in multi-agent collaboration among LLMs have enhanced their capabilities in advanced reasoning(Wang et al., 2023b), sophisticated planning(Yao et al., 2023; Sun et al., 2023), and memory enhancement(Wang et al., 2023a). These improvements enable LLMs to analyze medical data, formulate care plans, and recall patient history effectively(Tang et al., 2024). For instance, sophisticated planning allows LLMs to predict patient trajectories and prioritize care, potentially reducing wait times and improving clinical outcomes such as timely interventions(Du et al., 2023). However, significant challenges remain in applying multi-agent LLMs to clinical settings(Sharaf & Anoop, 2023). Their performance in complex clinical tasks, such as triage, often hovers around 60% accuracy. Many LLM-based methods are fine-tuned on specific data, limiting their generalizability across diverse clinical scenarios. Additionally, the lack of benchmark datasets hampers comprehensive comparisons between LLMs and human experts, complicating the validation of these methods' practical value.

We identify four major challenges hindering the effectiveness of LLMs in clinical triage as follows. 1) **Data Processing Speed**: The increasing number of patients necessitates rapid and efficient classification, requiring real-time data processing and decision-making. Traditional methods often face delays due to the need to gather diverse data for training and the extensive time required for data preprocessing, hindering timely clinical interventions. **2) Diversity in Clinical Documents**: Variability in patients' diagnostic histories and symptoms, as well as different writing styles and medical terminologies, complicates the classification process. **3) Contextual Understanding and Interpretability**: LLMs often struggle with the nuances of clinical medicine and, as black-box models, typically provide results without sufficient supporting evidence. This lack of transparency makes it difficult for clinicians to trust the outputs. **4) Boundaries Between Different ESI Levels**: Achieving high precision in distinguishing and classifying different urgency levels in clinical documents is crucial to prevent critical errors. This challenge involves accurately identifying the boundaries between ESI levels (1, 2, 3, 4, 5), which represent a hierarchy of urgency and require clear understanding and precise classification.

To address these challenges, we propose TRIAGEAGENT, a pioneering framework for multi-agent collaboration in clinical triage. This framework leverages LLM-based agents, enhanced with external tools and embedded medical knowledge. TRIAGEAGENT introduces several **key innovations**: retrieval-augmented generation for context-relevant evidence, a confidence score-based mechanism for precise decision-making, and an early stopping mechanism to improve time efficiency. These features enable effective information exchange and reliable interactions, making the framework ideal for zero-shot document classification and handling complex tasks without prior demonstrations. TRIAGEAGENT overcomes the limitations of traditional methods by enabling real-time decision-making, which is crucial for timely clinical interventions. By incorporating evidence support and robust confidence scoring, TRIAGEAGENT improves contextual understanding and interpretability, achieving high precision in ESI-level classification. The framework fosters a collaborative environment with dynamic, interactive debates among agents, refining and sharing information for a more accurate triage process. Experiments on three ESI test datasets demonstrate that TRIAGEAGENT significantly improves zero-shot performance with GPT-3.5-turbo and GPT-4, reducing discordance rates by up to 10.84% and 18.42%, respectively. Furthermore, we are the first to publicly release a clinical triage dataset that includes clinical notes alongside ESI levels, providing a valuable resource for research and practical application. By setting new benchmarks, we aim to advance clinical triage in both academic research and practical application.

2. Related Works

2.1. LLMs Applied in Healthcare Fields.

Large Language Models (LLMs) have recently seen significant advancements across various sectors, notably in healthcare (Ling et al., 2024; Bi et al., 2024; Nori et al., 2023; Bao et al., 2023). These models are increasingly used in medical applications, including text-based diagnostics(Ma et al., 2024), genetic analysis (Bi et al., 2024), pharmacy applications (Liu et al., 2023), and medical summary generation (Shaib et al., 2023). Current research on LLMs in healthcare primarily focuses on integrating external tools for enhanced clinical insights and refining models through instruction tuning. For example, GeneGPT (Jin et al., 2023) leverages Web APIs from the National Center for Biotechnology Information (NCBI) to access diverse biomedical information and then use GPT models for reasoning tasks. Additionally, methods described in (Zhang et al., 2024b; Singhal et al., 2022; Oniani et al., 2024; Kang et al., 2023) utilize instruction tuning with prompt design to adapt LLMs

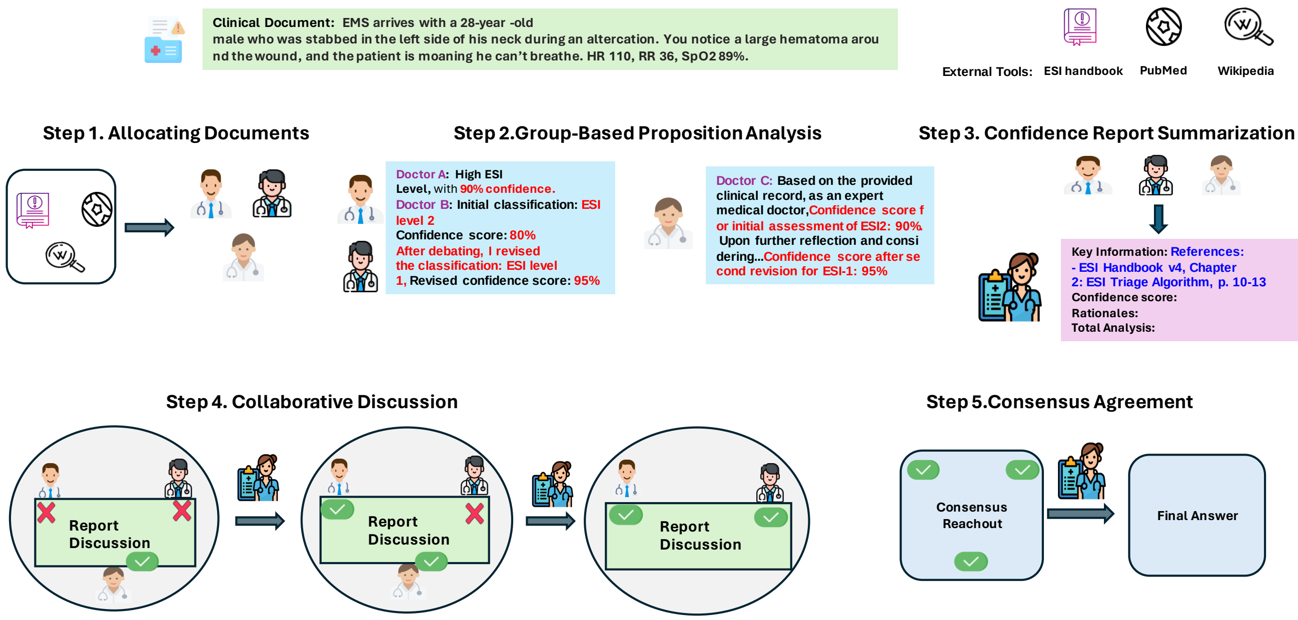


Figure 2. A diagram of our proposed dynamic multi-agent collaboration framework, TRIAGEAGENT, illustrates the reasoning process through five stages when given clinical documents with questions as input: 1). **Allocating Documents**: Patient clinical records are assigned to expert agents to initiate ESI discussions. 2).**Group-Based Proposition Analysis**: Agents are divided into two groups to combine **coarse** and **fine-grained** classification. 3). **Confidence Report Summarization**: A summarized report is created based on previous analysis, incorporating **confidence scores**. 4).**Collaborative Discussion**: Agents discuss the summarized report, iteratively revising it according to the key information, rationales, and confidence scores. In this step, an **early-stopping mechanism** is triggered by consistent high confidence scores with **evidence** across multiple rounds. The cross sign in the figure represents incorrect classifications, while the tick sign represents correct results proposed by the agents. 5).**Consensus Agreement**: The ultimate result is precise and thoroughly validated, highlighting the importance of collaborative decision-making.

for various healthcare tasks, including decision support, medical question answering, and disease diagnosis.

2.2. LLMs Applied in Autonomous Agent Systems

Research in both academic and industrial has concentrated on autonomous agents trained in isolated, self-contained environments with limited knowledge bases (Wang et al., 2024a). Significant advancements have been made with deploying LLM-based agents capable of independently sensing and decision-making, as detailed in (Yao et al., 2023; Xie et al., 2023; Zhou et al., 2023). The trend has shifted towards collaborative multi-agent systems, which enhance the capabilities of LLM agents through iterative feedback and teamwork, as discussed in (Xi et al., 2023; Wang et al., 2024b; Li et al., 2023). These systems, which mimic human learning and decision-making processes, involve agents assuming specific roles (Wang et al., 2024b; Hong et al., 2023) and engaging in effective communication (Qian et al., 2023; Wu et al., 2023; Li et al., 2023). Recent research has also explored improving agent performance through adversarial tactics like debates (Du et al., 2023; Liang et al., 2023; Xiong et al., 2023) and negotiations (Fu et al., 2023), with innovative frameworks where agents interact competitively (Liang et al., 2023) or negotiate roles as buyers and sellers (Fu et al., 2023).

3. Methodology

This section describes the framework of TRIAGEAGENT, which simulates the teamwork and problem-solving processes typically used in diagnosing and treating emergency patients. As shown in Figure 2, TRIAGEAGENT framework operates in five stages: 1) **Allocating Documents**: Patient clinical records are assigned to expert agents to initiate ESI discussions. 2).**Group-Based Proposition Analysis**: Agents are divided into two groups to combine coarse and fine-grained classification. 3). **Confidence Report Summarization**: A summarized report is created based on previous analysis, incorporating confidence scores. 4). **Collaborative Discussion**: Agents engage in discussions over the summarized report, iteratively revising it according to the key information, rationales and confidence scores. 5). **Consensus Agreement**: The ultimate result is precise and thoroughly validated, highlighting the importance of collaborative decision-making.

3.1. Allocating Documents

Given a patient's narrative clinical note $P = \{P_1, P_2, \ldots\}$ and a structured ESI handbook with different level references denoted as $R = \{R_1, R_2, \ldots\}$, the objective is to allocate relevant patient diagnostic and treatment documentation. This involves extracting hierarchically structured ESI-level categories for each patient using natural language query q. The allocation process can be summarized mathematically by defining a function f that represents the document allocation process: $f : (P, R, q) \to C$, where C is the set of hierarchically structured ESI-level categories. See Appendix B for details.

3.2. Group-Based Proposition Analysis

To improve the efficiency and accuracy of ESI-level classification, we have divided the agents into two groups and employed both coarse-grained and fine-grained classification techniques.

Group_1: Coarse and Fine-Grained Classification The first group comprises two agents. The first agent performs an initial, coarse classification of ESI levels into two broad categories: high (levels 1, 2, and 3) and low (levels 3, 4, and 5), with level 3 being flexible based on the situation. This initial classification can be represented mathematically as follows: $A_1 : P_i \to \{C_{\text{high}}, C_{\text{low}}\}$, where A_1 represents the first agent, P_i represents i_th patient record, and $\{C_{\text{high}}, C_{\text{low}}\}$ represents the high and low ESI categories. Subsequently, the second agent refines these broad categories into specific ESI levels:$A_2 : C_{\text{high}} \cup C_{\text{low}} \to \{1, 2, 3, 4, 5\}$, where A_2 is the second agent responsible for detailed analysis, breaking down the high and low categories into specific ESI levels. This two-step process reduces misclassifications and enhances the precision of the final classification.

Group_2: Direct Detailed Classification The second group consists of a single agent who directly performs detailed ESI-level classification on the patient records. This agent,denoted as A_3, is responsible for classifying each patient record P_i into one of the five detailed ESI levels: $A_3 : P_i \to \{1, 2, 3, 4, 5\}$.

Combining these two groups, this structured approach effectively addresses the boundaries between different levels in the ESI hierarchy, enhancing decision-making efficiency and accuracy by collaborating agents with diverse functions.

3.3. Confidence Report Summarization

In the Confidence Report Summarization stage, the summarizer agent A_s gathers and integrates previous analyses, including confidence scores, rationales, and supporting evidence from all role-playing agents (A_1,A_2 and A_3). This step consolidates the findings, ensuring the decision is well-supported and reliable. Each agent's analysis report is used to construct the summary prompt. Specifically, the debates from the agents are fused into the summarizer's prompt $Prompt_{rs}$.

The summarizer then generates a synthesized report by extracting key information and comprehensively analyzing the previous classifications provided by these agents. This process can be mathematically formulated as: $Repo = LLM(P, R, r_{rs}, prompt_{rs})$, where $Repo$ represents the synthesized report, P denotes the patient's clinical notes, R refers to the ESI handbook references, r_{rs} represents the role of the summarizer, and $Prompt_{rs}$ is the guideline prompt for the summarizer, which includes the analysis reports from (A_1,A_2 and A_3). The synthesized report is structured as follows: $Repo$=[key information; confidence score;rationale;total analysis]). This approach effectively combines insights from multiple agents, ensuring that the triage decision is based on comprehensive and thoroughly validated information. Consequently, this method enhances the accuracy of the clinical triage process while improving the interpretability and reliability of the decisions made.

3.4. Collaborative Discussion

Given the preliminary summary report $Repo$, the objective of the Collaborative Discussion stage is to engage all agents in multiple rounds of discussions to produce a final summary report that everyone accepts and agrees upon. Instead of voting mechanism (Tang et al., 2024), TRIAGEAGENT reflect on the classification results, reasoning, and confidence scores with evidence support provided by their peers during each round of discussions. For each agent A_i with an initial classification result C_i and confidence score S_i, the following process occurs: First, each agent A_i reviews the classification results C_j, reasoning R_j, and confidence scores S_j provided by every other agent $A_j (j \neq i)$. If $A_i$is persuaded by the reasoning and confidence scores from A_j, it updates its classification C_i to C_j and provides reasons for the modification.Specifically, if $S_j > S_i$ or R_j is convincing, then C_i is updated to C_j with explanations. Conversely, if A_i rejects the reasoning from A_j, A_i provides a justification for maintaining its original classification C_i. This occurs when $S_j \leq S_i$ or R_j is not convincing. Therefore, A_i needs to give the explanation of keeping C_i. To enhance efficiency, an early stopping mechanism is incorporated. If an agent consistently maintains its opinion with high confidence scores across multiple rounds, that agent can exit the discussion process early. This mechanism helps save time and improves decision-making efficiency. This iterative process continues until all agents reach a consensus or the early stopping mechanism is triggered. For details on the early stopping mechanism, see Appendix C.

Table 1. Statistics of the dataset

DATASET	VIGNETTES	NUMBER	EXPLANATION
TRAINING	218	218	218
TEST-1	72	72	0
TEST-2	72	72	0
TEST-3	72	72	0

3.5. Consensus Agreement

Finally, agents reach a consensus by reflecting on the answers, reasoning, and confidence scores provided by different agents, combined with their own domain knowledge. This collaborative process ensures the final decision is robust and well-supported by comprehensive analysis.

4. Experimental Setup

Datasets We collected patient cases based on the publicly available Emergency Severity Index (ESI) Handbook v4 (esi, Accessed: 2024-04-06) to construct a dataset for evaluating our TRIAGEAGENT framework. To our knowledge, we are the first to publicly release a clinical triage dataset that includes clinical notes and corresponding ESI levels for AI-triage tasks. This dataset serves as our benchmark for evaluating the effectiveness of our framework. Statistical information about this dataset is presented in Table 1. For more details about the dataset, see Appendix D.

Implementation We use GPT-3.5-Turbo and GPT-4(OpenAI et al., 2024) from OpenAI Service for all experiments conducted in zero-shot settings. The temperature is set to 0.7, top_k to 1.0, and the cache seed to 42. The maximum iteration number for our framework is set to 12, and the frequency penalty is set to 0.1. SCtr and SCons have an iteration number of 8 and a temperature of 0.7.

Baselines The performance of our TRIAGEAGENT framework is evaluated against several key baselines, including those utilizing LLM-based planning, tool usage, and retrieval-augmented generation.

- CoT (Kojima et al., 2023): integrates sequential step-by-step reasoning into the prompt. We applied CoT to our problem by incorporating it into the initial reasoning steps of our framework, ensuring a clear comparison in structured reasoning tasks.
- Self-contrast(Zhang et al., 2024a): improves stability and accuracy by contrasting different solving perspectives and summarizing discrepancies. We used Self-contrast to analyze different perspectives within our dataset, helping to identify and reconcile conflicting information.
- Self-consistency(Wang et al., 2023b): enhances Zero-shot and Few-shot CoT by generating predominant responses through multiple chain samplings. We implemented Self-consistency by generating multiple response chains and selecting the most consistent answers, providing a robust comparison for our framework.
- Exchange-of-thought(EoT)(Yin et al., 2023): enables cross-model communication and problem-solving integration. We applied EoT by facilitating communication between different models in our framework to enhance problem-solving capabilities.
- MedAgent(Tang et al., 2024): is a role-playing collaboration framework for medical scenarios using LLMs.We utilized MedAgent's methodology to simulate medical scenarios in our evaluations, allowing us to benchmark our framework against a collaborative multi-agent system.
- TRIAGEAGENT: Improves the *Self-Consistency* approach by incorporating dynamically updated confidence scores from various reasoning perspectives, enhanced by external resources as evidence. It strategically selects the top K most confident answers for critical assessment using internal and external data. Agents then proceed to a critical reflection phase with retrieval-augmented generation capabilities to refine choices and ensure minimal discordance and maximal coherence. This iterative process highlights the effectiveness of integrating multiple methodologies and specialized knowledge retrieval to boost decision accuracy and reliability.

Evaluation Protocol Our primary evaluation metric is ***Total Discordance***, which measures the percentage of queries incorrectly predicted by the model. A lower Total Discordance indicates better performance and is the most critical measure of our model's accuracy. This metric is paramount because it provides a comprehensive overview of the model's overall predictive accuracy across all types of queries. In addition to Total Discordance, we also assess the *Undertriage rate*, *Overtriage rate*, *Significant Undertriage rate*, and *Significant Overtriage rate*. These metrics are essential for evaluating the model's performance in categorizing the urgency of clinical events but are secondary to the overall accuracy reflected by Total Discordance. We further categorize input queries into five complexity levels (I-V) in descending order, with additional details and formulas provided in Appendix E.

5. Results

5.1. Main Results

Performance is evaluated by averaging results across three test sets. The backbone LLMs used are GPT-3.5, GPT-4, Llama-2, and Llama-3. Table 2 presents the main results, showcasing the superior performance of GPT-3.5 and GPT-4. Additional results for Llama-2 and Llama-3 can be found in Appendix F.

Table 2. Performance comparison of our completion on Emergency Severity Index Levels dataset. Results are averaged values of three independent clinic document classification predictions (%). The highest score (%) is highlighted in **bold**. Lower Total discordance represents higher performance. The median of three trials is reported when non-zero *temperature* is used. SCtr denotes the self-contrast prompting method, and SCons denotes the self-consistency prompting method. EoT denotes the Exchange-of-Thought prompting method.

SCENARIO	METHOD	TOTAL DISCORDANCE	UNDERTRIAGE	SIGNIFICANT UNDERTRIAGE	OVERTRIAGE	SIGNIFICANT OVERTRIAGE
GPT-3.5 -SUPERVISED	VANILLA	39.18	21.76	15.28	18.06	6.85
-ZERO-SHOT	MEDAGENT (W/HANDBOOK)	39.58	**5.56**	5.56	34.03	15.97
	COT(1-AGNT)	41.40	16.70	12.50	24.70	8.33
	SCTR(1-AGNT)	39.35	11.57	8.33	27.78	11.58
	SCONS(1-AGNT)	36.11	17.59	8.33	**15.74**	7.87
	EOT(4-AGNT)	36.81	11.57	6.94	15.28	9.72
	SCONS(4-AGNT)	34.72	5.56	**4.63**	27.31	10.65
	SCONS(4-AGNT) (W/HANDBOOK)	31.02	7.41	7.41	23.61	6.94
	SCONS(4-AGNT)+CONFIDENCE (W/HANDBOOK)	32.87	6.02	6.02	26.85	10.65
	TRIAGEAGENT (VANILLA)	34.72	**5.56**	**4.63**	27.31	10.56
	TRIAGEAGENT (W/HANDBOOK)	31.02	7.87	7.41	22.69	**5.56**
	TRIAGEAGENT (OURS)	**30.56**	6.94	6.48	24.54	9.72
GPT-4 -SUPERVISED	VANILLA	23.50	8.10	6.94	14.80	8.33
	VANILLA (W/HANDBOOK)	22.68	9.50	5.70	**7.10**	**1.90**
-ZERO-SHOT	MEDAGENT (W/HANDBOOK)	30.56	4.17	3.24	25.93	18.52
	COT(1-AGNT)	37.40	14.30	8.33	23.30	10.64
	EOT(4-AGNT)	29.86	9.03	5.56	20.83	12.50
	SCONS(4-AGNT)	29.63	11.11	7.87	18.06	8.33
	SCONS(4-AGNT) (W/HANDBOOK)	23.61	5.09	3.70	18.52	9.26
	SCONS(4-AGNT)+CONFIDENCE (W/HANDBOOK)	23.61	5.09	3.70	18.52	9.26
	TRIAGEAGENT (VANILLA)	29.63	11.11	7.87	18.06	8.33
	TRIAGEAGENT (W/HANDBOOK)	23.61	5.09	3.70	18.52	9.26
	TRIAGEAGENT (OURS)	**18.98**	**2.30**	**2.80**	17.10	8.80
HUMAN EVAL	**NURSING**	31.43	12.80	8.61	18.60	10.50

Performance Comparison with Baselines The performance comparison with the SOTA methods is shown in Table 2. TRIAGEAGENT framework surpasses traditional prompt engineering methods like Supervised Learning, Self-Contrast, and Self-Consistency, with gains of 9.25%, 8.79%, and 5.55%, respectively. Additionally, TRIAGEAGENT outperforms SOTA multi-agent frameworks (*EoT* and *MedAgents*) in a zero-shot setting on the GPT-3.5-turbo model, exceeding *MedAgents* by 9.02% and *EoT* by 6.25%.

Comparison with Complex CoT Methods Interestingly, performance can unexpectedly decline when using overly complex CoT methods. Simply stacking prompting methods without a clear logical sequence allocation or prioritization can inadvertently lead to hallucinations—erroneous outputs resulting from insufficient document comprehension and misunderstandings of medical terminologies. In contrast, our approach, which combines multi-agent role-playing with confidence assessments, effectively mitigates these issues. This demonstrates its potential as a more robust method for applying LLMs in clinical triage settings.

Comparison with Single-Agent Methods Methods such as *CoT*, *Self-Contrast*, and *Supervised Learning* overlook the crucial internal interactions among multiple LLMs. This lack of interaction prevents these methods from adaptively refining their responses with insights from other roles, leading to underperformance in triage question-answering scenarios. Consequently, the discordance rates for these methods hover around 19% on our datasets, highlighting a significant area for improvement.

5.2. Ablation Study

Our TRIAGEAGENT framework, which simulates a multi-role team collaboration, has been evaluated for its impact on final outcomes. Our ablation study focused on team optimization and integrating external resources; agents utilizing the Emergency Severity Index (ESI) handbook effectively reduced total discordance from 41.4%.

Team Optimization Our heterogeneous multi-agent collaboration framework involves multiple agents, each playing specific roles to derive the best possible outcomes collaboratively. We explored various architectural configurations to enhance team effectiveness and performance. By adjusting the number of role-specific agents while controlling for other variables, we found that a configuration of four agents optimally improves performance within the ESI dataset. This optimal configuration is detailed in Figure 3(a), illustrating how adjusting agent roles and numbers enhances the overall system performance.

External Resource Optimization In assessing various sources for supervising our model, we focused on the ESI Handbook, PubMed, and Wikipedia, as shown in Figure 3(c). Among these, the ESI Handbook proved the most effective due to its highly targeted content for ESI text classification. Its superior specialization and accuracy enhance model performance by ensuring relevant information and minimizing irrelevant data, thus improving search efficiency and reducing computational resource consumption. Additionally, We analyzed the optimal number of agents utilizing the ESI Handbook to maximize task efficiency, as illustrated in Figure 3(b).While increasing the number of agents accessing the ESI Handbook did not drastically change overall accuracy, it significantly boosted the model's effectiveness. This strategy enhances the agents' contextual understanding and interpretability of medical texts, optimizing the balance between performance improvement and resource utilization.

5.3. Case Study

Error Analysis Based on our findings, we conducted an expert evaluation to identify our model's key limitations and common issues. As shown in Figure 4, we categorized these errors into four major types: (1) **Lack of Document Understanding**: These errors arise from the model's insufficient grasp of essential medical knowledge or its failure to correctly link that knowledge to ESI levels, leading to misjudgments of clinical severity. (2) **Mis-retrieval of Domain Knowledge**: These errors occur when the model inaccurately retrieves irrelevant or mismatched information using external tools, compromising triage accuracy. (3) **Confidence-based Consistency Errors**: These errors occur when the model confidently provides contradictory responses to the same query, indicating confusion in understanding or applying knowledge. Additionally, situations where agents fail to reach a consensus and trigger the early stopping mechanism after multiple discussions are included in this category. Such errors result in falsely assured incorrect outcomes and may stem from inefficiencies in the model's internal processing or flaws in the early stopping mechanism, which can propagate incorrect information. (4) **Exchange of Information Errors**: These errors involve incorrect data transfer between agents, disrupting logical sequences and leading to wrong conclusions as agents change correct responses based on flawed information. See Appendix G for more details.

Table 3. Time efficiency on multiple agents (average seconds per test case)

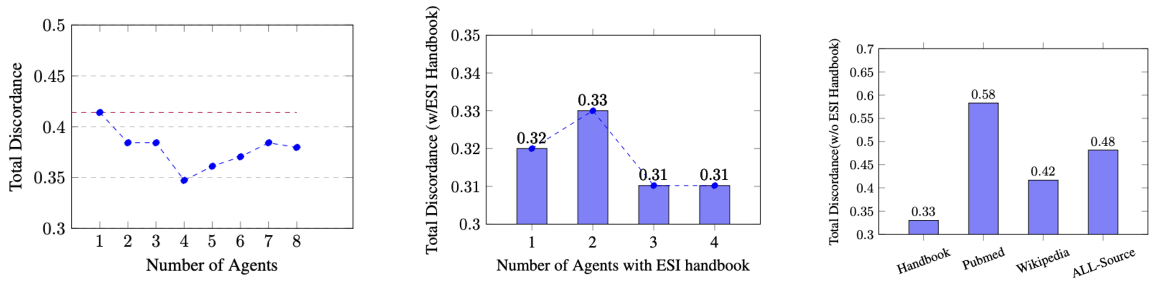
MODEL	TEST-1	TEST-2	TEST-3
CoT(GPT-3.5)	0.17	0.19	0.17
CoT(GPT-4)	0.16	0.19	0.17
SELF-CONSISTENCY(GPT-3.5)	0.21	0.23	0.24
THREE-AGENTS(GPT-3.5)	0.55	1.01	0.58
THREE-AGENTS(GPT-4)	0.57	1.00	0.59
FOUR-AGENTS(GPT-3.5)	1.28	1.36	1.41
FOUR-AGENTS(GPT-4)	1.31	1.56	1.53
FIVE-AGENTS(GPT-3.5)	1.55	1.56	1.49
SIX-AGENTS(GPT-3.5)	2.11	2.05	2.13
TRIAGAGENT(GPT-3.5)	1.30	1.45	1.52
TRIAGAGENT(GPT-4)	1.31	1.43	1.50

Time Analysis A potential concern with our framework is its time efficiency, which is crucial in emergency clinical triage. Table 3 compares the performance and time costs of different methods. While our multi-agent architecture slightly increases time costs compared to a single-agent model, this increase is minimal. Feedback from emergency department experts indicates it does not significantly impact overall efficiency. More importantly, it greatly enhances decision accuracy, which is essential for better patient outcomes and effective resource allocation.

Cost Analysis Another crucial aspect of our framework is its operational efficiency, particularly regarding API token usage costs. Table 4 evaluates the performance and costs for different configurations. While TRIAGEAGENT requires more API calls than a single-agent setup, slightly increasing costs, feedback from medical departments indicates that these costs are reasonable and manageable in real-world settings. Our system not only improves efficiency by 12.54% compared to standard human experts but also achieves an 18.54% higher performance compared to LLMs using CoT processing based on human experts.

6. Discussions

In this paper, we introduce a heterogeneous multi-agent collaboration framework called TRIAGEAGENT. Despite our efforts, the framework faces certain limitations due to the unique characteristics of the healthcare industry.



(a) (b) (c)

Figure 3. Ablation study results. (a), (b) and (c) show the impact of agent structure composition, optimizing the agent team with the ESI handbook and optimizing the agent team with various external resources, respectively. The lower *Total Discordance* value in this figure represents higher performance.

Table 4. Cost efficiency on ESI datasets classification

MODEL	OVERALL PERFORMANCE	#API CALLS
TWO-AGENT(GPT-3.5)	38.42%	324
THREE-AGENT(GPT-3.5)	38.42%	486
FOUR-AGENTS(GPT-3.5)	30.56%	648
FOUR-AGENTS(GPT-4)	18.98%	540
COT(GPT-4)	37.40%	216

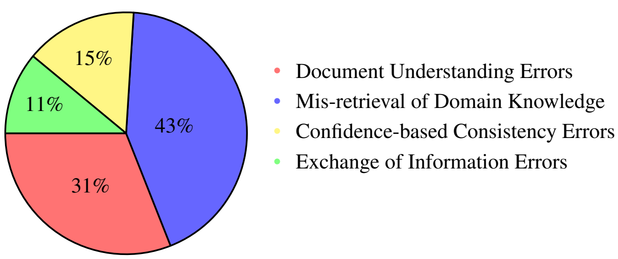


Figure 4. Ratio of different categories in error cases.

Limited Expert Evaluation Our research involves only three human experts, which limits the scope and depth of expert evaluation and may affect the generalizability of our findings in broader clinical settings.

Cross Institution Collaboration The complexity of emergency departments across specialties requires more sophisticated triage systems to address the nuances of patient conditions within the same urgency levels. This need arises from different institutions' varying conditions, operational protocols, and patient demographics.

Workflow Integration While our model provides a final triage decision, actual triage often involves multiple decision-making and information integration stages. Therefore, evaluating how well our system integrates into existing workflows and complements human-led emergency care processes is crucial.

Privacy Deploying our framework requires rigorous attention to privacy concerns and thorough training for clinical workers. Although it does not train on private data, processing clinical notes still carries the risk of exposing sensitive information. Ensuring compliance with privacy regulations such as HIPAA and GDPR is essential. Future framework versions should incorporate advanced data protection methods, such as encryption and differential privacy, to address these issues.

Time and Cost Efficiency TRIAGEAGENT aims to enhance clinical triage efficiency by automating the initial patient assessment stages, potentially reducing manual workload and waiting times. This could help lower occupational burnout and operational costs. However, relying on external technologies like OpenAI's API introduces additional expenses and dependency. A detailed cost-benefit analysis is essential to ensure the framework's sustainability and economic viability.

7. Conclusions

This paper introduces a novel heterogeneous multi-agent framework for clinical triage, using role-playing agents derived from LLMs in a multi-stage group chat setting. Experiments on three ESI-based clinical datasets show our framework significantly outperforms zero-shot baselines and experienced professionals. Human evaluations suggest future improvements in reducing document understanding errors, knowledge misretrieval, confidence consistency errors, and information exchange errors. After accepting our work, we will release our dataset and make it open source.

8. Acknowledgements

Our work is sponsored by the NSF NAIRR Pilot and PSC Neocortex, Commonwealth Cyber Initiative, Children's National Hospital, Fralin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and a generous gift from the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning.

References

Emergency severity index (esi): A triage tool for emergency department. Available online, Accessed: 2024-04-06. https://media.emscimprovement.center/documents/ESI_Handbook2125.pdf.

Bao, Z., Chen, W., Xiao, S., Ren, K., Wu, J., Zhong, C., Peng, J., Huang, X., and Wei, Z. Disc-medllm: Bridging general large language models and real-world medical consultation, 2023.

Bi, Z., Dip, S. A., Hajialigol, D., Kommu, S., Liu, H., Lu, M., and Wang, X. Ai for biomedicine in the era of large language models, 2024.

Castro, M. and Liskov, B. Practical byzantine fault tolerance. In *Proceedings of the Third Symposium on Operating Systems Design and Implementation*, pp. 173–186, USA, 1999. USENIX Association. ISBN 1880446391.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways, 2022.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate, 2023.

Fu, Y., Peng, H., Khot, T., and Lapata, M. Improving language model negotiation with self-play and in-context learning from ai feedback, 2023.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J. Metagpt: Meta programming for a multi-agent collaborative framework, 2023.

Jin, Q., Yang, Y., Chen, Q., and Lu, Z. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information, 2023.

Kang, M., Lee, S., Baek, J., Kawaguchi, K., and Hwang, S. J. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks, 2023.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners, 2023.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, Y., Zhang, Y., and Sun, L. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents, 2023.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. Encouraging divergent thinking in large language models through multi-agent debate, 2023.

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., White, C., Gu, Q., Pei, J., Yang, C., and Zhao, L. Domain specialization as the key to make large language models disruptive: A comprehensive survey, 2024.

Liu, Z., Wu, Z., Hu, M., Zhao, B., Zhao, L., Zhang, T., Dai, H., Chen, X., Shen, Y., Li, S., Murray, B., Liu, T., and Sikora, A. Pharmacygpt: The ai pharmacist, 2023.

Ma, D., Wang, M., Xiang, A., Qi, Z., and Yang, Q. Transformer-based classification outcome prediction for multimodal stroke treatment, 2024.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. Capabilities of gpt-4 on medical challenge problems, 2023.

Oniani, D., Wu, X., Visweswaran, S., Kapoor, S., Kooragayalu, S., Polanska, K., and Wang, Y. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines, 2024.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022.

Qian, C., Cong, X., Liu, W., Yang, C., Chen, W., Su, Y., Dang, Y., Li, J., Xu, J., Li, D., Liu, Z., and Sun, M. Communicative agents for software development, 2023.

Qin, Y., Hu, S., Lin, Y., Chen, W., Ding, N., Cui, G., Zeng, Z., Huang, Y., Xiao, C., Han, C., Fung, Y. R., Su, Y., Wang, H., Qian, C., Tian, R., Zhu, K., Liang, S., Shen, X., Xu, B., Zhang, Z., Ye, Y., Li, B., Tang, Z., Yi, J., Zhu, Y., Dai, Z., Yan, L., Cong, X., Lu, Y., Zhao, W., Huang, Y., Yan, J., Han, X., Sun, X., Li, D., Phang, J., Yang, C., Wu, T., Ji, H., Liu, Z., and Sun, M. Tool learning with foundation models, 2023.

Shaib, C., Li, M. L., Joseph, S., Marshall, I. J., Li, J. J., and Wallace, B. C. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success), 2023.

Sharaf, S. and Anoop, V. S. An analysis on large language models in healthcare: A case study of biobert, 2023.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Scharli, N., Chowdhery, A., Mansfield, P., y Arcas, B. A., Webster, D., Corrado, G. S., Matias, Y., Chou, K., Gottweis, J., Tomasev, N., Liu, Y., Rajkomar, A., Barral, J., Semturs, C., Karthikesalingam, A., and Natarajan, V. Large language models encode clinical knowledge, 2022.

Sun, H., Zhuang, Y., Kong, L., Dai, B., and Zhang, C. Adaplanner: Adaptive planning from feedback with language models, 2023.

Sánchez-Salmerón, R. et al. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing*, 60:101109, 2022. doi: 10.1016/j.ienj.2021.101109.

Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., and Gerstein, M. Medagents: Large language models as collaborators for zero-shot medical reasoning, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models, 2023.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J.-R. A survey on large language model based autonomous agents, 2024a.

Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., and Wei, F. Augmenting language models with long-term memory, 2023a.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models, 2023b.

Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, 2024b.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., and Gui, T. The rise and potential of large language model based agents: A survey, 2023.

Xie, T., Zhou, F., Cheng, Z., Shi, P., Weng, L., Liu, Y., Hua, T. J., Zhao, J., Liu, Q., Liu, C., Liu, L. Z., Xu, Y., Su, H., Shin, D., Xiong, C., and Yu, T. Openagents: An open platform for language agents in the wild, 2023.

Xiong, K., Ding, X., Cao, Y., Liu, T., and Qin, B. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate, 2023.

Yao, L., Leung, K., Tsai, C., Huang, C., and Fu, L. A novel deep learning-based system for triage in the emergency department using electronic medical records: Retrospective cohort study. *J Med Internet Res*, 23(12):e27008, 2021. doi: 10.2196/27008.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models, 2023.

Yin, Z., Sun, Q., Chang, C., Guo, Q., Dai, J., Huang, X., and Qiu, X. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication, 2023.

Zhang, W., Shen, Y., Wu, L., Peng, Q., Wang, J., Zhuang, Y., and Lu, W. Self-contrast: Better reflection through inconsistent solving perspectives, 2024a.

Zhang, X., Tian, C., Yang, X., Chen, L., Li, Z., and Petzold, L. R. Alpacare:instruction-tuned large language models for medical application, 2024b.

Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., and Neubig, G. Webarena: A realistic web environment for building autonomous agents, 2023.

Zhuang, Y., Yu, Y., Wang, K., Sun, H., and Zhang, C. Toolqa: A dataset for llm question answering with external tools, 2023.

A. Model Architecture Comparison with Previous Work

We compare work between TRIAGEAGENT and representative previous works as shown in Figure 5. We refer to our system as a **heterogeneous multi-agent** framework because each agent represents a different role, creating a role-playing heterogeneous structure. This structure allows us to simulate real-world collaboration among different roles, with each agent (or role) having its unique function and task, thereby improving the overall system's decision-making efficiency and accuracy. Additionally, the changing colors of nodes in the figure illustrate our dynamic heterogeneous nature, which represents agents' evolving perspectives based on different rounds of discussions. In the second row, the nodes represent agents at different time steps, arrows indicate the edges, and the color signifies the role of an agent. Among these works, TRIAGEAGENT is the only one demonstrating all four pivotal dimensions of LLM-agent collaboration: compatibility with multiple roles, an early-stopping mechanism, support for dynamic interactions, and agent team optimization.

B. Query Details

below:
query q: What is the ESI level of the following clinic record? Please give me a final unique answer after a second revision of your first proposed answer. You can learn from the Emergency Severity Index handbook v4. The Candidate answers: ESI-1,2,3,4,5. Double-check the ESI handbook and ask yourself again(two-round self-check)when you are sure about this ESI level classification before you give me the answer. Then classify the following medical record according to ESI level, candidate answers are ESI-1,2,3,4,5. Here is the record:

C. Early-stopping Mechanism Details

To enhance the efficiency of group chat discussions, we have implemented an improved early-stopping mechanism inspired by Byzantine Consensus theory. This approach requires at least $3p + 1$ agents to handle p faulty agents in a single communication round (Castro & Liskov, 1999). Additionally, we introduce a termination mechanism inspired by recent advancements showing that LLMs fine-tuned with Reinforcement Learning from Human Feedback (RLHF) can reach consensus after several rounds of debate (Du et al., 2023; Ouyang et al., 2022). Our mechanism relies on multiple rounds of high-confidence validation, terminating communication when agents consistently confirm their reasoning with high confidence. This global termination criterion reduces unnecessary computations and ensures that discussions conclude once a reliable answer is reached. Based on these principles, our early-stopping mechanism operates under the following conditions: 1). **Repetition of High Confidence Answers by a Single Agent**: If an agent repeatedly provides the same answer with high confidence across successive rounds, early-stopping is triggered for that agent. The agent will cease all interactions, stop transmitting or receiving data, and withdraw from the ongoing conversation. 2). **Repetition of High Confidence Answers by Multiple Agents**: When several agents consistently provide the same answer with high confidence during a discussion round, the dialogue can be terminated, and early-stopping is triggered. This dynamic, real-time stopping condition enhances the traditional Byzantine requirement by incorporating these specific conditions, ensuring efficient and accurate consensus in group discussions. These conditions collectively foster a comprehensive and adaptive termination criterion, responding to the consensus among the majority of agents (Yin et al., 2023) and prioritizing efficiency and accuracy in reaching conclusions.

D. Dataset Construction

The dataset was constructed by extracting relevant patient cases from the ESI handbook v4, focusing on a comprehensive range of clinical scenarios. Each case was carefully reviewed and labeled by professional human experts to ensure accuracy. Each case was carefully reviewed and labeled by professional human experts to ensure accuracy. The dataset is divided into a training set and three test sets (test-1, test-2, and test-3), with the training set containing 180 cases and each test set containing 72 cases. We maintained fixed proportions of ESI levels in the test sets as follows: ESI-1 (12), ESI-2 (20), ESI-3 (13), ESI-4 (12), and ESI-5 (15).

For the training scenarios, we have a total of 218 cases with the following distribution across ESI levels: ESI-1 (14), ESI-2 (92), ESI-3 (65), ESI-4 (22), and ESI-5 (25). The explanations in the training dataset are manually annotated by human experts. Detailed proportions of each ESI level in the training and test sets are provided in Table 1. The term ”VIGNETTES” refers to the concise case descriptions used for training and evaluation.

Each training and test set includes cases with detailed clinical notes and corresponding ESI levels. Although the test sets do not contain explicit explanations for each label, they have been accurately annotated by human experts based on the clinical information provided. The lack of explanations in the test set is due to the scarcity of medical data and the high cost of manual labeling. This highlights the advantage of our framework in reducing human labor costs and improving the efficiency of medical text classification.

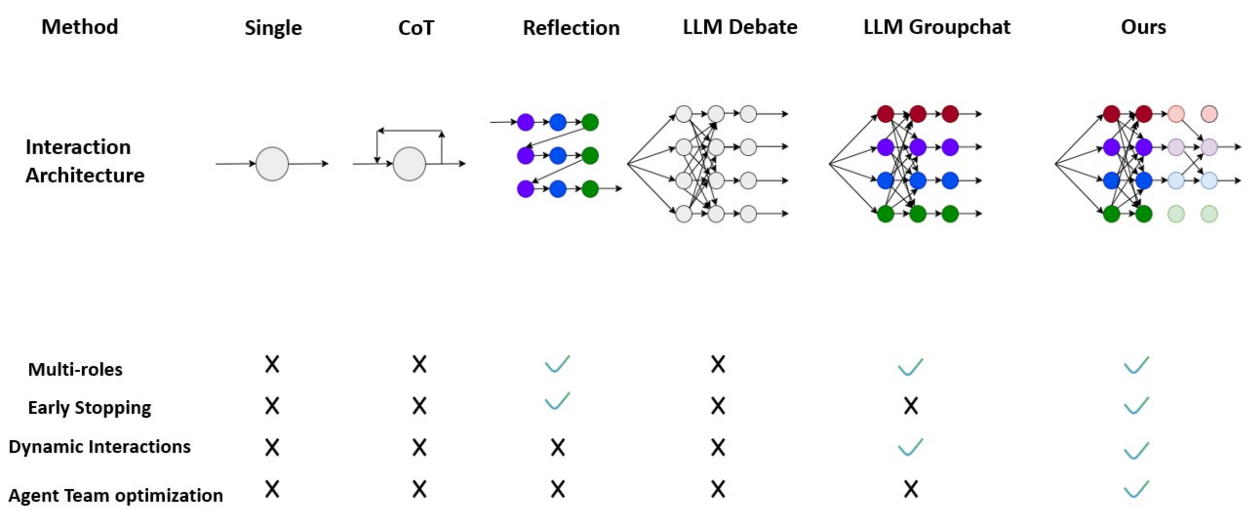


Figure 5. Topology Structure Comparisons of previous methods

E. ESI Hierarchy and Evaluation Metrics

E.1. Hierarchy of ESI levels

The ESI (Emergency Severity Index) levels classify medical events by urgency, from 1 (most urgent) to 5 (least urgent). Medical personnel can refer to the medical event in the ESI handbook with their medical experience to quickly categorize the current emergency medical event according to the ESI classification.

•ESI-1: Most Urgent; Immediate life-saving intervention

•ESI-2: High urgency; Potentially life-threatening, prompt attention necessary.

•ESI-3: Urgent; Requires multiple resources but not immediately life-threatening.

•ESI-4: Less urgent; Requires one resource; not immediately life-threatening.

•ESI-5: Least Urgent; No resources needed immediately;wait time is acceptable.

E.2. Evaluation Metrics

The *Total discordance* is calculated as the ratio of the total number of misclassified texts to the total number of texts, representing the overall error rate of the model. This metric is given by the formula:

$$\textbf{Total discordance} = \frac{\text{Total Misclassifications}}{\text{Total number of texts}} \tag{1}$$

where *Total misclassifications* is the number of queries incorrectly classified by the model, and *Total number of texts* is the total number of queries analyzed. The *Undertriage* rate is defined as the fraction of instances where the predicted label is greater than the true label, which is calculated as:

$$\textbf{Undertriage} = \frac{\text{Number of predictionsTrue\_labels}}{\text{Total number of texts}} \tag{2}$$

Similarly, the *Overtriage* rate is defined as the fraction of instances where the predicted label is less than the true label:

$$\textbf{Overtriage} = \frac{\text{Number of predictionsTrue\_labels}}{\text{Total number of texts}} \tag{3}$$

The *Significant Undertriage* rate captures the scenarios where the true label is critical (ESI *1* or*2*) but the prediction underestimates the urgency (predicted as *3*, *4*, or *5*):

$$\textbf{Significant Undertriage} = \frac{\text{Predicted-3, 4, or 5}}{\text{Total number of texts}} \tag{4}$$

Lastly, the *Significant Overtriage* rate addresses the instances where less critical labels (true label is 2, 3, or 4) are overestimated by the model (predicted as 1):

$$\textbf{Significant Overtriage} = \frac{\text{Predicted-3,4,or 5}}{\text{Total number of texts}} \tag{5}$$

Table 5. Performance of LlaMa-3-8B on our clinical triage task.

METHOD	ZERO-SHOT(COT)
TOTAL DISCORDANCE	**56.25%**
UNDERTRIAGE	3.48%
SIGNIFICANT UNDERTRIAGE	3.48%
OVERTRIAGE	52.78%
SIGNIFICANT OVERTRIAGE	35.42%
METHOD	FINE-TUNED
TOTAL DISCORDANCE	**27.78%**
UNDERTRIAGE	10.65%
SIGNIFICANT UNDERTRIAGE	7.64%
OVERTRIAGE	17.13%
SIGNIFICANT OVERTRIAGE	6.94%

F. Performance of Different Backbone LLMs

Using Different Language Models In our experiments, we utilized GPT-3.5-turbo [1], GPT-4 [2], their combination, and LLaMa-2-7B[3] and Llama-3-8B [4].When GPT-3.5-turbo and GPT-4 engaged in interactive debates on AI-Triage tasks, we observed performance improvements of 31.88% for GPT-3.5-turbo and 24.6% for GPT-4. In a zero-shot scenario, LLaMa-2-7B performed 17.025% worse than LLaMa-3-8B. We also fine-tuned LlaMa-3-8B using training scenarios from the ESI handbook to enhance its domain knowledge. Compared to GPT-4, which achieved a performance improvement of 23.50% on the same training dataset, LLaMa-3-8B showed a performance improvement of 27.78%, making it 4.28% worse than GPT-4.Detailed results for LLaMa-3 are presented in Table 5, demonstrating that its performance on clinical triage tasks was inferior to that of the GPT models. For details about the performance of LlaMa-3 on our task, see Table 5

G. Case Study

To better illustrate the issue, we have excerpted parts of the complete conversation as key examples to more effectively explain the corresponding types of errors.

•Type-1: Lack of Document Understanding

Admin (to chat_manager):
What is the ESI level of the following clinic record? Please give me a final unique answer after a second revision of your first proposed answer. The Candidate answers: ESI-1,2,3,4,5. Double-check the clinic notes and related medical knowledge and ask yourself again (two-round self-check) when you are sure about this ESL level classification before you give me the answer. Here is the record:" A healthy 10-year-old male is brought to the emergency department by his mother, who reports that her son has not moved his bowels for a week. He is complaining of 7/10 generalized abdominal pain, nausea, and lack of appetite. Vital signs: BP 107/66, HR 75, RR 20, T 98.6°F, SpO2 99%".

Agent_1 (to chat_manager):
Based on the clinical record provided, the most appropriate Emergency Severity Index (ESI) level for this patient would be ESI level 3. This decision is based on the patient's presenting complaints of abdominal pain, nausea, and lack of appetite, along with the vital signs indicating stable blood pressure, heart rate, respiratory rate, temperature, and oxygen saturation. While the patient's symptoms are concerning and require evaluation, there are no immediate life-threatening issues or unstable vital signs that would warrant a higher ESI level.
Confidence score: 70%.

Agent_2 (to chat_manager):
Based on a thorough review of the clinical record and consideration of the patient's presentation, vital signs, and symptoms, I would revise the ESI level for this patient to ESI level 2. The patient's complaints of abdominal pain, nausea, lack of appetite, and constipation for a week warrant urgent evaluation, but the stable vital signs indicate that there are no immediate life-threatening issues. However, the patient's symptoms require timely assessment and treatment, justifying a higher acuity level than initially considered.
Confidence score: 85%.

A*gent_3 (to chat_manager):*
Based on the information provided in the clinical record, a more appropriate Emergency Severity Index (ESI) level for this patient would be ESI level 2. The patient is a 10-year-old male with concerning symptoms of abdominal pain, nausea, lack of appetite, and constipation for a week. While the vital signs are stable, the duration of symptoms and the need for urgent evaluation justify a higher acuity level. This patient may require timely intervention and assessment to address the underlying cause of his symptoms, making ESI level 2 more appropriate in this case.
Confidence score: 90%.

Result_summarizer (to chat_manager):
After synthesizing the inputs from the Agent_1, Agent_2, and Agent_3, the final unique answer for the ESI level of the patient in the clinical record would be ESI level 2. This decision is based on the patient being a 10-year-old male presenting with concerning symptoms of abdominal pain, nausea, lack of appetite, and constipation for a week. While the vital signs are stable, the duration and nature

[1] https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/
[2] https://openai.com/index/gpt-4/
[3] https://llama.meta.com/llama2/
[4] https://llama.meta.com/llama3/

of the symptoms warrant urgent evaluation and potential intervention, justifying a higher acuity level.
Confidence score: 92%.

•Type-2: Mis-retrieval of Domain Knowledge

This example mis-retrieved the domain knowledge from *ESI Handbook v4, Chapter 2: ESI Triage Algorithm, p. 10-13.* Then this example was erroneously categorized as needing urgent intervention.

User_proxy (to chat_manager):
What is the ESI level of the following clinic record according to the Emergency Severity Index handbook? Please give me a final unique answer after a second revision from your first proposed answer. The Candidate answers: ESI-1,2,3,4,5. Double Check the clinic notes and related medical knowledge and ask yourself again (two-round self-check) when you are sure about this ESL level classification before you give me the answer. Here is the record:" This sounds really strange. A bug flew into my right ear while I was gardening. I tried to get it out by using a Q-tip. I just don't know what else to do, but this buzzing noise is driving me crazy," a 55-year-old female tells you. No previous medical history and vital signs are within normal limits.

Agent_1 (to chat_manager):
Based on the information provided and the Emergency Severity Index handbook, this patient's clinic note would be categorized as Low ESI level. The patient's condition is not life-threatening but requires medical intervention, which aligns with a lower ESI level.
References: - ESI Handbook v4, Chapter 2: ESI Triage Algorithm, p.10-13.

Agent_2 (to chat_manager):
The scenario describes a 55-year-old female who has a foreign object (a bug) in her ear, causing discomfort and a buzzing noise. She has no previous medical history and her vital signs are within normal limits. This is a non-life-threatening condition, but it does require medical intervention to remove the foreign object and alleviate the patient's discomfort.
The Emergency Severity Index (ESI) is a five-level emergency department triage algorithm that categorizes patients by both acuity and resource needs. ESI level 1 is for patients who need immediate life-saving intervention, while ESI level 5 is for patients with non-urgent conditions. Given the patient's condition, she would likely be classified as ESI level 4. This level is for patients with conditions that could potentially escalate to a more serious problem requiring emergency intervention if left untreated, but are not life-threatening. The patient's condition is not life-threatening, but it is causing her significant discomfort and could potentially lead to complications such as infection if not addressed.
The most alarming aspect of this issue is the potential for complications if the foreign object is not removed promptly. The use of a Q-tip could potentially push the bug further into the ear canal, causing more damage. It's also noteworthy that the patient's vital signs are within normal limits, indicating that she is not in immediate danger.
After a second revision, I still believe that the ESI level for this patient would be ESI level 4. The patient's condition is not life-threatening, but it does require medical intervention.
Confidence score: 90%.

Agent_3 (to chat_manager):
After reviewing the patient's condition and the Emergency Severity Index handbook, I would classify this patient as ESI level 4 with a confidence score of 90%. The patient's condition is not life-threatening but requires medical intervention to remove the foreign object from her ear and alleviate her discomfort. This aligns with the criteria for ESI level 4, which is for conditions that could potentially escalate to a more serious problem if left untreated but are not life-threatening.

Agent_2 (to chat_manager):
Based on the patient's clinical record and the Emergency Severity Index handbook, I would classify this patient as ESI level 4 with a confidence score of 90%. The patient's condition is not life-threatening but requires medical intervention to remove the foreign object from her ear and alleviate her discomfort. This aligns with the criteria for ESI level 4, which is for conditions that could potentially escalate to a more serious problem if left untreated but are not life-threatening.

Result_summarizer (to chat_manager):
After synthesizing the inputs from the Agent_1, Agent_2, and Agent_3, the consensus is that the patient should be classified as ESI level 4. This level is appropriate for patients with conditions that could potentially escalate to a more serious problem if left untreated but are not life-threatening. The patient's condition of having a bug in her ear is causing significant discomfort and requires medical intervention, but it is not life-threatening. Therefore, the final decision is ESI level 4.

•Type-3: Confidence-based Consistency Errors

This error might be caused by the early-stopping mechanism and confidence-score passing process.

Admin (to chat_manager):
What is the ESI level of the following clinic record? Please

give me a final unique answer after a second revision from your first proposed answer. The Candidate answers: ESI-1,2,3,4,5. Double Check the clinic notes and related medical knowledge and ask yourself again(two-round self-check)when you are sure about this ESL level classification before you give me the answer.Here is the record: "I slipped on the ice, and I hurt my wrist," reports a 58-year-old female with a history of migraines. There is no obvious deformity. Vital signs are within normal limits, and she rates her pain as 5/10.

Response 1 by chat_manager (initial response):
Based on the patient's description and clinical information provided, I recommend classifying this patient as ESI level 4. The injury mechanism and the pain level, along with normal vital signs, suggest a non-life-threatening condition. Confidence score: 90%.

Response 2 by chat_manager (second response after self-check):
Upon further reflection and review of the medical knowledge, I now assess the patient's condition as ESI level 3. Considering the possibility of a hidden injury due to the nature of the fall, a more cautious approach is warranted.
Confidence score: 92%.

•Type-4: Exchange of Information Errors

User_proxy (to chat_manager):
What is the ESI level of the following clinical record according to the Emergency Severity Index handbook? Please provide a final answer after a second revision from your initial proposed answer. The candidate answers are: ESI-1, 2, 3, 4, 5. Double-check the clinic notes and related medical knowledge before finalizing your decision. Here is the record: "A bug flew into my right ear while I was gardening. I tried to get it out by using a Q-tip. I just don't know what else to do, but this buzzing noise is driving me crazy," a 55-year-old female tells you. No previous medical history and vital signs are within normal limits.

Agent_1 (to chat_manager):
Based on the information provided and the Emergency Severity Index handbook, this patient's clinic note would be categorized as a low ESI level. The patient's condition is not life-threatening but requires medical intervention, aligning with a lower ESI level,(e.g., ESI level 5), with a confidence score of 80%.

Agent_2 (to chat_manager):
The scenario describes a 55-year-old female with a foreign object (a bug) in her ear, causing discomfort and a buzzing noise. She has no previous medical history, and her vital signs are within normal limits. This is a non-life-threatening condition but requires medical intervention to remove the foreign object and alleviate the patient's discomfort.
Given the patient's condition, she would likely be classified as ESI level 4. This level is for patients with conditions that could potentially escalate to a more serious problem if left untreated but are not life-threatening. After a second revision, I still believe that the ESI level for this patient is ESI level 4, with a confidence score of 90%.

Agent_3 (to chat_manager):
After reviewing the patient's condition and the Emergency Severity Index handbook, I would classify this patient as ESI level 4 with a confidence score of 90%. The patient's condition is not life-threatening but requires medical intervention to remove the foreign object from her ear and alleviate her discomfort. This aligns with the criteria for ESI level 4.

Result_summarizer (to chat_manager):
After synthesizing the inputs from the Agent_1, Agent_2, and Agent_3, the consensus is that the patient should be classified as ESI level 4. This level is appropriate for patients with conditions that could potentially escalate to a more serious problem if left untreated but are not life-threatening. The patient's condition of having a bug in her ear is causing significant discomfort and requires medical intervention, but it is not life-threatening. Therefore, the final decision is ESI level 4, with a consensus confidence score of 90%.

Explanation: In this example, the inconsistent classification by Agent_1 ("low ESI level" as ESI level 5) compared to the other agents (ESI level 4) could lead to confusion and incorrect final decisions, representing an Exchange of Information Error.