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

Anonymous EMNLP submission

Abstract

The global escalation in emergency department patient visits poses significant challenges to efficient clinical management, particularly in clinical triage. Traditionally managed by human professionals, clinical triage is susceptible to substantial variability and high workloads. Although large language models (LLMs) demonstrate promising reasoning and understanding capabilities, directly applying them to clinical triage remains challenging due to the complex and dynamic nature of the clinical triage task. To address these issues, we introduce TRIAGEAGENT, a novel heterogeneous multi-agent framework designed to enhance collaborative decision-making in clinical triage. TRIAGEAGENT leverages LLMs for role-playing, incorporating self-confidence and early-stopping mechanisms in multi-round discussions to improve document reasoning and classification precision for triage tasks. In addition, TRIAGEAGENT employs the medical Emergency Severity Index (ESI) handbook through a retrieval-augmented generation (RAG) approach to provide precise clinical knowledge and integrates both coarse- and fine-grained ESI-level predictions in the decision-making process. Extensive experiments demonstrate that TRIAGEAGENT outperforms state-of-the-art LLM-based methods on three clinical triage test sets. Furthermore, we have released the first public benchmark dataset for clinical triage with corresponding ESI levels and human expert performance for comparison.

1 Introduction

Emergency Departments (EDs) play a crucial role in the healthcare system by continuously assessing and prioritizing patients based on urgency and severity. This process, known as **clinical triage**, utilizes the **Emergency Severity Index** (ESI) as a standardized guide for decisions on rapid medical intervention, which is vital for prioritizing treatment and allocating resources. However, the growing number of patients poses significant challenges to the rapid and precise classification of cases, which is crucial for accurate ESI categorization. Currently, hospitals rely on human experts to review clinical notes and determine case urgency (as illustrated in Figure 1). Although effective, this manual method is time-consuming, labor-intensive, and burdensome for clinical staff. The increasing patient volume and complex triage process often lead to staff fatigue, diminishing accuracy and efficiency, and raising the risk of inconsistent classification or misdiagnosis.

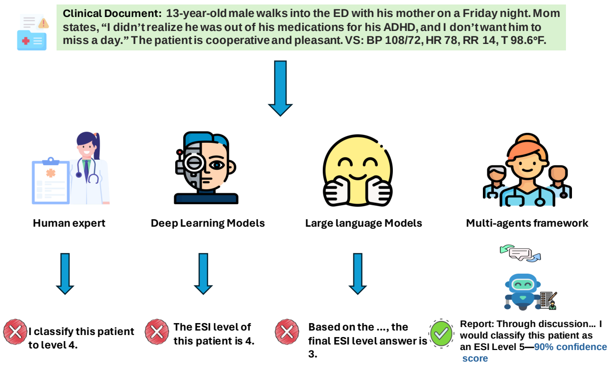


Figure 1: Workflow illustrations of clinical triage task.

Consequently, there is a high demand for AI methods to automate ESI classification. Traditional deep learning (DL) models (as illustrated in Figure 1) (Kojima et al., 2023; Yao et al., 2021; Sánchez-Salmerón et al., 2022) have assisted in clinical triage but often fall short due to the complex and dynamic nature of the task, which requires extensive labeled data and real-time adaptation. LLMs such as GPT (Kojima et al., 2023; OpenAI et al., 2024), Med-PaLM (Chowdhery et al., 2022), and Llama (Touvron et al., 2023) offer promising solutions with advanced text understanding capabilities, reducing time costs and errors by quickly interpreting and categorizing clinical documents. Additionally, LLMs can leverage external tools, such as knowledge base APIs (Qin et al., 2023; Zhuang et al., 2023), to enhance domain-specific knowledge, adaptability, speed, and accuracy. However, even with these tools, a single LLM may struggle with the complexities of emergency scenarios, such as diverse patient presentations and the need for immediate multi-disciplinary coordination (Chenais et al., 2023). Therefore, a multi-agent approach is necessary to improve performance by distributing tasks and assigning specialized roles to agents with diverse expertise.

Recent advancements have significantly enhanced multi-agent systems in areas such as reasoning (Wang et al., 2023b), sophisticated planning (Yao et al., 2023; Sun et al., 2023), and memory (Wang et al., 2023a). These improvements enable multi-agent LLMs to analyze medical data, formulate treatment plans, and recall patient histories more effectively (Tang et al., 2024). However, directly deploying multi-agent LLMs in clinical triage remains challenging due to their lack of optimization for triage-specific decision-making, resulting in accuracy levels around 60% (as illustrated in Table 2). This suboptimal performance stems from the intricate coordination required among agents and the need to design architectures that better utilize LLMs. Additionally, the lack of open-source benchmark datasets hampers comparisons with human experts, thereby affecting the practical credibility of multi-agent LLMs.

We identify four major challenges hindering clinical triage effectiveness. The first is **data processing speed**. The increasing patient volume necessitates rapid data processing and decision-making, yet traditional methods often face delays due to continuous data updates and extensive preprocessing, hindering timely clinical interventions. The second is **diversity in clinical documents**. Variability in patient histories, symptoms, writing styles and terminologies complicates the classification task. The third is **contextual understanding and interpretability**. The complexity of clinical contexts makes it challenging for models to accurately interpret information and provide transparent, evidence-based results, leading to a lack of clinician trust. The fourth is **boundaries between different ESI levels**. Precise differentiation and classification of ESI levels (1-5) are essential to avert critical medical errors. For instance, distinguishing between a level 1 (most urgent) and a level 2 (less urgent) case is vital, as it determines whether a patient receives immediate medical intervention or encounters a prolonged wait. Nevertheless, the subtle nuances between different ESI levels present considerable challenges to the clinical triage process.

To address these challenges, we propose TRIAGEAGENT, a novel heterogeneous multi-agent collaboration framework for clinical triage that leverages LLM-based agents enhanced with external tools and embedded medical knowledge. TRIAGEAGENT enables effective information exchange and reliable interactions among agents, making the framework ideal for zero-shot document classification and handling complex tasks without prior demonstrations. The **key innovations** include 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 enhance contextual understanding, interpretability, and precision in ESI-level classification. Additionally, the framework supports real-time decision-making and dynamic, interactive debates among agents, refining information for more accurate triage and improving the timeliness of clinical interventions.

Experiments on three ESI clinical triage test sets 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. Additionally, we are the first to publicly release a clinical triage dataset that includes clinical notes alongside ESI levels and human expert performance, providing a valuable resource for academic research and clinical practice. By setting new benchmarks, we aim to advance the field of clinical triage in both academic and practical applications. We will publicly release our code and dataset once the paper is published.

2 Related Work

2.1 LLM Applications in Healthcare

Large Language Models (LLMs) have recently experienced significant advancements across various fields, particularly in healthcare (Ling et al., 2024; Bi et al., 2024; Nori et al., 2023; Bao et al., 2023). These models are increasingly employed in medical applications, including text-based diagnostics(Ma et al., 2024), genetic analysis (Bi et al., 2024), pharmaceutical 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 to enhance clini-

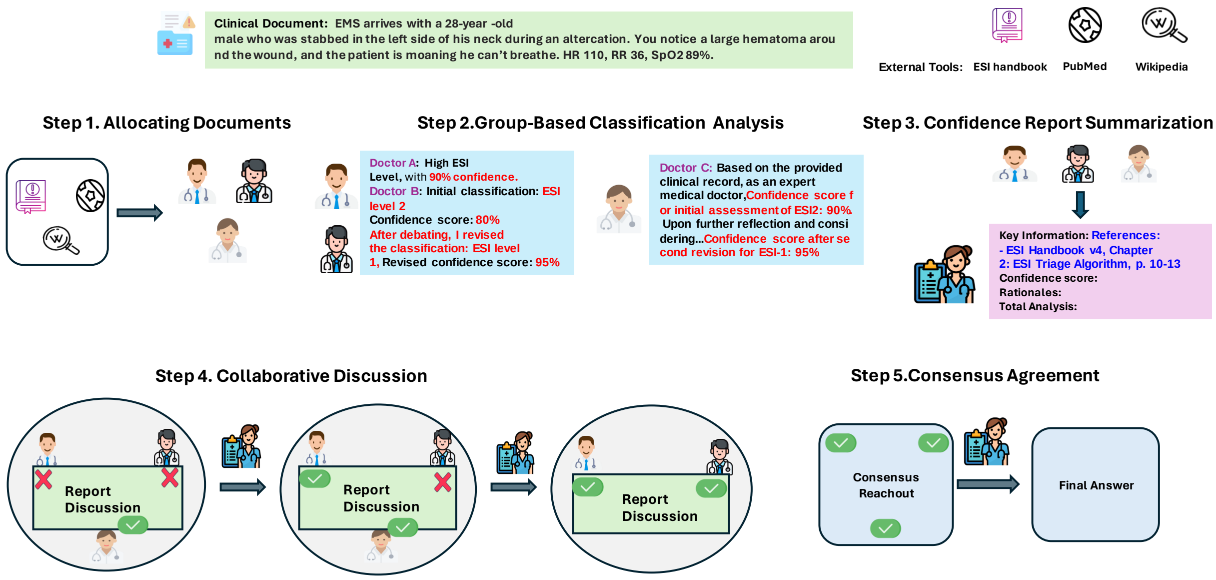


Figure 2: A diagram of our proposed heterogeneous multi-agent collaboration framework, TRIAGEAGENT, illustrates the reasoning process through five stages when given a clinical document as input. The five stages include 1) allocating documents, 2) group-based classification analysis, 3) confidence report summarization, 4) collaborative discussion, and 5) consensus agreement.

cal insights and refining models through instruction tuning. For instance, GeneGPT (Jin et al., 2023) leverages Web APIs from the National Center for Biotechnology Information (NCBI) to access diverse biomedical information and then employs GPT models for reasoning tasks. Additionally, the methods in (Zhang et al., 2024b; Singhal et al., 2022; Oniani et al., 2024; Kang et al., 2023) utilize instruction tuning combined with prompt design to adapt LLMs for various healthcare tasks, including decision support, medical question answering, and disease diagnosis.

2.2 LLM Applications in Auto-Agents

Research in both academia and industry has focused on autonomous agents trained in isolated, self-contained environments with limited knowledge bases (Wang et al., 2024a). Significant advancements have been achieved in 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 emulate human learning and decision-making processes, involving 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 such as 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 presents the TRIAGEAGENT framework, which simulates teamwork and problem-solving in diagnosing and planning treatment for ED cases. We explore the heterogeneous structure formulation of the framework, as introduced in Appendix A. We describe the five stages of TRIAGEAGENT's operation stages, illustrated in Figure 2 and detailed in the following procedures: First, **allocating documents**, where patient clinical records are assigned to expert agents to initiate ESI discussions. Second, **group-based classification analysis**, where agents are divided into two groups to conduct coarse and fine-grained classification. Third, **confidence report summarization**, where a summarized report is generated based on previous analyses, including classification results and confidence scores. Fourth, **collaborative discussion**, where agents engage in discussions over the summarized report, iteratively refining it according to key information, rationales, and confidence scores. Fifth, **consensus agreement**, resulting in an ultimate, precise, and thoroughly validated revised report, highlighting the importance of collaborative decision-making.

3.1 Document Allocation

Given a patient's narrative clinical note $P = \{P_1, P_2, \ldots\}$, a clinical natural language query q, and a structured ESI handbook with level references $R = \{R_1, R_2, \ldots\}$, this stage assigns documents to agents and initiates expert discussions, as illustrated in Figure 2 Step 1. For more details about query, see Appendix B.

3.2 Group-Based Document Classification

The classification stage is summarized as a function $f : (P, R, q) \rightarrow C$, where C represents the set of hierarchically structured ESI-level categories. To improve classification efficiency and accuracy, agents are divided into two groups, employing a direct and coarse-to-fine-grained classification, respectively.

Coarse-to-Fine-Grained Classification This group comprises two agents. The first agent, A_1, performs an initial coarse classification of the patient record P_i into two broad categories: high (levels 1, 2, and 3) or low (levels 3, 4, and 5) , with level 3 included in both. This can be represented as $A_1 : P_i \rightarrow \{C_{\text{high}}, C_{\text{low}}\}$. The second agent, A_2, then refines these broad categories into detailed ESI levels: if classified as C_{high}, the second agent selects from $\{1, 2, 3\}$; if C_{low}, it selects from $\{3, 4, 5\}$. This can be represented as: $A_2 : \{C_{\text{high}}, C_{\text{low}}\} \rightarrow \{\{1, 2, 3\}, \{3, 4, 5\}\}$. This two-step process reduces misclassifications and enhances precision. As illustrated in the left of step 2 in Figure 2, the first agent,A_1 (referred to as Doctor A), initially assigns a high ESI level with 90% confidence. A_2 (referred to as Doctor B) then refines the high-level category to ESI level 2 with an 80% confidence score. After rounds of discussion, Doctor B revises the classification to ESI level 1, achieving a revised confidence score of 95%.

Direct Fine-Grained Classification This group consists of a single agent, A_3, who directly assigns ESI levels: $A_3 : P_i \rightarrow \{1, 2, 3, 4, 5\}$. As illustrated in the right of step 2 in Figure 2, A_3 (referred to as Doctor C) initially assigns ESI level 2 with 90% confidence and later refines the classification to level 1 with 95% confidence after further consideration.

By combining the two strategies described above, TRIAGEAGENT effectively addresses the challenges of distinguishing boundaries between ESI levels and enhances decision-making efficiency and accuracy through this collaborative effort.

3.3 Confidence Report Summarization

In this stage, the summarizer agent A_s summarizes previous document classification results, including confidence scores, rationales, and supporting evidence from (A_1,A_2, A_3). This step consolidates the findings and uses each agent's analysis report to construct the summary prompt $Prompt_{rs}$, ensuring a well-supported and reliable decision. Additionally, debates among the agents are incorporated into the summarizer's prompt. The summarizer then generates a synthesized report by extracting key information and analyzing the previous classifications provided by the 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} is the role of the summarizer, and $Prompt_{rs}$ is the guideline prompt for the summarizer, including analysis reports from (A_1,A_2, A_3). The synthesized report is structured as follows: $Repo$=[*key information; confidence score; rationale; consolidated analysis*]. This approach effectively combines insights from multiple agents, ensuring that the triage decision is based on comprehensive and validated information. Consequently, this method enhances the accuracy interpretability and reliability of the clinical triage process. As depicted in Step 3 of Figure 2, the key information includes references from the ESI Handbook (e.g., ESI Handbook v4, Chapter 2: ESI Triage Algorithm, p. 10-13). The summarized report consolidates rationales and total analysis, ensuring that all relevant information is considered.

3.4 Collaborative Discussion

In this stage, agents engage in multiple rounds of discussions based on the synthesized summary report $Repo$ to refine their individual classifications. Unlike the commonly-used voting mechanism (Tang et al., 2024), TRIAGEAGENT critically reflects on the classification results, reasoning, and confidence scores, incorporating peer-provided evidence. Each agent A_i starts with an initial classification result C_i and confidence score S_i follows the following process: A_i reviews the classification results C_j, reasoning R_j, and confidence scores S_j from every other agent $A_j (j \neq i)$. If A_i is persuaded by A_j's reasoning or finds $S_j > S_i$, it updates C_i to C_j with explanations. Conversely, if A_i rejects A_j's reasoning or finds S_j lower or equal to S_i,it justifies keeping C_i. This iterative process continues until agents reach a preliminary consensus or the early stopping mechanism is triggered. In Step 4 of Figure 2, agents participate in collaborative discussions to resolve discrepancies and refine the report.

3.5 Early-stopping Mechanism

To enhance the efficiency of group chat discussions, we implement an early-stopping mechanism inspired by Byzantine Consensus theory (Castro and Liskov, 1999). This approach requires at least $3p + 1$ agents to handle p faulty agents in a single communication round. Additionally, our termination mechanism draws inspiration from advancements in LLMs fine-tuned with Reinforcement Learning from Human Feedback (RLHF), allowing consensus after several debate rounds (Du et al., 2023; Ouyang et al., 2022). The mechanism terminates communication when agents consistently confirm their reasoning with high confidence, thereby reducing unnecessary computations. It operates under two conditions: the first is repetition of high-confidence answers by a single agent: if an agent repeatedly provides the same answer with high confidence, that agent triggers early-stopping and exits the group discussions. The second is repetition of high-confidence answers by multiple agents: if all agents consistently provide the same answer with high confidence, the dialogue is terminated. This dynamic, real-time stopping condition enhances the traditional theory's efficiency, ensuring efficient and accurate consensus in group discussions. These conditions collectively foster an adaptive termination criterion, prioritizing efficiency and accuracy in reaching conclusions (Yin et al., 2023). In our case, we apply the early-stopping mechanism to each round of discussion of the TRIAGEAGENT.

3.6 Consensus Agreement

Finally, agents reach a formal consensus by integrating the refined answers, reasoning, and confidence scores from the collaborative discussion stage. This stage ensures all agents agree on a single classification, leveraging their combined domain knowledge to validate the final decision. This collaborative process ensures the final decision is robust and well-supported by comprehensive analysis. As shown in Step 5 of Figure 2, the final consensus is reached and the definitive classification is provided after all agents agree on the outcome.

Table 1: Statistics of the clinical triage dataset

Dataset	Training	Test-1	Test-2	Test-3
# of Docs	218	72	72	72

4 Experimental Setup

Dataset We construct a clinical triage dataset by collecting patient cases from the publicly available Emergency Severity Index (ESI) Handbook v4 (esi, Accessed: 2024-04-06). This dataset is designed to evaluate machine learning models and methods for categorizing ESI levels in medical documents. To our knowledge, this is the first publicly released clinical triage dataset that includes clinical notes and corresponding ESI levels necessary for triage tasks, serving as a benchmark for evaluating our framework's effectiveness. Since the patient cases are sourced from the official ESI Handbook, no de-identification is needed. The dataset is divided into a training set and three test sets (test-1, test-2, and test-3). The training set contains 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). Each test set contains 72 cases, maintaining fixed proportions of ESI levels: ESI-1 (12), ESI-2 (20), ESI-3 (13), ESI-4 (12), and ESI-5 (15). The dataset statistics are provided in Table 1. For more details, see Appendix C.

Implementation We use GPT-3.5-Turbo (OpenAI, 2024) and GPT-4 (OpenAI et al., 2024) from OpenAI for as our base models for the zero-shot experiments. Our TRIAGEAGENT framework utilizes the publicly open-sourced Autogen framework from Microsoft [1]. The temperature is 0.9, top_k is 1.0, and the cache seed is 42. The maximum number of iterations is 12, and the frequency penalty is 0.1. For SCtr and SCons, we perform 8 iterations with a temperature of 0.9.

Baselines The performance of the TRIAGEAGENT framework is evaluated against several state-of-the-art baselines, including methods that employ LLM-based planning, tool usage, and retrieval-augmented generation.

- Chain-of-thought (Kojima et al., 2023) integrates step-by-step reasoning into the prompt for LLMs. We implemented CoT on our clinical triage dataset as a baseline method for comparison.

[1] https://microsoft.github.io/autogen/

Table 2: Performance comparison of TriageAgent with baseline methods on the clinical triage dataset. The reported performance (%) in this table are averaged values from the three test sets in the dataset (Table 1). The highest performance is highlighted in **bold**. A lower total discordance represents a higher performance. CoT denotes chain-of-thought prompting, SCtr denotes self-contrast prompting, SCons denotes self-consistency prompting, and EoT denotes the exchange-of-thought prompting method.

Supervision	**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	**Human Experts**	31.43	12.80	8.61	18.60	10.50

- Self-contrast (Zhang et al., 2024a) improves stability and accuracy by contrasting different solving perspectives and summarizing discrepancies. We applied self-contrast on our clinical triage dataset to analyze and reconcile conflicting classification results as a baseline for comparison.
- 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 on our clinical triage dataset to generate multiple response chains and select the most consistent answers as a baseline for comparison.
- Exchange-of-thought (Yin et al., 2023) enables cross-model communication and problem-solving integration. We implemented EoT on our clinical triage dataset to facilitate communication between various agents as a baseline for comparison.
- MedAgent (Tang et al., 2024) is a role-playing collaboration framework for medical scenarios using LLMs. We utilized MedAgent's multi-agent framework on our clinical triage dataset as a baseline for comparison.
- TriageAgent utilizes dynamically updated confidence scores from various reasoning perspectives supported by external evidence to enhance the performance. It selects the top K most confident answers for critical assessment. Agents employ retrieval-augmented generation to refine choices, ensuring minimal discordance and maximal coherence. This iterative process integrates multiple methodologies and specialized knowledge retrieval to improve decision accuracy and reliability.

Evaluation Protocol Our primary evaluation metric is the ***total discordance rate***, which measures the percentage of incorrectly predicted queries. This metric is critical as it provides a comprehensive overview of the accuracy; a lower value indicates better performance. Additionally, we evaluate the *undertriage rate*, *overtriage rate*, *significant undertriage rate*, and *significant overtriage rate* to assess the model's performance in specific areas of clinical urgency categorization. While these metrics are important, the *total discordance rate* remains the primary measure of accuracy. Detailed definitions of the five ESI levels (I-V) and the evaluation metrics are provided in Appendix D.

5 Results

5.1 Main Results

We evaluate the performance of TRIAGEAGENT by averaging the results across the three test sets. The backbone LLMs in our experiments include GPT-3.5, GPT-4, Llama-2, and Llama-3. Table 2 presents the main results with a better performance from the GPT-3.5 and GPT-4 models. Detailed results for Llama-2-7B and Llama-3-8B are provided in Appendix E.

Performance Comparison with Baselines The performance of our framework compared to state-of-the-art (SOTA) methods is presented in Table 2. The TRIAGEAGENT framework surpasses traditional prompt engineering methods, including *supervised learning*, *self-contrast*, and *self-consistency*, with improvements of 9.25%, 8.79%, and 5.55%, respectively. It also outperforms SOTA multi-agent frameworks in a zero-shot setting using the GPT-3.5-turbo model, exceeding *MedAgents* by 9.02% and *EoT* by 6.25%.

Comparison with CoT Methods Performance can decline when employing overly complex Chain of Thought (CoT) methods. Simply stacking prompts without a clear, logical sequence can result in hallucinations—erroneous outputs caused by insufficient document comprehension and misunderstanding of medical terminologies. However, our approach, which integrates multi-agent role-playing with confidence assessments, effectively addresses these issues and demonstrates its potential as a more robust method for applying LLMs in clinical triage.

Comparison with Single-Agent Methods Methods such as *CoT*, *self-contrast* and *supervised learning* lack crucial interactions among multiple LLMs. This absence inhibits these methods from adaptively refining their responses, leading to suboptimal performance in triage question-answering scenarios. Consequently, their discordance rates average around 38.95% on our clinical triage dataset when using GPT-3.5 (as illustrated in Table 2), highlighting the need for improvement.

5.2 Ablation Study

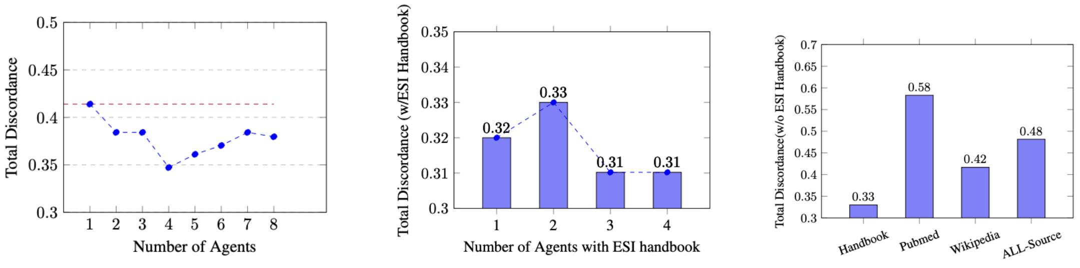
Our ablation study analyzes team optimization and external resource optimization. After determining the optimal structure, the TRIAGEAGENT framework simulates multi-role team collaboration, enabling agents to acquire the necessary capabilities to effectively accomplish triage tasks.

Team Optimization Our heterogeneous framework employs multiple agents, each with a specific role, to achieve optimal outcomes. We explored various configurations to enhance team performance. By adjusting the number of role-specific agents, we found that four agents provided the most optimized structure for performance on the ESI triage dataset. Figure 3(a) details this optimal configuration and demonstrates how adjusting agent roles and numbers enhances overall system performance.

External Resource Optimization We evaluate the ESI Handbook, PubMed, and Wikipedia for supervising our model (Figure 3(c)). The ESI Handbook proved to be the most effective, significantly enhancing model performance with its targeted clinical diagnostic guidelines. In contrast, PubMed and Wikipedia require processing extensive additional data. We also analyze the optimal number of agents using the ESI Handbook to maximize task efficiency (Figure 3(b)). Increasing the number of agents improved model effectiveness without significantly changing overall accuracy, enhancing the contextual understanding and interpretability of medical texts while optimizing resource utilization.

5.3 Case Study

Error Analysis Based on our findings, we conduct an expert evaluation to identify key limitations and common issues in our model. As shown in Figure 4, we categorize these errors into four major types. The first type of error is **lack of document understanding**. This type of error arises from insufficient medical knowledge or incorrect linking to ESI levels, leading to misjudgments of clinical severity. The second type of error is **misretrieval of domain knowledge**. Errors result from inaccurately retrieving irrelevant or mismatched information will compromise triage accuracy. The third type of error is **confidence-based consistency errors**. This type of errors is caused by confidently providing contradictory responses or failing



(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 better performance.

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

Table 4: Cost efficiency on ESI datasets classification

Model	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%	604
CoT (GPT-4)	37.40%	216

to reach a consensus, often due to internal inefficiencies or flaws in the early stopping mechanism, resulting in falsely assured incorrect outcomes. The last type of errors is **exchange of information errors**. This type of errors results from incorrect data transfer between agents, disrupting logical sequences and leading to erroneous conclusions. See Appendix F for more details.

Time Analysis Time efficiency is crucial in emergency clinical triage. Table 3 compares the performance and time costs of different methods. Our multi-agent architecture incurs a slight but negligible increase in time costs compared to a single-agent model. Feedback from emergency department experts indicates that our method does not significantly affect overall time efficiency but greatly enhances decision accuracy, which is crucial for better patient outcomes and resource allocation.

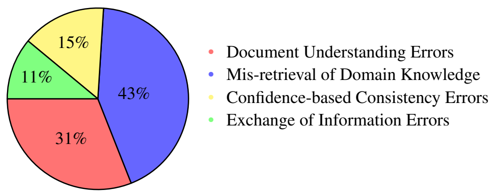


Figure 4: Ratio of different categories of error cases.

Cost Analysis Operational efficiency, particularly API token usage, is crucial for our framework. Table 4 compares the performance and costs of different configurations. While TRIAGEAGENT requires more API calls than a single-agent setup, feedback from clinical departments suggests these costs are reasonable. Our system improves efficiency by 12.54% compared to human experts manually classifying documents and achieves 18.54% higher performance than LLMs using CoT prompts operated by human experts.

6 Conclusions

This paper introduces a novel heterogeneous multi-agent framework, TRIAGEAGENT for clinical triage, utilizing LLM-based role-playing agents in a multi-stage group chat setting. This zero-shot, training-free, and interpretable framework comprises five significant stages. Experiments on clinical triage datasets demonstrate our framework significantly outperforms zero-shot baselines and experienced professionals. Case studies and human evaluations highlight areas for improvement, such as reducing document understanding errors and knowledge misretrieval. Future research can enhance the framework's efficiency by improving document comprehension and correcting domain knowledge retrieval errors. Upon acceptance, we will release our dataset as open source.

Ethics Statement

This research adhered to the highest ethical standards and best practices, as outlined in the ACL Code of Ethics. All experiments were conducted using publicly available datasets, thereby avoiding concerns related to privacy, confidentiality, or personal information. The datasets used are fully anonymized and have been vetted to ensure compliance with ethical guidelines. Additionally, we have carefully considered the broader impacts and potential applications of our work, ensuring that it does not inadvertently cause harm or misuse. Consequently, we believe this research is free from ethical issues.

Limitations

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

Limited Expert Evaluation Our research is limited by the involvement of three human experts, restricting the scope and depth of expert evaluation, which may impact the generalizability of our findings to broader clinical settings.

Cross-Institution Collaboration The complexity of emergency departments requires advanced triage systems to address patient conditions within the same urgency levels. This need arises from varying institutional conditions,protocols, and patient demographics.

Workflow Integration Our model provides a final triage decision, but actual triage often involves multiple decision-making stages. Thus, it's crucial to evaluate how well our system integrates into existing workflows and complements human-led emergency care.

Privacy Deploying our framework necessitates strict privacy measures and clinical worker training. Processing clinical notes can expose sensitive information, making compliance with HIPAA and GDPR.

Time and Cost Efficiency TRIAGEAGENT aims to enhance clinical triage by automating initial patient assessment, reducing manual workload, waiting times, burnout and operational costs. However, using external technologies like OpenAI's API adds expenses and dependencies. A detailed cost-benefit analysis is essential for sustainability and economic viability.

References

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

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation.

Zhenyu Bi, Sajib Acharjee Dip, Daniel Hajialigol, Sindhura Kommu, Hanwen Liu, Meng Lu, and Xuan Wang. 2024. Ai for biomedicine in the era of large language models.

Miguel Castro and Barbara Liskov. 1999. Practical byzantine fault tolerance. In *Proceedings of the Third Symposium on Operating Systems Design and Implementation*, pages 173–186, USA. USENIX Association.

G Chenais, E Lagarde, and C Gil-Jardiné. 2023. Artificial intelligence in emergency medicine: Viewpoint of current applications and foreseeable opportunities and challenges. *J Med Internet Res*, 25:e40031.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. Metagpt: Meta programming for a multi-agent collaborative framework.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

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

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, Carl Yang, and Liang Zhao. 2024. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Zhengliang Liu, Zihao Wu, Mengxuan Hu, Bokai Zhao, Lin Zhao, Tianyi Zhang, Haixing Dai, Xianyan Chen, Ye Shen, Sheng Li, Brian Murray, Tianming Liu, and Andrea Sikora. 2023. Pharmacygpt: The ai pharmacist.

Danqing Ma, Meng Wang, Ao Xiang, Zongqing Qi, and Qin Yang. 2024. Transformer-based classification outcome prediction for multimodal stroke treatment.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. 2024. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines.

OpenAI. 2024. Gpt-3.5 turbo documentation. `https://platform.openai.com/docs/models/gpt-3-5-turbo`. Accessed: 2024-06-15.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool learning with foundation models.

Chantal Shaib, Millicent L. Li, Sebastian Joseph, Iain J. Marshall, Junyi Jessy Li, and Byron C. Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success).

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, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera 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. 2022. Large language models encode clinical knowledge.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2023. Adaplanner: Adaptive planning from feedback with language models.

Rocío Sánchez-Salmerón et al. 2022. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing*, 60:101109.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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. 2024a. A survey on large language model based autonomous agents.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023a. Augmenting language models with long-term memory.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024b. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration.

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.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, and Tao Yu. 2023. Openagents: An open platform for language agents in the wild.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate.

LH Yao, KC Leung, CL Tsai, CH Huang, and LC Fu. 2021. 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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024a. Self-contrast: Better reflection through inconsistent solving perspectives.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2024b. Alpacare:instruction-tuned large language models for medical application.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. Webarena: A realistic web environment for building autonomous agents.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools.

A Model Architecture Comparison with Previous Work

We compare TRIAGEAGENT with 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 simulates 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. The changing colors of nodes in the figure illustrate our dynamic heterogeneous nature, representing agents' evolving perspectives based on different rounds of discussions. In the second row, nodes represent agents at different time steps, arrows indicate the edges, and colors signify the roles of the agents.

Node A node represents an agent at a specific time step, each with a unique role in a heterogeneous role-playing structure. This setup simulates real-world collaboration, enhancing decision-making efficiency and accuracy by assigning distinct functions and tasks to each agent. Additionally, the changing node colors in the figure highlight our framework's dynamic nature, illustrating how agents' perspectives evolve as they assimilate and process new information through subsequent discussion rounds.

Edge Edges represent the communication channels between nodes during multi-agent collaboration, illustrating how information flows between agents and through the system. In our LLM-agent-based feed-forward network, these edges show how agents share information across different time steps to generate the final answer for the task query.

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. 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 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.

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.

C.1 Recruitment and Payment

The human experts assisting us in the classification of medical documents are professional experts who voluntarily participated in our comparative study as collaborators. We did not provide them with any additional payments or benefits.

C.2 Instructions Given to Participants

We invited professional human experts to serve as human annotators. The full text of the instructions given to participants is: "This study aims to evaluate the effectiveness of a generative AI model in predicting ESI levels and compare its performance with traditional nurse triage. Your participation will help us understand the potential of AI in augment-

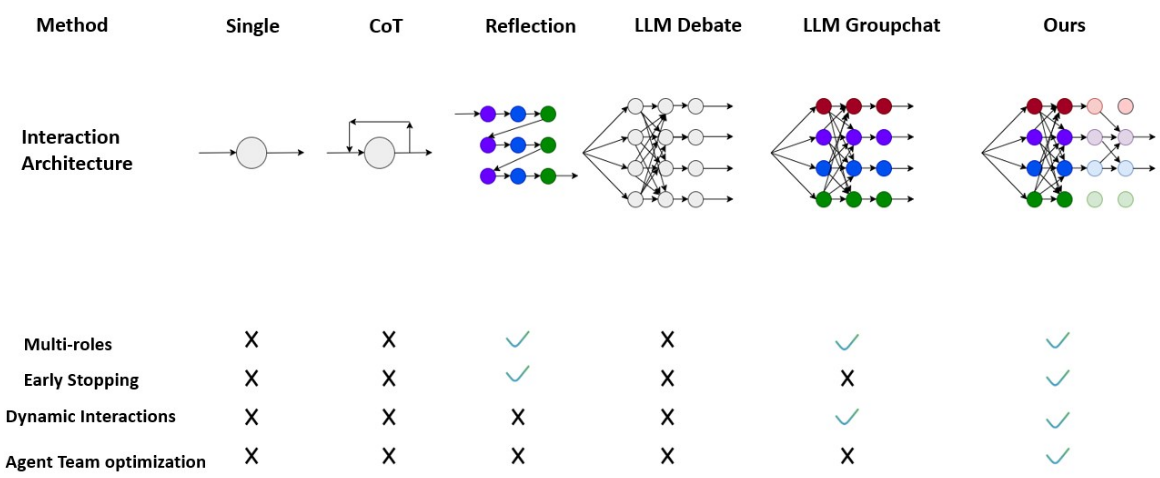


Figure 5: Topology Structure Comparisons of previous methods

ing emergency department workflows and improving patient care. All test results are de-identified and will only be used for the purposes of this research study. Please do not look up answers or use any additional resources to complete the test as that can negatively impact the validity of this study."

D ESI Hierarchy and Evaluation Metrics

D.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.

D.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:

$$\textbf{Total discordance} = \frac{\text{Total Misclassifications}}{\text{Total number of texts}} \quad (1)$$

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:

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

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

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

The *Significant Undertriage* rate captures the scenarios where the true label is critical (ESI *1* or2) but the prediction underestimates the urgency (predicted as *3*, *4*, or *5*):

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

Lastly, the *Significant Overtriage* rate addresses the instances where less critical labels (true label

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%

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

is 2, 3, or 4) are overestimated by the model (predicted as 1):

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

E Performance of Different Backbone LLMs

Using Different Language Models In our experiments, we utilized GPT-3.5-turbo [2], GPT-4 [3], their combination, and LLaMa-2-7B[4] and Llama-3-8B [5].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

F 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.

[2] https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

[3] https://openai.com/index/gpt-4/

[4] https://llama.meta.com/llama2/

[5] https://llama.meta.com/llama3/

•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%.

Agent_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 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.